



THE UNIVERSITY OF
BUCKINGHAM

Automation of Violent Activity Recognition Utilising CCTV Video Data

School of Computing at the University of Buckingham United
Kingdom

By

Matthew Marlon Gideon Parris

A thesis submitted for the Degree of Doctor of Philosophy in
Computer Science to the School of Computing at the University of
Buckingham

August 2024

Student ID: 1804564

Abstract

In this era, security trends identified violence as a significant issue plaguing society globally. Statistics depicted alarming thresholds for violence, establishing itself as a momentous challenge for homeland security and defence institutions, predominantly in schools and other public locations. The advent of state-of-the-art closed-circuit television (CCTV) surveillance solutions exists to aid in limiting the manifestations of violence and its impact. However, most institutions need proper analysis mechanisms that lead to prevention, apprehension, or conviction in a timely fashion. Manually monitoring and collectively analysing anthropometric data generated by CCTV surveillance devices proved impractical and time-consuming, and its outcome increases the complexity of identifying violent behavioural patterns as substantial evidence. Despite innovative CCTV sensor improvement, the impact of adequately analysing vast amounts of CCTV data adds to the monitoring challenge. This thesis proposed the amalgamation of the "You Only Look Once version five medium" model (YOLOv5m) as activity recognition and Three Dimensional Convolution Neural Network Single level (3DCNNsl) activity recognition, two state-of-the-art artificial intelligence models incorporating weight embedding procedures to identify primitive stages of violence and weapons artefacts. The approach integrates classification support to confirm the existence of specific weapon objects (knives, bladed instruments, clubs, and guns) of interest belonging to a specific class of violence (beating, shooting, stabbing). It also validates the presence (primitive stages of violence) of violent classes by utilising the existence of weapons belonging to its category group to infer the activity outcome. Utilising classification support concepts to validate the existence of primitive stages of violence enhances the classification outcome of violent activity recognition with robust results. This thesis commenced by conducting a two-stage literature investigation to satisfy the research objectives, which disclosed the state-of-the-art 3DCNNsl at stage one and the YOLOv5m framework for activity with artefact recognition towards violence at stage two. The proposed one-stage (simultaneously performing object localisation and classification) solution combines the models' processing, reducing the impact of their architectural

limitations. 3DCNNsl facilitates behavioural pattern classification, generically associating sub-class labels suggesting the presence of violence at high accuracy. In addition to 3DCNNsl, YOLOv5m architecture serves two functions: operating in an activity recognition capacity, fortifying 3DCNNsl activity output, and detecting artefacts, which establish the presence of weapons, enhancing the action classification and overall accuracy. The thesis optimised the deep learning model selections by identifying violence in scenarios and validating its presence through a redundant weapon artefact classification weight embedding procedure. The concept allows the classification of violence in its primitive stages before its impact escalates to lethal outcomes. The proposal extensively reviewed its operations via transfer learning in multiple fusion scenarios to identify the most optimal strategies to realise the research objective. The evaluation dataset utilised in this thesis encompassed a selection of samples accumulated via the University of Central Florida (UCF) dataset and several social media forums. The violent action samples reflect several multifaceted real-world scenarios representing sporadic accelerated motion attributes in various environments, which aids in reducing the risk of dispensing biased results and affecting the model's robustness. The proposal disclosed three contributory elements, which reflect the following;

1. Conducted performance testing of two known machine learning techniques (YOLOv5m and 3DCNNsl) in independently recognising violent and non-violent activities in CCTV video footage.
2. Demonstrated violent activity recognition performance in such videos when both machine learning techniques operate in tandem.
3. Implemented performance enhancement by further incorporating threat object detection in the previous combined solution.

Contribution one disclosed the effectiveness of YOLOv5m activity recognition at 74% and

the state-of-the-art 3DCNN at 75%, conceding high misclassifications utilising data with and without augmentations and resolution modifications. The operations emphasised the obligation to explore alternative processing measures to alleviate the disadvantages of the two machine learning models. Contribution two emphasised the effectiveness of fusion enhancement techniques via decision-level voting at 85.20% over 3DCNNsl and YOLOv5m activity recognition. As a validation strategy, the operations incorporated surplus data encompassing 50 samples designed to enhance the classification complexity. The approach rigorously appraised the operations, thus confirming its applicability. Contribution three showcased the amalgamation of fusion’s activity recognition and the power of object detection to establish its effectiveness in concatenating weight embedding. The experiments maintained data consistency similar to contribution two. Analysis disclosed the dominance of fusion incorporating threat object detection at 88.20% over 3DCNNsl, YOLOv5m activity recognition, and fusion without threat object enhancement.

The results underscore the robustness of the proposed method, which has proven its classification competence, particularly in scenarios with surplus data, from an overall accuracy perspective. While the proposal debates the efficiency of individual processing compared to fusion without support, the research endeavour accentuates the effectiveness of integrating classification redundancy through weight embedding to suggest the presence of artefacts confirming the occurrence of violent actions. The findings highlight the effectiveness of the proposed method without artefact processing at 85.20% while incorporating threat object support analysis concatenating weapons (knife, club, gun in the videos) improved the accuracy to 88.20%. This evidence substantiates the solution’s robustness, fulfilling the research objectives to conclude the investigations¹.

¹**Keywords:** Anomalous Detection, Activity Recognition, Motion Detection, Violence Recognition, Human Activity Detection

Dedication

The achievement of the title Doctor is humbly devoted to the following:

- **Mother Olinda Maria Maraj:** Mother, dearest, this lifetime achievement represents a gift of sacrifice as being the first member of all generations to accomplish this ever-challenging task. With mixed emotions, acknowledging this auspicious occasion is a momentous task due to Granddad and Grandmother's recent passing.
- **Aunt Eleanor Gray:** Thanks for creating the essential opportunity to be in the UK and for bestowing support and contributions in every way possible. Please accept this gift as a projection of a grateful heart.
- **Aunt Peta Bain:** When the journey became unstable and faith diminished in every way possible, the encouraging directives and steadfast support prevented further hindrances towards the end of this journey. Thanks for making this aspect of the journey bearable and possible.
- **Aunt Tracy Chase:** Jah Sent, an angel sent by God, described the effort, devotion, and unflinching support which fortified the importance of completing this task with spiritual enlightenment. This gift of being called Doctor is undoubtedly fitting for such an angel, utterly grateful.
- **Aunt Dianne Dumas:** For demonstrating an unwavering determination to stabilise the thoughts of success. For love demonstrated via support/contributions through **Thibaut** and for believing in concepts foretold by grandmother. This Gift of Dr is fitting to commemorate the thoughtfulness disclosed.
- **Aunt Gail Da'Costa:** For pushing the mindset towards achieving this task to encourage

the younger generations. The support and words expressed cannot go unnoticed via **Aunty Joyce** in Trinidad and Tobago and **Uncle Kirk** in the UK. With gratitude, please receive this humble gesture of success as doctoral proof that anyone can achieve with a specific mindset.

- **The brothers Mario and Lorenzo Maraj.** The assistance and love as siblings were displayed from a great distance, no matter how small it may have seemed. The efforts proved instrumental in turning a lifelong goal into a reality. This gift of success represents the evidence of an honorary achievement, which is fitting to demonstrate a grateful heart.

- **Offspring: Kaís Parris.** This achievement represents a lifetime of love and **great sacrifice**, demonstrating a tangible element as motivation that anything is possible with the right mindset. Just believe, and the job is halfway complete with God as the source of strength!

- **Nieces: Jena-Marie (Jen) Gissippi, Caroline (Caro), and Esperanza (Espards) Maraj.** The title of Doctor serves as motivation to achieve unattainable dreams. No excuses; starting and finishing are critical factors for life, so take heed!

(Matt)

(Matthew Marlon Gideon Parris)

© 8th-August-2024

Acknowledgment

Firstly, all glory and praises to the most high god! Only by grace was this venture possible. Let thy precious name shed light in exaltation through **Jesus Christ (Jesus-Always-Helps (Jah))**.

Special thanks to the dean of the School/Faculty of Computing UoB, **Professor Harin Sellaheewa**, who commented on exuding tenacity to achieve from inception. The humanitarian efforts demonstrated and directives ushered in the opportunity to conduct this endeavour.

To the Programme Director/Research Lead of MSc Innovative and MSc Applied Computing, **Dr Hisham Al-Assam**. Special thanks for the profound dedication and all directives at the Doctoral level.

Thanks to Programme Director/Lecturer **Dr Mohammad Athar Ali** for assisting at this level.

Special thanks to all UoB Staff/Members of the School/Faculty of Computing for their efforts towards this journey, especially during COVID-19. Considering those trying times, gratitude is fitting at this stage.

Special thanks to the auxiliary advisors **Professor Ihsan Lami/Dr Maysson Ibrahim** for bestowing encouragement and directives at the time of need.

Special thanks to **my research colleagues** specifically **Dr David Traoré**, for demonstrating kindness/guidance towards understanding Artificial Intelligence via MATLAB.

Special acknowledgements to **Dr Kelenne V. and Linette Tuitt, Martha Randoo, Helmise Morales, Carol Chadband, and Jaizarno Parris** for encouraging/contributing resources towards the completion of this venture.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 21 |
| 1.1 | Summary of the Research Motivation | 24 |
| 1.2 | Summary of the Research Questions | 28 |
| 1.2.1 | Summary of the Main Research Question: | 28 |
| 1.2.2 | Summary of the Sub-Research Questions: | 28 |
| 1.3 | Overview of the Aim and Objectives | 29 |
| 1.4 | The Contributions of the Thesis | 30 |
| 1.5 | Overview of the Thesis Structure | 31 |
| 2 | Context and Fundamental Knowledge | 33 |
| 2.1 | Overview of Data Pre-processing and Blob Analysis | 36 |
| 2.1.1 | Summary of the Class of Activity Template (CoAT): | 37 |
| 2.1.2 | Blob Analysis Applied on the Class of Activity Template (CoAT): . | 38 |
| 2.2 | Summary of Cross-Validation Split | 39 |
| 2.3 | Overview of Software and Library Packages | 40 |
| 2.4 | Overview of Classification Network Operations | 41 |
| 2.4.1 | Overview of Kernel Stride and Padding Operations: | 42 |
| 2.4.2 | Summary of RGB Images Evaluation: | 43 |
| 2.4.3 | Summary of RELU Layer Operations: | 44 |
| 2.4.4 | Summary of Pooling Layer Operations: | 45 |
| 2.4.5 | Overview of 1st Object Detection Processing via RCNN: | 46 |
| 2.5 | Background of YOLOv5 | 49 |
| 2.5.1 | Overview of YOLOv5 Activity Recognition Perspectives: | 49 |
| 2.6 | Background of 3DCNN | 52 |
| 2.6.1 | Overview of 3DCNN Activity Recognition Layers: | 53 |

| | | |
|----------|---|-----------|
| 2.6.2 | Overview of Activity Recognition Processing Components: | 57 |
| 2.7 | Overview of Classification Evaluation Measures | 57 |
| 2.8 | Summary of Classification Issues: Over/Under and Good Fitting | 60 |
| 2.9 | Literature Review | 62 |
| 2.9.1 | Stage-1: Object Detection for Violent Activity Mechanisms: | 65 |
| 2.9.2 | Stage 2 Investigations: for Violent Activity Recognition: | 70 |
| 2.10 | Conclusion | 72 |
| 3 | Research Methodology | 73 |
| 3.1 | YOLOv5m Activity Recognition Class Description | 75 |
| 3.1.1 | YOLOv5m/3DCNN Activity Recognition Dataset Grouping: | 75 |
| 3.2 | 3DCNN Activity Recognition Class Description | 77 |
| 3.3 | Experimental Protocol: | 78 |
| 3.3.1 | Overview of YOLOv5m Activity Recognition Setup: | 79 |
| 3.3.2 | Overview of 3DCNN Activity Recognition Setup: | 80 |
| 3.4 | Evaluation Approach for Standalone Models | 81 |
| 3.5 | Conclusion | 81 |
| 4 | Evaluating YOLOv5m/3DCNNsl | 83 |
| 4.1 | Experimental Setup | 84 |
| 4.1.1 | Overview of Experimental Conditions for YOLOv5 (Phase One): | 84 |
| 4.1.2 | Overview of Experimental Conditions for 3DCNN (Phase-2): | 85 |
| 4.1.3 | Summary of YOLOv5m Experimental Setup: | 85 |
| 4.1.4 | Summary of 3DCNNsl Experimental Setup: | 87 |
| 4.2 | How YOLOv5 Operates | 88 |
| 4.2.1 | YOLOv5 Activity Recognition Limitations: | 89 |
| 4.3 | How 3DCNN Operates | 90 |
| 4.3.1 | 3DCNN Activity Recognition Limitation Breakdown: | 91 |

| | | |
|----------|--|------------|
| 4.4 | Overview of YOLOv5m and 3DCNNsl Results Analysis | 91 |
| 4.4.1 | Summary of YOLOv5m Precision, Recall and mAP: | 92 |
| 4.4.2 | Summary of 3DCNNsl Precision, Recall and mAP: | 94 |
| 4.5 | Overview of YOLOv5m/3DCNNsl Discussions | 96 |
| 4.5.1 | YOLOv5m Metrics and Confusion Matrix Discussions: | 96 |
| 4.5.2 | 3DCNNsl Metrics and Confusion Matrix Discussions: | 97 |
| 4.5.3 | Research Question Discussions Evaluating YOLOv5m/ 3DCNNsl: | 97 |
| 4.5.4 | Summary of Operational Challenges: | 100 |
| 4.5.5 | Summary of 3DCNNsl Operational Challenges: | 104 |
| 4.6 | Conclusion | 107 |
| 5 | Merging 3DCNNsl/YOLOv5m for Robust Violent Activity Recognition | 109 |
| 5.1 | Motivation for the Fusion Enhancement Approach | 110 |
| 5.1.1 | Overview of the Fusion Rationale: | 114 |
| 5.1.2 | Definition of Decision Level Fusion Protocols: | 115 |
| 5.1.3 | Experimental Setup of Fusion: | 118 |
| 5.2 | Fusion Scheme-1: Utilising 12-Frame Processing | 119 |
| 5.3 | Fusion Scheme-2: Processing the Entire Video | 120 |
| 5.4 | Fusion Scheme-1 Results: Utilising 12-Frame Processing | 121 |
| 5.4.1 | Definition of Column Title per Outcome: | 122 |
| 5.4.2 | Definition of Confidence Thresholds for Stabbing: | 122 |
| 5.4.3 | Fusion Scheme-1 12-Frame Results for Stabbing24.avi: | 126 |
| 5.4.4 | Fusion Scheme-1 12-Frame Stabbing37.avi Results: | 127 |
| 5.4.5 | Fusion Scheme-1 12-Frame Fencing27.avi Results: | 127 |
| 5.4.6 | Fusion Scheme-1 12-Frame Overall Accuracy Evaluation: | 129 |
| 5.4.7 | Fusion Scheme-1-12 Frames per Sample Results 50 Samples: | 131 |
| 5.5 | Fusion Scheme-2 Results: Processing the Entire Video | 132 |
| 5.5.1 | Summary of Confidence for Full Video Classification Effectiveness: | 133 |

| | | |
|----------|---|------------|
| 5.5.2 | Fusion Scheme-2 Whole Video Evaluation on 10-Test Samples: | 134 |
| 5.5.3 | Fusion Scheme-2 Whole Video Evaluation on 50 Test Samples: | 135 |
| 5.6 | Fusion Scheme-1 and Fusion Scheme 2's Discussions | 137 |
| 5.7 | Conclusion | 138 |
| 6 | Merging Action Recognition/Object Detection for Violence Recognition | 140 |
| 6.1 | Overview of YOLOv5m as Artefact Object Detection | 141 |
| 6.1.1 | YOLOv5m Artefact Object Detection Experimental Setup: | 141 |
| 6.1.2 | Evaluating YOLOv5m Artefact Object Detection | 143 |
| 6.2 | How YOLOv5m Artefact Detection Works | 146 |
| 6.2.1 | YOLOv5m Action Recognition Artefact using 12-Frame Stabbing8.avi: | 148 |
| 6.2.2 | YOLOv5m Action Recognition Artefact using 12-Frame Stabbing48.avi: | 149 |
| 6.2.3 | YOLOv5m Action Recognition Artefact using 12-Frame Fencing27.avi: | 149 |
| 6.3 | How Fusion Activity Recognition Work using Artefact Support | 149 |
| 6.3.1 | Embedding YOLOv5m Activity Recognition Artefact Fusion Weights: | 153 |
| 6.3.2 | Fusion using Artefact Decision Level Protocol Embedding: | 155 |
| 6.3.3 | Fusion Scheme-1 12-Frame Artefact Support (on 10-Test Samples): | 160 |
| 6.3.4 | Fusion Scheme-1 12-Frame Artefact Support (on 50-Test Samples): | 161 |
| 6.3.5 | Fusion Scheme-2 Whole Video Artefact Support (on 10-Test Samples): | 162 |
| 6.3.6 | Fusion Scheme-2 Whole Video Artefact Support (on 50-Test Samples): | 163 |
| 6.4 | Artefact Operational Discussions | 164 |
| 6.4.1 | Fusion Scheme-1 12-frame Discussions on Artefact Processing: | 165 |
| 6.4.2 | Fusion Scheme-2 Whole Video Discussions on Artefact Processing: | 166 |
| 6.5 | Conclusion | 167 |
| 7 | Conclusion | 169 |
| 7.1 | Research Questions' Assessment | 170 |
| 7.2 | Contributory Factors | 173 |

| | | |
|-------|---|------------|
| 7.2.1 | Introduction to publications: | 174 |
| 7.3 | Limitations of the Study | 174 |
| 7.4 | Future Research and Recommendations | 175 |
| | References | 179 |
| | Appendix | 215 |
| 1 | Overview of the Demographic Influence of Violence | 215 |
| 2 | Overview of YOLOv5 Processing | 217 |
| 3 | The Methodology | 227 |
| 4 | Appraising YOLOv5 and 3DCNN | 237 |

List of Figures

| | | |
|------|---|-----|
| 1.1 | CCTV Monitoring Creates Data Volumes, Source: [3], [5], [6], [7], [8],[9] | 24 |
| 1.2 | Knife or sharp instrument offences: [12],[13] and [14] | 25 |
| 1.3 | Lethal Events Source [15], [16], [17], [18], [19], [20], [21], [22] 01-10-2022. | 26 |
| 2.1 | Blob Analysis Operations via MATLAB. | 38 |
| 2.2 | Cross-Validation Dataset Split Procedures. | 40 |
| 2.3 | Kernel Processing Operations. | 41 |
| 2.4 | (a) Kernel Processing Operations. | 42 |
| 2.5 | (b) Padding with Stride Parameter Set as 1/2. | 43 |
| 2.6 | Classifying an RGB Image. | 44 |
| 2.7 | RELU (Rectified Linear Activation Function) Operations. | 44 |
| 2.8 | Max Pooling Operations. | 45 |
| 2.9 | R-CNN 2-Stage Processing Operations Source: [97]. | 46 |
| 2.10 | R-CNN Processing Operations Source: [97]. | 47 |
| 2.11 | R-CNN Ground-truth Overlapping and Intersection Over Union | 47 |
| 2.12 | R-CNN Dispenses Added Offset Value for Prediction Support Source: [103]. | 48 |
| 2.13 | YOLO's Intersection Over Union Processing Output. | 49 |
| 2.14 | YOLOv5 Object Detection Processing Source: [116]. | 51 |
| 2.15 | 3-Dimensional Input Compared to 2-Dimensional Classification Input. | 53 |
| 2.16 | 3DCNN Proposed Architecture. | 55 |
| 2.17 | 3DCNN Soft-Max Cross-Entropy Log Loss Process. | 56 |
| 2.18 | Graphical Representation of a Good Data Fitting Operations Model. | 61 |
| 4.1 | Rational for the Proposed Fusion: YOLOv5's Misclassification Performance. | 89 |
| 4.2 | Validating 3DCNN's Limitations via its Class Status/Accuracy Output. | 92 |
| 4.3 | YOLOv5m Separate Object Processing Complications. | 101 |

| | | |
|-----|--|-----|
| 4.4 | YOLOv5m's Complications with Minute Objects. | 102 |
| 4.5 | YOLOv5m Separate Object Processing Complications. | 103 |
| 4.6 | Combining Classes to Reduce YOLOv5m Class Misrepresentation. | 104 |
| 4.7 | Combining Classes to Reduce YOLOv5m Class Misrepresentation. | 105 |
| 4.8 | Combining Classes to Reduce YOLOv5m Class Misrepresentation. | 105 |
| 5.1 | Illustration of Fusion Operations via Step-1-5. | 113 |
| 5.2 | Decision Level Fusion Protocol Operations. | 117 |
| 5.3 | Fusion 12-Frame Confidence Processing using 1-Sample (Stabbing91.avi). . . | 124 |
| 5.4 | 10-Test Sample Nuance of Fusion Special Classification Cases. | 125 |
| 6.1 | YOLOv5m Artefact Object Detection Results for Image 1-6. | 144 |
| 6.2 | YOLOv5m Artefact Object Detection Results for Image 7-12. | 145 |
| 6.3 | YOLOv5m Artefact Weighted Values. | 147 |
| 6.4 | Proposed Fusion with Artefact Decision Level Weight Enhancement. | 152 |
| 6.5 | YOLOv5m Artefact Weighted Values | 154 |
| 6.6 | Fusion Scheme-1 12-Frame Stabbing24.avi Activity Unknown Cases. | 165 |
| 6.7 | Fusion Scheme-1 12-Frame Stabbing37.avi Artefacts Special Cases | 166 |
| 6.8 | Fusion's Dominance Over Individual Processing | 168 |

List of Tables

| | | |
|------|--|-----|
| 3.1 | YOLOv5m Subclass Violent and Non-Violent Description. | 76 |
| 3.2 | 3DCNN Activity Recognition Subclass Description. | 78 |
| 4.1 | Summary of Activity Recognition Experimental Conditions. | 84 |
| 5.1 | Decision Level Protocol Operations. | 115 |
| 5.2 | Column Description of the Results. | 122 |
| 5.3 | Examples of Confidence Level for Stabbing | 123 |
| 5.4 | Fusion Scheme-1 12-Frame Evaluation via stabbing24.avi. | 127 |
| 5.5 | Fusion Scheme-1 12-Frame Evaluation via stabbing37.avi. | 128 |
| 5.6 | Fusion Scheme-1 12-Frame Evaluation via fencing27.avi. | 128 |
| 5.7 | Fusion Scheme-1 Performance Evaluation for 12 Frames per 10-Samples. . . | 131 |
| 5.8 | Evaluating Accuracy/Confidence on 50-Samples at 12-Frames per Sample. . | 132 |
| 5.9 | Performance Evaluation for 10-Samples in Whole Video Processing. | 135 |
| 5.10 | Performance Evaluation for 50-Samples in Whole Video Processing | 136 |
| 6.1 | YOLOv5m Artefact Activity Recognition on Stabbing8.avi. | 148 |
| 6.2 | YOLOv5m Artefact Activity Recognition on Stabbing48.avi. | 150 |
| 6.3 | YOLOv5m Artefact Activity Recognition on Fencing27.avi. | 150 |
| 6.4 | Definition of Weighted Class & Score Values. | 155 |
| 6.5 | Artefact Decision Level Fusion. | 157 |
| 6.6 | Fusion Scheme-1 Artefact Results for 10-Samples at 12-Frames per Sample. . | 160 |
| 6.7 | Fusion Scheme-1 Artefact Results for 50-Samples at 12-Frames per Sample. . | 161 |
| 6.8 | Fusion Scheme-2 Artefact Results for 10-Samples in Whole Video Processing. | 162 |
| 6.9 | Fusion Scheme-2 Artefact Support for 50-Samples in Whole Video Processing. | 164 |
| 6.10 | Fusion Scheme-1 & 2 Artefact Processing Dominance. | 167 |

List of Abbreviations

1. 3DCNN: Three-Dimensional Convolution Neural Network
2. 3DCNNsl: Three-Dimensional Convolution Neural Network Single Level
3. 3Dsc: 3DCNN Score
4. ACC: Overall Accuracy
5. AGG: Aggressor
6. AI: Artificial Intelligence
7. AR: Activity Recognition
8. BBC: BBC British Broadcasting Corporation
9. BLOB/blob: Region or object of interest within an image or images/ video data
10. BL: Blood
11. CCTV: Close Circuit Television
12. CNN: Convolution Neural Network
13. CoAT: Class of Activity Template describing actions of interest for this research
14. CM: Confusion Matrix
15. CPU: Central Processing Unit
16. COCO: Common Objects in Context
17. CSP: Cross-Stage Partial
18. DL: Deep Learning
19. DNN: Deep Neural Network
20. DDR3: Double Data Rate Three
21. EX#/EXP/exp: Experiments Deployed

- 22. FN: False Negative
- 23. FP: False Positive
- 24. FS: From Scratch
- 25. Gb: Gigabyte Of RAM (Random Access Memory)
- 26. GHz: Gigahertz, Measures Computer Processing Speed
- 27. GPU: Graphic Processing Unit
- 28. HMDB51: Human Motion Database Dataset 51
- 29. HND: Hand
- 30. i7: Intel computer processors group name
- 31. IoT: Internet of Things
- 32. J-IMDB: Joint-internet movie datasets
- 33. K: KTH action datasets
- 34. K-W/KW: Knife-Weapon
- 35. K-D/KD: Knife-Deploy
- 36. LIRIS: Laboratoire d'informatique en image et systèmes d'information
- 37. LSTM: Long short term memory network
- 38. mAP: Mean average precision
- 39. MATLAB: Matrix laboratory software
- 40. MATHWORKS: Mathematical computing software
- 41. Mb: Megabyte, measures the amount of data personal computers can store
- 42. MHz: Megahertz, a unit of alternating current (ac)

- 43. M1: Apple m1 series of arm-based systems-on-a-chip (socs)
- 44. Neutral: Non-violent human activity
- 45. OD: Object detection
- 46. PAnet: Path aggregation network operations
- 47. PC/PC's: Personal computer/personal computers
- 48. PT: Pre-trained
- 49. RELU: Rectified linear activation function
- 50. R-CNN: Region-based convolutional neural network
- 51. ROI: Region of interest relative to important blobs features
- 52. RQ: Research questions
- 53. STAB: Stabbing class
- 54. SOTA: State of the art
- 55. TN: True negative
- 56. TP: True positive
- 57. TPU: Tensor processing unit
- 58. UCF: University of central Florida
- 59. VGG: Visual geometry group-utilising 16 or 19 layers
- 60. VICT: Victim class person receiving injuries from violent in this research
- 61. VRWA: Violent-real world-actions
- 62. YAML: Yet another markup language for writing configuration files.
- 63. YOLO: You only look once

- 64. YOLOv: Yolo version/higher the number equates to the latest version
- 65. YOLOv5: Yolo version 5
- 66. Ysc: YOLOv5 score

Declaration

The research efforts entitled Violent Behavioural Pattern Recognition, which utilises object detection and activity recognition with a fusion mechanism for classification enhancement, are the product of a personal endeavour. Included are citations that aid in supporting the context of the thesis from additional sources, but such are not directly related to affect their conclusions.

This declaration unequivocally confirms that no part of this research effort, as presented in this thesis, has been submitted elsewhere or conflicts with another thesis for any degree, diploma, or other qualification at the University of Buckingham or any other academic institution.

Author: Matthew Marlon Gideon Parris _____

Chapter 1

Introduction

The study of human behaviour proved a complex phenomenon that has amazed researchers over the millennium. The correlation of individual behavioural patterns varies considerably, especially when comparing the accelerated propensity of such normal or abnormal human motions from its primitive stage (start, middle and end). Those varying actions inspired security concepts with capabilities that capture such complexities through innovative sensor devices. Vision sensors such as closed-circuit television (CCTV) devices are considered appropriate monitoring tools that can manually substantiate and focus on abnormal action features with the collective support of human aid. In this era, such tools evolved mainly to alleviate security issues by tracking violent human actions to limit the manifestation of violent corollaries in society. Although these powerful surveillance CCTV devices are at the forefront when monitoring nefarious human activities, further analytical aid is required to mitigate law enforcement's deficiencies competently with the apprehension of collaborating subjects. As a result, criminals evade proper convictions as the manual CCTV data collective measures lack appropriate activity analysis that substantiates crucial evidence with precision accuracy ratings. Most of the CCTV images attain low-level gradient intensities with poor resolution, negatively impacting the aggregation of pertinent details that can

provide valuable conviction results. It is immensely challenging for humans to manually discern abnormal activities accurately in scenarios where scenes contain low-level gradient intensities. However, manual discernment issues have a high possibility of being solved by implementing artificial intelligence (AI) mediums.

With the availability and accessibility of AI super-processing, it can produce details of an object’s instantaneous pre-movement based on anthropomorphic patterns in dismal locations in real-time. Such state-of-the-art AI solutions attain capabilities that can identify homogeneous and heterogeneous attributes in behavioural patterns, concluding an object’s speed, trajectory, and velocity. AI solutions can achieve formidable superhuman tasks by monitoring anthropometric comparisons and analysis, but some areas are unscaled regarding research. Object detection as an individual processing pipeline can detect the possibility of weapon objects during the scenario, and the activity detection pipeline corroborates the resonance of violent human activities. The previously mentioned approaches proved formidable for complex detection scenarios, but their advantages are ostensible when surveying the literature.

Nevertheless, object detection and activity recognition methods experience challenges as individual processing mechanisms with framework limitations concerning the risks of producing erroneous results. Some classification misrepresentations plagued object detection methods because of the model’s convolutional capabilities. That issue occurs when evaluating the propensity of microscopic objects (mainly knives, guns, and sharp objects), specifically when their optical flows accelerate with intensity across multiple video frames. Most activity detection models deploy multiple scaling techniques that enhance image dimensionality during convolution. This procedure exhausts the processing resources during the generalisation stages amidst the dense, fully connected layers and negatively impacts real-time performance.

Moreover, a distinct detection singularity emerged that evaded the focus of researchers as

they are yet to explore the performance impact of consolidating both practices as a single pipeline that accurately and efficiently discerns abnormal behavioural patterns circumventing violence.

To address the previously mentioned concerns, the thesis proposed the aggregated recognition of AI mediums as a reliable application that accurately and efficiently recognises violent activity utilising real-time CCTV sensors. The investigations targeted the first state-of-the-art mechanism, YOLO (you only look once, YOLOv5), as the object detection tool that discerns distinctive weapon features in addition to the violent activity classes. YOLO's detection limitations (concerning minute objects) improved by passing a combination of the regions of interest (violence) within the image scenery and weapon artefacts as input in a feed-forward operation for convolution, thus producing a classification output. Identifying and combining the objects this way through pre-processing and blob analysis enhances YOLO's accuracy ratings tremendously, thus enhancing its ability to classify a much larger object.

The second state-of-the-art mechanism entails deploying a 3DCNN (3-dimension convolution neural network) activity recognition technique with the aptitude to discern anthropometrics relative violence and abnormality in human behaviour. Excessive computational resources bear significance in processing data and can negatively impact real-time operations. Negating such impact issues increases the 3DCNN processing capabilities by applying layer modifications. From a parallel standpoint, this reduces image dimensionality challenges and stabilises the complexity of the class features amidst the dense layers during generalisation. The proposed joint operation integrates a fusion mechanism of the model's output in its final stages, producing precision accuracy results at its highest. It also reveals two possible methods of applying the YOLO mechanism. The first method constitutes sectioning weapon objects and separately feeding that as input together with the activity recognition blob features for classification. The other approach entails pre-processing both

weapon and activity, merging those regions of interest to act as one object entirely; this technique defies YOLO's limitation of detecting and classifying minute blobs.

1.1 Summary of the Research Motivation

Violence persistently impedes society, effectuating lethal outcomes predominantly when parties engage in criminal intentions, substances, and differing opinions in communication, domestic disputes, discrimination, and religion [1] and [2]. These scenarios assist in being the catalyst of severe injuries resulting in psychological distress by the death of loved ones involved [3] and [4]. Committing violence produces negative consequences; however, if technologically monitored, innovative CCTV sensors can be deployed to pre-empt the risks of attacks relative to those outbursts through automated surveillance. The idea reduces the impact to its minimum or nullifies the altercation completely. These malicious attacks reported in Figure 1.1 produced vast data volumes as UK citizens are ranked high for being

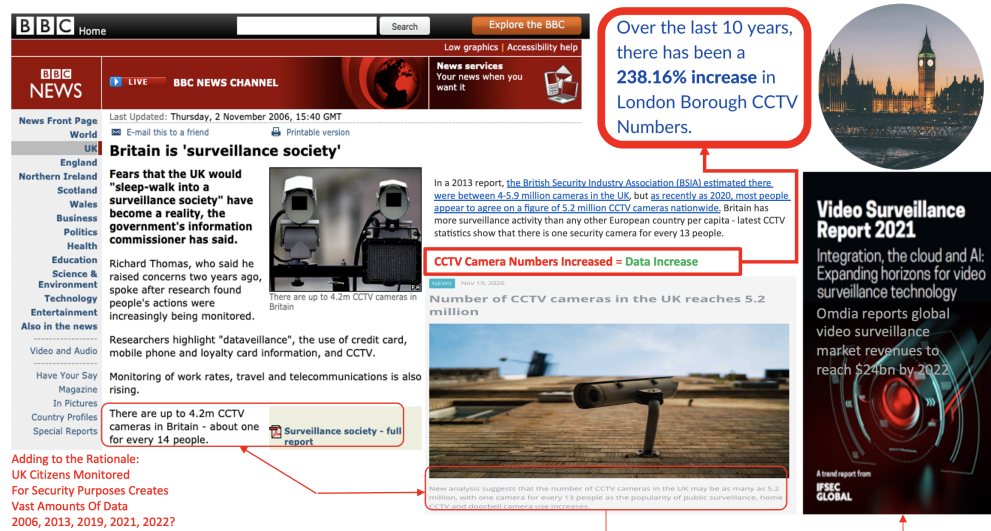


Figure 1.1: CCTV Monitoring Creates Data Volumes, Source: [3], [5], [6], [7], [8],[9]

constantly monitored. Another concern surrounded the deficiency in simultaneously analysing the magnitude of data that induces an insignificant number of convictions for involving parties [7], [10], [9], [11], and [5]. Authors [12],[13] and [14] presented statistics on violence in Figure 1.2 illustrating the number of violent offences fluctuating from zero to 25,000 over 13 years from March 2011 to Mar 2023. The motive behind this illustration accentuates the high number of violent instances occurring at 21,456 for assaults with injury and with intentions to enact bodily harm in dark blue, which substantiates the investigative notions. The evidence outlined the fluctuation of violence from 2011 to 2023. The data showed

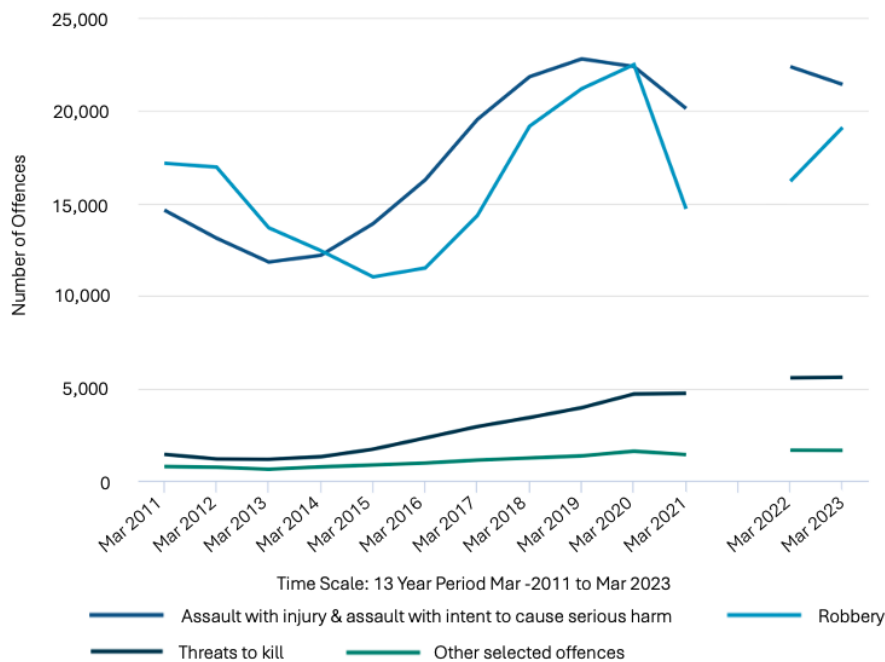


Figure 1.2: Knife or sharp instrument offences: [12],[13] and [14]

19,126 robberies in turquoise blue, differentiating in instances by 2,330 compared to assaults with an injury. The fluctuation in offences depicted threats to kill in navy blue at 5,599 instances, with other selected offences at 1640 as the lowest category. The other offences indicate violence by sexual assault or rape, attempted murder, and homicide. Police

recorded an increase of 3% in crimes involving a knife or sharp instrument (knife-enabled crime) at 50,833 offences in June 2023 compared to June 2022's 49,435 offences in England and Wales, excluding Greater Manchester. The gap in analysis between March 2021 and 2022 reflects a data adjustment made before the year ending March 2020 for police forces, who recently implemented a national data quality improvement service (NDQIS) tool, which calculated the statistics for England and Wales. The data from March 2011-2019 proved the seriousness of this escalating issue as the fluctuation of violent patterns illustrates an increase in the proportion of offenders as a direct link to the COVID-19 pandemic.

The significance concerning the statistical overview fortifies the rationale for the research proposal as the issue of violence is unmanageable, with a steady rise in fluctuations in offences for the period March 2021 [14], 2023 [12] and 2024 [13]. The rationale for the proposal accentuates the severity of violence relative to human demographics per category and the weapons used in Appendix 1.1. The rationale behind this aspect of the statistical



Figure 1.3: Lethal Events Source [15], [16], [17], [18], [19], [20], [21], [22] 01-10-2022.

analysis proved that violence is not partial to one specific group, as the demographic in-

fluence relative to homicides showed no boundaries from the ethnic sense. However, it signified the rising impact of the number of occurrences. The authors [15], [19], [20], and [16] fortified the last analysis in Figure 1.3 with heinous reports afflicting human lives (from ordinary folks to parliamentarians) during that specific year only, yet violence perpetuates. The statistical data relating to the categories of violence indicated that an effective solution is urgently needed to mitigate such instances. As the need for a solution became apparent, concatenating the evidence and analysis of the literature proved crucial. The investigations revealed complex techniques utilising hybrid feature extraction and generalisation procedures. Although those hybrid mediums present a measure of significance in applicability, the challenge of pre-empting violence persists. The previously mentioned approaches cannot accurately discern violence to distinguish its generic state (whether the activities are violent or not with its sub-class) as output. Additionally, other problem factors relative to heavy computing resource dependency during processing impact the trade-off between real-time performance and accuracy, and this tenacious challenge can potentially affect their approach negatively.

As analysis of past violent trends intensified, the disclosure of unscaled domains with limited knowledge of a proper solution emerged. Such discoveries included artificial intelligence mediums that can significantly reduce the impact of current security challenges, thus propelling technological advancements in this domain. Exposing object detection and activity recognition as individual processing pipelines disclosed the potential to detect the pre-existence of weapon objects during violent scenarios. In this regard, the activity detection pipeline corroborates the resonance of abnormal human activities relative to anthropometrics, and object detection focuses on lethal weapons. To fully understand their (the AI models) capabilities, thorough investigations into existing solutions proved crucial to gain insight into their processing performance, architecture, and critical limitations before proposing the fusion approach. The following research questions highlight the objectives/ intended operations and their stages to commence the task of activity recognition.

The questions facilitate the investigation of the architectures and demonstrate a profound analysis of the model's superiority via simulations. Subsequently, a presentation of the proposed technique aids in solidifying its superiority for real-time processing and robust results.

1.2 Summary of the Research Questions

The research objectives encompassed a main section as a link to the proposed solution and sub-questions requiring a measure of investigation to disclose the feasibility of the operations. The guidelines are as follows,

1.2.1 Summary of the Main Research Question:

Research Question-1: Can violent activity and weapons (bladed instruments, knives, guns) be recognised in CCTV videos?

1.2.2 Summary of the Sub-Research Questions:

Research Question-2: What is the impact of deploying pre-processed (modified data) data and no pre-processing (data without modification)?

Research Question-3: What is the model's processing impact based on performance if the training data or sample size increases?

Research Question-4: Can multi-class activity as neutral (non-violent) or violent be generalised?

Research Question-5: Is object detection superior to activity recognition and vice versa in predicting violence?

Research Question-6: How can object detection enhance activity recognition's output

through fusion?

1.3 Overview of the Aim and Objectives

The research aims to develop an innovative violent activity detection mechanism for real-time applications that can accurately and efficiently recognise violence in CCTV data. A series of objectives facilitated the realisation of the overall aim below.

1. Conduct an extensive investigation into the state-of-the-art YOLOv5 object detection and 3DCNN activity recognition and authenticate the mechanisms for effectiveness.
2. Acquire an object detection dataset and construct the first R-CNN (region base convolution network) object detection model to evaluate the possibility of weapons- (bladed instruments, guns) detection and violent activity recognition in CCTV videos with high-performance capabilities.
3. Evaluate the applicability of the state-of-the-art YOLO and 3DCNN for performance efficiency
 - (a) Evaluate the impact possibilities of deploying pre-processed data modified to enhance feature selections (augmentation, contrast alterations, cropping, zooming, shearing and rotations of image features of interest).
 - (b) Evaluate the impact possibilities of deploying raw data containing no pre-processing, resolution, or illumination enhancements.
 - (c) Evaluate the impact of increasing the training data sample sizes.
 - (d) Evaluate the impact of deploying (from-scratch) random hyper-parameter options, bias, and weights.

- (e) Evaluate the impact of applying pre-training weights and biases.
4. Evaluate model superiority by generalising multi-class activity as neutral (non-violent) or violent.
 - (a) Evaluate whether object detection is superior to activity recognition and vice versa.
 5. Present the proposed fusion technique utilising object detection with the activity recognition framework relative to previous simulations and discoveries.
 - (a) Evaluate the effectiveness of the proposed fusion operations through simulations, observations, and analysis.
 - (b) Evaluate the effectiveness of the operations before fusion, after fusion, and after fusion supported by YOLO artefact detection enhancement.
 - (c) Hypothesise the results and formulate conclusions.

1.4 The Contributions of the Thesis

The following summary emphasises contributory elements towards validating the effectiveness of the proposed fusion operations.

1. **Chapter-3 to 4** Conducted performance testing of two known machine learning techniques (YOLOv5m and 3DCNNsl) in independently recognising violent and non-violent activities in CCTV video footage.
2. **Chapter-5** Demonstrated violent activity recognition performance in such videos

when both machine learning techniques operate in tandem through a decision-level fusion operation.

3. **Chapter-6** Implemented performance enhancement by further incorporating threat object detection in the previous combined solution. The idea applied weight value embedding to suggest the presence of lethal weapon objects/artefacts, thus enhancing the outcome.

1.5 Overview of the Thesis Structure

The thesis structure summarises each chapter, further informing the reader on a brief reflection of the subject matter. The idea also manages the topics with clear operational directives at every stage. The structure is as follows:

Chapter-1 Introduction: The chapter introduces the issue of violence and its impact on society, providing the motivation, aim and objectives for the research efforts.

Chapter-2 Context & Fundamental Knowledge: Demonstrates essential knowledge required to comprehend the technical operations, followed by an introduction into the phases of the literature review concepts.

Chapter-3 Research Methodology: The section justified the quantitative approach through experiments using volumes of violent data collected from benchmark repositories and social media forums. The chapter also considers the experimental protocol and evaluation strategy to measure performance.

Chapter-4 Evaluating 3DCNNsl & YOLOv5m for Activity Recognition: The section presented evaluations for YOLOv5m as activity recognition and 3DCNNsl operations derived from the experimental setup with a focus on stabbing as the violent

class. The chapter also fortified the rationale for proposing fusion by presenting processing challenges affecting both models and discussions on operational procedures limiting the impact of such challenges.

Chapter-5 Fusing YOLOv5m & 3DCNN for Reliable Activity Recognition: The section introduced concepts of the proposed fusion by combining the operations of YOLOv5m as activity recognition and 3DCNNsl and applying additional configurations to foster performance efficiency. At this level, an evaluation of the fusion results demonstrated model superiority with operational discussions disclosing weapon artefacts as additional elements detected in the violent scenes. The idea of artefacts is applied via embedding to facilitate performance enhancements.

Chapter-6 Utilising Object Recognition for Improving Activity Recognition: The section emphasises YOLOv5m object detection via weapon artefact enhancement. The concept combined weapons with the target activity objects that suggest the preempting of violence. The idea further enhances the classification status and accuracy scores. The approach evaluated violence from a frame-by-frame approach to consider scenarios limited by the availability of test data. The chapter also investigated utilising the entire video in scenarios that reflect an abundance of test data with a discussion on model superiority.

Chapter-7 Overall Conclusions: The chapter provided overall conclusions to validate the fulfilment of the objectives with clear indications of performance contributions towards achieving the aim.

Chapter 2

Context and Fundamental Knowledge

Chapter 2 provides essential information regarding object detection and activity recognition concepts. At this stage, the primary aim is to highlight key elements that facilitate the processing of human activity towards anticipating such actions as violent activity. The chapter also describes the data pre-processing concepts, computational library packages, and classification networks with various mathematical computing procedures and graphical representations required to validate the overall operations. The concept of artificial intelligence considers operations of machine learning and deep learning systems to achieve the tasks of activity recognition and object detection. Machine learning, described by [23], is a division of artificial intelligence applied to generalise a specific task utilising its cognitive sense derived from experience without requiring direct programming. However, Deep learning in [24] is a machine learning division employing artificial neural networks that comprehend the complexity of data patterns through computer vision. Activity recognition and object detection formulate essential domains to achieve substantial results in several tasks. Artificial intelligence follows cognitive tasks such as supervisor learning,

unsupervised learning, and reinforcement learning. Supervised learning in [25] derived from machine learning encompasses training an algorithm that uses manually annotated labelled datasets to suggest correct classes. Unlike supervised learning in [25], where data is labelled, Unsupervised learning in [26] operates from an unlabeled data perspective that identifies features of interest and relations of objects within the data without the need for guidance. The idea facilitates scenarios when data proves a challenge to acquire.

Contrarily, reinforcement learning discussed in [27] incorporates agent learning to generate decisions based on environmental interactions. The idea integrates a general reward, which transfers into learning strategies towards recognising actions. Indicating the objects of interest encompasses learning processes suggested in [28], which utilises a classification task to identify the type of action performed in a video or image sequence. Activity recognition facilitates the enactment of educating systems to identify human action through classification in a video or sequence of actions within image data [29]. Object detection discussed in [30] employed through classification networks extracts spatial features relative to a desired object from images. The idea incorporates accumulating enormous volumes of data identified by specific category labelling for training. The data endures a feature extraction process through techniques such as blob analysis in [31], which educates the artificial intelligence on the contours and distinct characteristics of an object of interest. The operations endure training demonstrating relationship attributes among features, activities, and objects of interest. Subsequently, during the inference stages, the artificial intelligent model is redirected to data that it had not seen to project its predictions based on the actions or activities it recognises similar to [29] and [30]. The applicability of such systems facilitates the detection of unwanted, abnormal or gesturing actions, pedestrian detection and traffic management through surveillance sensor devices. With insight into artificial intelligence, chapter two expounds further on necessary operations to achieve high performance.

Chapter two incorporates ten sections to present an overview of the motive, concepts and

mechanisms applied during development. The structure at this stage is as follows. Section 2.1 emphasised fundamental concepts concerning Object Detection (OD), Activity Recognition (AR) data pre-processing, and blob analysis. Section 2.2 details the importance of data validation split procedures. Section 2.3 encompasses deploying software applications and library packages that facilitate complex processing operations. Section 2.4 highlights various classification neural network concepts and their importance. Section 2.5 presents a background of YOLOv5 object detection (you only look once) operations. Section 2.6 provides a background of 3DCNN activity recognition processing components. Section 2.7 presents an overview of measures to evaluate classification operations. Section 2.8 accentuates graphical representations depicting the classification's true processing capabilities. Section 2.9 presents an investigative analysis disclosing the applicability of other available techniques compared to the state-of-the-art for object detection and activity recognition. Finally, Section 2.10 presents a conclusion of the entire chapter.

To fully understand the context of the research operations, one must first appreciate artificial intelligence as a construct of computer vision that trains a computer to decipher the visual interpretations of the world [32]. The concept considers the similarity of displaying many images towards educating toddlers to recognise objects. This notion is somewhat the same when training the machine; however, the computer's cognitive sense discussed in [33] involves the computational processing of numbers. Image representations projected as 2-dimensional arrays of numbers called pixels allow computer processing to categorise specific number values to perceive such objects as items of interest [34]. By utilising millions of digital images generated from IoT (internet of things) devices to train a computer, the possibility of accurately classifying objects through classification algorithms is astounding. Some concepts require a measure of understanding to achieve the full complement of the operation and encourage performance efficiency.

2.1 Overview of Data Pre-processing and Blob Analysis

Data pre-processing describes the stage of preparing (cleaning and formatting) the raw data to meet the required standards given the specifications of the artificial intelligence architecture. The operations entail acquiring a prodigious amount of data and conforming such video files to a respective size with substantial representations of the relevant objects of interest (violent activity, weapons, knives occurring across the temporal plane) in the image scenery. The data acquisition generated raw files online from various social media platforms with variations in image dimensions and other anomalies that adversely affect the artificial intelligence processing state. Considering pre-processing, the issue reflects better representations of the objects in the image scenery, simultaneously reducing the video file duration and negating object redundancy (removing objects with similar characteristics in multiple images). The investigation disclosed [35], an open-source tool that conforms the previously mentioned issues to the required data specification and dimensionality. Blob analysis outlined by [36] is the term that describes the technique of locating and annotating the regions of interest (the object) relative to its coordinates reflecting violence in the image scenery. The technique separated object/s of interest from other unimportant elements in an image by creating a demarcated bounding box around its boundaries. Other data generation processing techniques consist of augmentation enhancements. This operation generates variations of the clean data by shearing (slanting the image to the left of the right), rotation (rotating the image to an aspect ratio of 0.001%, 45%), flipping (clockwise/anti-clockwise, left, right), contrast alterations (Grayscale/ images without colour), illumination (brightening/ darkening), distortion (introducing salt & pepper noise) and blurring. Blob analysis techniques assist the object detection algorithm by generating bounding boxes around the regions of interest. The approach demonstrated in [37] and [38] utilises the outline of multiple objects in the image to infer that its label is present during classification.

2.1.1 Summary of the Class of Activity Template (CoAT):

To understand the motive for blob analysis, one must grasp the characteristics of importance relative to violent actions and their range of erratic patterns. A class of activity template (CoAT) described by [37] emphasises anthropometrics that distinguishes its action significance from other features in an image. Its purpose is to identify abnormal behaviour indicators relative to the specificity of objects, their motion, their acceleration, and their trajectory [39], [40] and [41]. The idea identifies actions symbolising abnormal human-to-human violent altercations to cause severe injuries [40]. These actions (beating, stabbing, fighting, and shooting) reflect the human gait with or without weapon objects depending on the class category relative to 1-human versus 1-human, 1-human vs many and many individuals vs 1-human. The class of activity template decomposes the context of violent actions to present a holistic perspective of the features of significance extracted during the blob analysis phase. Its attributes emphasised by [42] denote a stabbing action as slicing, penetrating, wounding, or causing any grievous bodily harm with sharp or bladed objects. According to [43], Humans that physically strike others with or without an object to cause physical or psychological harm conveyed beating actions. Author [44] described a shooting action as the act of pointing or discharging lethal barrelled weapon/s of any description (handgun, shotgun or taser) from which any shot, bullet or other projectile missiles are dispensed with high speed to endanger life intentionally. Authors [45], [46], and [47] projected non-violent human actions as any human action accepted and embraced by society relative to its governing laws. It is necessary to highlight the target categories relative to violent and non-violent actions to demonstrate the applicability of datasets during the developmental stages. Although non-violent actions are not the focal point of this research, their definition and significance show the model actions that are not violent to enhance the classification’s performance during inference. Non-violent blob features of this class will assist the model in generalising and distinguishing violent from non-violent features during processing.

2.1.2 Blob Analysis Applied on the Class of Activity Template (CoAT):

We sectioned the class-of-activity-template (CoAT) to demonstrate how colliding trajectories and motion features of the object within the images can be tracked through blob analysis. The features representing the pre-violent action leading up to the start, middle and end of a potential attack can be measured and classified accordingly across (temporal plane) several image frames [47], [48], [49], [50], and [51]. The processing approach disclosed in [52], [38], [53] is paramount as it reduces the proficiency of the entire operation if blob selections neglect standards and consistency during training. The concept circumvents transfer learning initiatives (modifying a relevant model already trained on another task to process the current activity recognition task for prediction) that facilitate high-accuracy results and reduce time spent during development analysis. The algorithm must

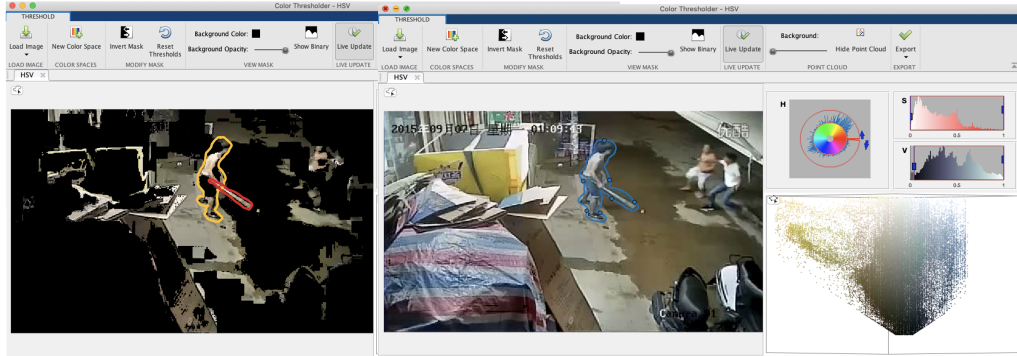


Figure 2.1: Blob Analysis Operations via MATLAB.

understand the object/s of interest size variation and its coordinates. The blob analysis technique assists in identifying and annotating the class of activity template within the images to present the object's specification to the processing architecture during training [54], and [55]. The approach in Figure 2.1 discerns pictorial values through colour space sectionation, sorting the foreground and background as pixel properties as an illustration of blob analysis concerning [56], [57], and [58]. Refining the class's spatiotemporal regions re-

duces image noise (unwanted details in image) by applying colour space thresholding and morphological dilation (object expansion, erosion, reduction) processing techniques [59], [60], [61], [62], and [63]. The concept of blob analysis cannot extract the spatiotemporal activity details of an object’s accelerated trajectory and boundaries in static images. However, [36], [64], and [65] confirmed the possibility across the temporal plane when processing the optical flow of actions in several frames.

2.2 Summary of Cross-Validation Split

Following the data augmentation, which generated more data and pre-processing (data cleaning) concepts, applying the cross-validation methods validates accuracy performance after every epoch (data traversing forward and backward as one iteration). This process guides the operations towards generating the highest output value and systematically highlights whether the training is on target, ensuring that the model learns effectively and accurately. The cross-validation hold-out technique similar to [66] and [67] splits the given dataset into smaller subsections and utilises a combination of the subsections to evaluate the model’s overall performance. The data sectioning ratio in Figure 2.2 relative to [68] and [69] consists of an 80:20 splits reflecting 60% training, 20% validation, and 20% for the testing stages. Alternative splitting concepts conveying 70:30 [70], 55:45 [71], 60:40, and 50:50 [72] exists to allow further analysis. Those previous splitting concepts negatively impact the training if the model is not trained sufficiently on significant ratios of data. The literature approach of 80:20 in [73] provides sufficient data to allow the model to generalise the construct of violence over time effectively. The technique prevents the model from over-fitting, where it excels at classifying the samples in the training set but experiences generalisation challenges when classifying objects on unused data [74]. The 60% training data is applied to train the model to generalise hidden features/patterns relative to the object in the data. In each epoch (iterations of the data in a forward and

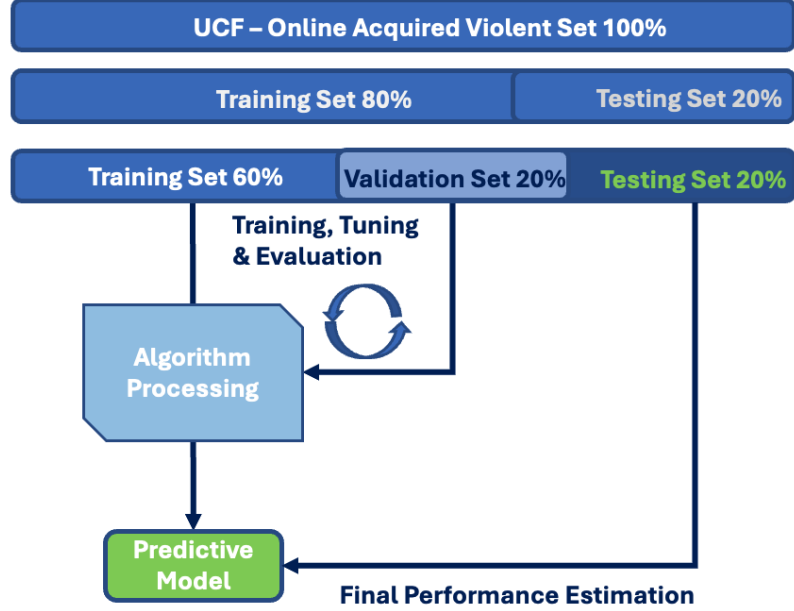


Figure 2.2: Cross-Validation Dataset Split Procedures.

a backward pass during training), the same training data is fed to the neural network repeatedly in batches, and the model continues the reiteration of the learning cycle of relevant features of the objects in the dataset. The 20% section of validation data is separate from the training as it is applied to fine-tune hyper-parameters (specific option values applied to fine-tune the model’s accuracy) and configurations to assist the model’s performance.

2.3 Overview of Software and Library Packages

Software tools such as MATLAB and Python facilitate the computational dispensing of the results as output. MATLAB’s sophisticated programming functions in [75] assist with visualising data anomalies requiring attention through scripting. Python allows intuitive programmable syntax control and advanced functionality, thus simplifying processing outputs. Its libraries (various pre-compiled codes) provided the framework for initialising

artificial intelligence models utilising well-defined operations without complex deployment. Python's library package contains over 200 modules consisting of scripting documentation, memo templates, configuration data files, and class definitions to aid AI with its computations [76]. These packages ranged from TensorFlow, Matplotlib, Pandas, NumPy, SciPy, Scikit-Learn and PyTorch. Most library packages are applied to reduce the iterative task of writing and re-writing script files to executive a combination of commands. The processing packages search for a particular library and execute its pre-scripted functions to interpret the input-output commands of the classification operations accordingly.

2.4 Overview of Classification Network Operations

As a mathematical approach, classification detailed in Figure 2.3 amalgamates two functions to produce a third function, which occurs at the input stage using a kernel to generate a feature maps output [33] and [77]. A classification neural network has many layers that

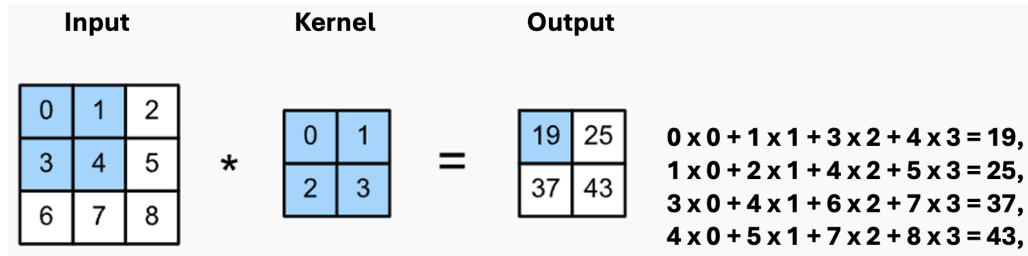


Figure 2.3: Kernel Processing Operations.

generalises different image features as the processing deepens from an input-output perspective [50]. Processing Filters (kernels) compute the values of each training image at specified resolutions. Author [78] disclosed how all output results of each operation act as the input for the next layer. These layered kernels commence by processing simple features relative to brightness and edges. Subsequently, it increases its processing capabilities to discern more complex features unique to the nature of the object [79]. Its processing com-

mences at the top left corner of the image, traversing from left to right on a 1-pixel column per interval until the filter edge reaches the edge/end of the image [80]. A single kernel is defined as 2 pixels high and 2 pixels wide (rows, columns) and can vary according to the system's operations [81]. According to [82], differently sized filters will detect differently sized features in the input image, resulting in differently sized feature maps. A common practice applies a 3×3 , 5×5 or even 7×7 sized kernel for larger input images [83].

2.4.1 Overview of Kernel Stride and Padding Operations:

The neural network's stride parameter is responsible for modifying the movement across the width of the input image. The model's processing reflects whole integers instead of fractions to alleviate processing issues when generating its output [84]. Therefore, if the stride parameter reflects 1 with no padding, the 2×2 kernel's operations on an array input of 6×6 will move to one pixel at a time, resulting in 5 positions, producing a 5×5 output identical to Figure 2.4's illustration. Stride value modification promotes processing

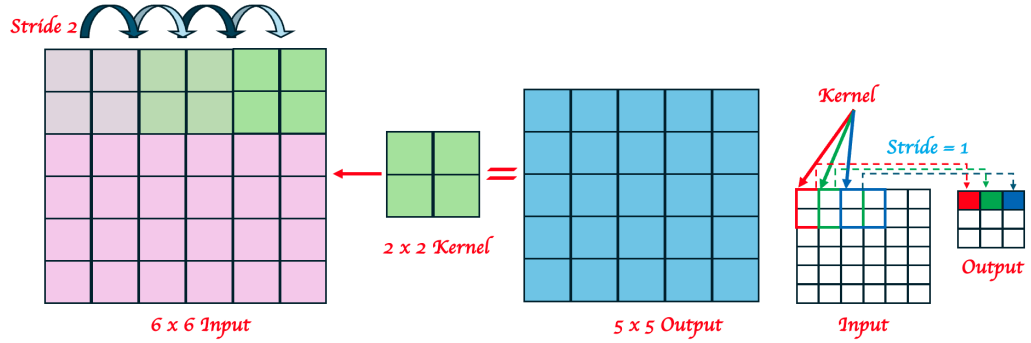


Figure 2.4: (a) Kernel Processing Operations.

efficiency depending on the input image size and the required task at hand. Its processing reduces the generalisation complexity and the computational memory required to generate an output, thus creating a smoother convergence due to a smaller volume of array values [85]. If the stride reflects a processing value of two as described in Figure 2.5 with a $2 \times$

2 kernel on a 6 x 6 input array with padding, the output will result in a 4 x 4 output. A reflection of padding in [86] allows the creation of deeper neural networks by controlling image border details. Padding extends the image borders by adding pixel details during classification, and this presents the kernel with more area to optimise the analysis of the images [87] and [83]. Therefore, setting padding to zero means that each pixel value added will reflect zero. In contrast to the padding setting 1, a single-pixel border of zeros is applied to the image.

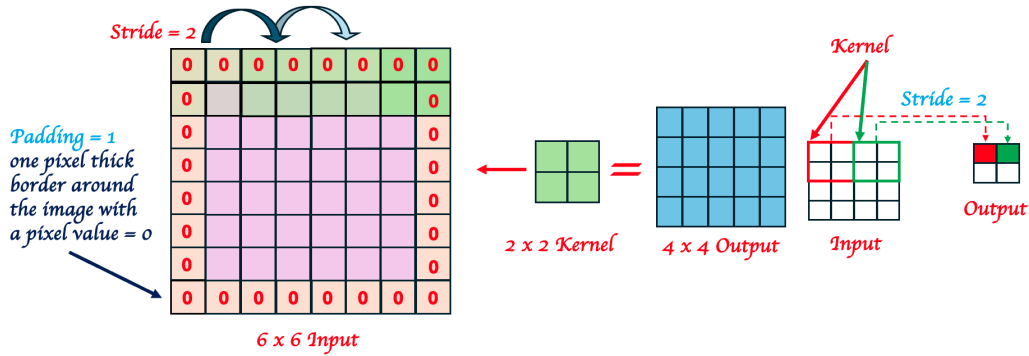


Figure 2.5: (b) Padding with Stride Parameter Set as 1/2.

2.4.2 Summary of RGB Images Evaluation:

Completing the classification processing requires the inter-operation of several specialised hidden layers within a hierarchy. Therefore, the first layers detect lines, curves, and edges. As the operations reach more profound processing components within the hierarchical layers, high-level complex shapes and colours relative to an object or objects are generated (body parts, knife sections). A digital image incorporates 3-distinct elements: its width, height relative to its pixelated values, and its colours in 3 channels: red, green, and blue [88]. Because the input image represents three distinct colour channels, the operations apply three filters to compute their values simultaneously. The concept emphasised in [89] enables the convolving 3x3x3 kernel demonstrated in Figure 2.6 to process images with a

depth reflecting the output of lines and edges from the RGB colour channels.

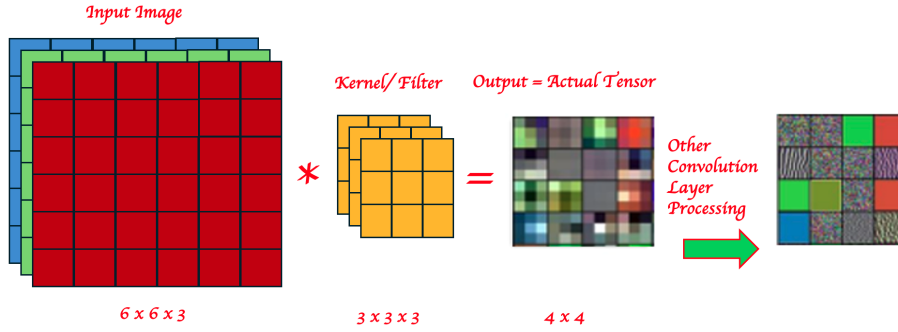


Figure 2.6: Classifying an RGB Image.

2.4.3 Summary of RELU Layer Operations:

Another classification aspect entails the rectified linear activation function in Figure 2.7 as a piece-wise linear operation that outputs an input directly in red if it is positive. Otherwise, it dispenses an output of zero as per the blue arrow. During classification, the rectifier

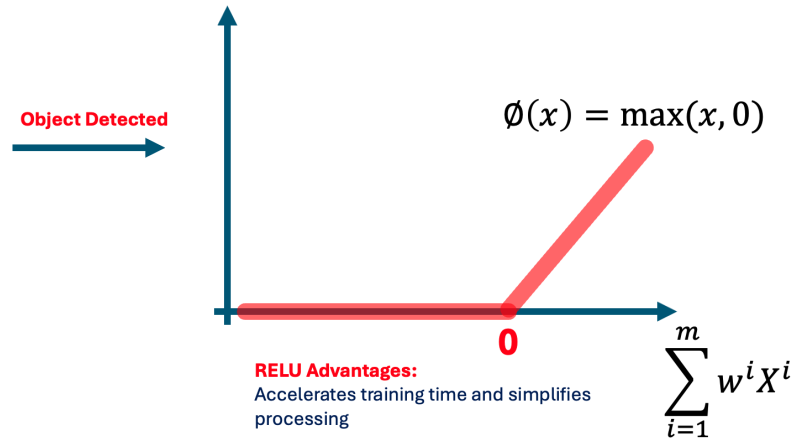


Figure 2.7: RELU (Rectified Linear Activation Function) Operations.

function increases non-linearity in the input data to remove all dark-shaded elements within

the image. In [90]’s demonstration, the processing retains only those details carrying a positive value; this reflects a grey-coloured tensor as its output. Some ubiquitous techniques within the literature apply the sigmoid activation function, the logistic function, or the hyperbolic tangent to improve their performance based on the framework and the research objectives.

2.4.4 Summary of Pooling Layer Operations:

Classification emphasised in [91] generates high computations between its layers, which increases the complexity of the processing. If the latter is unregulated, as per [92], it intensifies the tensor dimensions substantially and negatively decreases the processing speed and capability of the entire operation. The processing solves the issue by exercising a down-sampling max pooling operation in [93], which reduces the input feature map and the volume of computing parameters required for its processing. Pooling has no computing parameters defined in [94]; this operates by sliding a kernel algorithm (window) over its input and selects only the max value as the output. The operation accentuated in [95] creates shorter training times and regulates over-fitting if the function is aligned too close to a limited set of data points. As pooling works independently on each depth slice of the feature maps generated from the RGB input, it simultaneously increases the network’s operational efficacy. Other techniques can suffice relative to the application and framework

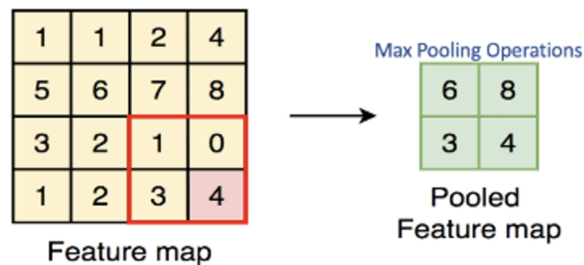


Figure 2.8: Max Pooling Operations.

deployment. However, figure 2.8 max pooling operations depicts a ubiquitous approach that computes its operations with a 2x2 window, a stride of 2 and no padding.

2.4.5 Overview of 1st Object Detection Processing via RCNN:

RCNN demonstrates the concept of object detection as the first processing medium relative to the previously mentioned classification sections. This section briefly discusses its object detection framework and attributes utilising ubiquitous static images online. Figure 2.9 and 2.10 disclose fundamental concepts of RCNN's processing in [96] regarding a two-

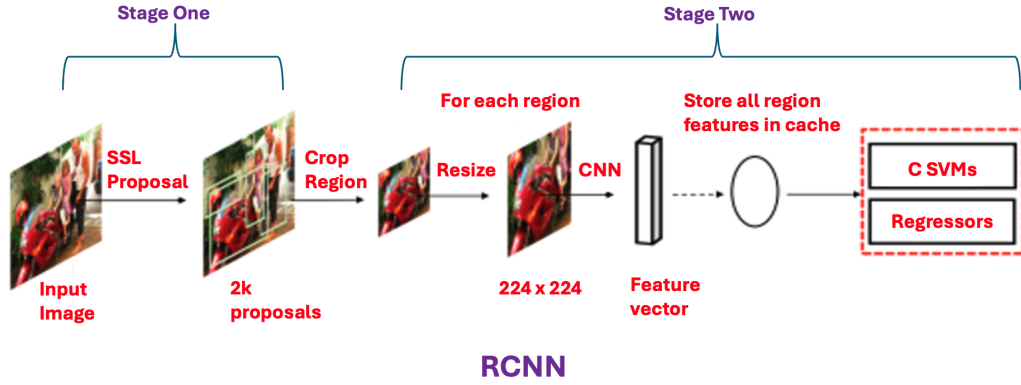


Figure 2.9: R-CNN 2-Stage Processing Operations Source: [97].

stage object detection technique. Its framework illustrated in [78] emphasises the proposal generation region operations, feature extraction processing and classification operations. The RCNN object detection processing in [98] utilises a fixed selective search approach to generate region proposals using a greedy search algorithm, which recursively aggregates feature candidate similarities into large ones and produces a result. The RCNN algorithm extracts 2000 proposal regions from the image at the input image at 1 in Figure 2.10. The processing highlighted in [97] accepts the features from step-1/2 for feature extraction and classification processing. RCNN computes relevant features at step-3 generating a 4096-feature vector of object estimations for classification by a support vector machine utilising

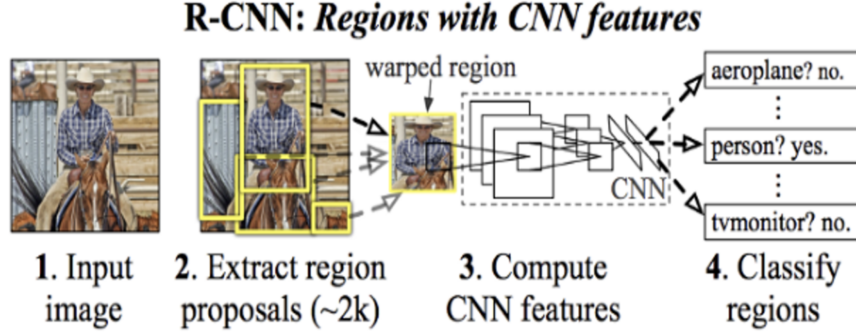


Figure 2.10: R-CNN Processing Operations Source: [97].

regression at step-4 similar to [99]. RCNN in Figure 2.11 demonstrates its ability to predict additional offset proposal values that enhance the operations accuracy output by precisely adjusting the bounding boxes around the coordinates of the regions of interest in [97]. The application of non-maximum suppression is highlighted at this stage as it plays a significant role during classification. This value quantifies the degree of overlapping between the boun-

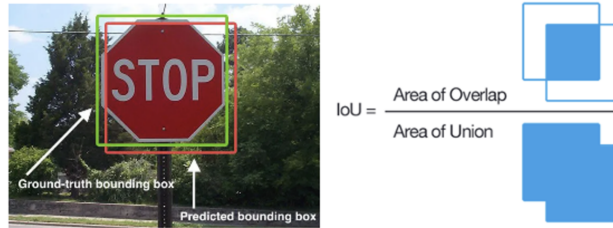


Figure 2.11: R-CNN Ground-truth Overlapping and Intersection Over Union

-ding boxes of the ground truth (hand-labelled bounding boxes of the test dataset specifying object coordinates in the image) of the class data and the predicted class regions made by the model [100]. The intersection over union overlapping operations suppresses insignificant estimations with its computations to evaluate its accuracy on the dataset. Authors [101] and [102] defined the process as a single score operation for each classified object.

R-CNN's pipeline in Figure 2.12 discloses its object detection capabilities and emphasises its constraints when processing objects within static images. Linked to Figure 2.12's architecture, the model is affected by real-time constraints as it computes 2000 proposal regions with a convergence of the results, which generates an output between 40 to 50 seconds per image [103]. The previously mentioned limitation intensifies the model training operations demonstrated in [98] and [104] by applying a fixed selective search algorithm that restricts further feature learning and frequently encourages impractical feature candidate proposals with large feature map generations. With an understanding of the fundamental concept of classification, Figure 2.13 presents an overview of YOLOv5's non-maximum suppression predictive ability with bounding box encapsulation and confidence scores that suggest poor, good, and excellent processing conditions. The green square demonstrates the significance

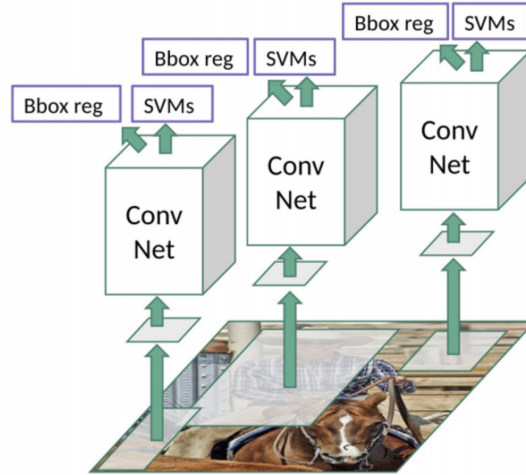


Figure 2.12: R-CNN Dispenses Added Offset Value for Prediction Support Source: [103].

of the classification accuracy compared to the ground truth data in red regarding poor, good and excellent performance. The background of the 3DCNN model provides insight into its processing as the state-of-the-art activity recognition from the class label prediction process. The operation undertakes identical input-output label classification operations

employing a 3-dimensional input for its processing layers.



Figure 2.13: YOLO's Intersection Over Union Processing Output.

2.5 Background of YOLOv5

Discussions on YOLOv5 architecture in this section highlight its processing operations as a formidable approach for object detection. Although multiple techniques exist for object detection (mask [105], vehicle [106] and head [107] detection), the YOLOv5 framework strongly applies to the proposed approach. The aim is to present the classification concept in the possibility of objection detection and activity recognition before the literature review. The idea fortifies the reader's knowledge of classification mechanisms as the proposed strategy for activity recognition before presenting investigations that led to the model selections. To commence, demonstrating the YOLOv5 individual activity recognition method provided systematic processing for the proposed model.

2.5.1 Overview of YOLOv5 Activity Recognition Perspectives:

YOLOv5 operations in [108] encompass 3-primary classification layers relative to its backbone processing network centred on CSPDarknet53. The approach utilised another 15-processing layer comprising the neck and 3-defining the classification detection comprising the head in [109]. Classification networks like CSPDarknet53 in [110] separate the base layer activations into two sections, merging them through a cross-stage hierarchy to en-

hance further its generalisation (learning violent action features) process. Further details via Appendix 2 concerning YOLOv5 input-output processing accentuate its classification operations. The essence of YOLOv5 demonstrated in Figure 2.14 accepts input and processes it, utilising multiple layers to derive an output result. CSPL (cross-stage partial connections) bottleneck layers are applied to regulate parameters that affect computational processing loads, thus smoothing the transitional flow of data directly by skipping specific layers. Author's [111] outlined how a 1x1 convolution layer achieves this by decreasing erroneous error details within the data to increase the performance of YOLOv5 additional layers. A spatial pyramid pooling (SPP) layer at stage two accepts this data as the input and normalises the regions of interest that contain size and scale variations. Its processing sections the input data into several feature map grids. According to [112], it independently computes a max pooling operation to each grid to retain varying scales and gradient intensities of the object/s features. SSP constitutes three individual pyramid levels at differing scales with specific pooling that increases the feature processing for minute object details. YOLOv5 feature fusing neck applies a path aggregation network operation (PANet) like [113], which combines the ROI attributes of the data in the adjacent layers to facilitate higher prediction results. Another CSPL convolution operation at stage four normalises those activation tensor block values. Following this is a 1x1 filter at stage five that generates activations containing similar spatial attributes with distinctive channels. Those operations improve the feature processing performance by minimising dimensionality issues and adding non-linearity to the tensor output with a RELU non-linear activation function. The model introduces up-sampling at stage seven to maximise the spatial resolution of activations by merging low-quality layer feature maps with others of higher quality in adjacent layers. Author's [114] suggested that its application improved the model's prediction accuracy on objects with multiple sizes and scale ratios. A concatenation technique at stage eight described by [115] applied another refining process, further improving the clarity of the object of interest by combining appropriate gradient intensity details of the

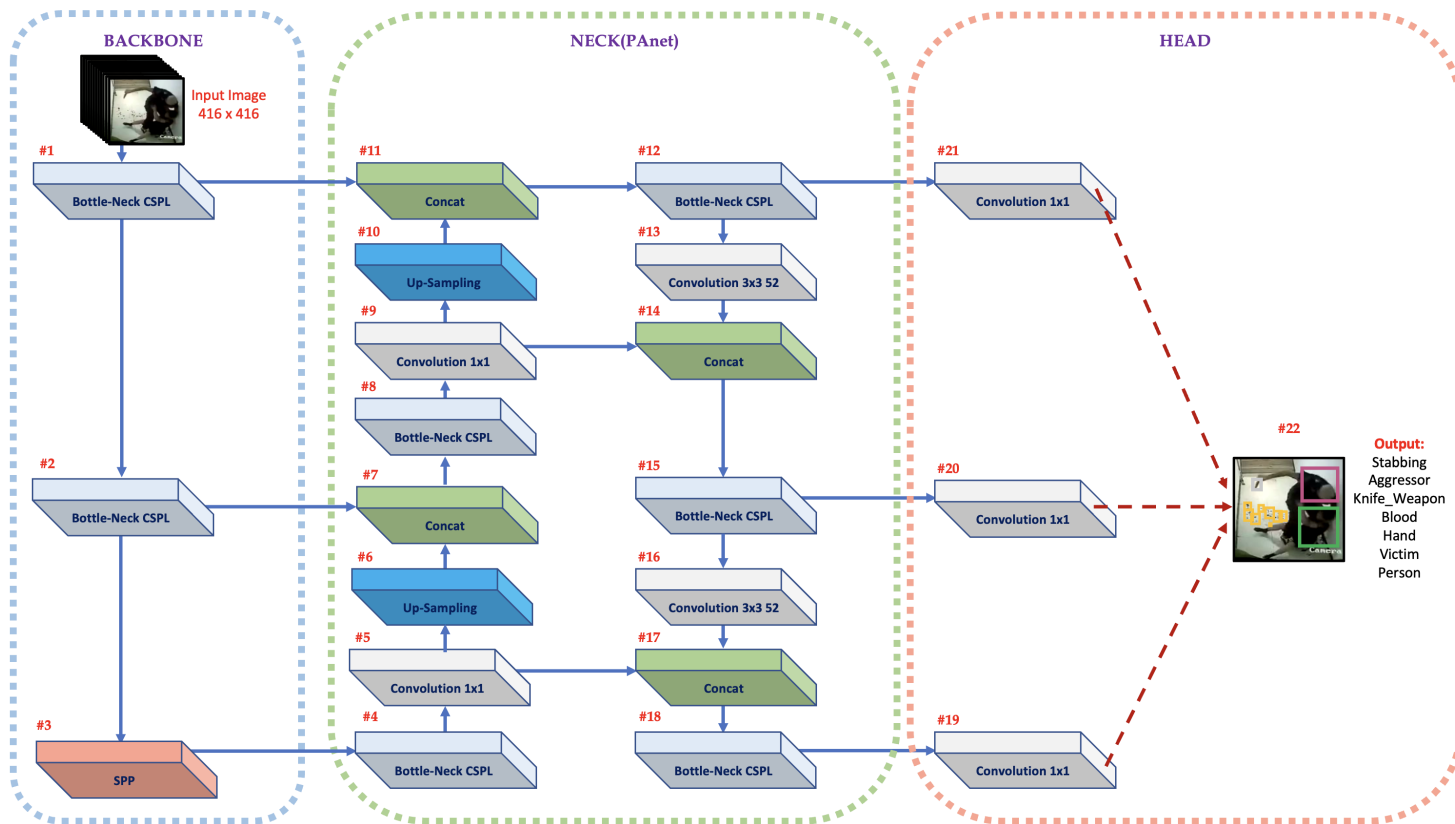


Figure 2.14: YOLOv5 Object Detection Processing Source: [116].

tensors. The data traverse additional classification, up-sampling, concatenation, and bottlenecking from stages nine to eighteen to enhance the feature fusion operations. Within the classification head, the function computes a final object detection prediction output based on the feature representations generated from the neck’s operations. Discussed in [108], it applied a feature extraction layer which processed the objects, scales and sizes. An additional function layer within the head combines those extracted features to produce refined tensors of the objects of interest. At stages nineteen to twenty-one, these final layers applied a grid cell over each image with a series of bounding box scores and coordinates for each cell. As discussed in [106] and [110], the function determined the prediction’s probability relative to the object of interest at stage twenty-two.

2.6 Background of 3DCNN

It is paramount for the reader to understand the core aspects of activity recognition and its 3D kernel towards 3-dimensional processing. The 3DCNN model is applied to predominantly detect medical imaging and complex human actions with high accuracy [117]. The technique entails processing blob features demonstrated in images via [118] and [119] in a cube across a sequence of spatiotemporal frames utilising dense classification layers. The 3DCNN kernel discussed in [120] slides in 3 dimensions and dispenses a feature map of width (how broad the region of interest is), height (the height of such regions), and depth by several channels (visual colour spectrum as red, blue, and green). The depth identified in [104] and [121] plays a significant role in the 3-dimensional structure towards regulating the growth of the feature map block. The significance of the growth feature mentioned in [122] and [123] negatively impacts the operation’s real-time processing if improperly configured. Its depth mentioned in [118] has a specific configuration value applied to designate a selective number of image frames for classification. This operation computes feature representations of each primary colour channel to produce a final output. The critical concept accentuates its importance, as it requires a degree of experience attained through multi-

ple experiments to produce a balanced hyperparameter value. The depth regularisation highlighted in [124], [125] and [118] reduces the impact of the computational load required during the generalisation stages, which increases the real-time processing simultaneously. Figure 2.15 classification provides context into 2D(a) and 3D(b)-dimensional data input to bridge the reader's understanding of the idea.

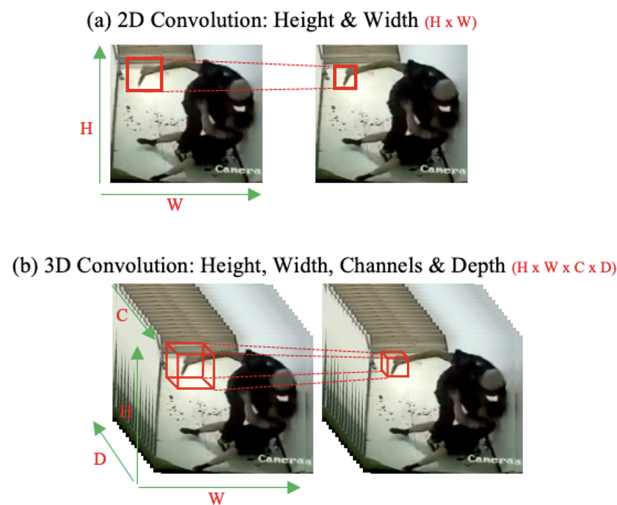


Figure 2.15: 3-Dimensional Input Compared to 2-Dimensional Classification Input.

2.6.1 Overview of 3DCNN Activity Recognition Layers:

Figure 2.16 illustrates the concepts of activity recognition with a further breakdown of its operations in Appendices 2.2 to Appendices 2.3.2. Classification commences from input stage zero into Cov3d, the first classification layer at stage one. The clean data reflect an image size of (32x32x12x32) with a height and width of thirty-two relative to the model's dimensional requirements, with a depth of 12 to regulate the growth of the tensor block. The depth parameter is paramount as it reduces the impact of application memory and computation overload. A batch of 32 samples aligned with [126] is configured during training to regulate the accuracy output. At stage two, data from stage one is fed to this layer

for processing with similar image size specifications via an activation layer. Its operation defined in [127] refines the activity object’s edges and line pattern details within this tensor block. At stage three, the Cov3d-1 layer, the data undertakes another convolution conversion utilising similar dimensional specifications to accentuate the violent activity blob objects. The operations applied a mathematical max pooling function at stage four to reduce the tensor map dimensionality to 11x11x4x32.

The concept accentuated in [128] alleviates the computational complexity between adjacent layers. Stage five Cov3d-2 conducts another classification processing to refine those high-level features at stage four. Its computations generated a larger tensor map size of 11x11x4x64. Upsizing and downsizing image dimensionalities mentioned in [129] during convolution reduces its computational requirement, keeping its symmetrical specifications in each layer. At stages six to seven, another activation layer, Act-1 and Con-3, refines the violent activity features utilising similar image dimensions at stage #5. The operations applied another Max Pooling-1 layer at stage eight to reduce the computational complexity and the data’s dimensionality to 4x4x2x64. Following the latter, a Conv3d layer at stage nine, Act-2 layers at stage ten and a Conv3d layer at stage eleven refine the violent features further. Its operations generate more features at a dimension of 4x4x2x128.

The model applies another Max Pool-2 at stage twelve to reduce computational complexity and dimensionality to 2x2x1x128 relative to the previous layers. Following is another dropout layer at stage thirteen with similar size specifications. This operation regulates the neurological processing mentioned in [80] between layer nodes on the network. The function alleviates over-fitting and biased results by dropping neurons that become excessively dependent on other reinforced input features. The operation feeds forward the data towards Cov3d-6 at stage fourteen, Act-3 at stage fifteen and Cov3d-7 at stage sixteen. Its computations alter the image dimensions to 2x2x1x256 as output. At this level, 256 high-level features are refined further and fed forward towards a final Max-Pool-3 at stage seventeen. Max pooling reduces the computational processing and image dimen-

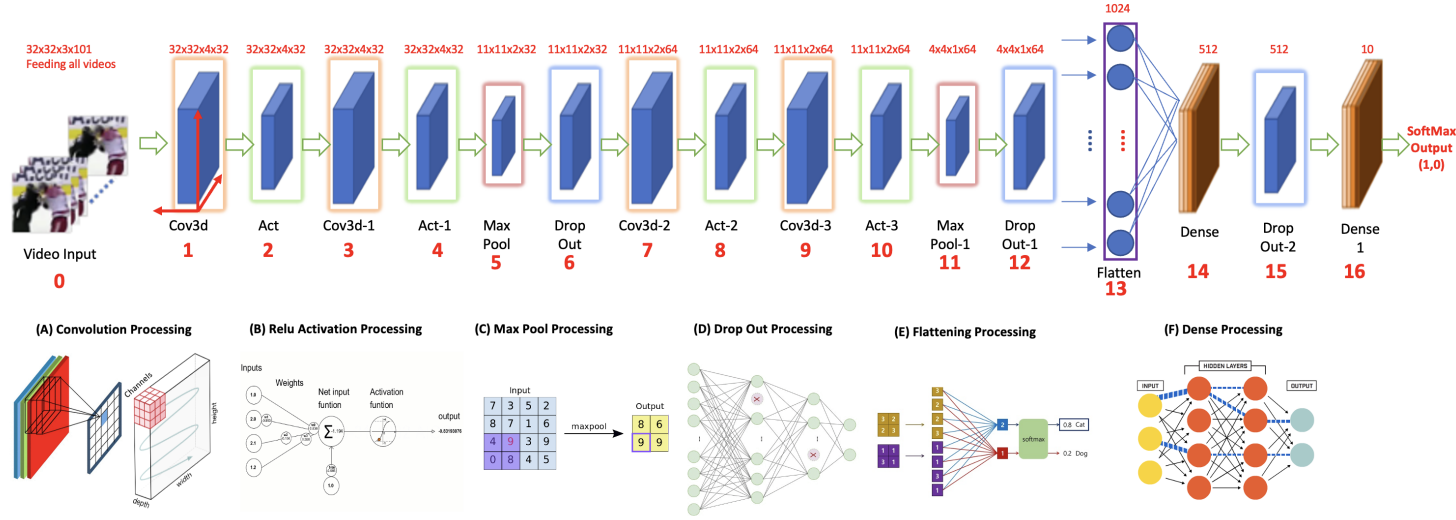


Figure 2.16: 3DCNN Proposed Architecture.

-sions to $1 \times 1 \times 1 \times 256$. The operations utilised stage seventeen's data as input for another Drop Out-1 at stage eighteen. Its operations implement regularisation to encourage robust predictions and further reduce the risks of over-fitting. The network utilises a flattening procedure at stage nineteen to reshape the cube-like output from stage eighteen into a 1-dimensional vector that retains the spatial and temporal attributes of the violent activity features similar to [117]. The data reformation procedures defined in [130] emphasised the format requirements for the densely connected, fully interconnected layers. The operations applied a dense layer at stage twenty to generate a label mapping sequence for the activity input data from stage nineteen. Its logic discussed in [131] and [132] corresponds to feature relationships and determines the probability scores of its predictions through linear and non-linear computations.

The notion reinforced by [132] formulated the actual concept of activity recognition. Another Drop Out-2 layer induces regularisation and feeds output to a final Dense-1 Layer at stage twenty-one for classification. At stage twenty-two, the Soft-Max function defined in [133] maps probability scores with class labels, interpreting the confidence of the 3DCNN's predictions to conclude its processing. The final Soft-Max fully connected classification layer utilises Soft-Max and Cross Entropy loss functions that ensure the probability of predictions. Those layers aggregate features generalised during convolution identical to

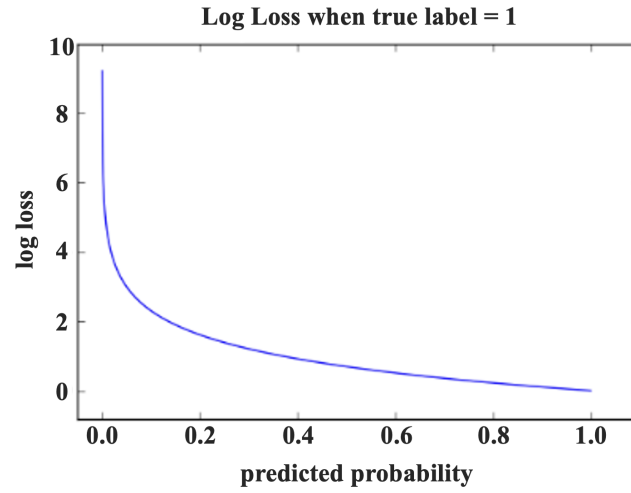


Figure 2.17: 3DCNN Soft-Max Cross-Entropy Log Loss Process.

[134], which constructs a universal depiction of the object or objects of interest. The Soft-Max function reflected in [135] interprets the output values as scores (turns the flattened vector values into a flattened vector of values that sum to one), where input values undergo transformations that denote probabilities as output with a range between 1 and 0. In this demonstration, the data complexity hindered the processing from generating sufficient values for the x,y curve zero intersection point, which demonstrates the starting position of the processing. Log loss or cross-entropy loss outlined in [136] is applied to evaluate the classification performance of the probability output as its predictions diverge from the

ground-truth label as the score increases. Therefore, having a log loss of zero or closest to zero with a predicted value of 1 is ideal. Nevertheless, an example of the log loss value demonstrated in [136] suggests .012 with low probability scores is considered insignificant. Figure 2.17 illustrates an ideal log loss performance for activity recognition where lower log loss constitutes high accuracy results towards one hundred per cent.

2.6.2 Overview of Activity Recognition Processing Components:

Like YOLOv5, other mechanisms exist that perform activity recognition in the literature as the state-of-the-art [65], [137] and [138]. With that knowledge, the investigations considered 3DCNN architecture as the state-of-the-art component to satisfy the objectives at this stage. In this section, classification connotations provided context relative to 3DCNN from an individual visual perspective, illustrating its processing capability as the second element applied to fortify the proposed fusion technique. 3DCNN detailed in Appendices 2.2 incorporates several classification processing layers, thus transforming the CCTV data from a 3-dimensional input stage into activity recognition at the output.

2.7 Overview of Classification Evaluation Measures

With an understanding of 3DCNN's operations, evaluation techniques played a significant role in appraising the actual performance. The measuring components accentuated in [139] include graphical depictions and the confusion matrix's application to reflect the output's interpretation. Author [140] projected graphical depictions of the results, encouraging visual interpretations of the findings. Moreover, the confusion matrix outlined in [141] measures overall classification performance. The approach followed [142] utilising combinations of 4-predicted values that discern the integer one as a positive occurrence or a lesser value to zero as a negative occurrence. The confusion matrix in [143] 's processing facilitated binary (2-class) and multi-class (greater than two classes) classification relative to processing tasks. Its classification process considers important computational equations

reflecting the mean average precision, accuracy, false discovery rate, sensitivity/ recall, the f1-score, and specificity. Further details in Appendices 2.3.1 to Appendices 2.3.3 expanded the concepts of the evaluation metric discussions below to fortify the understanding of binary and multi-class assessment techniques.

- (i) **TP/Recall/Sensitivity (TPRS):** Demonstrates results of all classes that are truly positive, how many were labelled correctly [144]. Sensitivity (TP, Recall) denotes:

$$TPRS = \frac{TP}{TP + FN}$$

- (ii) **TN/Specificity:** The ability of a model to correctly classify a class as weapons or not [145]. Specificity denotes:

$$Specificity = \frac{TN}{FP + TN}$$

- (iii) **Fall Out:** The proportion of incorrect predictions incorrectly identified as correct predictions. The idea refers to the probability that a false alarm will be raised [146]. FP (false positive) or Fall Out Rate denotes:

$$FalsePositive(FP) = \frac{FP}{FP + TN}$$

- (iv) **FN/False Negative Rate/Miss Rate:** The proportion of correctly predicted cases that were incorrectly classified as incorrect [141]. False Negative Rate denote:

$$FalseNegativeRate = \frac{FN}{TP + FN}$$

- (v) **Precision/Positive Predictive Values:** Of all correct labelled predictions, how many

are correct? [147]. Precision denote:

$$Precision = \frac{TP}{TP + FP}$$

(vi) **False Discovery Rate:** The number of classes that are not weapons in nature but are identified as weapon objects [148]. False Discovery Rate denotes:

$$FalseDiscoveryRate = \frac{FP}{FP + TP}$$

(vii) **F1-score:** The total amount of low False Positives and False Negatives [149] and [144]. F1-score denotes:

$$F1 - score = \frac{2x(Precision \times Recall)}{(Precision + Recall)}$$

(viii) **Accuracy:** The overall performance of predictions correctly classified [150]. Accuracy denote:

$$Accuracy = \frac{TP + TN}{(TP + FP + FN + TN)}$$

(ix) **Average Precision (AP):** The average in this instance computes results for each class utilising several other metrics, including the Precision, Recall, IoU, the Precision Recall-Curve (graphical representation of the precision against the recall values) and the area under the PR curve (AUC). The idea encompasses generating the model's prediction score, evaluating its precision, and recall outcome status with the confusion matrix for each class object. The metric condenses the Precision/Recall curve to 1 numeric value. Its output is usually high when the precision and recall values are high and considered low when both are low relative to the range of confidence threshold results. The numeric output described in [151] ranges between 0 and 1. It takes the area under the precision/recall curve by applying the integral function of the recall

values ranging from 0 to 1, where r is the recall, p is the precision at specific values with a summation of the precision values $p(r)$ [152]. Computing the average precision is as follows:

$$AveragePrecision(AP) = \int_{r=0}^1 p(r) dr$$

(x) Mean Average Precision(mAP):IoU=0.5 The performance is measured by computing the average relative to the AP of every class to analyse convolution classification accuracy. Here, (n) is the number of classes for each class that (i) and $(AP-i)$ considers. The idea represents the AP value for the $(i\text{-th})$ class across different IoU thresholds relative to Figure 2.11. Calculating the final mAP value specified in [151] produces an average of all mAP scores per class. The computation of the Mean Average Precision (mAP) is as follows:

$$MeanAveragePrecision(mAP) = \frac{1}{n} \sum_{i=1}^n AP_i$$

Appreciating the model evaluation techniques, the outlook on performance establishes a comprehension of the results regarding positive operations or otherwise. The outlook on performance is as follows.

2.8 Summary of Classification Issues: Over/Under and Good Fitting

At this level, performance issues relative to over-fitting, under-fitting and context of a good fit accentuate evaluation concepts as per previous discussions. Also, emphasising significant progress and insignificant performance demonstrate essential items for deep learning models. Graphical depiction scripts are implemented during development to highlight the visual outlook of these issues to convey the operation's significance using the matplotlib library packages [153], [154] and [155]. The analysis of these metrics plays a crucial role when

visually observing the processing. The curvature of the graphical depictions expressed in [156] and [157] determines whether the model requires further hyper-parameter fine-tuning (tuning the model's option features), more training data, or architectural adjustments to reduce the negative impact on its processing. These evaluating factors discussed in [158], [159], and Appendices 1.5 expound on whether the issue of over-fitting or under-fitting is present when dispensing the final output. The idea of an insignificant performance is presented at this stage to enhance one's awareness of factors that can limit the processing efficiency

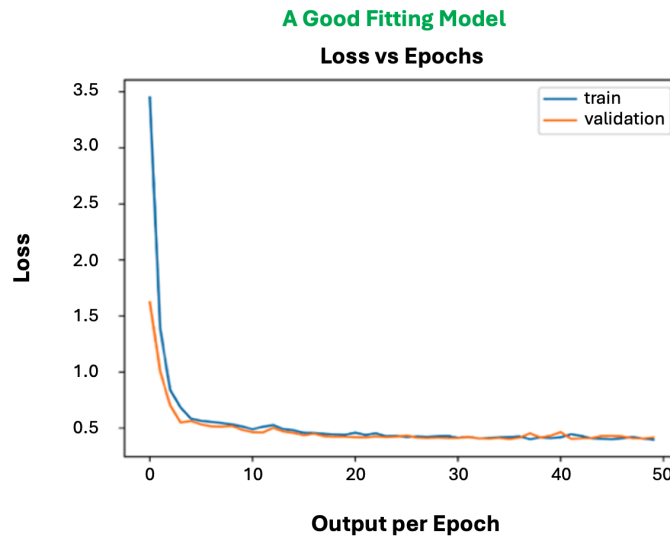


Figure 2.18: Graphical Representation of a Good Data Fitting Operations Model.

of the model. Moreover, the idea of a good fit defines positive processing results. The output determines the mitigation strategy when experiencing fluctuations by appreciating the previously discussed graphical depictions. Reducing model complexity to align with the task and integrating large datasets with sufficient feature patterns can mitigate over-fitting and under-fitting issues. A good fit demonstrated in Figure 2.18 displayed a graphical representation of a well-fitted operation reflecting high-performance processing. The visual

reflection of the operation coincides with [160], [161] and [162] as a balanced distribution of the values facilitating a gradual descent of the training in blue followed by the validation outputs in orange. The graph illustrates the operations at high loss up commencing the processing and gradually descending, which projects how well the model generalises for each epoch iteration.

2.9 Literature Review

Activity recognition is a prominent concept that facilitates tasks involving predicting human motion relative to acceleration, velocity and trajectory. Before artificial intelligence concepts, authors in [163], [164], [165] and [166] explored heuristic and statistical methods concerning relationship and pattern knowledge to recognise activities as traditional measures. Other researchers in [167], [168] and [169] incorporated decision trees concerning random forests to simplify processing input features and relations of target variables. Authors [170], [171] and [169] incorporated the power of support vector machines (SVM) to compute high dimensional activity recognition data in classification challenges. Activity recognition via [172], [173] and [174] incorporated K-Nearest Neighbors to achieve classification effectiveness by locating the k-nearest data points and assigning the most common label to suggest the presence of the activity. Some approaches assumed independence amid features to disclose the classification state utilising Naive Bayes techniques [175], [176] and [177]. A more chronological data and time-series approach in [178], [179] and [180] employed activity recognition with hidden Markov models (HMM) to achieve classification effectiveness. Authors [181], [173] and [182] attempted a Gaussian mixture models (GMM) clustering technique to process the activity recognition data's probability distribution towards activity classification.

Further investigations revealed a template pattern-matching procedure in [183], [184] and [137] that achieved classification by generating templates of activities from sample data

and matching new attributes to classify activities. Authors [185], [186] and [187] attempted ensemble bagging and boosting techniques via combining multiple classifiers to enhance the activity recognition performance outcome. Some investigations integrated signal processing techniques via Fourier and wavelet transform, thus altering time and frequency signals relative to accelerometer data to convey actions [188], [189] and [190]. The authors in [191], [192] and [193] applied graph-based methods to pinpoint probabilistic relationships among variables. Activity recognition in [169], [194] and [195] applied clustering, which segments the data into clusters relative to the similarity in relationships to unveil activity patterns without the application of labelled data. Other techniques in [196], [197] and [198] finetuned hyperparameters to generate rules through natural selection and genetic processing. More futuristic approaches in [199], [200], [201] and [202] explored the applicability of sensors base measures to conclude the human gait in various forms.

Like activity recognition, object detection traditional methods in [203] encompasses HAAR feature-based concepts as a cascade approach of complex classifiers to effectively generate an image's edges, lines and gradient boundaries in real-time. Comprehending the detection process within an image in [203], [204] and [205] relates to feature-based detectors and descriptor algorithms working in tandem to establish edge boundaries accentuating its relationship resonance. Descriptor tools such as the Local Binary Pattern LBP [206], Histogram Of Gradient HOG [207], Scale-Invariant Feature Transform SIFT [208], Speeded-Up Robust Features SURF a combination of both [209], Binary Robust Invariant Scalable Key-points BRISK [210], Binary Robust Independent Elementary Features BRIEF [211], and Gradient Location and Orientation Histogram GLOH [212], match features between different images. Detectors concerning Harris corner detection [213], Oriented Fast And Rotated Brief ORB [214], and Features from accelerated segment test FAST [215] obtain key repeatable attribute points to present the critical attributes of an object within an image.

Some authors incorporated a sliding window technique across an image at several ratio

scales, extracting features from each window and implementing machine learning via support vector machines SVMs or decision trees to achieve its classification [216], [217], and [218]. On the other hand, a selective search in [98], [219], and [220] incorporated a region proposal classification scheme, which creates candidate object regions by combining region and pixel similarity regarding resolution, size, gradient colour, colour textures, object size, and its dimensional shape. Some researchers investigated colour histograms and textures-based solutions depending on texture patterns and colour proportions for object identification and classification [221], [222], and [223]. Traditional object detection model-based techniques implemented in [224], [225], and [223] relied heavily on applications coupled with the object’s geometric properties regarding boundary detection and matching shapes to achieve its classification outcomes.

As available techniques for the proposed solution through the literature investigations became apparent, most research works generally applied object detection and activity recognition to classify actions. Ubiquitous evidence proved that the gist of achieving a high-performance activity prediction solution circumvents anthropometric notions specific to normal (neutral) human actions. Several research efforts presented the activity classification individually rather than disclosing its generic status. The evidence in the literature proved that human activities lacked the holistic representation as non-violent actions or violent behavioural patterns before aligning the specifics of its action class (whether the action is violent or not and then aligning its subclass status towards a stabbing or rough playing). Most works lack representations of pre-empting violent attacks that identify victims, aggressors, and weapon possibilities deployed during the altercations. The literature investigations also disclosed extensive research towards neutral human action classification without affiliation to violent outbursts. The previous notion led the investigations towards individual processing mechanisms, which facilitate violence processing and promote robustness for the proposed concept. The approach commenced by exploring the possibility of developing a reliable human activity model that accurately and efficiently pre-empts

violent activity in CCTV videos as the primary prospect.

At inception, the investigations reflect innovative artificial intelligence frameworks in 2-stages to evaluate the true capability of the formidable processing techniques. Stage-1 explored innovative concepts of object detection towards violent action and weapon artefact classification. Following Stage-1 is Stage-2, which determines the state-of-the-art techniques concerning activity recognition. Implementing this specific investigative structure determines whether object detection can efficiently perform activity recognition before integrating supplementary activity recognition mechanisms. The approach fortifies the rationale for the proposed robust fusion concepts and their processing. The review is as follows.

2.9.1 Stage-1: Object Detection for Violent Activity Mechanisms:

The investigative task disclosed a human key-point pose prediction and action classification in raw images as a proposition for action detection utilising RCNN as the first object detection mechanism [226]. Their proposal attained a 70.5% mean average precision with an input object proposal using RCNN towards neutral action classification issues on the PASCAL VOC dataset. The approach considers activity detection tasks containing less complex movement. However, the complexity of tracking anthropometric key points during violent altercations could adversely impact the overall processing due to the sporadic nature of the motions relative to the ROI's speed, trajectory, and velocity. Because RCNN's framework generates 2000-proposal regions at a rate of 45-50 seconds per image for object detection, its applicability towards real-time operations proved challenging as it slows processing speed. Further investigations presented an adaptive RCNN concept that processed typical contextual action cues using R*CNN [227]. They applied several proposal regions for classification while maintaining the ability to localise the action detection. Their approach achieved 90.2% mean average precision with 1 second per iteration for training and 0.4 seconds on testing per image using GPU on PASCAL VOC and Berkeley attributes

of people dataset. Their efforts emphasised classification significance relative to its GPU processing and inference speed for real-time applications. However, they employed neutral actions (jumping, phoning, playing an instrument, reading, riding a bike, riding a horse, running, taking a photo, using a computer, walking) that reflect limited cross-correlation of body parts with a high sporadic motion to evaluate classification complexity during inference. Because of the importance of human life in real-world scenarios, the demand for performance mechanisms satisfying the dependency ratio between total accuracy and speed is high.

Author [228] introduced a collaborative effort utilising mask region-based classification network (Mask RCNN), key-point detection, and long short-term memory (LSTM) for action detection relative to punching and kicking. They achieved 93.4% as the highest accuracy on 40,423 frames at a split ratio of 80:20 generated from Weizmann (containing 90 video samples) and KTH (with 2391 sequences) datasets of neutral human action. They introduced a third fabricated dataset of 273 videos sectioned into 90-Boxing, 90-Kicking, and 93-neutral video samples, in which they extracted the object’s temporal key-point mask details for processing. Compared to author [227], author [228]’s approach facilitates activity recognition; however, its limitation reflects similarities relative to RCNN. Violence’s complexity and sporadic nature potentially intensify generalisation issues if human-body correlations escalate. Also, the computational resources required for its operation are exceptionally high. As discussed in Section 2.4.5, RCNN’s processing complexity adversely impacts computational resources. This notion led to the investigation towards [229], which confirmed the drawbacks of region-based approaches utilising RCNN. The author developed a Fast-RCNN (fast region-based classification network) technique emulated from RCNN, achieving a mean average precision of 66.9%. The approach applied a single-stage training algorithm to detect object proposals and refine their spatial locations in 0.3 seconds (excluding the object proposal time) using several test samples derived from PASCAL VOC 2012. Fast-RCNN attained high significance for object detection relative to weapon de-

tection in static images; however, its operations require reinforcements to perceive objects containing spatiotemporal trajectories over high speed relative to humans holding bladed instruments. Also, its framework must improve towards real-time efficacy due to heavy computational storage dependency when scaling image sizes.

Author [216] proposed Faster-RCNN as a region proposal network (RPN) concept that generates premium, nearly cost-free region proposals and predicts object boundaries with scores simultaneously using Fast R-CNN for detection. They achieved this by alternating the selective search process with RPN for object detection. Their efforts dispensed a mean average precision accuracy of 75.9% with 300 proposals per image on a selective sample size of 10k images derived from PASCAL-VOC-2007 and 2012-MS-COCO. In this instance, RPN's shared concept demonstrated a high capability for object detection relative to weapon artefacts. Its object detection design validated its operations for typical human actions in non-violent scenarios. However, the technique introduces high risks concerning robust performance towards the criticality of pre-empting lethal scenarios relative to human life. Its capability to generate 300 proposals per image is a critical output risk concerning the depletion of the computational resource; nevertheless, its processing capability exceeded RCNN's 2000 proposal efforts. The discovery of [230] 's method "you only look once" (YOLO) is the first milestone in the state-of-the-art object detection for this domain. The authors applied YOLO as a human object detector on the Pascal VOC dataset and achieved a mean average precision score of 63.4% accuracy at 45 frames per second. YOLO's processing divides violent image data into a grid system where each cell detects objects within itself at high speed and accuracy. Although the state-of-the-art is fast, its processing is limited when detecting small objects conveying high acceleration, velocity, and trajectory. Further investigation disclosed several versions of YOLO's architectures that produce high-efficiency performance ratings for object detection.

Further investigations into [231] and [232] reveal LIRIS non-violent human activity recog-

nition concepts utilising YOLO. They focused on classifying the complexity of non-violent human actions and detected each human object's localisation. Their ADSC-NUS-UIUC team achieved a precision of 41%, an f1-score of 53%, and a recall of 74% on the LIRIS dataset comprising 55,298 images of activity containing 828 actions in their dataset. Their approach facilitated activity recognition, but the YOLO base framework cannot detect highly accelerated motions of bladed objects, in addition to the combined complexity of violent actions and the criticality of attaining robust outputs. Author [113] applied YOLO version two (YOLOv2) as a pedestrian detection system, achieving accuracy increments ratios of 9.03%, 6.37% and 5.91% on 14999-samples containing multiple pedestrians per image in the KTH dataset. YOLOv2 proved fast for object detection on neutral data; however, it suffers from classification issues relative to low recall compared to YOLO's base version. This issue escalates in conditions where the correlated patterns of humans or small objects convey highly accelerated motions. YOLO's version three (YOLOv3) in [redmon2018yolov3] improved on YOLOv2. They applied a modified framework to facilitate a high-speed human detection approach, achieving a mapping result of 57.9% on the COCO dataset. The approach facilitated human object detection; however, the model exacts a high demand for application memory and classification challenges when processing smaller objects. The investigations disclosed [233] 's YOLO version four (YOLOv4). They developed a weapon detection solution towards identifying violent crimes. They focused on detecting deadly weapons such as handguns and knives, utilising custom-trained models circumventing the YOLOv4 Darknet framework for single and multi-class classification. Their single-class approach achieved a mAP of 77.78% accuracy, whilst their multiple-class endeavour dispensed up to 100% accuracy on a section of the Open Images V6 dataset exceeding 3000 images. Their approach facilitates small weapon detection from a static perspective with high performance. Although the approach proved promising, the framework demands high application memory processing and experiences challenges when processing low-level CCTV image resolution with high noise ratios and fluctuating

luminous intensities of the scenery.

Author [234] developed a pedestrian detection and suspicious activity recognition tracking technique utilising the YOLO version five (YOLOv5) on a pedestrian dataset containing 600 videos of student behaviour relative to cheating, stealing lab devices, disputes involving minor scuffling scenarios. They attained a mAP of 96.12% towards object detection. Their work facilitates object detection but needs proper analysis of small weaponry with high motion and excessively violent correlated human interactions. There are limited official publications on YOLOv5. The investigations into [235] disclosed a suspicious activity trigger system (SATS) that automatically triggers a suspicious activity alert message when such actions are detected. They applied YOLOv6 to detect human objects and determine whether the actions were suspicious or not concerning property intrusions. Their solution triggers an alarm, sending notifications for evasive actions in the context of suspicious activities. They attained a mAP of 96.6% accuracy, utilising a split ratio of 80% training and 20% for validation and testing on 1000 images of humans from the Google Open Image dataset. Their efforts disclosed merit towards real-time object detection for weapons and suspicious actions; however, its applicability towards the complexity of human-to-human violent interactions with artefacts requires further analysis to determine its impact. YOLOv6 processing lacks framework flexibility with stability issues compared to YOLOv5's ability to facilitate the processing of more enormous frame proportions with high-resolution video inputs.

Author [236] applied YOLO version seven (YOLOv7) towards anomalous activity detection using a control room alarm technique. They collected video data from online platforms and applied a split ratio of 60% for training and 40% for validations. They achieved 65-75% mAP, disclosing that a significant volume of data is necessary for high performance. The approach demonstrated merits towards object detection for activity recognition as a new object detection medium. However, it lacked the robust performance requirement due

to the significance of pre-empting the loss of human life. Authors' [116] YOLOv7 and [237] YOLOv8 are new techniques with limited research prospects within the literature to evaluate their feasibility. The review disclosed [238] 's YOLO version nine (YOLOv9) and [239] 's YOLO version ten (YOLOv10) as the state-of-the-art for object detection and activity recognition. However, no significant publications detailed the evaluation of violent activity detection with weapon artefacts from a theoretical standpoint. The context of YOLO's model selection in Chapter 3 disclosed the significance of the current research. Exploring the state-of-the-art object detection possibilities for enhanced classification with minimal disadvantages was necessary to expose activity recognition innovations in Stage 2 below. The idea is to evaluate the pros and cons of performance compared to YOLO individual processing discoveries.

2.9.2 Stage 2 Investigations: for Violent Activity Recognition:

The investigations commenced Phase 2 of the literature review by reviewing state-of-the-art activity recognition methodologies relative to the previous discussions. Several innovative propositions substantiated the capability of YOLO object detection performance. However, an appropriate concept that nullifies vulnerable classification issues regarding the movement of small objects with a high trajectory, velocity and acceleration is yet to be determined. The earliest concept of activity recognition considered [240] 's wearable activity motion sensors approach that detected neutral actions in motion relative to the distribution of the human gait. They developed a novel approach for automatic activity recognition relative to multi-sensor data via an offline adaptive-hidden Markov model. The technique detected commonly performed non-violent actions rather than complex violent scenarios. They examined human subjects with digital button sensors generating activity data by monitoring a gyroscope's axis orientation signals, the accelerometer motion and location from a GPS. Their activity recognition efforts dispensed an f1-score of 0.98% as the highest output for non-violent actions compared to complex, violent actions. The framework's core

operation introduced complex configurations to capture the heterogeneous violent features. Violent real-world scenarios require robust processing systems that supersede accelerometer sensors relative to their ambiguous nature and the range of violent sporadic motions.

The investigations disclosed [119]. They applied a concatenation of the state-of-the-art 3DCNN and LSTM as a Bi-Directional LSTM solution for early action frame prediction relating to motion patterns and object appearances as a modern automated technique. The authors achieved 0.68% accuracy on the UCF crime dataset toward activity recognition. However, the technique required improvement relative to feature learning and extractions at the input stages. The impact introduced an adverse effect that hindered the model's inference capability due to overlooking pertinent complex processing anomalies encompassing violent activity features during training. Its operations increase the risk of biased results.

Further investigations revealed [241], who integrated a mixed classification Resnet-50 regression block approach on the UCF-101 dataset based on feature and model fusing for specific targets. They achieved an activity recognition correct rate of 71.07% at a processing speed of 200 fps (frame per second) with complex configurations. Authors [242] developed a hybrid 3DCNN HAR model for activity recognition actions of KTH and J-HMDB datasets. They demonstrated state-of-the-art performance compared to baseline methods, with an accuracy output of 78% on KTH and 90% on J-HMDB from a non-violent perspective. They experienced image dimension challenges that adversely affected their generalisation operations. The issue negatively impacts the model's robust processing capability as a critical component of preventing the loss of human life. Author [243] proposed a hybrid technique including 3D-CNNs and optical flow-gated networks for violent activity recognition as the second milestone. They obtained an accuracy of 87.25% on the RWF dataset containing 2000-sample videos with complex script configurations. The idea at this stage was to establish state-of-the-art activity recognition that allows framework integration flexibility for further development.

2.10 Conclusion

To summarise Chapter Two, essential knowledge provided context into the research area to assist the reader with understanding the technical jargon and processing operations required for this research thesis. The chapter provides context into blob analysis, cross-validation, library packages, and an introduction to classification models as crucial data handling and pre-processing measures that encourage high performance during the development stages. To fortify the processing concept, an introduction to the background of the YOLOv5 and 3DCNN provided an overview of the evaluation measures to achieve processing efficacy. It was necessary to demonstrate the context of classification issues to emphasise the importance of attaining a well-fitted model. Conducting the literature review in two stages facilitates the difference in the object detection and activity recognition approaches. The literature's first stage presented YOLO as object detection, with a second stage disclosing 3DCNN. Achieving the milestones allowed the assessment of the experimental methodology in Chapter 3 to formulate processing efficacies between both models. Disclosing the methodology is as follows.

Chapter 3

Research Methodology

Chapter 3 accentuates the experimental approach towards developing an innovative violent activity detection solution, which recognises violence accurately and efficiently within CCTV data. The research methodology reflects a series of experiments that expose the technical operations required to align Section 1.3 objectives with the developmental phase. The investigations commenced with a quantitative approach aligning [244]’s approach towards observing violent scenarios affecting society through statistical analysis relative to collecting and processing significant volumes of data. The concepts emphasise data patterns that aid model learning through convolution to infer results dispensed via simulations; this validated the significance of the proposal’s contributions. The operations designed encompass the effectiveness of YOLOv5m and 3DCNN as diverse artificial intelligence mechanisms from an individual processing perspective. The idea of YOLOv5m as the state-of-the-art for object detection facilitated the classification of lethal weapons pertinent to the class of activities within the scenery of static frames. In contrast, the literature review disclosed the possibility of the YOLO base model as an activity recognition solution using non-violent data across a sequence of frames. The possibility of a YOLO base version in a non-violent capacity allowed the integration of YOLOv5m in a similar ac-

tivity recognition context, utilising multiple classes conveying violent actions. The strategy reduced integration challenges via programming and provided a method to merge weapon and action objects portrayed in a sequence of frames.

Fortifying the processing encompasses the implementation of 3DCNN as the state-of-the-art for activity recognition. The concept validated the existence of the action's generic category, and this inferred whether such leads to potential violence, non-violent attributes, or individual sub-classes reflecting stabbing, beating, fighting, or shooting. The idea entails observing the significance of YOLOv5m and 3DCNN by implementing activity recognition fine-tuning modifications that established model consistency and high performance. Applying the precision, mAP and recall metrics allowed the evaluation of the task to determine the factual processing accuracy discussed within Section 2.7 and Section 2.8. Chapter 3 is divided into five sections to illustrate the stages required to achieve the methodology towards pre-empting lethal violent scenarios. Section 3.1 commenced by providing an overview of dataset processing operations in alignment with research question-2's objectives in section 1.2.1. Given the previous data procedures, a projection of the experimental protocol in Section 3.2 aids in formulating prospects to attain the results in alignment with the research initiatives discussed in Section 1.3.

Moreover, details concerning YOLOv5m's framework in Section 3.3 and 3DCNN in Section 3.4 provide the evaluation strategy to foster effectiveness. Finally, a summary of discussions disclosed the alignment of the research endeavours in Section 3.5 to conclude the chapter. Acquiring significant volumes of relevant data containing quality resolutions of real-world stabbing data, with its pre-start duration, proved challenging to locate as a prerequisite. The fact that the nature of the research depended on data predominantly concerning individuals being fatally injured or losing their lives intensified the previously mentioned challenge during the data acquisition process. Because of the criticality of human life, an additional challenge presented itself if unrealistic data repositories were

employed. The issue escalates biased false positive computations dispensing erroneous outcomes. Additionally, the challenge intensified the processing complexity by distorting the model’s ability to interpret valid attributes of violence via stabbing compared to rough playing. The research investigations in Section 2.9 disclosed multiple bench-marked dataset repositories that facilitated activity recognition from a neutral, non-violent human action perspective. The data acquisition task selectively sourced violent samples via online social media forums and avoided items containing poor resolutions to establish a balance of violent samples compared to the availability of non-violent samples in ubiquitous repositories. The raw data acquisition approach entails downloading significant violent action samples for YOLOv5m and 3DCNN during the training stages. The strategy produced a balanced range of motions as a significant sample size to enhance the models’ individual processing/learning initiatives regarding the architectural difference.

3.1 YOLOv5m Activity Recognition Class Description

Completing the data acquisition operations previously discussed ensured that the classes obtained matched the context of the class of activity template specified in Section 2.1.1. The approach ensured that the action categories clearly distinguished between what constitutes violent and non-violent classes before commencing development. Applying 16 classes in Table 3.1 of a balanced ratio of violent and non-violent categories assisted in accentuating the operation’s significance during YOLOv5m’s development. The significance of the sixteen classes assists in optimising the processing and encourages effective classification in alignment with the aim and objectives.

3.1.1 YOLOv5m/3DCNN Activity Recognition Dataset Grouping:

With an understanding of the data classes, a grouping strategy sectioned the samples to reflect a new real-world violent action detection dataset (RWVAD). After the augmentation process, a 560-sample dataset endured further segmentation into a balanced ratio of 280

| Class | Labels | YOLOv5 Class Label Description |
|----------------------------|--------------------|---|
| Violent Classes | | Classes used to indicate the presence of or violent activity |
| 0 | Aggressor | Individual/s about to commit or committing an act of violence |
| 1 | Blood | Indication of injuries sustained during violent altercations |
| 2 | Knife-weapon | Indication of a knife used for violence or is in the image scenery |
| 3 | Stabbing | Pre-violent/ violent action has occurred/ is continuing |
| 4 | Sword | Indication of a bladed object (fencing mainly/ can occur in stabbing) |
| 5 | Knife-Deploy | An indication of the Stabbing Posture or Gesturing |
| 6 | Hand | Indication of a knife being held to initiate a stabbing |
| 7 | Victim | Indication of person/s being attacked/receiving injuries |
| Non-Violent Classes | | Classes used to indicate Non-violent Activity |
| 8 | Fencing | sport actions used in action similarity experiments |
| 9 | Person | Indication of additional person/s involved in or around the attack |
| 10 | Discussion-WGI | Persons chatting Whilst Giving an Item to another person |
| 11 | Discussion-WOBoard | Persons chatting whilst Writing On black or white Board |
| 12 | Discussion-ppl | People/Person chatting casually or formally |
| 13 | Item-passed | Person handing off an item to another person |
| 14 | Writing-On-Board | Person/s Writing on black or white board |
| 15 | Background-Images | No object of interest in the scenery (enhances inference ability) |

Table 3.1: YOLOv5m Subclass Violent and Non-Violent Description.

samples per the generic class between violent and non-violent actions with varying duration, resolutions, and dimensionality. Each subclass category reflects 35 videos per class for consistency in the generic classes. The intentional integration of action similarity samples supported the evaluation of non-violent/violent actions containing high similarity in characteristics, such as fencing and stabbing. The rationale behind the deliberate grouping strategy challenges

the classification operation for both models as a prerequisite towards thoroughly validating the approach. The class objects significance within Table 3.1 and Table 3.2 constitute object detection and from an activity recognition sense to convey the difference in YOLOv5m and 3DCNN input sequence. For example, the stabbing class label in Table 3.1 is an object for YOLOv5m, which conveys one object demonstrated across a sequence of frames for a given duration between 10-12 seconds. Contrarily, the 3DCNN stabbing object in Table 3.2 represents the entirety of the video to signify the action from a class label classification perspective.

77

The real-world violent action detection dataset (RWVAD) was strategically separated into RWVAD1st and RWVAD2nd. This separation was not just a technical detail but a key factor in facilitating the comparison between the two models (3DCNN/YOLOv5m). Grouping the data in this fashion facilitated superiority investigations into YOLOv5m as an activity recognition model and the state-of-the-art 3DCNN activity recognition. The distinction in the datasets' naming convention also plays a crucial role, providing a clear way to identify the data designed for YOLOv5m's framework as RWVAD1st dataset, as opposed to the data designed for 3DCNN's architecture as RWVAD2nd during development. Although the architectures differ, as discussed in Appendices 3.1 and Appendices 3.2, consistency in alignment with the model's processing proved paramount as standards. Utilising the same data derived from the acquisition procedures via social media sources and [245]'s repository adds value to the operation's effectiveness.

3.2 3DCNN Activity Recognition Class Description

Like YOLOv5m's class overview in Table 3.1, it was necessary to investigate 3DCNN action similarity processing utilising classes emphasised in Table 3.2 to establish its true processing abilities. The findings proved the fulfilment

of the action similarity objectives by implementing RWVAD2nd conforming its details relative to Appendix 3.2 to facilitate 3DCNN's architecture.

| # | Class Label | 3DCNN Class Label Description |
|-----------------------------|-----------------------|---|
| Non-violent(neutral) | | Generic Category: Indication of normal human actions |
| 0 | Cutting-in-Kitchen(C) | Indications of food preparation in a kitchen as a neutral class |
| 1 | Nun-chucks(N) | Indications individual using nun-chuck alone in a non-violent manner |
| 2 | Fencing(Fe) | Actions relative to the fencing sport/ For Action Similarity experiments |
| 3 | Sumo-wrestling(Su) | Similarity relative to the wrestling sport |
| 4 | Walk-with-dog(W) | Actions relative to person/s walking dog/s |
| 5 | Knitting(K) | Actions relative to person/s sitting/ knitting |
| Violent Classes | | Generic Category: 1 vs 1, many vs 1, 1 vs many, group violence |
| 6 | Fighting(Fi) | Striking with arms/legs to cause harm / For Action Similarity Experiments |
| 7 | Beating(B) | Striking with object to cause bodily harm |
| 8 | Shooting(Sh) | Use of projectile weapon/s to cause human endangerment |
| 9 | Stabbing(St) | Use of bladed/sharpened instrument/s to cause bodily harm |

Table 3.2: 3DCNN Activity Recognition Subclass Description.

3.3 Experimental Protocol:

The application of YOLOv5m activity recognition with 3DCNN allowed the generation of essential results to satisfy the research objectives relative to Appendix 1.3 and 1.4. The notion emerged from the understanding of the data operations and YOLOv5m as object detection concepts discussed in Appendix 1.5. The research approach focused on a series of tasks to demonstrate the practicality of artificial intelligence models in achieving the aim and objectives. These tasks include.

Task-(1) Explore YOLOv5m object detection locating human stabbing objects/weapon artefacts using RWVAD1st dataset to satisfy the aim and objectives via research question-1 to 3 and 5 in section 1.2 and Section 1.3

Task-(2) Evaluate 3DCNN activity recognition ability to determine the generic status of actions between violent and non-violent classes. Following those results, investigations of 3DCNN from an individual perspective established subclasses such as stabbing, beating, fencing, and shooting. The operations satisfied the aim and objectives via research question-1 to 5 in section 1.2 and Section 1.3.

Task-(3) The applicability of YOLOv5m supports the enhancing of 3DCNN for activity recognition relative to model superiority. The operations satisfied the aim and objectives via research question-1 to 6 in section 1.2 and Section 1.3.

3.3.1 Overview of YOLOv5m Activity Recognition Setup:

The operations commenced by downloading YOLOv5m object detection files from the MacBook command line terminal adhering to explicit standards detailed in Appendix 3.3 concerning [108]’s online repository. Following those initiatives, blob analysis techniques applied section regions of interest (ROIs) concerning violent action features from the RWVAD1st dataset. From a sequence of frames, a blob extraction task sections specific features conveying violence. The concept accentuated regions of interest (ROIs) utilising bounding boxes to establish the coordinates of weapons or the activity object’s spatiotemporal location within the data during training stages. Blob analysis proved beneficial as it also reduced the classifications high processed demand on the memory by eliminating redundant image frames in the RWVAD1s dataset. The breakthrough promoted further augmentation procedures that generated surplus data as both models are heavily dependent on enormous volumes of data to generalise the concept of violence and encouraged robust operations. The RWVAD dataset endured pre-processing to dispose of unwanted noise (unwanted ac-

tions/objects) to facilitate YOLOv5m and 3DCNN input standards. The idea reduced the negative processing impact that affects the real-time result convergence. Augmentation operations assisted in generating more data (from initially 200 to 560 raw samples) via rotations left and right 30-90% at 1-5%, shearing ranges between 0.001 to 5%, grey-scale between 1 to 10%, cropping between 1 to 10%, Gaussian noise at 5%, and contrast manipulation between 1 to 10%. The data sectioning reflects 80% training, 20% validation and 20% exempted from the RWVAD1 dataset for testing as the ratio to match ubiquitous approaches within the literature review. The split operations via [246] applied 448 samples for training (112 videos for validation) and 112 videos for testing. For demonstration purposes, ten additional samples facilitated the testing procedures to project the model’s capability from an informed perspective. Those efforts ensured that a significant amount of data facilitated the cross-validation operations to establish the model’s aptitude via training. YOLOv5m operations incorporated an M1-silicon chip Metal 3, Ventura-13.2.1 operating system, 32-processing cores, 64-giga byte Mac-Book Pro computer with the availability of a graphical processing unit (GPU) for real-time operations and central processing unit (CPU) for mundane tasks.

3.3.2 Overview of 3DCNN Activity Recognition Setup:

Unlike YOLOv5m, 3DCNN’s framework classifies action class labels instead of regions of interest processed with bounding boxes during the blob analysis stages. The framework undertakes a different setup encompassing some similarities relative to YOLOv5m. 3DCNN’s operation utilised the PyTorch platform and Python previously mentioned in Section 2.3. Acquiring author [118]’s framework via the online GitHub repository allowed the modification of its layers and hyper-parameter fine-tuning options to meet the processing requirements of the proposed fusion concepts. The operations mirrored cross-validation split procedures to reflect YOLOv5m’s ratio to maintain configuration consistency utilising the RWVAD2nd dataset. Like YOLOv5m, 3DCNN’s architecture required explicit

standards discussed in Appendix 3.4 to encourage efficiency.

3.4 Evaluation Approach for Standalone Models

Validating YOLOv5m and 3DCNN concerning artefact and activity recognition incorporated the confusion matrix, precision, recall and mean average precision (mAP) as measures to estimate the model's classification performance discussed in Section 2.7 and Appendix 2.3. The idea established the true nature of the operations and its robust state at this level. The confusion matrix summarised the volume of predictions accurately and inaccurately classified per action class from a True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN) perspective. Its operations presented an overview of the classes that challenged the models' classification capabilities. The precision was adopted to estimate the correct proportions of all class objects predicted by the model. At this stage, applying the recall aided in quantifying the number of accurate predictions classified relative to all positive classifications generated. Applying the mean average precision provided a critical understanding of the models' classification state with a score projection range between 0–1. Classification scores closer to the range of 1 insinuate high accuracy performance, and low scores suggest the opposite. Analysis of the operations involved observing performance ratings exceeding a 50% margin. The significance of the threshold insinuated how accurate the models were when producing scores above the 50% range.

3.5 Conclusion

The methodology described in this chapter emphasised critical discussions regarding the dataset acquisition process and the importance of applying explicit violent classes to train YOLOv5m and 3DCNN. The significance of grouping new data to satisfy the intricacies of object detection and activity recognition frameworks towards promoting processing ef-

iciency proved crucial to the overall task. A projection of the experimental protocol demonstrated results generation operations utilising the experimental setup linked to the objectives. A clear overview of the evaluation methods at this level established the validation procedures. The methodological approach standardised the objectives to evaluate the effectiveness of individual processing in Chapter 4.

Chapter 4

Evaluating YOLOv5m/3DCNNsl

Understanding the research methodology facilitated the next step, which involved an investigation of the functionality of YOLOv5 and 3DCNN's individual activity recognition processing to generate the necessary data for analysis. Individually evaluating the models' operations fortified the activity recognition concepts previously mentioned regarding actual performance. At this level, the results presented evidence endorsing the proposed fusion strategy through experimental investigations and observations. The chapter is organised into 6-sections to define the outcome with conclusions. A definition of YOLOv5 operations provided context in Section 4.2, followed by 3DCNNsl in Section 4.3. Section 4.1 discusses the experimental setup facilitating high performance. Section 4.4 outlined the results and observations through analysis. Further discussions on the operational challenges between YOLOv5/3DCNNsl fortified the proposed fusion concepts in Section 4.5. Finally, a conclusion in Section 4.6 projects the evaluated objectives.

4.1 Experimental Setup

With the understanding of YOLOv5 and 3DCNN activity recognition, the next step involved creating an experimental setup to evaluate the individual approaches to establish the possibility of alternative processing enhancement. The evaluation approach considers conditions in two phases' detailed via Appendix 4.5 that evaluate the architecture's limitations when discerning real-world altercations. Without essential processing knowledge, the operations can drastically induce possibilities of biased results, further threatening the operation's processing. Phase-1 describes the YOLOv5 experimental setup, whereas Phase-2 follows similar conditions for 3DCNN. Table 4.1 expounds on these conditions.

| # | Experiment Conditions | Impact Experiment Definition |
|---|--|--|
| 1 | No Pre-processing/ No Background Images | Experiment Contains No Data pre-processing or enhancements or background image support. |
| 2 | No Pre-processing/ With Background Images | Experiment Contains No Data pre-processing or enhancements But, contains background image support. |
| 3 | With Pre-processing/ No Background Images | Experiment contains data pre-processing enhancements But, no background image support. |
| 4 | With Pre-processing/ With Background Images | Experiment Contains Data pre-processing enhancements and has background image support. |

Table 4.1: Summary of Activity Recognition Experimental Conditions.

4.1.1 Overview of Experimental Conditions for YOLOv5 (Phase One):

Standards detailed in Appendix 4.5.2 considered the efficiency and feasibility of multiple versions of YOLOv5's architecture following Table 4.1's conditions to validate the rationale driving the model selection. The investigations considered pre-trained and from-scratch operations utilising the RWVAD1st dataset to determine performance superiority. The objective satisfied concepts of recognising violent actions in CCTV videos via research question-1, the impact of data modifications in questions-2/3 and evaluating performance superiority via question-5 in section 1.2.1.

4.1.2 Overview of Experimental Conditions for 3DCNN (Phase-2):

At this phase, the single-level approach proved its processing capability to obtain the generic status of the activity by utilising script configurations. Single-level network operations reduce processing complexity via input layers, which receive the data and output layers that produce the results. Moreover, simulations utilising single-level 3DCNN approaches mirrored Table 4.1 for processing, focusing on action similarity conditions using the RWVAD2nd dataset. To adhere to developmental standards, specifics in [247]’s platform supported the script initiation process previously highlighted in Section 2.3 and Section 3.1.1. The rationale behind Phase-2 fulfilled research questions 1 to 4, focusing on research question 5 in Section 1.2.1 from a generic standpoint.

4.1.3 Summary of YOLOv5m Experimental Setup:

Defining the structured phases aligned the fashioning of 12 pre-trained and 12 from-scratch experiments to evaluate research questions 1 to 5 via section 1.2.1, which exposes framework feasibility and superiority. Through transfer learning investigations, the results disclosed the applicability of YOLOv5m compared to other versions in Appendix 4.5.2 discussions. The operations required script reconfiguration through programming precepts to align the model’s processing with action classes previously specified in Section 3.1 utilising the new RWVAD1st dataset. The application of mean average precision with a subscript threshold of 0.5 denoted the significance of the outcomes. The subscript defined class elements at varying intersections over union (IoU) thresholds above 50% accuracy. The measure identifies accuracy’s of importance and insinuates the prediction outcome. That procedure considered scores above the 50% margin as significant, and scores below the margin are deemed insignificant. The measure quantifies ground truth bounding boxes with predicted bounding boxes as high integers insinuate high performance. Incorporating precision gauges to measure how frequently YOLOv5m accurately recognises positive class occurrences, based on all other instances, is anticipated to be positive. Employing recall metrics accurately

measured YOLOv5m's ability to classify favourable circumstances regarding all the ground truth instances. To validate performance, variations in hyper-parameter options specified in Appendix 4.5.3 are applied to regulate the model's overall performance. Pre-processing each video file to reflect a max duration between 5-15 seconds at 30fps (frames per second) sacrificed critical actions within the data but fostered processing speed in real-time. The operations presented the opportunity to specify multiple objects of interest, suggesting the pre-start of any attack from its inception. Two strategies emerged as proposed methods to achieve this.

Strategy-1: Separating Activity and Weapon Blob During Blob Analysis:

Advantage: Increasing the number and variation of object of interest as blobs for training intensifies classification proficiency to discern the suitable object class.

Disadvantages: Separately specifying violent objects of interest drastically affected the model's processing performance. The approach intensified the complexity of distinguishing multiple regions from unwanted background objects conveying high acceleration, high velocity, and trajectory. Utilising graphic unit processing (GPU) processing can facilitate such issues in previous discussions.

Strategy-2: Combining Activity and Weapon Blob During Blob Analysis:

Advantage: Amalgamating objects of interest during blob analysis to reflect a combined object of heterogeneous traits increased classification performance. The approach reduced computational resources required for classification with fewer objects of significance to generalise during convolution.

Disadvantages: Like Strategy-1, Strategy-2 displayed significant potential for rapid performance with graphic unit processing (GPU) integration.

With knowledge of YOLOv5m experimental setup strategies, Strategy-2, as the chosen option, proved effective in reducing the generalisation complexity and promoting high classification results.

4.1.4 Summary of 3DCNNsl Experimental Setup:

Following the YOLOv5m setup, a projection of 3DCNN approaches enact single and multi-level operations to attain processing efficiency. The idea examined single-level neural networks with a construct that allows single input and output layers. Contrarily, multi-level, like single-level, apply multiple hidden layers to promote efficiency. With insight on multi-level and its limitations detailed in Appendix 4.5.4, the idea entailed fashioning 3DCNN single-level impact experiments to fulfil research questions-4 to 6 in 1.2 using the RWVAD2nd dataset. Integrating surplus data conveying real-life videos in RWVAD2nd dataset aided in deliberately challenging research questions 2 and 3 in section 1.2 like YOLOv5m setup for 3DCNN. Detailed in Appendix 4.5.5, sample variations reflecting 160-videos of 2-violent and 2-non-violent (neutral) class categories supported the project’s current life-cycle and evaluate the generic and multi-class prediction operations. The rationale behind the selection criterion considers the assessment of violent actions emulating heterogeneous and homogeneous properties within the hardware’s processing capacity utilising identical setup standards as YOLOv5m. The prospect maintained configuration consistency to eradicate biased results between YOLOv5m/3DCNN’s operations. Boosting 3DCNN’s efficiency required framework adjustments (adapt 3DCNN to 3DCNNsl’s processing) to recognise sporadic violent actions via transfer learning to encourage robust classification. 3DCNNsl evaluation incorporated analysis of combinations of graphical projections, accuracy, recall, precision, and confusion matrix estimations. During development, two methods emerged in Appendix 4.5.6 that generalised the action status to satisfy research questions 1 to 5 in section 1.2. Those methods are as follows.

- o **Method-1:** Entails the categorisation of the dataset to reflect the generic status from the dataset level.
- o **Method-2:** Entails scripting the generic statuses with subclass labels at the output level. The concept entails categorising the dataset to emulate the generic class status from the dataset level.

Considering the two approaches, method 2 proved its feasibility by regulating complex data categorisations with fewer computations. The idea positively impacted 3DCNNsl operations, reducing the computational dependency on the hardware resources, thus allowing real-time processing speeds.

4.2 How YOLOv5 Operates

Appendix 3.3.1 analysis proved that the spatiotemporal flow of violence combined with weapon artefacts in static images is unlikely. Considering several simulations with object detection, processing incorporating YOLOv5m object detection from an individual processing perspective is required to achieve the overall objectives. Further investigations disclosed [248] utilising YOLO’s base version for activity recognition applying LIRIS’s non-violent action dataset. The authors implemented the YOLO base version as activity recognition using sequences of video frames to denote the spatiotemporal regions of interest detailed in Appendix 3.3.1. They tracked the action’s coordinates across time with a bounding box approach encapsulating regions of interest within each image frame. The discovery presented the opportunity to implement YOLOv5 as activity recognition towards satisfying research question-1 in section 1.2.1. The prospect promoted model integration with reduced memory resources required for operations compared to object detection. The idea extrapolated in Appendix 3.3.1 discusses the distinction between object detection and activity recognition from the input stage with the following limitations.

4.2.1 YOLOv5 Activity Recognition Limitations:

Investigating the actual classification potential for activity recognition enacts YOLOv5 to detect challenging classes purposely selected during inference in alignment with [248]. The operations projected the model's inference capability utilising mean average precision as the performance metric on 8-designated images at a balanced class ratio between violent

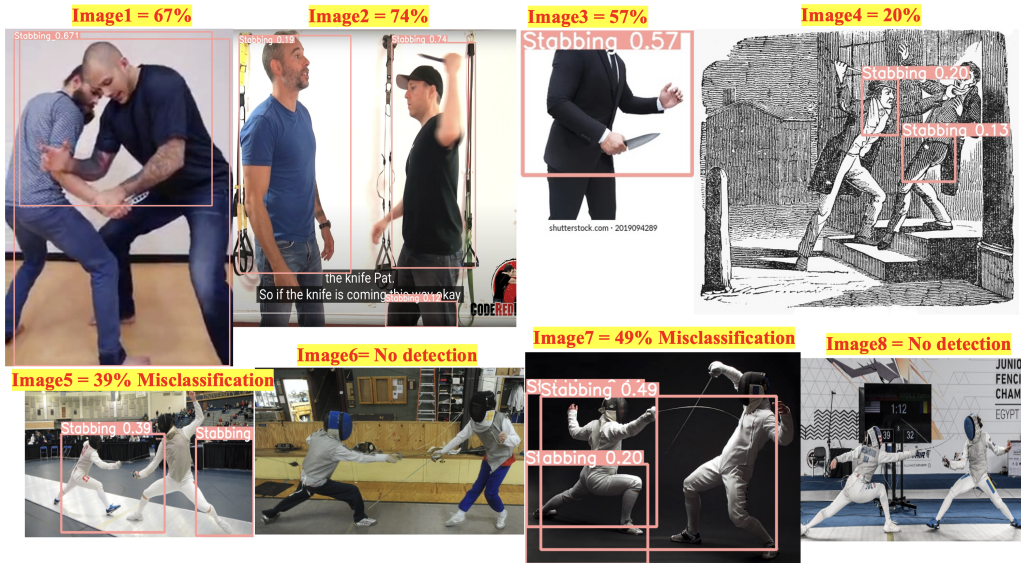


Figure 4.1: Rational for the Proposed Fusion: YOLOv5's Misclassification Performance.

and neutral samples in alignment with Appendix 4.2. The predictions project the designated class's label and its highest classification outcome per image to show its processing capability. YOLOv5m produced 67% on Image-1 in Figure 4.1 as the highest intersection over union outcome, 74% on Image-2, 57% on Image-3 and 20% on Image-4 for the stabbing class. For the non-violent fencing class, YOLOv5 dispensed 39% on Image-5, 40% on Image-7 and no score or prediction for Image-6 and Image-8. YOLOv5 activity recognition operations disclosed evidence of classification limitations, suggesting the need for a more refined approach. Figure 4.1 evidence proved YOLOv5's true capabilities when anticipating

and distinguishing stabbing activities. The results exposed severe class misclassification between the stabbing/aggressor and the victim.

Further analysis validated its processing limitations for activity recognition on two classes with homogeneous actions (stabbing & fencing), that is, actions conveying similar traits with different outcomes. The evidence fortified the notion and critical need for a profound solution. The discoveries of YOLOv5m limitations prompted the evaluation of 3DCNNsl before considering alternative measures.

4.3 How 3DCNN Operates

Insight into the state-of-the-art 3DCNN for activity recognition processing within Section 2.6 aligned its operational prospects similarly to YOLOv5. Compared to YOLOv5 frame-by-frame processing, 3DCNN processes the entire duration of video files to establish the probability of category labels. The model produces probability accuracy scores affiliated with class labels via convolution during inference. With knowledge of 3DCNN's processing in Section 3.3.2, a root folder containing videos representing actions of a suitable volume acts at the datasets for its processing. That specific folder plays a vital role in identifying the relevance of the generic and individual categories relative to violent/non-violent actions. With the concept specified in Appendix 2.6.1, 3DCNN generally takes 16 videos as input and applies the classification operations from a 3-dimensional perspective, outputting the object's height, width, depth, and channels. The input-output process specified in Section 2.6 conforms to the output associating probability score with class labels as the final score. The rationale behind 3DCNN considers the state-of-the-art to identify violence and evaluate its processing to satisfy research questions 1-4 in section 1.2.1. Further investigations into 3DCNN's performance outlined its flaws, given the fragility of human life during violent scenarios, before analysing its effectiveness compared to YOLOv5. The idea eliminated anomalies associated with category misclassifications, presenting knowledgeable

assessments concerning the true processing impact in conditions that reflect homogeneous actions, focusing on stabbing.

4.3.1 3DCNN Activity Recognition Limitation Breakdown:

Appreciating concepts of an ideal process in Appendix 4.3 fortifies the general understanding of good performance in contrast to 3DCNN’s limitations. As 3DCNN is state-of-the-art for activity recognition tasks, a deliberate selection of closely related actions proved necessary to evaluate its discerning abilities in complex lethal scenarios relative to the criticality and fragility of human life. Considering 3DCNN without applying additional tuning or script enhancements as an individual processing model determines its actual processing, like YOLOv5’s operations. The idea considers generic categories reflecting violent/non-violent actions with an individual representation of sub-class statuses and accuracy scores of either stabbing/or fencing. Concluding the processing appraisal meant evaluating architectural limitations by presenting performance results from the previously mentioned experiments. The operations accentuated the previously mentioned alerts projected for the generic action status as violent activity and its subclass stabbing, with a high accuracy of 75%. Simultaneously, 3DCNN flagged the generic class status as neutral or non-violent

human activity and its subclass as fencing at 62% accuracy in Figure 4.2. Although the results appeared promising, the classification processing outlined in Appendix 4.3 to 4.4 suggest the model could not efficiently discern anthropometrics, especially heterogeneous action objects bearing spatiotemporal relevance.

4.4 Overview of YOLOv5m and 3DCNNsl Results Analysis

With insight into the model’s experimental operations, evaluating YOLOv5m and 3DCNNsl’s results proved crucial to establishing the classification potential for pre-empting violence. The appraisal considers the outcomes utilising accuracy, precision, and recall

```

3DCNN Prediction Operations Completed
|||||
*****
Generic Class:___Non-Violent (Neutral) Human Activity Detected ___

Activity Sub-Class: _____( Fencing Activity Detected )_____
Accuracy: Fencing = 62%

|||||
*****
WARNING ALERT: ➔
Generic Class: _____!! VIOLENT HUMAN ACTIVITY DETECTED!! _____

Activity Sub-Class: !!!!! STABBING ACTIVITY DETECTED !!!!!
Accuracy: Stabbing = 75%

```

Figure 4.2: Validating 3DCNN’s Limitations via its Class Status/Accuracy Output.

outlined in Appendix 4.5; this aligns the target objectives with the experiments. Accuracy denoted via the mean average precision with a subscript of 50% (mAP_{0.5}) represents the overall classification metric. The threshold establishes the significance of the outcomes versus poor performance. Additionally, the mean average precision metric with precision and recall projects the model’s effectiveness. The results structure coincides with the research questions, followed by the confusion matrix representations for YOLOv5m/3DCNNsl. The result analysis is as follows.

4.4.1 Summary of YOLOv5m Precision, Recall and mAP:

The analysis commenced by investigating the possibility of research questions 1-3 and 5 in section 1.2.1 utilising YOLOv5m and evaluating its classification state. The notion projected clear performance indications that determined how the research initiatives were satisfied regarding violent objects of interest via the impact experiments. Appendix 4.6-A analysis proved that #8 experiment 21 maintained its superiority overall with a precision score of 0.85, recall of 0.82 and a mean average precision subscript set to 50% at 0.85. The

research questions convey the following.

A. Fulfilling Research Question-1 with YOLOv5m: In all simulations, YOLOv5m demonstrated its ability to recognise violent activity/weapons in CCTV videos above a threshold of 50%. Considering Table 4.1, Appendix 4.4's number 1-8 from scratch and pre-trained processing accentuated all mAP_0.5 scores above the previously mentioned thresholds. The operations fulfilled research question-1's objective in section 1.2 by achieving the highest performance from #8 pre-trained experiment-21 with a precision score of 0.85, recall of 0.82 and a mAP_0.5 of 0.85. Analysis proved that experiment 21 maintained superiority over all other experiments.

B. Fulfilling Research Question-2 and 3 via YOLOv5m Pre-Trained Operations:

The experimental investigation projected the performance impact when utilising pre-processed data and no pre-processing (data without modification) via Appendix 4.6-B. The tasks were analysed from scratch first, followed by a pre-trained perspective to emphasise the contrast between both methods via performance. Analysis of pre-trained operations via #5 experiment-12 proved the fulfilment of research questions 2 and 3 in section 1.2. The operations generated a precision score of 0.77, recall of 0.72 and a mAP_0.5 of 0.75 that superseded all from-scratch operations on every level.

C. Fulfilling Research Question-5 via YOLOv5m: The operations demonstrated superiority by analysing all from-scratch versus pre-trained performance outputs relative to experiments with and without enhancements aligned with Table 4.1. In the previous analysis in section B, research questions 2 and 3 above projected pre-trained operations as superior. However, #7 experiment-9 procedures proved that pre-processing with background image support positively impacted the YOLOv5m classification state via a precision score of 0.83, recall of 0.81 and a mAP_0.5 of 0.83 compared to experiment-21.

4.4.2 Summary of 3DCNNsl Precision, Recall and mAP:

Considering YOLOv5m’s results, 3DCNNsl data analysis engaged similar contexts to project performance increments for 2-neutral and 2-violent classes to satisfy research questions 1 to 5 in section 1.2. Like YOLOv5m, the outputs denote subsections A, B, and C above to reflect the fulfilment of the research objectives in chronological order. Detailed in Appendix 4.6, Table 4.5A, B, and C illustrates individual scores and overall processing accuracy to emphasize the model’s predictions and validate its processing superiority. The integration of the confusion matrix appraised error rates to endorse 3DCNNsl’s classification operations. From this perspective, all experiment subcategories reflect stages of no pre-processing, pre-processing and action similarity data, which convey real-world conditions utilizing action classes discussed in Table 3.2. The analysis commenced in the following section.

A. Fulfilling Research Question-1 Via 3DCNNsl Containing No Pre-Processing:

In this group, evaluations on action results reflect differentiation in characteristics with no pre-processing in 6 experiments focusing on stabbing violence. Implementing the class selection per experiment assesses the processing impact of 2-violent and 2-neutral actions, which satisfy research question-1 in section 1.2. Appendix 4.7, Table 4.5-A presents class distinctions in performance. The performance distinctions relate to combined operations individually and action similarity as a measure to challenge the model’s discerning abilities. The evidence validated the efficiency of individual classification on violent actions containing no pre-processing support exceeding the 75th percentile. Analysis in Appendix 4.7, Table 4.5-A highlighted 0.94% for stabbing as the highest performance rating overall to satisfy the objective.

B. Fulfilling Research Question-2/3 via 3DCNNsl Classes with Pre-Processing:

Considering Group-A’s results, the evaluation considered the pre-processing context to identify robust results via Group-B impact experiments and satisfy the research ques-

tions focusing on the stabbing class. Appendix 4.7, Table 4.12-**B** data emphasized dissimilarity between combined processing, individual processing and violent actions. The enhancements were deliberately applied to challenge 3DCNNsl’s discerning abilities. The results outlined an improvement in individual precision and recall scores, which impacted 3DCNNsl’s overall classification state irrespective of the complexity of the violent/neutral classes. Though stabbing in Appendix 4.7, Table 4.5-**A** experiment #5 projected the highest violent prediction score at 94%, unstable individual precision and recall insinuated that the models attained a high misclassification rate. Stabbing at 0.88 depreciated between 6-13% in Appendix 4.7 and Table 4.12-**B** but achieved stable individual metrics scores with higher overall ratings. The evidence proved that pre-processing strategies positively impacted the operations towards robust results compared to the context of no pre-processing. Further analysis on Group-B disclosed prediction results exceeding the 70th percentile ratio for all violent classes therein Appendix 4.7, Table 4.12-**B** compared to the dissimilarity in Appendix 4.7, Table 4.5-**A**. The evaluation fulfilled research questions-1-3 with further analysis in Appendix 4.7, Table 4.12-**B** to validate result interpretations.

C. Fulfilling Research Question-2/3 via 3DCNNsl Containing Action Similarity: Evaluating the data conveying pre-processing with action similarity occurred by accumulating the necessary insight into Group B’s experiments in previous discussions. The investigations validated performance on stabbing and neutral activities with identical attributes but distinct actions. In this instance, 3DCNNsl processes violence and identical traits of non-violent action to challenge the model’s classification. Appendix 4.7, Table 4.19-**C** emphasized variations in performance between individual classifications and overall accuracy operations. Analysis proved action similarity conditions impacted 3DCNNsl’s prediction state in circumstances where data samples convey homogeneous attributes. Most individual violent classes ranked above the 75th percentile, with fighting and shooting at 100%. Because stabbing sporadically alternates

between fighting, beating and back to stabbing, 3DCNNsl experienced classification challenges. Those limitations reflected sporadic fluctuations between precision (false positive indicators) and recall (false negative indicators), thus suggesting stability issues. Moreover, stabbing generated the lowest score of 63% due to its highly sporadic nature and ability to alternate between beating, fighting, and fencing attributes. The evaluation fulfilled research questions-1 to 3 in section 1.2 with further analysis in Appendix 4.7, Table 4.19-C to validate the result interpretations. With an appreciation for the results, a projection of discussions into the operational anomalies presents the pros and cons of the findings. The discussions are as follows.

4.5 Overview of YOLOv5m/3DCNNsl Discussions

With insight into pre-empting violence via YOLOv5m Phase-1 strategy and 3DCNNsl's Phase-2 method, discussions on overall performance emphasize the technique's true processing impact at this level. The approach is necessary as it accentuates the operational and developmental challenges with risk factors that could impede pre-empting violence in real-world conditions. Finally, further discussions align the mitigation strategy concerning operational and developmental risks to alleviate the challenges entirely or reduce their impact to a minimum. The discussions are as follows.

4.5.1 YOLOv5m Metrics and Confusion Matrix Discussions:

The results implied that the YOLOv5m classification state improved from scratch and pre-trained experiments. The model's precision and recall outcomes increased when more data was applied, as demonstrated in Appendix 4.9.1 to Appendix 4.9.2. The evaluations projected pre-trained operations as the superior approach compared to from-scratch methods. Referencing evidence in Appendix 4 provided the context to validate the processing, which satisfied research questions-1-3 and 5 in section 1.2.1.

4.5.2 3DCNNsl Metrics and Confusion Matrix Discussions:

3DCNNsl analysis presented classification limitations when discerning interchanging violent conditions. Unstable precision and recall scores emphasised the erratic nature of this model type regarding the construct of violence. By fashioning 3DCNNsl’s simulations with multiple types of challenging violent actions, emulating real-world scenarios aids in demonstrating the model’s true abilities while observing its robust performance. The evidence suggested that the stabbing class intensified 3DCNNsl’s generalisation ability due to its alternating gaits, increasing the risks of misinterpretation between fighting, beating, and fencing. 3DCNNsl operations dispensed high outcomes via the pre-processing approach, which satisfied research questions-1 to 3 in section 1.2.1 summarised in Appendix 4.9.2 and Appendix 4.9.3. However, Appendix 4.7 ’s results projected heavy misclassification outcomes utilising the entirety of the videos and its processing ranking compared to YOLOv5m, which employed a frame-by-frame operation. With insight into the metric performance for YOLOv5m and 3DCNNsl, a re-evaluation of the research questions proved necessary to assess the fulfilment of the objectives. The re-evaluation procedure reflects the following.

4.5.3 Research Question Discussions Evaluating YOLOv5m/ 3DCNNsl:

With an idea of the previous metrics, section 1.2.1 re-evaluated the fulfilment of the research objectives to demonstrate framework feasibility.

1. Fulfilling Research Question-1; (Can violence/weapons be recognised?):

The tasks facilitated the recognition of violent activity and weapons (bladed instruments, knives) in CCTV videos by conducting YOLOv5m (from scratch/pre-trained) and 3DCNNsl (single and multi-level) experiments. The thesis proved this by fabricating conditions to challenge and evaluate the model’s interpretation of violent alterations from a real-world perspective. Convolution models facilitated the prediction of violence by processing raw data via the input stage during training operations.

Achieving the input objective encompasses feeding raw data through programmable scripts fortified by Appendix 4.10.

2. Fulfilling Research Question-2; (What is the Impact/Data Modification?):

Understanding the results previously discussed, the operations evaluated the impact of pre-processing and no pre-processing enhancements to satisfy research question-2 in section 1.2.1. Appendix 4 's results proved the models' capability to pre-empt violence and establish its significance relative to non-violent behavioural patterns. Although YOLOv5m/3DCNNsl established dis-similarities in anthropometric variations, their operations dispensed high fluctuations via previous metric evaluations. Action similarity operations dispensed a 7% decrease in performance, yet applying a surplus of violent samples further increases its potential. The evidence proved the positive impact on performance regarding pre-trained, pre-processing or pre-processing with action similarity techniques via Appendix 4 's results. The continuous fluctuation indicated the model experienced challenges during individual classification, specifically via the stabbing class. The high complexity of violent actions and the small sample size applied in training influenced the stability of precision and recall processing. Further details in Appendix 4.10 fortify the previous notions.

3. Fulfilling Research Question-3; (What is data impact if sample increase?):

Evaluations into the impact of sample size and parameter increase determine realistic performance thresholds by establishing the ability to pre-empt violence in CCTV videos via the pre-processing context. By increasing the data volumes, the evidence proved that the models dispense steady signs of performance increase in accuracy, precision, and recall values irrespective of the fluctuating individual scores. With the ability to recognise violence in CCTV videos, the operations assessed 3DCNNsl's performance impact if the sample size increased. Appendix 4.13 to Appendix 4.16 proved that if the volume of violent samples increased during training, the overall accuracy and individual performance improved above the 80th percentile. More-

over, the individual model experienced challenges when processing data containing homogeneous action attributes, specifically between stabbing and fencing.

4. Fulfilling Research Question-4; (Can actions be generalised?): With insight into research question-3 above, the operations satisfied research question-4 in section 1.2.1 by categorising violence/neutral non-violent activities within the configurations. Because the operations combined the outcomes of 2-architectures, the programmed scripts applied the classification once in the final layers, which emulated Figure 4.2 results. A limitation emerged where the simultaneous implementation of the programmable configurations for YOLOv5m and 3DCNNsl in the final layers hindered real-time results. The previously mentioned issue introduced unknown risks resulting in slow processing memory deprivation regarding the hardware’s processing capability. The idea behind the programmable scripts provided auxiliary classification support aiding in distinguishing the generic and subclass status between non-violent/violent actions in any scenario.

5. Fulfilling Research Question-5; (Can model superiority be determined?): Understanding research question-4’s result, an evaluation of the models’ processing superiority determined the status of research question-5 in section 1.2.1. Detailed in Appendix 4.16, YOLOv5m proved its processing superiority over 3DCNNsl outcomes regardless of the changes in environmental conditions relative to the gradient luminosities and environmental scenery. Although YOLOv5m/3DCNNsl proved incapable of truly discerning homogeneous attributes of violent actions, the misclassification evidence therein in Appendix 4.17 validated YOLOv5m’s superiority over 3DCNNsl when processing fencing27.avi. Since action similarity conditions complicated the classification, YOLOv5m identified 3- scenarios via frames #52, #55 and #56 that projected fencing’s status towards fulfilling research question-5 in section 1.2.1. Moreover, the evidence validated the need for the proposed fusion concept, as both models succumb to complexity in the action similarity conditions. In that sense, auxiliary

support through the proposed fusion approach can mitigate misclassifications and maintain robust accuracy performance regardless of the conditions.

4.5.4 Summary of Operational Challenges:

With crucial insight into model superiority, projecting several operational challenges requiring attention proved crucial to promote efficiency. To commence the discussion, a projection of YOLOv5m challenges with a proposed solution emphasised context that limits the impact of its issue to a minimum. Subsequently, 3DCNNsl disclose its additional challenges detailed in Appendices 4.17 to 4.19. Investigating YOLOv5m's processing exposes its framework intricacies from an object detection and activity recognition viewpoint. The crucial insight is derived by fortifying knowledge in both areas through auxiliary artificial intelligence courses. The strategy entails devoting considerable time towards gaining practical experience via programming. At this level, emphasis on the challenges influencing YOLOv5m's operations and its mitigation strategy proved the model's effectiveness, with additional details in Appendix 4.18. The overview demonstrates further issues impacting the operations or the project's timeline. Following the identical operations utilising the same eight test images in section 4.2.1, is a nuance of the YOLOv5m incorporating separate objects, solidifying the processing effectiveness notions regarding classification limitations. Figure 4.3 accentuates multiple instances of YOLOv5m processing limitations regarding class misrepresentations within the data. YOLOv5m on Image 1 generated five inaccurate predictions, two stabbing outcomes at 19% and 32%, two knife weapon predictions at 22% and 32%, and one aggressor class at 10%. The model processing separate objects on Image 2 created ten misclassifications. Those outcomes reflect two victim classes at 12% and 9%, two stabbing at 32% and 0.19, three aggressors at 7%, 9%, and 34% and three knife weapon objects at 10%, 15%, and 32%. Image three disclosed more promising outcomes at four misclassifications regarding the aggressor class at 19%, knife deploy at 39%, knife weapon at 49%, and stabbing at 20%. Image 4 declined in processing with three misclassifications;

this showed knife weapon at 15% and two stabbing class instances at 12% and 9%. Image 5 dispensed one inaccurate victim outcome at 6%. Image 6 displayed two incorrect stabbing instances at 19% and 14%. Finally, Images 7 and 8 produced no results to insinuate an insignificant classification attempt. The outcomes on Images 7 and 8 provided room for open interpretations where it failed at the task or made a correct classification as violence is absent in the fencing class. The following discussions elaborate further on the limitations and the possibility of misinterpretations.

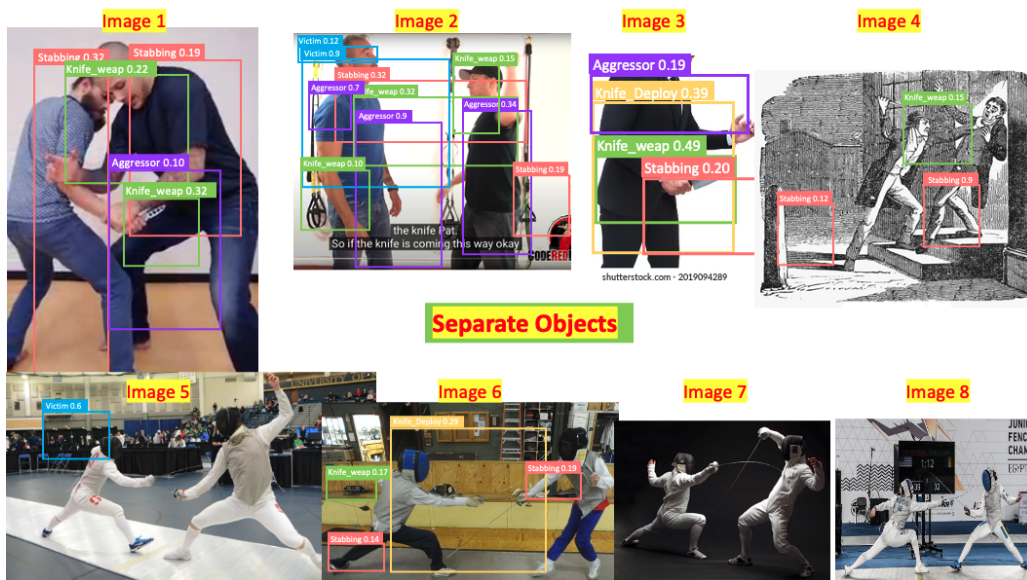


Figure 4.3: YOLOv5m Separate Object Processing Complications.

A. YOLOv5m Processing Limitation on Small Objects Discussions: YOLOv5m revealed multiple flaws in processing small spatiotemporal objects containing high sporadic acceleration, trajectory, and velocity. Violent scenarios conveying minute weapons proved problematic for the model to classify seconds before the climax of a lethal attack. The analysis projected increased processing complexity during inference if diminutive weapons undergo separation from each stabbing action object. The application of small

human body parts and artefacts representing the stabbing class object/s contributed to the intricacy of the classification operation. The processing misconstrued the relevance of essential objects with background elements, thus hindering the object’s classification. Figure 4.4 highlights the intricacy issue of specifying multiple-minute artefacts separately to represent the actions of stabbing scenarios.

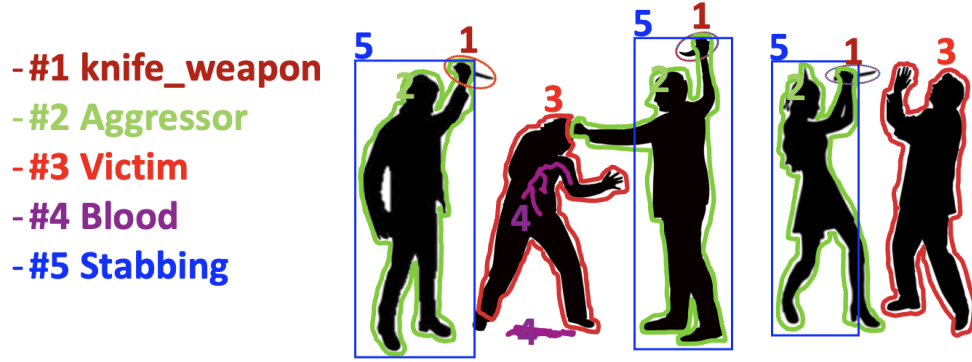


Figure 4.4: YOLOv5m’s Complications with Minute Objects.

B. *Solution to A: YOLOv5m Processing Limitation on Small Objects: With insight into YOLOv5m’s processing, a contribution emerged by proposing combining weapon artefacts and the aggressor with relevant objects to suggest the disposition of a stabbing scenario as a prominent solution. The proposed idea reduced the complexity between generalising overlapping and correlated human features during training. The operations projected a stable increase in performance by amalgamating each class during inference stages, as illustrated in Figure 4.6. The concept suggests that if an aggressor is present, a knife and hand posturing class are also available to contribute to pre-empting the stabbing scenario, thus enhancing its classification. The proposed rationale established clear distinctions of bounding boxes between the focus stabbing class, aggressor, victim, and other artefacts. The idea increased the combined objects’ significance relative to violence as one class with supporting categories. To validate the

theory, further experiments incorporating the same eight image test samples applied in section 4.2.1 confirmed YOLOv5m's processing utilising the proposed combining of objects approach. Figure 4.5 evidence proved the processing superiority of the proposed amalgamation of objects approach compared to separately processing each object to

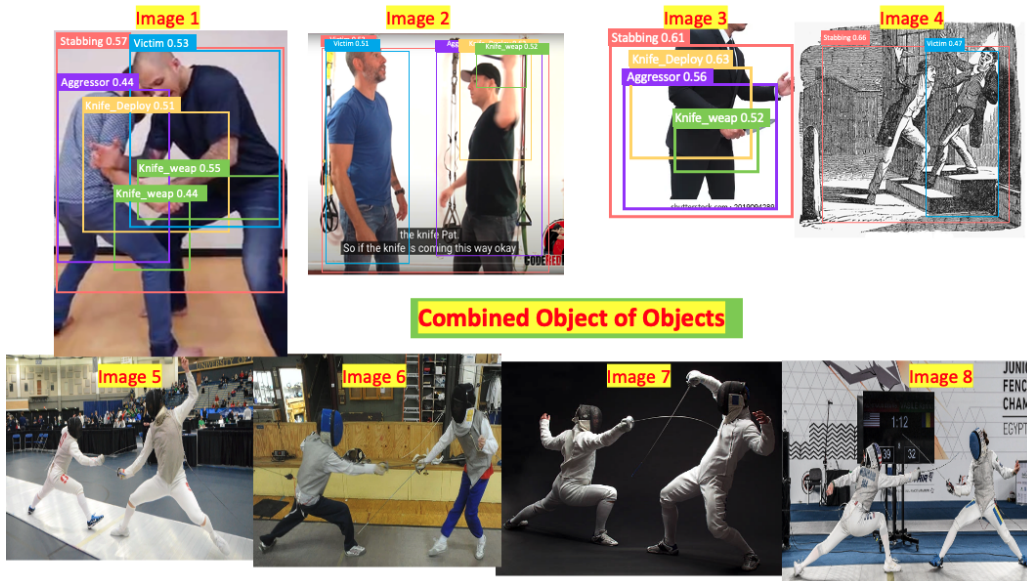


Figure 4.5: YOLOv5m Separate Object Processing Complications.

reveal violence. YOLOv5m produced six outcomes with four accurate instances. YOLOv5m correct classifications on Image 1 produced the stabbing class at 57%, victim at 53%, knife deployed at 51%, and the knife weapon at 55%. The two misrepresented classes reflect aggressor at 44% with an additional prediction for knife weapons at 44%. Image 2 created five accurate instances at 53% for the stabbing class, the aggressor achieved 55%, the knife deployed produced 62%, the victim achieved 51%, and the knife weapon at 52%. The evidence in Image 3 showed four accurate classifications where stabbing achieved 61%, knife deployed produces 63%, aggressor at 56%, and knife weapon at 52%. Image 4 dispensed two instances, a correct prediction for stabbing at 66% and an

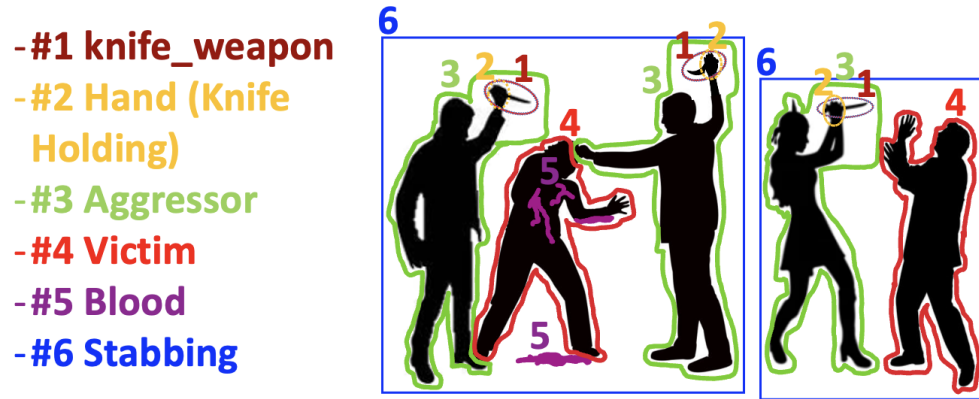


Figure 4.6: Combining Classes to Reduce YOLOv5m Class Misrepresentation.

inaccurate reflection of the victim at 47%. Images 5 to 8 produced no results to suggest the model's decisions on violence in a non-violent domain. The proposed approach decreased its class misclassification issues that harmed the operations as the object's size increased, thus positively impacting inference. As a by-product, the concept added robustness in processing scenarios containing minute weapons with sporadic acceleration, high velocity, and object trajectory.

4.5.5 Summary of 3DCNNsl Operational Challenges:

Discussions on 3DCNNsl operational challenges similar to YOLOv5m utilising the same data nullify operation discrepancies. Because YOLOv5m's issues were remedied (raw data acquisition, pre-processing, applying two classes), the operations avoided additional alterations at this stage. With an unbalanced layer configurations issue, the original 3DCNN struggled to distinguish between fencing/stabbing scenarios via action similarity. To emphasise 3DCNN's processing limitations, a nuance of its processing issues in Figure 4.7 displayed high classification performance at 100% for all predicted classes (WalkWithDog 100%, Knitting

100%, Fighting 100%, and Shooting 100%). 3DCNN's results utilising the same approach

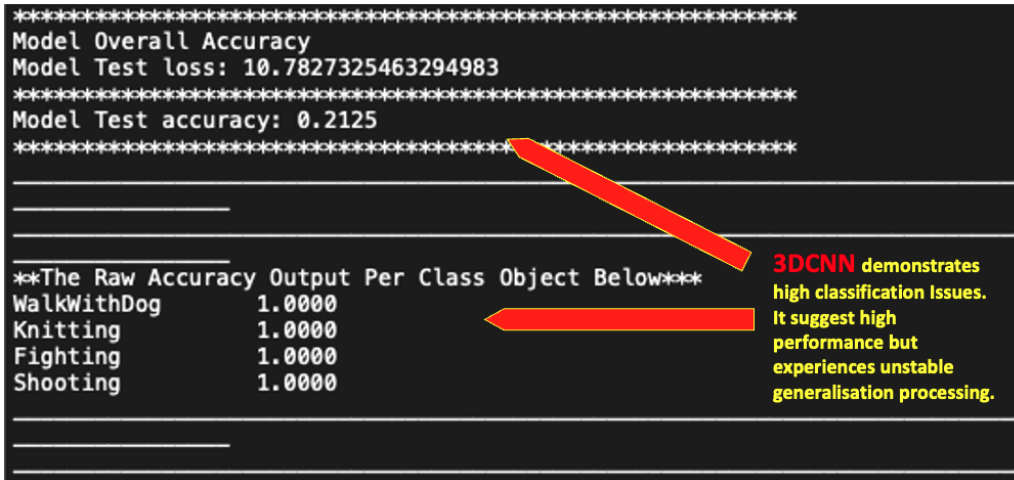


Figure 4.7: Combining Classes to Reduce YOLOv5m Class Misrepresentation.

disclosed in Section 4.1.4 displayed a low accuracy score of 21% and a high test loss at 10.7827% suggesting an insignificant classification performance. The low test loss outcomes insinuate robust performances. Compared to YOLOv5m, 3DCNNsl cannot facilitate the

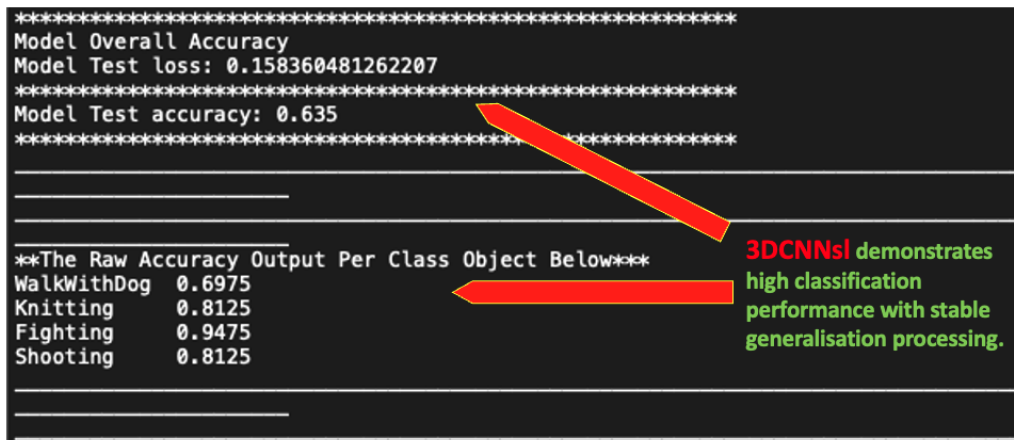


Figure 4.8: Combining Classes to Reduce YOLOv5m Class Misrepresentation.

coordinates of objects via bounding boxes; however, it's structure enabled category labels and scores to insinuate its prediction state, utilising the entire duration of the videos.

3DCNN's classification operations improved employing the remodelling strategy via 3DCNNsl in Figure 4.8. The approach utilising the identical concept in Section 4.1.4 generated realistic outcomes reflecting WalkWithDog at 69%, Knitting 81%, Fighting 94%, and Shooting 81%. The model projected its robustness with an accuracy of 63% and a low loss outcome at 0.15%, taken from Appendix 4.7.2, Table (A) 3DCNNsl impact experiments with no pre-processing. The evidence of 3DCNNsl compared to 3DCNN proved its processing stability at this level. Like YOLOv5m, discussions accentuated 3DCNN's main challenges and mitigation strategy to promote efficiency with additional factors in Appendix 4.19 to validate the approach of violence recognition. 3DCNNsl's discussions are as follows.

- A. **3DCNN Heavy Resource Demand Discussions:** The significant memory demand of 3DCNN severely hampers its processing capabilities. Despite the CPU functioning, 3DCNNsl encountered processing latency during training, stretching the durations from hours to weeks. This issue was particularly pronounced when processing raw violence containing parallel accelerated motion. In contrast to YOLOv5m, 3DCNNsl's configurations reverted to CPU processing due to a platform package mismatch that affected PyTorch on new silicon chip GPU-compatible MacBook Pro computers.
- B. *** Solution to A: 3DCNN Heavy Resource Demand:** Reducing the data dimensionality standards from 414x414 to 320x240 mitigated 3DCNNsl's issues discussed in (A). The operations considered pre-processed salient action frames regarding violence, only emulating variations in background scenery to fortify its effectiveness. The approach significantly reduced the data volume for processing with prolonged file durations from 5-10 minutes to 5-15 seconds per video file. 3DCNNsl produced smoother computational outputs, reducing the sporadic interruptions caused by the MacBook Pro's CPU. Though the operations projected a significant reduction in training and inference times, the future intentions consider GPU support to eradicate 3DCNNsl sporadic

lagging. 3DCNNsl reduces computational overload by adopting additional processing support, facilitating action similarity irrespective of the parallel action scenarios. Like YOLOv5m, monitoring tasks engage the PyTorch community for integration strategies towards solving the CPU reverting issues within the configurations between the MacBook m1 MPS device and Torch-vision.

4.6 Conclusion

Chapter 4 underscored the importance of the experimental setup in validating the performance outcomes. This setup, which served as the plot to generate valid results, confirmed the rationale for alternative measures to foster robust processing. The setup disclosed YOLOv5m/3DCNNsl's limitations via multiple experiments discussed in Appendix 4, thus fulfilling the objectives. Those operations accentuated YOLOv5m's superiority over 3DCNNsl from the frame-by-frame level. Given the task of violent activity recognition from an individual perspective at this level, comparing the outcomes with other state-of-the-art (SOTA) approaches in the literature proved impractical due to the type of data applied to solve the aim and objectives. Most solutions achieved activity recognition in a non-violent sense. Several efforts applied violent activity recognition, disclosing the activity in motion and conveying the lethal impact instead of the pre-stages and generic status of such actions. In multiple scenarios, the data used demonstrated easily identified violent situations on complex hybrid models compared to investigating the complexity of the actions in homogeneous conditions. The significance of reviewing the literature's solutions acts as a guide as opposed to formulating comparisons. Because of the extent of the investigations, the results within the appendix act as an overview of the operations to solidify the approach if further details are required. The experimental operations primarily focus on the possibility of achieving activity recognition and the model's limitations. The chapter proposed several methods towards promoting further classification enhancement by combining weapon artefacts/stabbing classes for YOLOv5m and reducing 3DCNNsl processing

complexity and the model's limitations. The idea relates to lessons learnt discussions in Appendix 4.20. The effectiveness of the actual performance encouraged proceedings to Chapter 5. The following chapter emphasises theories behind the proposed fusion and its effectiveness towards violence.

Chapter 5

Merging 3DCNNsl/YOLOv5m for Robust Violent Activity Recognition

Chapter 4 investigated the automation of violent activity recognition centred on video analysis using 3DCNNsl prediction models. It also investigated YOLOvm5 concerning the recognition of relevant objects in violence. This chapter outlines the need for the proposed fusion, which builds on the previous operations discussed in Chapter 4. It explores fusion strategies using two models to enhance the system's prediction accuracy and satisfy research question-5's tasks of superiority amid models in section 1.2.1. The concept investigated 3DCNNsl/YOLOv5m activity recognition to evaluate its effectiveness compared to the proposed fusion, mainly on 2-samples. The chapter investigates the complexity of fencing and stabbing as actions relatively challenging to discern as opposed to utilising easily classified actions with high-performance outcomes. The idea added a measure of significance to the proposal's rationale, emphasising the classification effectiveness using challenging action samples. Chapter 5 comprises seven sections that justify fusion's activ-

ity recognition prospects via a decision-level approach. Section 5.1 outlines the motivation for the fusion concepts concerning 2-fusion scheme scenarios. Following the motive, the operations evaluated the first fusion scheme using only 12-frames-per-video samples to establish its effectiveness in Section 5.2. Section 5.3 highlights fusion scheme-2 concerning the efficacy of processing the entire video using surplus data instead of applying 12-salient frames. Following those sections, the analysis entails results of all operations in Section 5.4 and 5.5 to evaluate the proposed fusion robustness. Section 5.6 provided further discussions into the effectiveness of fusion scheme-1/2, followed by the conclusion in Section 5.7 to end the chapter.

5.1 Motivation for the Fusion Enhancement Approach

Appreciating the processes and results demonstrated in Chapter 4, Chapter 5 discusses the rationale concerning fusion to support the operations towards encouraging robust results. The context of fusion has potential in various applications concerning processing as it joins the computations of multiple classifier models to provide a final output [249] and [250]. Several classifier-based and arithmetic fusion methods concerning [251], [252], [253], [254], [255], and [256] exist to generate the outcomes. With knowledge of the individual models' processing limitations in Chapter 4, the operations focused on computational simplicity, which targeted the following fusion approaches as a plot to reduce the computational load to further enhance the effectiveness at the final stages.

1. **Arithmetic Averaging Fusion:** Computes an average on multiple values generated by multiple classifiers at the score level to produce a single value as the final prediction outcome.
2. **Decision-Level via Majority Voting:** Offers even weights to the decisions generated by several classifiers using a majority voting class-association prediction

system, which records the highest number of votes as the final result.

Arithmetic Averaging Fusion: The score-level arithmetic averaging fuses multiple classifier outcomes, thus creating a final value reflecting its overall prediction. It generates a final value per [254], [255] utilising simple computations that reflect summation, average, median, minimum, maximum, and the product. The approach achieved 99.25% utilising a 3-dimensional 3D-face-ear for human recognition on face recognition grand challenge via the University of Notre Dame collection F, 3Dfaceear datasets [257]. Those authors solved unimodal bio-metric systems and 2D-bio-metric problems regarding occlusion and illumination. The score-level fusion discussed in [258] demonstrated its capability in 4-stages to recognise human activities. The authors used pre-processing conversions with 2-hybrid classifiers and score-level fusion strategies to generate 95% accuracy via the UCF-ARG dataset. Other works exploited the score-level approach to build a combiner classifier to facilitate multi-modal biometric user authentication [259]. They used bio-metric sensors as input, and their fusion algorithm achieved a true positive rate of 99.15% and a true negative rate of 99.28%. Although arithmetic at score-level proved effective in previous discussions, [252] outlined the approach as lacking robustness where the scores vary by a logarithmic factor, thus intensifying the computational load with possible risks that add classification complexity to the current model's operations at the output stages.

Decision-Level Fusion: With knowledge of the arithmetic score-level processing approach, the operations investigated the possibilities of decision-level fusion in alignment with [260]. Those authors achieved 98.22% utilising CaffeNet and GoogleNet by amalgamating features in the model's last layer for VHR image classification trained on the ImageNet dataset. Others, such as [261], examined YOLOv4 activity detection and 3DCNN fine-tuned decision level on UCF crime and the Microsoft COCO dataset. The authors achieved suspicious activity recognition circumventing the In-

ternet of Things in intelligent city security with 94.21% accuracy. Moreover, [262] employed a fusion of depth camera and inertial sensors by applying feature-level fusion techniques and decision-level fusion to combine the outcomes from 2-classifiers. The authors achieved 2-23% recognition rate improvements on the Berkeley MHAD dataset to improve human action recognition. Because of the decision-level's fusion efficacy in the previous discussion, the operations investigated the fusion approach to pool YOLOv5m activity recognition and 3DCNNsl computations at the output stage. The idea promotes fusion simplicity, creating a mutual evaluation process of the action classes. The decision-level strategy approach followed [263] multi-modal driving behavioural algorithm at 96.57% accuracy rating, [264] localisation activity recognition at 98.2%, and [265] decision-level facial recognition fusion at 81%. Those concepts rely heavily on majority voting systems, which process outcomes to devise a routine multi-modal decision relative to [266], [267], and [268]. Although decision-level operations discussed in [269] and [270] deteriorate while representing inconsistent, ambiguous statistical data, it reduces computational loads as per [271], this improves the effectiveness of the current model's processing in scenarios with fewer class categories. Utilising decision level at this level creates a single outcome generated from YOLOv5m activity recognition and 3DCNNsl by incorporating majority voting techniques. The idea fuses all associated decisions to represent binary outcomes reflecting violent or non-violent actions. Facilitating decision-level fusing through programmable configurations merged each model's (YOLOv5m/3DCNNsl) final layer of several decisions to generate a robust result as the outcome. Figure 5.1 below illustrates the proposed decision-level fusion concept by training the model at step 1 and processing the actions through steps 2 to 6. The proposed fusion above encompasses 6-stages, which convert video data from the input stage at steps 1 to 5, utilising a layered convolution combined with decision-level configurations to achieve an outcome. The following details itemise fusion's stages and its critical operations.

Step-1 - Model Training: The dataset acquisition process explored social platforms to accumulate a substantial volume of data containing the pre-start to violence concerning shooting, stabbing, fighting, and beating. The models assisted in establishing action differentiation during the final classification stages combined through programming for training operations.

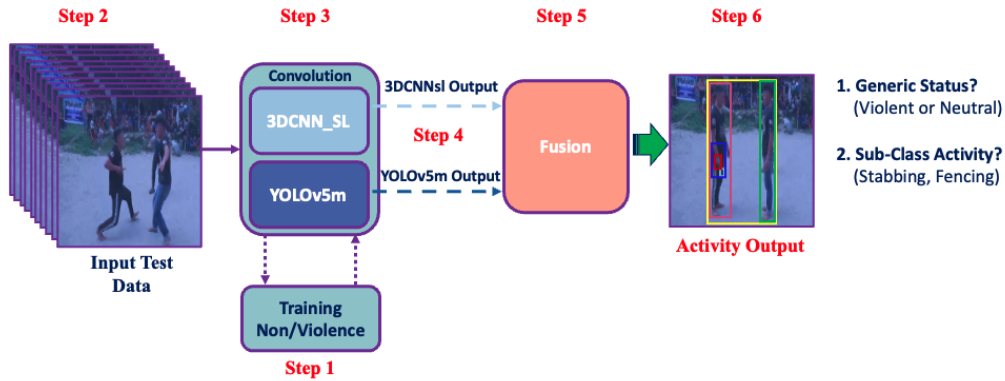


Figure 5.1: Illustration of Fusion Operations via Step-1-5.

Step-2 -Pre-Processing: Following Step-1, the operations engaged software (Robo-Flow and HandBreak) to align the raw data with the model’s input standards. At this stage, software tools manipulated the testing data to emulate no pre-processing, pre-processing, and action similarity experiment conditions as a measure to validate the model’s effectiveness.

Step-3 –YOLOv5m/3DCNNsl activity recognition Convolution: Following Step-2, Step-3 presents the data for convolution operations. By applying convolution, the models accept the data and generalised distinctions between violent/non-violent actions across several processing layers. 3DCNNsl modifications allow the generations of several outcomes relative to the generic status and the subclass label stage in one go to achieve robust

classification at this stage.

Step-4 –YOLOv5m/3DCNNsl activity recognition Output: Following Step-3’s convolution, YOLOv5m activity recognition/3DCNNsl generates individual outcomes at Step-4, emulating its classification during inference. The convolution processing dispensed results projecting non-violent/violent categories with sub-class labels (fencing or stabbing).

Step-5 –Fusion: Following Step-4, Step-5 encompasses a decision-level fusion; this aggregates class associations from Step-4 and applies a majority vote process, which establishes the outcome.

5.1.1 Overview of the Fusion Rationale:

The rationale for the proposed fusion support limits the crucial misclassifications exposed during the individual assessment of YOLOv5/3DCNN towards pre-empting violence. Because of those misclassification issues, the operations applied YOLOv5m activity recognition and 3DCNNsl with decision-level support to create an enhanced outcome. 3DCNNsl’s processing establishes the actual nature of the activity concerning the generic status (whether the actions are violent or not) and the action’s subclass (if it is violence, what type of violence). YOLOv5m activity recognition substantiates 3DCNNsl’s processing by acting as a supporting plot that establishes the activities’ resonance. The idea mitigates misclassification challenges relative to fencing27.avi in Appendix 4.40, which demonstrates classification instances of the non-violent sample as violent actions. The outcomes proved 3DCNNsl unreliable operations in the 3DPred column and YOLOv5m activity recognition via the YoloPreds output column. Those operations outlined episodes of the model’s classification effectiveness on complex real-world samples concerning identifying the action’s correct nature via pre-processed data. YOLOv5m activity recognition experienced adverse effects when processing violence, mainly in circumstances where the actions portrayed

high acceleration trajectory and velocity. By applying alterations towards amalgamating YOLOv5m activity recognition and 3DCNNsl output through decision-level fusion, the model drastically improved its robust accuracy regardless of the class of violence. Another fundamental factor involves reducing programming challenges via developing decision-level combined with YOLOv5m activity recognition/3DCNNsl. Decision-level via fusion facilitates high and low data volume developments by promoting fewer integration challenges. The idea maintains processing consistency, which enhances the value of its predictions.

5.1.2 Definition of Decision Level Fusion Protocols:

It is essential to define conditions that consider the outcome of 3DCNNsl/YOLOv5m activity recognition via class association to achieve a desired outcome regardless of the action type and luminous intensities of objects in the image scenery. At this level, the decision-level protocols' intentional design bears partiality towards positive categories, which suggests the presence of the generic status (is the action violent?) and its subclass (type? stabbing?). Promoting bias for the positive classes as an intentional flag of violence in

| # | YOLOv5m Outcome | 3DCNNsl Outcome | Decision-Level Protocol Outcome | |
|---|------------------|------------------|---------------------------------|------------------|
| 1 | Positive | Positive | YOLOv5m | Positive |
| 2 | Positive | Negative | YOLOv5m | Positive |
| 3 | Negative | Positive | 3DCNNsl | Positive |
| 4 | Activity Unknown | Positive | 3DCNNsl | Positive |
| 5 | Activity Unknown | Negative | YOLOv5m | Activity Unknown |
| 6 | Positive | Activity Unknown | YOLOv5m | Positive |
| 7 | Negative | Activity Unknown | 3DCNNsl | Activity Unknown |
| 8 | Activity Unknown | Activity Unknown | YOLOv5m | Activity Unknown |
| 9 | Negative | Negative | YOLOv5m | Negative |

Table 5.1: Decision Level Protocol Operations.

negative cases act as a redundant contingency measure and allow validation procedures by a manual operator towards reducing the occurrence of lethal actions. The operations favoured the previous idea of aligning the fundamental objectives behind the research pro-

positional, which pre-empts violence as a focus to reduce its impact using YOLOv5m as a bias technique, due to its robustness in Chapter 4. Incorporating additional validation monitoring support reduces the impact of violence, thus fortifying the classification operations. Table 5.1 introduces possibilities to establish the outcomes. In previous discussions, those operations opted for an approach that applies a biased operation towards the positive (violence exists) class and activity unknown (actions the model is not trained to identify) outcomes over negative (non-violent action) outcomes. The idea differentiates the pre-empted attacks in violent and non-violent actions in any conditions. Table 5.1 Decision-level reasoning shadows the following conditions to create results.

1. if **both models dispensed positive** (produced violent outcomes) predictions, decision-level fusion suggests that **the outcome is positive**.
2. if **one of the models predicted a positive** (produced violent outcomes) and the other produced a **negative** (produced negative/inaccurate result) output, decision-level fusion **selects only the positive outcome** to suggest violence as the outcome.
3. If one of the models predicts a **negative** (produced negative/inaccurate result) and the other predicts a **positive** (produced violent) outcome, decision-level **selects the positive** prediction as the outcome.
4. If one model predicts **activity unknown** (produced unknown outcome), and the other predicts **positive** (produced violent outcome), decision level fusion selects the model with the **positive prediction** as the outcome.
5. If one model predicts **activity unknown** (produced unknown outcome), and the other predicts **negative** (produced negative/inaccurate result), decision level

fusion selects the **activity unknown** prediction as the outcome.

7. If **both** models **produced activity unknown**(produced unknown) as outcomes, decision level fusion **implies the outcome is activity unknown**.
8. If both models predict **negative** outcomes (produced negative/inaccurate results), decision-level processing infers the outcome is **negative** and reiterates its operations to reproduce an appropriate outcome. In this scenario, the process recycles, reiterating the fusion tasks to avoid errors. The possibility of protocol #7 proved rare. Thus, the operation continually implies appropriate outcomes. Prospectively, further configurations applied mitigated such cases as a contingency strategy should this situation become apparent.

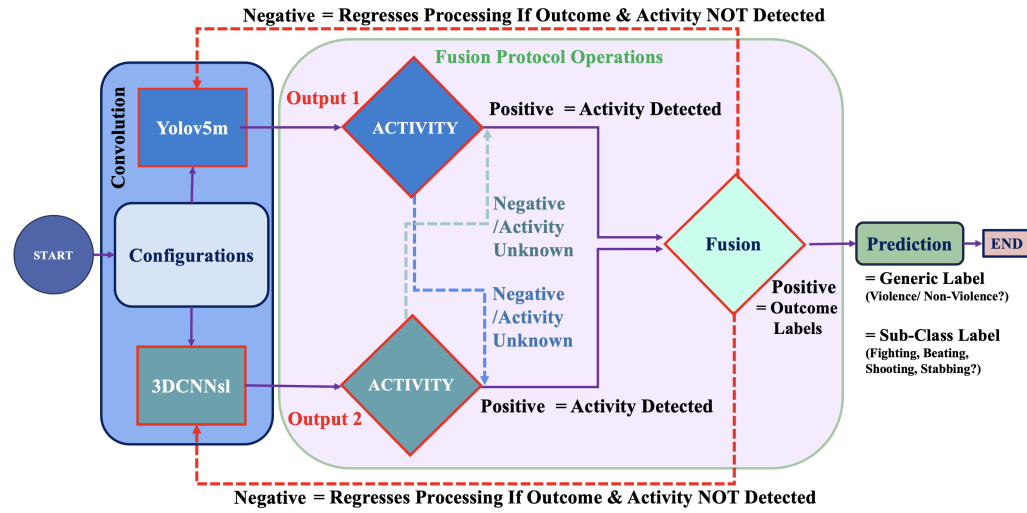


Figure 5.2: Decision Level Fusion Protocol Operations.

Figure 5.2 illustrates the operational concept of the proposed fusion protocol above as a visual nuance of the conditions and the expected outcome.

5.1.3 Experimental Setup of Fusion:

Implementing the fusion operations meant considering the final layers processing of 3DCNNsl and YOLOv5m activity recognition to achieve optimal performance by employing 2-sets of decision-level constraints. The rationale for fusion constraints aids in establishing the decision level's practicality, which positively impacts the computing resources, especially in instances emulating sporadic increment and decrement of input data. Initialising the processing detailed in Chapter 4, commence the proposed fusion utilising the following protocol conditions.

1. **Fusion Scheme-1:** Decisions of the activity recognition models (YOLOv5m and 3DCNNsl) from selected frames (first 12 frames of pre-stages of violence) are fused individually at the frame level using the decision-level strategy explained above. A majority voting process among the resulting 12 decisions at the frame level generates a final decision for the video/activity. The idea concluded the operation's processing impact using only 12 frames to represent a video and simulate scenarios involving a decrement in input data.
2. **Fusion Scheme-2:** Like Fusion Scheme-1 decision-level processing, the operation at this stage considers the entire video duration at a frame rate of 30fps(frames per second) with an interval window of 10-12 seconds of data. The idea concluded the operation's processing impact by utilising more data to simulate scenarios involving an increment in input data to represent a video instead of only 12-action frames.

The difference between the fusion scheme conditions is that Fusion Scheme-1 evaluates

the model’s effectiveness when processing fewer data, and Fusion Scheme-2 examines the model’s impact using more data. The operational difference between the fusion schemes outlines their ability to provide robust outcomes regarding computational efficacy consistently. The following provides a further breakdown of the fusion schemes’ intricacies to accentuate the fusion operations.

5.2 Fusion Scheme-1: Utilising 12-Frame Processing

Appreciating the fusion motivation above, a projection of Fusion Scheme-1’s concept concludes research question-5 model superiority objective in section 1.2.1. YOLOvm5, applied as an activity recognition model, recognises violence by assigning violent actions to individual frames in a video. Contrarily, 3DCNNsl architecture recognises violent activity by analysing the entire duration of the video and assigning violence to multiple frames to achieve its violent classification outcome. Both operations are combined to evaluate Fusion Scheme-1 by re-configuring 3DCNNsl’s conventional whole video processing to mirror a similar frame-by-frame approach like YOLOv5m activity recognition. The rationale concerning the frame-by-frame idea allows the simultaneous grouping of sequential frames to establish class associations per model. Specifying robust class association outcomes aids the smooth transitioning of decision-level operations at this level. The approach integrates only 12 salient frames, suggesting the pre-start (mainly leading to violence) during the (violent start) attack period per video, focusing on violence as the positive class. The operations strongly considered the approach in alignment with the research proposal as violent actions occur predominantly during the pre-start of the video samples. With this insight, the development task incorporated non-violent samples to deliberately confuse the model’s operations concerning action similarity to establish processing superiority. The idea dispensed 12 decisions per video as 3DCNNsl/YOLOv5m activity recognition outcomes. The idea is to investigate suitable fusion techniques from the 12-frame video perspective to determine effective strategies towards combining the outcomes to foster accuracy score

improvements. Because of its computational effectiveness, the operations integrated the decision-level technique instead of arithmetic processing to emphasise the fusion aggregation prospects. The operations produced the decision-level's frame-by-frame outcome by integrating programming scripts to define its class association amid 3DCNNsl/YOLOv5m activity recognition. Decision-level fusion operations consider the first decision from 3DCNNsl and the first decision from YOLOv5m activity recognition on frame #1, which fuses it for a single outcome at the frame level. The identical operation applies for the remaining 11 frames, which generate 12 fused action outcomes. The decision level finalises its outcome by employing majority voting procedures to imitate the fusion protocols discussed via section 5.1.2. The overall concept defines the processing effectiveness and the model's robustness in scenarios lacking data.

5.3 Fusion Scheme-2: Processing the Entire Video

Fusion Scheme-1 evaluated the models in a frame-by-frame operation incorporating only the first 12 salient frames of violence to represent a video. The idea determined efficacy in cases lacking data during processing to satisfy research question-5 superiority objectives in section 1.2.1. The tasks were built on Fusion Scheme-1 by evaluating activity recognition performance using the entire video as Fusion Scheme-2 of 10 videos to conclude processing superiority. As YOLOv5m activity recognition recognises violence by assigning violent actions to individual frames on a frame-by-frame basis, the operations employed the entire video duration of each sample to attain an overall accurate outcome. The operations achieved this by applying configurations that passes each video sample for frame extraction at a frame rate of 30fps(frames per second) with a duration interval of 10-12 seconds. The application of surplus data represents the entire duration of violence in a video from its primitive stage (pre-start) to its end per sample. Supplying the models with surplus data in this context presented opportunities to stimulate robust outcomes. Considering 3DCNNsl/YOLOv5m activity recognition from a frame-by-frame perspective, the operations

explored the same decision-level techniques applied in Fusion Scheme-1 to produce a single outcome per video. 3DCNNsl/YOLOv5m activity recognition creates more outcomes because of the video 30-frame rate of 10-12 seconds per sample. Like Fusion Scheme-1 in the previous discussion on duration interval, the decision level creates class associations for each frame. The technique applies Fusion Scheme-1's majority voting processing to fuse the class associations, creating a single class outcome. Evaluating processing effectiveness employing surplus data proved crucial to cross-analyse section 5.1.3 fusion protocols to conclude research question-5's processing superiority in section 1.2.1. By understanding the fusion scheme techniques above, the operations determine the actual processing impact by computing overall accuracy confidence to reflect decision-level fusion's overall performance per model relative to [272], [273], [274], [138], [275], [276], [277], and [278]. Achieving the overall accuracy task incorporates computing the ratio of correctly classified samples divided by the total number of samples. The previous computation is multiplied by one hundred, exposing the operation's performance outcome as a percentage value. The following algorithm emphasises the computation of activity recognition's overall performance. To summarise the formula, "arOA" represents the overall accuracy of activity recognition, where the total number of correctly labelled videos denotes "nCLvids". The total number of videos suggests the "TnVids" factor, where "100%" represents the percentage computation and the outcome.

$$arOA = \frac{nCLvids}{TnVids} \times 100$$

5.4 Fusion Scheme-1 Results: Utilising 12-Frame Processing

The operations demonstrate results for YOLOv5m activity recognition, 3DCNNsl, and fusion to satisfy research question-5 processing superiority between the models in section 1.2.1. Column title definitions provide context to the output towards evaluating the

results, which defines the outcomes and their significance per model. Following the title description is an overview of the confidence threshold possibilities to validate the proposed fusion outcomes.

5.4.1 Definition of Column Title per Outcome:

Distinguishing the model’s performance status meant defining the relevance of the following column titles in Table 5.2 in alignment with section 5.1.2 proposed fusion constraints to express the outcome’s significance.

| # | Column Name | Description |
|--------------------------------------|----------------|---|
| 1 | Fr | Number of frames from the 12-frame only processing operations |
| 2 | Video | Number of Videos used in the whole video operations |
| The Actual Class | | |
| 3 | Correct_Class | The action’s true class label as the ground truth sample |
| 3DCNNsl Individual Processing | | |
| 4 | 3D-Gen | Generic class predicted is the action Violent or Non-Violent? |
| 5 | 3DPreds | Sub-Class label discloses the action type |
| 6 | 3D-OA | Computing 3DCNNsl’s Overall Accuracy |
| 7 | 3D-Conf | Discloses 3DCNNsl’s classification confidence in its decision outcome |
| YOLOv5m Individual Processing | | |
| 8 | YoloPreds | Sub-Class label predicted during processing |
| 9 | YO-OA | Computing YOLOv5m’s Overall Accuracy |
| 10 | YO-Conf | Discloses YOLOv5m’s classification confidence in its decision outcome |
| Proposed Fusion Results | | |
| 11 | Fusion_Class | 12 Sub-Class label output from decision level processing |
| 12 | Fusion_Results | Decision Level fusion applies generic, and subclass label its outcome |
| 13 | F-OA | Computing Fusion’s Overall Accuracy |
| 14 | F-Conf | Discloses Fusion’s classification confidence in its decision outcome |
| 15 | Confidence | Displays how confident the model is in its decision |

Table 5.2: Column Description of the Results.

5.4.2 Definition of Confidence Thresholds for Stabbing:

Appreciating the output description discussed above, it proved necessary to accentuate various planes of confidence similar to [279], [280], [281], and [282], which signifies the

model’s classification effectiveness from the video level. The following details imply the model’s aptitude and confidence towards making the right decision using only 12 frames per video. Confidence indicates the number of positive instances predicted for the stabbing class and negative instances for fencing and activity unknown categories. At this level, the confidence adopts a high, medium, and low threshold, which signifies the number of violent positive and non-violent negative instances produced during the classification stages. A nuance of a high confidence threshold count indicates the highest positive prediction ratios ranging between 10-12 instances, with a lower bound for the negative category outcomes (via the subscripts a, b, c). High confidence instances also indicate the model’s decision-level effectiveness as favourable results.

Table 5.3: Examples of Confidence Level for Stabbing

| # | Confidence | Positive (Stabbing) | Negative (Fencing) | Negative (Activity Unknown) |
|-----------------|------------|------------------------|-----------------------|--------------------------------|
| 1 ^a | High | 12 | 0 | 0 |
| 2 ^a | High | 11 | 0 | 1 |
| 3 ^a | High | 10 | 1 | 1 |
| 4 ^b | Medium | 8 | 2 | 2 |
| 5 ^b | Medium | 9 | 2 | 1 |
| 6 ^c | Low | 7 | 1 | 4 |
| 7 ^c | Low | 6 | 1 | 5 |
| 8 ^c | Low | 4 | 2 | 6 |
| 9 ^c | Low | 2 | 1 | 9 |
| 10 ^c | Low | 0 | 0 | 12 |

^a The highest confidence range, the higher the stabbing instance between 10-12, the stronger the classification confidence is towards stabbing to suggest violence.

^b Instances ranging between 8-9 for stabbing outcomes and indicating low instance ratios of 1-2 for fencing and activity unknown to suggest violence.

^c The lowest classification confidence range indicating low ratios for stabbing with high outcomes for fencing and activity unknown to suggest violence. This is rank as an insignificant classification performance.

A medium confidence threshold denotes positive prediction ratios between 8-9 instances with a sparse indication of the negative classes between 1-2 outcomes. Medium confidence also discloses the presence of classification challenges adversely affecting the model's ability to produce accurate outcomes. The episode of medium confidence necessitates more training samples with additional architecture fine-tuning to encourage suitable outcomes. Conversely, low confidence indicates an insignificant prediction operation; this exposes low positive instances between 0 and 7, with high representations of negative categories as outcomes. Low confidence signifies that the models experienced operation discrepancies relative to processing anomalies surrounding the data samples. Sporadic low confidence outcomes also suggest that the hyper-parameter option features require higher fine-tuning levels to achieve suitable/stable outcomes. Option

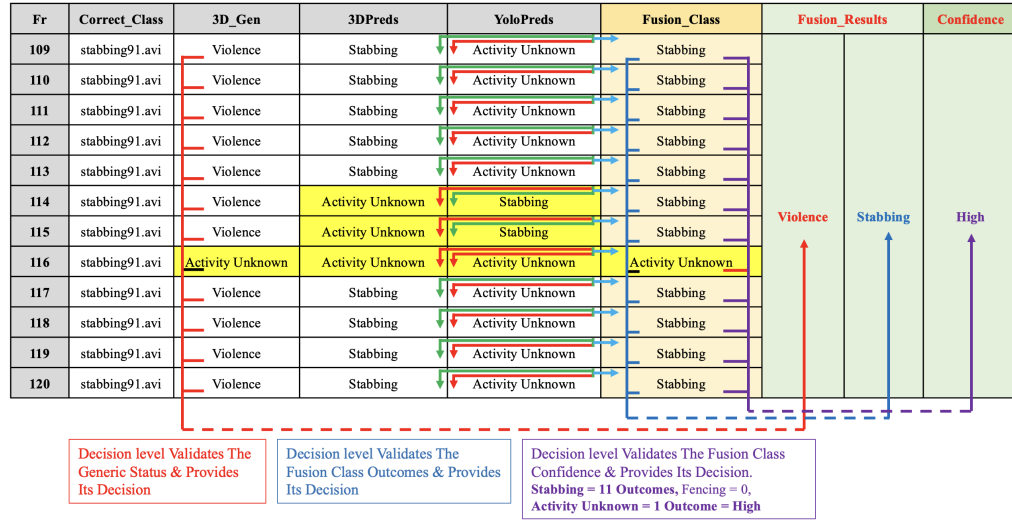



Figure 5.3: Fusion 12-Frame Confidence Processing using 1-Sample (Stabbing91.avi).

fine-tuning refers to regulating the learning controls to enhance prediction effectiveness. Table 5.3 provides nuances of previous discussions of the confidence levels to illustrate the concepts of the model's range of outcome possibilities. The operations evaluated scenarios with surplus data, suggesting a specific confidence ratio of 68+% as high prediction out-

comes to accommodate the data increase, with a ratio of 34-67% for medium confidence and 0-33% for low confidence operations. Fusion Scheme-1 considers overall accuracy and confidence concerning prediction correctness instead of representing accuracy score outcomes from an individual sample approach. The concept measures the models' aptitude in its decision strategy, producing outcome consistency and the effect of processing samples represented by a few frames to mimic scenarios lacking data from an overall performance level. Figure 5.3 clarifies nuances of decision level 12-frame processing employing 1-stabbing91.avi video sample, which emulates the concept of the model's outcomes regarding its confidence outputs. Primarily, the operations validated the proposed fusion confidence discussed above by disclosing exceptional cases utilising a 10-random test samples operation, each at 12 frames per video, to establish effectiveness. The 10-test sample experiment incorporated a ratio of five fencing and five stabbing actions; this produced ten single outcomes. A demonstration of 3-exceptional action similarity cases in Figure 5.4 concern-



| # | Sample# | Stabbing | Fencing | Activity Unknown | Fusion Outcomes | | Confidence |
|----|----------------|----------|---------|------------------|------------------|------------------|------------|
| 1 | Stabbing8.avi | 12 | 0 | 0 | Violence | Stabbing | High |
| 2 | Stabbing24.avi | 12 | 0 | 0 | Violence | Stabbing | High |
| 3 | Stabbing37.avi | 12 | 0 | 0 | Violence | Stabbing | High |
| 4 | Stabbing48.avi | 12 | 0 | 0 | Violence | Stabbing | High |
| 5 | Stabbing75.avi | 12 | 0 | 0 | Violence | Stabbing | High |
| 6 | Fencing11.avi | 0 | 12 | 0 | Non-Violence | Fencing | High |
| 7 | Fencing12.avi | 0 | 12 | 0 | Non-Violence | Fencing | High |
| 8 | Fencing27.avi | 0 | 3 | 9 | Activity unknown | Activity unknown | Low |
| 9 | Fencing32.avi | 0 | 12 | 0 | Non-Violence | Fencing | High |
| 10 | Fencing38.avi | 0 | 12 | 0 | Non-Violence | Fencing | High |

Figure 5.4: 10-Test Sample Nuance of Fusion Special Classification Cases.

-ing Figure 5.3 's confidence procedures provided context into the idea. Moreover, the operations employed the 10-test samples approach to illustrate those abnormal instances where the proposed fusion exhibited its processing effectiveness on complex action similarity classes with suitable outcome responses specifically for Fencing27.avi. The results

insinuated the absence of violence denoted by zero; fencing’s presence occurred in three instances, with nine activities with unknown outcomes. The nine unknown instances of activity occurred because the model struggled to establish its classification of actions that it cannot identify based on its training, therefore responding with a suitable low-confidence outcome to reflect its decisions. Following the proposed fusion’s case demonstration concerning Figure 5.4 ’s outlined samples, the operations integrated 50-random samples to thoroughly scrutinise and validate fusion’s processing impact like the 10-sample evaluation operations incorporating 12-frames per video.

5.4.3 Fusion Scheme-1 12-Frame Results for Stabbing24.avi:

With knowledge of Figure 5.3 ’s fusion outcomes, assessments determine its effectiveness on stabbing24.avi to demonstrate notable results employing 12-frames to represent a video sample. Because the primary focus is on 2-action classes only, a demonstration of Fusion’s prediction power emphasised the processing anomalies affecting 3DCNNsl/YOLOv5m activity recognition misclassifications. Frames #36-47 evaluation disclosed 3DCNNsl capability to discern stabbing24.avi efficiently and confidently. YOLOv5m activity recognition created 12 unknown activity instances, failing to identify stabbing and accurately producing a low confidence threshold to emulate its decisions. The low confidence issue links to excessive pre-processing procedures, which degrade the video’s resolution and further hinder the object’s classification. Table 5.4 results demonstrated Fusion’s decision effectiveness by accurately discerning the action as violent, reflecting its subclass label as stabbing with high confidence. The findings depicted low confidence thresholds for YOLOv5m activity recognition relative to its insignificant results. Although the proposed Fusion relies heavily on individual processing (3DCNNsl/YOLOv5m activity recognition), it generated 12-high confidence cases irrespective of YOLOv5m activity recognition’s 12-misclassification errors. The model satisfied research question-5 in section 1.2.1 by demonstrating the proposed Fusion’s processing superiority over 3DCNNsl/YOLOv5m activity recognition with high

confidence considering Table 5.4.

5.4.4 Fusion Scheme-1 12-Frame Stabbing37.avi Results:

Comparing fusion’s results above, YOLOv5m activity recognition improved by producing 5-accurate and 7-inaccurate outcomes in Table 5.5. YOLOv5m activity recognition failed unknown classification cases, and low confidence does not affect the final fusion decision because it considers accurate responses between individual models. Analysis proved 3DCNNsl’s effectiveness with 12 accurate outcomes identifying the action’s stabbing status with high confidence via frames #72-83. The proposed fusion processing dominated 3DCNNsls/YOLOv5m activity recognition by generating 12 accurate responses with high confidence to solidify its violence decision. The evidence proved the proposed fusion’s processing effectiveness as expected, which satisfied research question-5 processing stability objective in section 1.2.1.

| Fr | Correct_Class | 3DPreds | YoloPreds | Fusion_Class | Fusion_Results | | Confidence |
|----|----------------|----------|------------------|--------------|----------------|----------|------------|
| 36 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | Violence | Stabbing | High |
| 37 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 38 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 39 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 40 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 41 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 42 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 43 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 44 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 45 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 46 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 47 | stabbing24.avi | Stabbing | Activity Unknown | Stabbing | | | |

Table 5.4: Fusion Scheme-1 12-Frame Evaluation via stabbing24.avi.

5.4.5 Fusion Scheme-1 12-Frame Fencing27.avi Results:

A representation of the results disclosed stabbing attributes for the non-violent class to illustrate fusion’s power, as most test samples validated the proposed fusion’s process-

| Fr | Correct-Class | 3DPreds | YoloPreds | Fusion_Class | Fusion_Results | | Confidence |
|----|----------------|----------|------------------|--------------|----------------|----------|------------|
| 72 | stabbing37.avi | Stabbing | Stabbing | Stabbing | | | |
| 73 | stabbing37.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 74 | stabbing37.avi | Stabbing | Stabbing | Stabbing | | | |
| 75 | stabbing37.avi | Stabbing | Stabbing | Stabbing | | | |
| 76 | stabbing37.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 77 | stabbing37.avi | Stabbing | Stabbing | Stabbing | Violence | Stabbing | High |
| 78 | stabbing37.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 79 | stabbing37.avi | Stabbing | Stabbing | Stabbing | | | |
| 80 | stabbing37.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 81 | stabbing37.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 82 | stabbing37.avi | Stabbing | Activity Unknown | Stabbing | | | |
| 83 | stabbing37.avi | Stabbing | Activity Unknown | Stabbing | | | |

Table 5.5: Fusion Scheme-1 12-Frame Evaluation via stabbing37.avi.

ing effectiveness on stabbing in action similarity conditions. The rationale validates the proposed fusion confidence stability and its aptitude to produce accurate results on stabbing features in non-violent conditions. Analysis on fencing27.avi in Table 5.6 confirmed YOLOv5m activity recognition’s confidence in its decis-

| Fr | Correct_Class | 3DPreds | YoloPreds | Fusion_Class | Fusion_Results | | Confidence |
|----|---------------|----------|------------------|------------------|------------------|------------------|------------|
| 48 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | | | |
| 49 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | | | |
| 50 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | | | |
| 51 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | | | |
| 52 | fencing27.avi | Stabbing | Fencing | Fencing | | | |
| 53 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | Activity Unknown | Activity Unknown | Low |
| 54 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | | | |
| 55 | fencing27.avi | Stabbing | Fencing | Fencing | | | |
| 56 | fencing27.avi | Stabbing | Fencing | Fencing | | | |
| 57 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | | | |
| 58 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | | | |
| 59 | fencing27.avi | Stabbing | Activity Unknown | Activity Unknown | | | |

Table 5.6: Fusion Scheme-1 12-Frame Evaluation via fencing27.avi.

-ions over 3DCNNsl. 3DCNNsl’s low confidence validated its inability to discern fencing correctly. YOLOv5m activity recognition analysis outlined nine cases where activity unknown tags were applied to insinuate its low confidence and uncertainty of the action’s status. Although fencing and stabbing actions sometimes convey homogeneous qualities, the model’s intense training should generate high-confidence decisions to discern action dissimilarity. Because the similarities between stabbing and fencing intensified in this sample,

3DCNNsl provided 12 inaccurate low-confidence instances for fencing. The decision-level protocols maintained fusion’s processing effectiveness by designating correct activity unknown replies for unidentified action cases. Subjectivity toward stabbing as the positive class is granted only in cases where the models’ (3DCNNsl/YOLOv5m) predictions are accurate. In such cases, fusion protocols monitor these bounds by validating the correct class ground-truth data with the actual predictions, thus incorporating accurate replies in sync with its confidence level. The evidence on fencing27.avi 12-frame operations verified the proposed fusion confidence, which satisfied research question-5 activity recognition processing superiority in section 1.2.1.

5.4.6 Fusion Scheme-1 12-Frame Overall Accuracy Evaluation:

The anomaly discussed above proved the overall accuracy of fusion’s processing superiority compared to 3DCNNsl/YOLOv5m activity recognition, which incorporates the same 10-samples applied in Fusion Scheme-1. Fusion Scheme-2 activity recognition’s overall accuracy formula considers the number of correctly classified samples divided by the total number of frames from all videos (10 videos x 12 frames), multiplied by 100%. The operations disclosed overall accuracy for all 12-frame videos by considering misclassifications for all video samples instead of computing misclassifications for a single video. Confidence threshold ratios discussed in section 5.4.2 played an integral role in determining high, medium, and low outcomes to validate the models’ aptitude.

Incorporating the 10-sample dataset defined in Table 5.7, 3DCNNsl disclosed an overall accuracy of 40.83% utilising 120 frames, with 49 correctly classified instances in total, zero activity unknown, and zero instances of the fencing class. Also, 3DCNNsl dispensed 71 inaccurate predictions, with 50 incorrectly classified stabbing instances, including 21-activity unknown cases where the model could not accurately discern violence within the actions. 3DCNNsl recorded no inaccurate instances of fencing, thus indicating its ability to differentiate the fencing class amid violent actions with a medium confidence ranking.

3DCNNsl depicted classification instances per class, emphasising its capability to detect violence. YOLOv5m activity recognition overall accuracy decreased by 0.7% at 33.33%, with 40 correctly classified samples. YOLOv5m demonstrated its ability to correctly identify 3-instances of the non-violent fencing class and 37 instances in the stabbing class. YOLOv5m, at this stage, produced 80 incorrect predictions in total, with no inaccurate cases for the fencing class. The model dispensed 46 inaccurate stabbing instances, including 34-activity unknown cases, with a low confidence ranking. The findings indicated that the model experiences challenges discerning violence using a few frames as the dataset.

Fusion demonstrated its dominance by generating an overall accuracy of 52.50%, disclosing its superiority lead by 11.67% over 3DCNNsl and by 19.17% over YOLOv5m activity recognition. At this level, fusion produced 63 correctly classified instances with no activity unknown cases and three predictions for the non-violent fencing class. Fusion proved its ability to identify violence with robust ratings compared to YOLOv5m's processing. Though fusion proved effective, the model dispensed 57 inaccurate predictions with zero inaccurate cases for the non-violent fencing class and 9-activity unknown outcomes. Fusion misclassifications increased by 2% at 48 cases compared to YOLOv5m activity recognition at 46 cases. The outcomes insinuate that the proposed fusion experience processing challenges linked to the insignificant size of the dataset during 3DCNNsl and YOLOv5m activity recognition operations and their ability to generalise violence adequately.

Moreover, fusion demonstrated its dominance and capability to differentiate between stabbing and fencing over 3DCNNsl and YOLOv5m activity recognition. Fusion's total column at 63 correct cases depicts a medium prediction rating projecting fusion's decision certainty, thus simultaneously satisfying research question-5's superiority between models in section 1.2.1. Using inadequate sample sizes leads to invariance (creates the same results, regardless of system fine-tuning), which scientifically limits thorough evaluations of fusion's processing power. However, surplus data comprising 50 samples (50 videos x 12 frames) establishes the models' processing capabilities in the following evaluation.

| Dataset | Performance | | Correctly_Classified_Instances | | | | Incorrectly_Classified_Instances | | | | Confidence-Threshold |
|---|---------------|---------------------------------|--------------------------------|-----------|------------------|-----------|----------------------------------|-----------|------------------|-----------|----------------------|
| | Model | Overall-Accuracy | Fencing | Stabbing | Activity Unknown | Total | Fencing | Stabbing | Activity Unknown | Total | |
| 10 Samples at 12-Frames Per Sample = 120 Frames | 3DCNNsl | 40.83% | 0 | 49 | 0 | 49 | 0 | 50 | 21 | 71 | Medium |
| | YOLOv5m | 33.33% | 3 | 37 | 0 | 40 | 0 | 46 | 34 | 80 | Low |
| | | 52.50% | 3 | 60 | 0 | 63 | 0 | 48 | 9 | 57 | Medium |
| | Fusion Led By | 11.67% Superiority Over 3DCNNsl | | | | | | | | | |
| | | 19.17% Superiority Over YOLOv5m | | | | | | | | | |

Table 5.7: Fusion Scheme-1 Performance Evaluation for 12 Frames per 10-Samples.

5.4.7 Fusion Scheme-1-12 Frames per Sample Results 50 Samples:

Considering the 10-test video evaluation discussed above, the operations validated the proposed fusion effectiveness by utilising 50 random samples of equal fencing and stabbing ratios (50 videos x 12 frames) in a similar context. Employing the surplus 50 samples 600 frame action similarity dataset at this stage allowed the models to act realistically as a plot to eradicate the risks of bias processing. 3DCNNsl via Table 5.8 produced an overall accuracy of 66.50% with 400-correctly classified instances. The model generated 203 correctly classified cases of the non-violent fencing class, with 197 instances for stabbing and zero for the activity unknown category. 3DCNNsl, at this stage, demonstrated its capability to recognise violence in complex action scenarios. The model generated 200 inaccurate classifications via the activity unknown class in Table 5.8. The outcome suggests that the actions processed contain no fencing or stabbing attributes, with no direct incorrect classifications for both classes, ranking medium confidence for its decisions. YOLOv5m activity recognition improved its overall accuracy performance by 36.83% at 70.16% compared to Table 5.7. The model produced 421 correctly classified instances, exceeding 3DCNNsl's 400 outcomes.

The analysis recorded 211 cases of fencing and 210 of stabbing for YOLOv5m, which demonstrated its ability to classify violence accurately in complex conditions with high confidence in its decisions. The model dispensed 179 inaccurate predictions via the activity unknown category without directly misclassifying the fencing and stabbing categories. YOLOv5m demonstrated its processing effectiveness with no indication of processing challenges via incorrect classification of the target classes using the same 50-sample dataset. Like the 10 test sample discussions, the proposed fusion proved its prediction superiority by dispensing an overall accuracy of 83.83%. Fusion

validated its prediction superiority over 3DCNNsl by 17.33% and YOLOv5m activity recognition by 13.67%. The model dispensed 503 correctly classified instances, with 260 cases of the fencing class and 243 for stabbing. Fusion produced 97 inaccurate predictions directly linked to unknown activity cases without traces of fencing/stabbing attributes. Although 3DCNNsl/YOLOv5m activity recognition demonstrated classification challenges in some scenarios, the proposed fusion maintained its realistic outcome consistency, proving its processing flexibility and stability via the correctly classified outcomes. Fusion produced high classifications for fencing/stabbing with fewer misclassifications compared to the other models. The findings indicate that by applying the proposed fusion processing on surplus data, the effectiveness improved as a superior approach over the previously mentioned models to satisfy research question-5’s model dominance in section 1.2.1. With an appreciation for fusion’s realistic 12-frame processing, further evaluations of the model’s power using whole video duration samples disclosed its overall accuracy performance using more data in the following discussions. The operations considered the same ten samples with similar column output titles taken from Fusion Scheme-1 to disclose Fusion Scheme-2’s effectiveness and maintain data and experiment consistency.

| Dataset | Performance | | Correctly_Classified_Instances | | | | Incorrectly_Classified_Instances | | | | Confidence-Threshold |
|--|---------------|--|--------------------------------|------------|------------------|------------|----------------------------------|----------|------------------|-----------|----------------------|
| | Model | Overall-Accuracy | Fencing | Stabbing | Activity Unknown | Total | Fencing | Stabbing | Activity Unknown | Total | |
| 50 Samples at 12-Frames Per Sample = 600 Frames | 3DCNNsl | 66.50% | 203 | 197 | 0 | 400 | 0 | 0 | 200 | 200 | Medium |
| | YOLOv5m | 70.16% | 211 | 210 | 0 | 421 | 0 | 0 | 179 | 179 | High |
| | Fusion Led By | 83.83% | 260 | 243 | 0 | 503 | 0 | 0 | 97 | 97 | High |
| | | 17.33% Superiority Over 3DCNNsl | | | | | | | | | |
| | | 13.67% Superiority Over YOLOv5m | | | | | | | | | |

Table 5.8: Evaluating Accuracy/Confidence on 50-Samples at 12-Frames per Sample.

5.5 Fusion Scheme-2 Results: Processing the Entire Video

Appreciating Fusion Scheme-1’s capability of employing the 50-videos-12-frames-per-sample dataset encourages another strategy involving building on the idea to substantiate fusion’s power. The idea incorporated the entire video as surplus data for Fusion Scheme-2. The results also measure the impact of processing superiority between models from an overall accuracy perspective.

Like Fusion Scheme-1 in the previous evaluation, the task considered all misclassifications for all samples instead of evaluating each sample individually. The individual video misclassification concept detours the research objectives towards considering single-shot accuracy scores. Hence, the concept avoided the individual video approach to maintain the proposal's time constraints. The models proved their effectiveness by computing the overall accuracy, which considers the number of correctly classified samples divided by the total number of frames from all videos, multiplied by 100%. The investigations commenced with Fusion Scheme-1's 10-samples incorporating the entire video duration, followed by the same 50-sample dataset featuring the entire duration. Fusion scheme-1's 10-samples full video duration dataset considered a frame rate of 30fps(frames per second) at a 10-second interval (10-vids x 30-fps x 10-secs) to establish the model's decision-making effectiveness on 3000-frames. The 50-sample entire video duration dataset considered a frame rate of 30fps(frames per second) varying in duration between 10-12 seconds. The data features 10-video samples at 11-seconds (10-vids x 30-fps x 11-secs), 20-video samples at 10-seconds (20-vids x 30-fps x 10-secs) and 20-video samples at 12-seconds (20-vids x 30-fps x 12-secs) to establish decision effectiveness on 16,500-frames. The idea returns 1-outcome computed by the activity recognition overall accuracy formula in alignment with [272], [273], [274], [138], [275], and [276]. The formula reflects the "arOA" value representing activity recognition's overall accuracy, with "nCLvids" indicating the total number of correctly classified samples. The computation applies the total number of samples via "TnVids", with a "100%" value denoting the percentage computation which indicates the outcome. The following expression illustrates the formula used to generate activity recognition's overall confidence per model to establish processing superiority.

$$arOA = \frac{nCLvids}{TnVids} \times 100$$

5.5.1 Summary of Confidence for Full Video Classification Effectiveness:

Like section 5.4.2 discussions on confidence, the operations applied identical measures similar to [279], [280], [281], [282], [283] and [284] to accommodate the surplus data, which constitutes boundaries to describe the models' decision certainty. High confidence indicates the highest classification ratings regarding positive predictions above 68% confidence. High confidence indicates favourable outcomes and the model's prediction effectiveness. Medium confidence signifies confidence ratios between 34-67%, suggesting the presence of processing challenges affecting the effectiveness of the

operation. Medium confidence episodes require additional training samples with fine-tuning architecture to generate significant results. Conversely, low confidence indicates insignificant results, reflecting a 0-33% ratio. Sporadic low confidence also suggests that the models experienced generalisation challenges necessitating hyper-parameter option fine-tuning and sample pre-processing to achieve favourable results. Option fine-tuning encompasses tweaking the models' generalisation controls to promote effectiveness. The following demonstrates Fusion Scheme-1's 10-test sample results employing the whole video duration to emphasise the range of outcome possibilities concerning overall accuracy.

5.5.2 Fusion Scheme-2 Whole Video Evaluation on 10-Test Samples:

To establish the models' overall accuracy, the operations validated fusion whole video processing using the approach discussed in section 5.4.6 on ten videos totalling 3000 frames. Table 5.9 analysis disclosed an overall accuracy of 82.43% for 3DCNNsl using the whole duration of the video, with a 41.60% difference compared to its efforts in section 5.4.6 and 15.93% in section 5.4.7. 3DCNNsl produced 2,473 correctly classified instances, with 1275 cases suggesting fencing, 1198 instances for stabbing and zero outcomes for the activity unknown class. 3DCNNsl, at this stage, proved its ability to discern the target actions in violent and non-violent scenarios accurately. Although 3DCNNsl's performance improved with surplus data, the operations produced 527 inaccurate predictions disclosing zero outcomes for fencing, one inaccurate prediction for stabbing and 526 activity unknown cases. The findings suggest that the model improved significantly because of surplus data, with high confidence in its decisions regardless of the 527 inaccurate cases. YOLOv5m activity recognition produced an overall accuracy of 79.50%, improving by 46.17% compared to its 33.33% efforts in section 5.4.6 and 9.34% at 70.16 in section 5.4.7. The model correctly classified 2,385 instances, with 1105 predictions suggesting the presence of the non-violent fencing category. At this level, the stabbing class generated 1280 violent predictions with no activity unknown cases. The findings disclosed a 2.93% depreciation in performance by YOLOv5m at 79.50% compared to 3DCNNsl's processing at 82.43%. The operations projected an anticipated outcome because of the difference in architectures and the complexity of similar action activities in various scenarios.

Regarding YOLOv5m 615 inaccurate predictions, the model projected 102 inaccurate instances of the non-violent fencing class and 513 inaccurate cases for activity unknown. Though YOLOv5m 615 inaccurate predictions appeared high, stabbing recorded zero inaccurate outcomes with high

confidence in its decisions. The findings suggested that YOLOv5m experiences challenges discerning the true nature of fencing against actions it cannot recognise due to its training procedures. At this level, fusion proved its dominance by dispensing 85.20% overall accuracy for whole video operations. Fusion's operations increased by 32.70% compared to its 52.50% efforts in section 5.4.6 and 1.33% in section 5.4.7. Fusion dominated 3DCNNsl by 2.77% and YOLOv5m activity recognition by 5.70%. The findings reflected 2556 correctly classified instances, with 1228 cases of the non-violent fencing class and 1328 outcomes for stabbing in bold text. The correctly classified results proved fusion's capability to identify the target classes in violent and non-violent scenarios. Although fusion proved formidable in its processing, the operations produced 444 inaccurate predictions, with 102 cases suggesting fencing and activity unknown each and no direct outcomes for stabbing. The evidence validated fusion's processing effectiveness with high confidence in its decisions regardless of the 444 inaccurate predictions utilising surplus data.

| Dataset | Performance | | Correctly_Classified_Instances | | | | Incorrectly_Classified_Instances | | | | Confidence-Threshold |
|--------------------------------------|---------------|--------------------------------|--------------------------------|-------------|------------------|-------------|----------------------------------|----------|------------------|------------|----------------------|
| | Model | Overall-Accuracy | Fencing | Stabbing | Activity Unknown | Total | Fencing | Stabbing | Activity Unknown | Total | |
| 10 Whole Video Samples = 3000 Frames | 3DCNNsl | 82.43% | 1275 | 1198 | 0 | 2473 | 0 | 1 | 526 | 527 | High |
| | YOLOv5m | 79.50% | 1105 | 1280 | 0 | 2385 | 102 | 0 | 513 | 615 | High |
| | | 85.20% | 1228 | 1328 | 0 | 2556 | 102 | 0 | 102 | 444 | High |
| | Fusion Led By | 2.77% Superiority Over 3DCNNsl | | | | | | | | | |
| | | 5.70% Superiority Over YOLOv5m | | | | | | | | | |

Table 5.9: Performance Evaluation for 10-Samples in Whole Video Processing.

5.5.3 Fusion Scheme-2 Whole Video Evaluation on 50 Test Samples:

Like section 5.4.7 using 50 samples at 12 frames per sample, the same 50 samples employing the entire video duration established realistic outcomes. Table 5.10 whole video analysis revealed a decrease in 3DCNNsl's processing by 13.38% at 69.05% compared to Table 5.9 whole video processing at 82.43%. Increasing the surplus data concerning real-world action similarity scenarios impacted the model's processing because it intensified the complexity of differentiating violence from non-violent actions. Predictably, the model dispensed realistic outcomes matching the complexity of the action similarity data in real-world conditions. 3DCNNsl correctly classified 11,394 instances, with 5503 cases of the non-violent fencing class and 5891 cases suggesting the presence of stabbing violence. The evidence

suggests that the model experienced challenges processing the complexity of action similarity utilising an increment in surplus data via 50 samples. 3DCNNsl dispensed 5,106 inaccurate activity unknown cases with no direct misclassifications for fencing/stabbing. Although 3DCNNsl experienced classification challenges, receiving only activity unknown misclassifications suggests that the model maintained high confidence in accurately classifying violence in the total column (11394 instances).

| Dataset | Performance | | Correctly_Classified_Instances | | | | Incorrectly_Classified_Instances | | | | Confidence-Threshold |
|---|---------------|---------------------------------|--------------------------------|-------------|------------------|--------------|----------------------------------|----------|------------------|-------------|----------------------|
| | Model | Overall-Accuracy | Fencing | Stabbing | Activity Unknown | Total | Fencing | Stabbing | Activity Unknown | Total | |
| 50 Whole Video Samples = 16,500 Frames | 3DCNNsl | 69.05% | 5503 | 5891 | 0 | 11394 | 0 | 0 | 5106 | 5106 | High |
| | YOLOv5m | 67.86% | 5211 | 5987 | 0 | 11198 | 1348 | 0 | 3954 | 5302 | Medium |
| | | 84.75% | 6853 | 7131 | 0 | 13984 | 1082 | 0 | 1434 | 2516 | High |
| | Fusion Led By | 15.70% Superiority Over 3DCNNsl | | | | | | | | | |
| | | 16.89% Superiority Over YOLOv5m | | | | | | | | | |

Table 5.10: Performance Evaluation for 50-Samples in Whole Video Processing

YOLOv5m activity recognition demonstrated a decline in performance by 11.64% at 67.86% in Table 5.10 compared to its 79.50% efforts in Table 5.9. Predictably, like 3DCNNsl's processing, YOLOv5m demonstrated a similar depreciation in performance linked to the action similarity complexity via the 50-sample dataset. YOLOv5m activity recognition correctly classified 11,198 instances, with 5211 cases suggesting the fencing class presence and 5987 for stabbing violence. The model recorded 5,302 inaccurate predictions with an increase in misclassifications of 1348 for fencing, 3954 for activity unknown and zero for stabbing. YOLOv5m depreciation in performance reflected medium confidence in its decision because of several misclassifications and the action similarity complexity of the 50 video data. The findings revealed a decrease in performance for the proposed fusion by 0.45% at 84.75% compared to its 85.20% efforts in Table 5.9 using ten whole video samples. Predictably, like 3DCNNsl and YOLOv5m, the complexity of action similarity in the 50 samples dataset depicted a similar effect. At this level, fusion produced 13,984 correctly classified cases, with 6853 cases suggesting the presence of the fencing class and 7131 for stabbing. The model dispensed 2516 inaccurate classifications with 1082 misclassifications for fencing in bold text, zero for stabbing and 1434 reflecting activity unknown cases. Unlike YOLOv5m's 1348 misclassification, fusion improved its outcomes with high confidence in its decisions. The evidence proved that by incorporating the proposed fusion with surplus data, its effectiveness and dominance improved over 3DCNNsl by 15.70% and YOLOv5m activity recognition at 16.89% to validate the research

question-5 objectives in section 1.2.1.

5.6 Fusion Scheme-1 and Fusion Scheme 2's Discussions

The primary goal of the fusion schemes is to enhance the model's predictive power and accuracy instead of generating individual video sample outcomes. This concept evaluates the models' decision-making confidence, ensuring consistent and effective outcomes in scenarios with varying data availability. The complexity of actions like stabbing and fencing often leads to ambiguous interpretations, necessitating additional data to strengthen the model's discerning strategy. Multiple datasets serve as a tool to rigorously evaluate the complexity of violence rather than specifying classification efficiency on easily identifiable class samples.

The fusion scheme ideas avoided simple classification tasks as an informed decision by focusing on two complex samples only because of the sporadic nature of violence. The complexity of fencing and stabbing encouraged classification difficulty, which provides the opportunity to evaluate fusion's effectiveness honestly and avoid the risks of producing erroneous results in real-world conditions. Investigating alternative processing possibilities by incorporating average arithmetic methods endorsed the proposed fusion's processing before considering alternative solutions, which increased the classification anomalies and lengthened the duration of results convergence. However, alternative solutions were applied to evaluate the fusion thoroughly. Analysing the max score arithmetic as other methods proved unrealistic and biased as its operations produced identical high scores from 3DCNNsl and YOLOv5m activity recognition. With insight into the max score arithmetic impact, the procedure abandoned further analysis because it negatively affected all operations.

Contrary to the alternative solutions, processing surplus data increased the misclassification outcomes, thus impacting the individual model's overall performance. However, the analysis showed minimal impact via the proposed fusion efficiency, maintaining its confi-

dence consistency above 80%. Accuracy score as a solution relative to individual samples and frame-level processing considered the ratio of inaccurate prediction conceded per frame during processing [285], [286], and [287]. Exploring accuracy per sample and frame as a solution significantly detoured the research objectives to establish its full potential. The operations avoided the accuracy approach because it exceeded the proposal’s time constraints; this primarily answered another research question, which investigates the model’s ability to accurately classify violence in a single shot via a single frame perspective. The findings showed several instances where the models experienced classification challenges that impacted their decision-making capabilities because of the homogeneity of the datasets but maintained their processing robustness on surplus data. Analysis accentuated Fusion Scheme-2’s utilising whole video surplus data over Fusion Scheme-1 employing 12-frame-per-video from 10 and 50-sample perspectives. One reason surrounds the limitations of the 3DCNNsl/YOLOv5m activity recognition processing ability when employing complex action similarity data. Providing surplus data to the fusion models increased their confidence, boosting their robustness in action similarity conditions.

5.7 Conclusion

Chapter 5 evaluated 3DCNNsl/YOLOv5m activity recognition capabilities compared to the proposed fusion concerning action similarity datasets in 2-fusion schemes. The chapter evaluated the proposed fusion’s overall accuracy by applying surplus data reporting realistic outcomes compared to smaller datasets employing fewer samples. Like discussions on state-of-the-art comparisons in Section 4.6, the same outlook applies at this level. Because of the dataset and the pre-stages of violence, it proved impractical to thoroughly compare the outcomes with other state-of-the-art solutions. The concept implemented two state-of-the-art models (YOLOv5m and 3DCNNsl) to compare those simulations and Chapter 5’s operations. Chapter 5 fusion demonstrated dominance of 52.50% over 3DCNNsl at 40.83% and YOLOV5m activity recognition at 33.33% to quantify Fusion Scheme-1.

Fusion dominance increased to 83.83% over 3DCNNsl at 40.83% and YOLOV5m activity recognition at 33.33% by employing a 50-sample-12-frames-per-sample dataset as surplus data. The findings validated fusion's superiority at 85.20% over 3DCNNsl capabilities at 82.43% and YOLOV5m activity recognition at 79.50%, incorporating surplus data via the 10-sample-whole-video dataset. Fusion proved its robust processing with 84.75% overall accuracy over 3DCNNsl at 69.05% and YOLOV5m activity recognition at 67.86% via the 50-sample-whole-video dataset.

Fusion outlined its consistency and stability in Fusion Scheme-1 and 2 by dispensing high confidence thresholds instead of fluctuating outcomes between low and medium for the other models. The chapter proved the viability of Fusion Scheme-2 operations toward high confidence, demonstrating robustness involving an increment and decrement in complex action conditions in the data. The proposed fusion at this level demonstrated its robustness and aptitude to classify the complexity of the actions and create a suitable response in Figure 5.4. The idea emphasises an increase in performance but at the cost of efficiency by utilising a few frames to suggest the resonance of a violent class. The approach can potentially introduce errors during blob analysis due to human fatigue caused by long durations of intense object encapsulation, which positively impacts the real-time results. The efficiency cost projected a trade-off to performance, signalling the need for alternative measures to increase the model's speed of discerning the actions via a graphical unit processing approach. The model experienced efficiency challenges resulting from a PyTorch package issue discussed in Section 4.5.4. The integration of hardware facilitating graphical processing unit operations attains the computational power to mitigate challenges regarding efficiency to counter the speed of the model's result convergence. With knowledge of fusion's performance, the operations evaluated an alternative strategy in Chapter 6, which enhances the operations further specific to the artefacts of interest and their existence within the scenery.

Chapter 6

Merging Action Recognition/Object Detection for Violence Recognition

Chapter 5 implemented fusion by modifying 3DCNNsl/YOLOv5m activity recognition's final layer outputs via fusion strategies. Chapter Six's operations incorporated weapons in scenarios by exploiting YOLOv5m object detection dominance to advance the classification status of fusion activity recognition. The chapter propositioned weapon detection operations denoted as artefacts specific to its violent class, which combined its output with the proposed fusion discussed in Chapter 5. The idea validated the existence of violence relative to violence focusing on stabbing. The focus on stabbing maintained the proposal's life cycle without negatively impacting its competition time. This chapter features five sections to substantiate the proposed fusion's idea incorporating YOLOv5m activity recognition artefact power. The operations commenced by presenting the effectiveness of YOLOv5m artefact object recognition to illustrate the potential of utilising static images containing specific artefacts of interest for fusion's development in Section 6.1. Section 6.2

focuses on the fundamental operations of YOLOv5m activity recognition artefact processing. Section 6.3 emphasises essential fusion stages using artefact support to satisfy research question-6 in 1.2. The section also expresses the findings via the results dispensed from each model. Essential discussions on operational anomalies express performance issues in Section 6.4. Finally, Section 6.5 outlines the findings to conclude the chapter.

6.1 Overview of YOLOv5m as Artefact Object Detection

YOLOv5m object recognition considers objects of interest in static frames using bounding boxes, which encapsulate and identify targets in the region of interest. The strategy at this stage emphasises its effectiveness as an informed decision before exploring alternative artefact processing measures within violence. The idea exploits the identical frame-by-frame processing strategies as YOLOv5m activity recognition. However, YOLOv5m activity recognition considers frame sequences constituting an action over a spatiotemporal period, which utilises bounding box techniques to identify action objects of interest in a scenario. To evaluate YOLOv5m artefact object detection processing, the concept employed 12 randomly selected static images in the following experimental setup to investigate its performance outcomes.

6.1.1 YOLOv5m Artefact Object Detection Experimental Setup:

Maintaining experiment consistency and eradicating biased results is crucial during training. The object detection model incorporated Chapter 5's data to reduce consistency risks. Implementing blob analysis specified the following classes relevant to violence to facilitate YOLOv5m Artefact Object Detection operations to evaluate its capability.

Aggressor Class: Signifies perpetrator/s exists (suspicious actions leading up to the beginning of violence) The aggressor validates/indicates the attacker's existence in

motion to execute stabbing actions.

Knife-Weapon Class: Signifies the existence of weapons in a scenario before the action of violence commences. The knife-weapon class validates/indicates the existence of a weapon (bladed/knife objects) and posturing before committing stabbing.

Knife-Deployed Class: Signifies that a stabbing action is pending. The knife-deployed validates/indicates the existence of knife weapon/s (bladed objects) before and after committing the act. This class suggests the presence of a bladed instrument (**weapons**) used in the scenery for violence.

Hand Class: Validates the existence of the knife-deployed action class. The hand validates/indicates the existence of knife weapons (**bladed objects**) in the hand/s of the aggressor before the attack. This class acts as a redundancy to confirm that the aggressor is indeed holding/using a weapon to commit an act of violence.

Blood Class: Signifies that individual/s sustained injuries. The existence of blood validates/indicates injuries sustained, and medical attention is required. Its significance specifies the presence of blood/injuries, which usually occur after the pre-empting duration of the action. The idea ensures that a solution is available to prevent the loss of human life, further indicating the need for urgent medical attention in such cases.

Victim Class: Signifies the target person/individual's about to receive/receiving the injuries. The victim category validates/indicates the existence of stabbing and highlights the individuals receiving lethal injuries as the target.

Stabbing Class: Signifies the action object that suggests a pending stabbing action for object detection. The stabbing object for object detection validates/indicates

the existence of stabbing but cannot determine the action over time.

People Class: Signifies individuals within the scenario for object detection purposes only that suggest a pending stabbing action. The people object for object detection validates/indicates the existence of an individual within the scenario as a measure to enhance the object detection processing to facilitate distinctions between normal actions and stabbing.

Crucial modifications via programming assist the model in focusing on 12 static images derived from online social media platforms to consider the abovementioned artefacts. The idea establishes YOLOv5m's object detection classification possibilities from an informed decision concerning violent weapon artefacts on randomly selected images. Like Chapter 5 's experimental setup, the processing of the operation quantifies YOLOv5m's artefact object detection effectiveness on violence in the following results.

6.1.2 Evaluating YOLOv5m Artefact Object Detection

YOLOv5m artefact object detection operations produced multiple outcomes relative to objects of violence in images 1-6 in Figure 6.1 and 7-12 in Figure 6.2, a total of 12-static samples. The idea projects examples of the power of object detection's performance before considering its integration. Image-1 in Figure 6.1 analysis demonstrated 4-correctly classified instances, 2-hand artefacts at 0.50% and 0.82%, knife-deploy at 0.82%, and knife-weapon at 0.75%. Image 2 dispensed 3-correct incidents where the hand artefact achieved 0.65% misclassification and 0.47% for the actual hand object. The analysis projected an accurate classification for knife-deployed at 0.37%. Although the model produced fluctuating results, its classification distinction outlined the targets to prove its capability. Image-3 dispensed 1-accurate classification at 0.58% for the stabbing artefact with lower scores for persons at 0.45% and 42%, aggressor at 0.33%, knife weapon at 0.44%, hand

at 0.38% and victim at 0.35%. Like Image-1, Image-4 produced 0.29% for knife-deploy, 0.55% and 0.39% for the hand artefact as three accurate classifications. Image-5 dispensed accurate instances of 0.81% for stabbing and 0.52% for the person's artefact. Analysis showed that the model overlooked the actual perception of stabbing because of the image pre-processing operations. Image-6 produces no artefacts as its rationale evaluated the context of violent artefact in non-violent actions relative to object detection. The idea provided crucial insight into YOLOv5m artefact object detection ability to differentiate

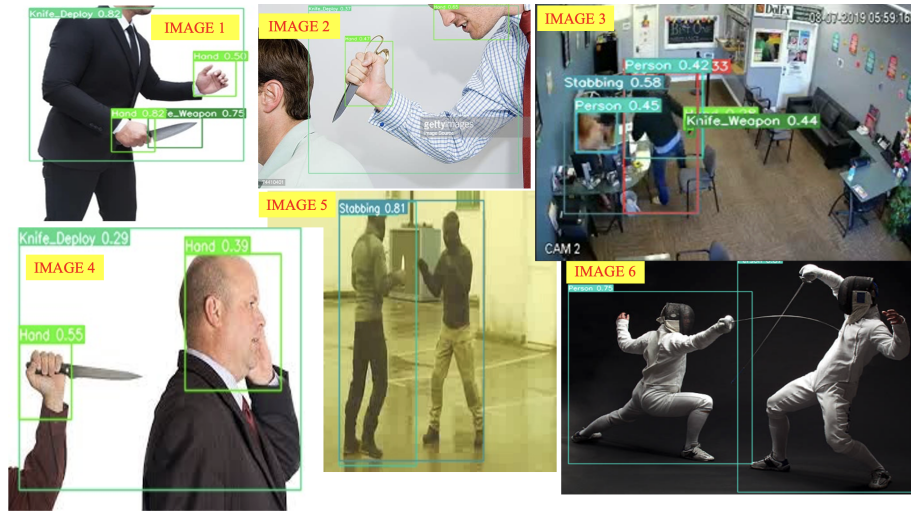


Figure 6.1: YOLOv5m Artefact Object Detection Results for Image 1-6.

the resonance between fencing/stabbing. Image-7 in Figure 6.2 dispensed 0.28% for the stabbing artefact, 0.26% for a person, 0.31% for knife-deploy and 0.42% for the hand object. The analysis depicted 2-stabbing instances at 0.28% and 0.74% for Image-8. Like Image-8, Image-9 produces 2-artefact outcomes, one at 0.88% for stabbing and 0.65%. The model produced one score indicating stabbing at 0.46% for Image 10. Although Images 8 to 10 produced fewer artefact outcomes, the classification operations proved the artefact processing's significance in establishing the object's existence. Image-11 dispensed 3-artefact instances, 0.27% for knife-weapon, 0.33% for knife-deploy, and hand at 0.39%

and 0.70%, respectively. Finally, Image-12 produced nine instances of the artefacts to prove the effectiveness of the artefact operations at this level. The findings disclosed 0.78% and 0.64% for the stabbing artefact, 0.74% for the aggressor, 0.54% for knife-weapon, 0.69% for knife-deploy, 0.59% and 0.21% for the hand artefact, with 0.52% for the victim. The model produced multiple instances of lower scores; however, it identified the **main artefact** representing violence to prove the viability of the artefact processing for research question-6 in section 1.2.1.

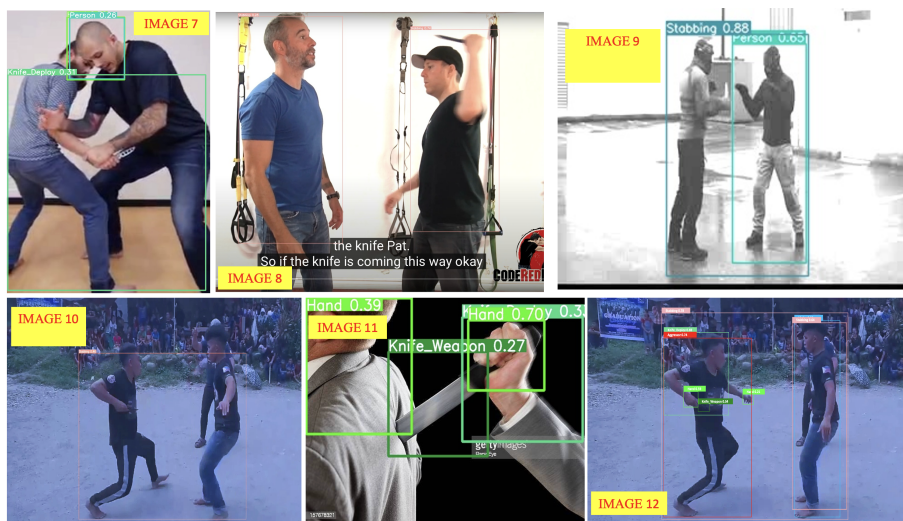


Figure 6.2: YOLOv5m Artefact Object Detection Results for Image 7-12.

With knowledge of YOLOv5m as activity recognition and fusion's superiority in Chapter 5, the chapter proposed a contributory fusion element exploiting YOLOv5m's object detection bounding box encapsulation capabilities in the context of YOLOv5m activity recognition. Similar blob analysis techniques integrated across a series of frames facilitated YOLOv5m object detection utilising static image processing. YOLOv5m object detection cannot facilitate activity recognition in the context of unrelated objects in static image processing. The frame sequence approach encompassing related actions in images allows the complete representation of action/artefacts across spatiotemporal periods. Remodelling

of the YOLOv5m activity recognition input sequence via programming accepts the data dispensing artefacts synchronised with its activity objects across a series of frames during inference. The concept enhances activity recognition utilising classification redundancy, which validates artefact presence in sync with its activity object over a spatiotemporal period at the output stage.

6.2 How YOLOv5m Artefact Detection Works

As Chapter 5 discussed the classification of stabbing objects over spatiotemporal boundaries, Chapter 6 enhances the classification by employing YOLOv5m activity recognition capability to identify stabbing action and artefacts relative to the knife-weapon, aggressor, victim, posturing, hand positioning, and knife deployed across a series of frames [288], [289], [290]. The objects discussed above represent fundamental attributes of artefact association, which aids in establishing existing pre-empted stabbing conditions to enhance action classification. Appreciating Chapter 5's fusion discussion, the concept integrated YOLOv5m activity recognition to target spatiotemporal actions and artefacts to enhance its predictions similar to [291], [292], [293]. The idea behind artefact prediction enhances activity recognition by fusing weapon/activity object outcomes, proving the existence of violence because weapons linked to stabbing actions are present in a scenario. Implementing further classification redundancy to support activity recognition enhances the processing's effectiveness before stabbing attacks, thus reducing the violent possibilities.

Achieving object detection's power by utilising the YOLOv5m activity recognition model incorporated final layer programming instructing YOLOv5m activity recognition to identify specific actions and artefacts depicting the pre-stages of violence. The analysis disclosed the model's ability to produce results on stabbing incorporating datasets processed in Chapter 5's operations. Considering blob analysis, specific object boundaries facilitate YOLOv5m activity recognition, which enables learning operations. To maintain exper-

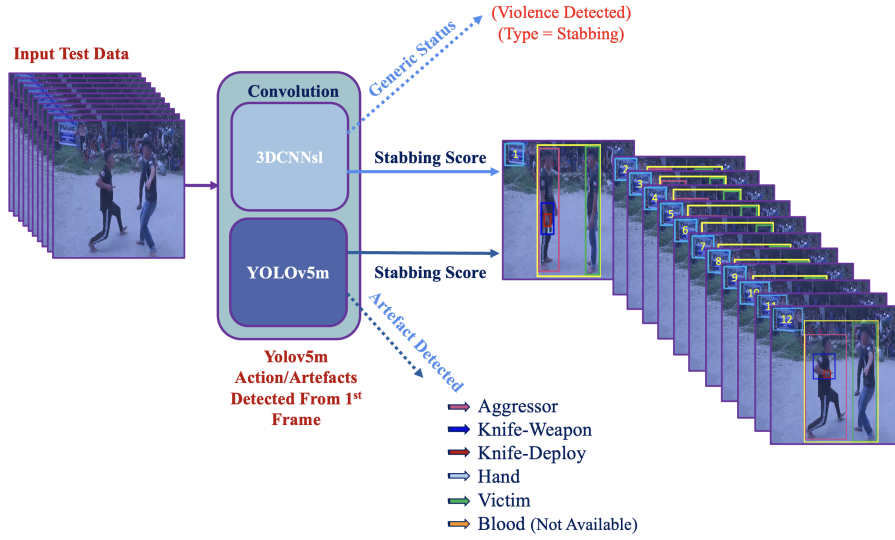


Figure 6.3: YOLOv5m Artefact Weighted Values.

-iment consistency, the operations implemented Chapter 5’s decision level fusion demonstrated in Section 5.1 and its protocols in section 5.1.2 to generate relevant data in alignment with [260], [261], [262], [263], [264] and [265]. Figure 6.3 demonstrates 3DCNNsl/YOLOv5m activity recognition artefact processing as a nuance of its framework before surveying alternative fusion approaches. The illustration emphasises the YOLOv5m activity recognition artefact processing, clarifying object detection class label dispensing capabilities for activity recognition. YOLOv5m activity recognition artefact processing dispenses results with clear distinctions of the objects in bounding boxes concerning their spatiotemporal location in the scenario. The operations apply no detection field outcomes to compensate for instances lacking the presence of weapons within the image scenery. YOLOv5m activity recognition artefact processing dispenses outcomes from the first frame, thus indicating the artefacts’ existence in violence. Incorporating Chapter 5’s 10-sample dataset encompassing a balanced ratio of fencing and stabbing actions disclosed the model’s aptitude, which justified the reduction in score biases by maintaining sample consistency. The approach investigates YOLOv5m artefact activity recognition using 12-salient frames

per video sample, establishing its effectiveness before considering alternative datasets and fusion adjustments. Incorporating similar Fusion Scheme-1 and 2 conditions employing Chapter 5’s 10-sample dataset aided in maintaining experiment consistency. At this level, YOLOv5m artefact activity recognition expressed its capability in 3-examples, which validated its viability.

6.2.1 YOLOv5m Action Recognition Artefact using 12-Frame Stabbing8.avi:

YOLOv5m activity recognition artefact processing confirmed its capability from an informed perspective on stabbing8.avi in Table 6.1. The analysis emphasised the presence of artefacts in frames #1-12 without the blood label. YOLOv5m activity recognition artefact processing proved its viability towards satisfying research question-6 activity recognition enhancement objectives in 1.2.

| Fr | Correct_Class | 3DPreds | YoloPreds | Fusion_Class | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | Blood |
|----|---------------|------------------|-----------|--------------|-----------|--------------|----------------|------|--------|-------|
| 1 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 2 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 3 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 4 | Stabbing8.avi | Activity Unknown | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 5 | Stabbing8.avi | Activity Unknown | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 6 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 7 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 8 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 9 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 10 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 11 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 12 | Stabbing8.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |

Table 6.1: YOLOv5m Artefact Activity Recognition on Stabbing8.avi.

6.2.2 YOLOv5m Action Recognition Artefact using 12-Frame Stabbing48.avi:

Table 6.1 findings demonstrated 2-instances lacking the presence of the artefacts in frames #17-18 action. No detection fields (ND) insinuate the obscuring of objects from the camera sensor's field of view. The analysis emphasised YOLOv5m artefact activity recognition's robustness, dispensing ten accurate labels suggesting violence in this sample. The outcomes satisfied objective-6 object detection towards enhancing action recognition in 1.2.

6.2.3 YOLOv5m Action Recognition Artefact using 12-Frame Fencing27.avi:

YOLOv5m activity recognition artefact evidence presented no results because Table 6.3 fencing27.avi represents non-violent action. The model depicted signs of classification challenges discerning the non-violent status by producing 9-activity-unknown instances with 3-correctly classified fencing outcomes. Although fencing27.avi dispensed no artefact results, the idea demonstrates examples of the model's processing capability even in action similarity conditions. YOLOv5m activity recognition artefact processing proved its viability towards satisfying research question-6 activity recognition enhancement objectives in 1.2.

6.3 How Fusion Activity Recognition Work using Artefact Support

With insight into Chapter 5 's fusion and YOLOv5m artefact activity recognition discussed above, Chapter Six proposed an alternative contributory fusion strategy via enhancing activity recognition's overall accuracy. The concept incorporates YOLOv5m activity recognition artefact processing to identify groups of target objects with an embedded weight values operation, which confirms the existence of violence. At this stage, merging Chapter 5 's fusion operations with YOLOv5m activity recognition artefact processing fostered enhancements utilising the same 10 and 50-sample

| Fr | Correct_Class | 3DPreds | YoloPreds | Fusion_Class | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | Blood |
|----|----------------|------------------|------------------|--------------|-----------|--------------|----------------|------|--------|-------|
| 13 | Stabbing48.avi | Activity Unknown | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 14 | Stabbing48.avi | Activity Unknown | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 15 | Stabbing48.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 16 | Stabbing48.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 17 | Stabbing48.avi | Stabbing | Activity Unknown | Stabbing | ND | ND | ND | ND | ND | ND |
| 18 | Stabbing48.avi | Stabbing | Activity Unknown | Stabbing | ND | ND | ND | ND | ND | ND |
| 19 | Stabbing48.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 20 | Stabbing48.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 21 | Stabbing48.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 22 | Stabbing48.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 23 | Stabbing48.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |
| 24 | Stabbing48.avi | Stabbing | Stabbing | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | ND |

Table 6.2: YOLOv5m Artefact Activity Recognition on Stabbing48.avi.

| Fr | Correct_Class | 3DPreds | YoloPreds | Fusion_Class | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | Blood |
|----|---------------|---------|------------------|------------------|-----------|--------------|----------------|------|--------|-------|
| 25 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 26 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 27 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 28 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 29 | Fencing27.avi | Fencing | Fencing | Fencing | ND | ND | ND | ND | ND | ND |
| 30 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 31 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 32 | Fencing27.avi | Fencing | Fencing | Fencing | ND | ND | ND | ND | ND | ND |
| 33 | Fencing27.avi | Fencing | Fencing | Fencing | ND | ND | ND | ND | ND | ND |
| 34 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 35 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 36 | Fencing27.avi | Fencing | Activity Unknown | Activity Unknown | ND | ND | ND | ND | ND | ND |

Table 6.3: YOLOv5m Artefact Activity Recognition on Fencing27.avi.

datasets as a plot to maintain data and experiment consistency discussed in section 5.4.2. The following process summarises the fusion embedding concept with weighted values to facilitate research question-6 points in 1.2. Figure 6.4 emphasises the fusion enhancement concept further by illustrating redundancy to confirm the presence of violent action objects, identifying artefacts, and computing an improved overall accuracy result. Given the existence of artefacts within the scenarios, the rationale behind the weight enhancement fusion processing validates pending violence by introducing classification redundancy.

1. **Process-#1** separates objects of interest dispensed from individual model outcomes after the starting point
2. **Task-#2** validates the presence of positive class labels (violence) in the outcomes before assessing the presence of artefacts. If the negative class labels exist, fusion-1 applies its results as the outcome.
3. **Operation-#3** confirms the artefact group presence from #2's processing and applies a weighted score. Artefact grouping relates to its significance to violence, which represents a specific weight value that triggers the enhancement programming upon its classification at process #3. If the artefact object groups are absent at operation #3, the processing employs fusion-1 output as the result.
4. **Process-#4** fuses label outcomes derived from process #3 and generates a score with their weight values reflecting the existence of the artefact groups.
5. **Task-#5** avoids biased results by regulating process #4's overall accuracy score at 100%. If the scores record thresholds below 100%, the processing accepts the results as the outcomes.
6. **Process-#6** applies subtraction adjustments to regulate the outcomes, which simultaneously represents artefacts should process #5 score exceed the 100% threshold. The idea regresses the summation of the artefact category applied at process #4,

which regulates its values at a 100% threshold.

7. **Process-#7** generates the outcome from the stages at the end. Fusion at this level produces a generic status (violence or not), its subclass (stabbing or fencing), and overall accuracy with artefacts that triggered the score enhancement. Artefact validation introduces classification redundancy, which promotes robust predictions, in contrast to Chapter 5's fusion, which contains action objects only.

152

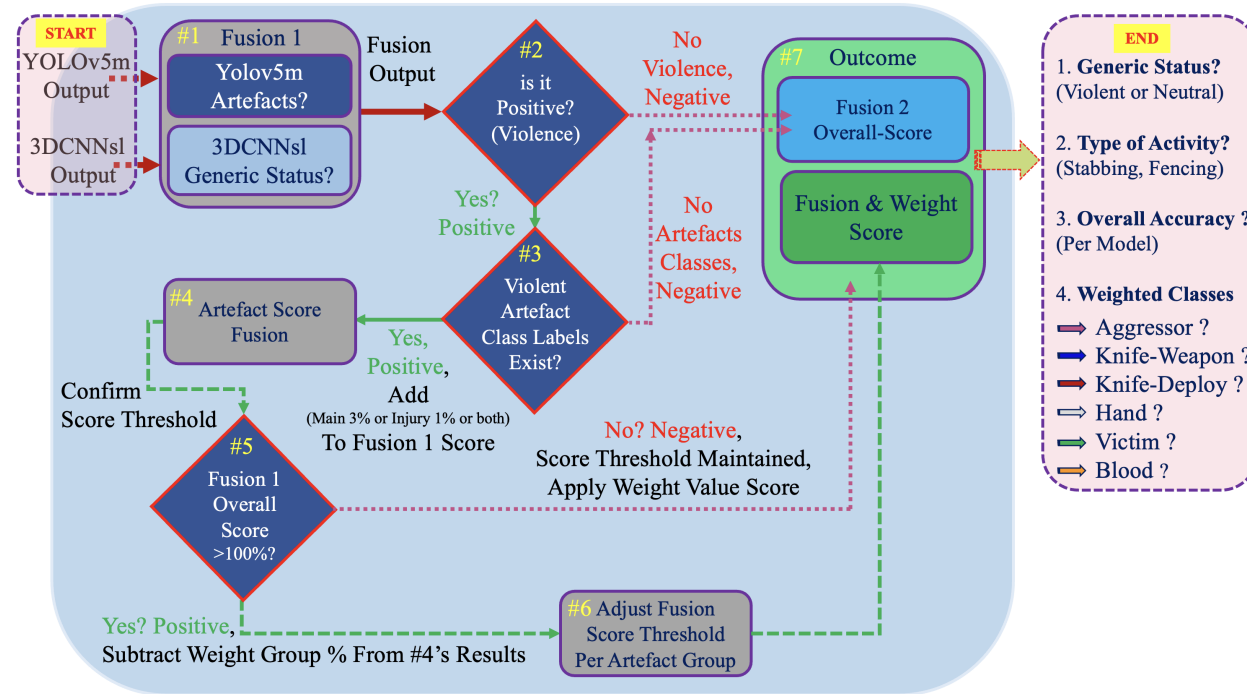


Figure 6.4: Proposed Fusion with Artefact Decision Level Weight Enhancement.

Demonstrating the proposed overall accuracy enhancement fusion operations discloses the prospect of embedding the weight values concept. The following Figure 6.5 illustration fortifies discussions on Figure 6.4, representing the significance of the weight embedding operations in 6-stages from start to end.

Step-1 Training: Trains the model to recognise homogeneous differences between violent/non-violent actions and identifies the presence of artefact objects.

Step-2 Input Test Data: Prepares the test data samples for the convolution and inference stages.

Step-3 YOLOv5m activity recognition artefact and 3DCNNsl Convolution: Initialises the convolution operations on the input data, generating the critical results

Step-4 Action Output: Applies configuration that dispenses action objects relative to generic and sub-classes for 3DCNNsl and action objects with artefact status for YOLOv5m activity recognition artefact processing.

Step-5 Fusion & Weights: Determines the action status and embeds weight scores to boast the final overall accuracy score via programming.

Step-6 Final Output: The model's programming dispenses the generic status (violent or not), its sub-classes, the overall accuracy, artefacts detected and confidence level of the operations, whether high, medium, or low.

6.3.1 Embedding YOLOv5m Activity Recognition Artefact Fusion Weights:

Appreciating Section 6.3 artefact operations further clarified the weighting process to simplify the weight delegation concept to represent specific artefact groups. Defining the weight value's significance in Table 6.4 emphasises specific artefacts discussed in Section 6.1.1 via

YOLOv5m activity recognition artefact processing for investigation. The concept incorporates integers representing the artefacts' existence during inference. If specific groups of artefacts are detected, a weight value is added to the proposed fusion's overall accuracy value, boosting its outcome. Conversely, the model preserves the original overall accuracy if no artefact objects exist per Section 6.3's discussions. The following integer and weighted value representation improves the classification, enhancing the proposed fusion's overall accuracy and confidence during inference.

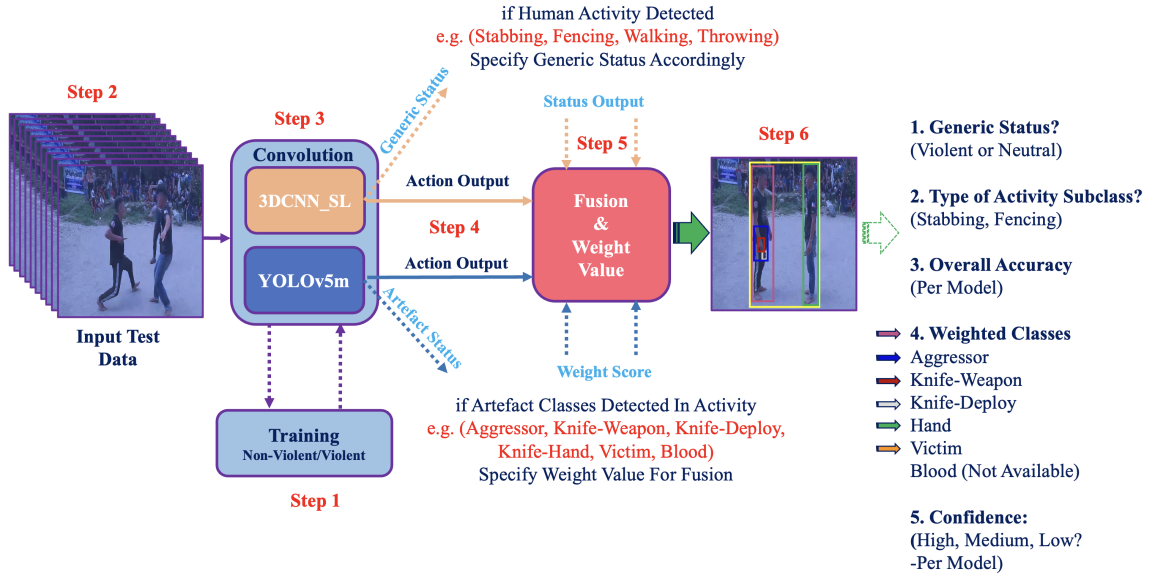


Figure 6.5: YOLOv5m Artefact Weighted Values

Specifying weights by grouping artefacts suggests the pre-stage importance of stabbing actions relative to violence. A weighted score of 0.03 is applied if YOLOv5m activity recognition artefact operations identify the **main artefacts** in action; this represents 3%. A weighted score of 0.01 is applied if YOLOv5m activity recognition artefact processing identifies artefacts concerning the **injury indicators**; this represents 1%. The previous weight values meet the criteria that signified the presence of artefacts during inference as a plot to enhance the proposed fusion's output. At this weight evaluation level, the strategy

| # | Weighted-Classes | | Weight-Value | Description |
|---|--------------------------|---|--------------|--|
| 1 | Main Artefacts | Aggressor (Committing the act) | 0.03 | If all elements are present in a scenario, weighted value will be applied to the stabbing accuracy. The idea validates the pre-empting of violence by detecting specific action attributes before an attack. |
| 2 | | Knife-Weapon (Aggressor has a weapon) | | |
| 3 | | Knife-Deployed (Aggressor with weapon displays stabbing posturing) | | |
| 4 | | Hand (Validates aggressor holding a weapon) | | |
| 5 | Identifies Injury | Victim (Person being attacked) | 0.01 | If this class is present in a scenario, the weighted value will be applied to the stabbing accuracy. The idea validates the victim sustaining injuries, immediate assistance required |
| 6 | | Blood (Victim received injuries) | | |

Table 6.4: Definition of Weighted Class & Score Values.

explored alternative integers experimentally (0.01-0.10) as a plot to evaluate the chosen values' effectiveness in representing the artefacts. The literature investigations suggested avoiding alternative weights reflecting (0.04-0.10) as they adversely affected the idea of increasing the final scores (beyond 100%) to an unrealistic outcome, thus intensifying biased outcomes. The concept, applied during development, validated instances of generating biased results.

6.3.2 Fusion using Artefact Decision Level Protocol Embedding:

Like Chapter 5 decision level fusion protocols, the operations followed the same Fusion Scheme-1 and 2 procedures (**Fusion Scheme-1: 12-salient frames representing a video sample processing, Fusion Scheme-2: processing the entire video**) to preserve experiment consistency and reduce bias outcomes. The idea evaluates weight values discussed above combined with YOLOv5m activity recognition artefact processing against Chapter 5 operations using the same 10 and 50-sample datasets towards establishing effectiveness. Modifying YOLOv5m activity recognition final output layers through pro-

gramming identifies distinct artefacts in Steps-1 to 3 via Figure 6.5. At this level, the strategy adapted Section 5.1.2 protocols, which considered specific conditions to identify cases of positive (artefacts exist) and negative (artefacts absent) actions. Table 6.5 conditions integrate the presence of artefacts as a redundancy measure to capture primitive stages of violent actions. After combining weight enhancements, applying these conditions ensured that the model dispensed the action’s generic/subclass label and fusion overall accuracy. Considering the artefact, decision-level protocols maintain partiality towards the positive action category. The intentional design suggests the presence of the generic status (action violent or not), its subclass (type), its overall accuracy fusion score after the artefacts support, the weight classes identified and its confidence ratio. Conversely, negative cases suggest the absence of violence and artefacts in the actions. Also, the bias design previously discussed intentionally maintains realistic scores below 100% and flags violence in negative cases as a contingency measure to be validated by a manual operator. The idea reduces escalating violent scenarios in its primitive stages. The strategy involves subtracting the artefact percentage identified during prediction from the fusion score to regulate the 100% score threshold, maintaining the significance of the artefact’s presence in the overall accuracy. The previously discussed strategy maintains a realistic outcome should the fusion score exceed the 100% threshold. Utilising Chapter 5 ’s overall accuracy formula, the overall accuracy calculates the correctly classified samples ratio divided by the total number of samples, multiplied by 100% to maintain a percentage value in line with [272], [273], [138], [275], [276], [277], and [278]. To summarise the formula below, ”arOA” represents the activity recognition overall accuracy where the total number of correctly labelled videos denotes” nCLvids”. The total number of videos represents ”TnVids”. The integrations of ”100%” reflect the percentage computation and the outcome.

$$arOA = \frac{nCLvids}{TnVids} \times 100$$

Although the approach introduces redundancy by applying a subtraction computation to the overall accuracy and returning the values, the tactic implies that Table 6.4 artefact process is combined with the fusion outcomes to represent score enhancements and validate its presence. The approach aligns the fundamental objectives behind research question-6 in 1.2, that is, to pre-empt violence by identifying artefacts of relevance (utilising object detection’s power) to enhance the classification outcome and reduce its impact. Introducing artefact enhancement possibilities in Table 6.6 aligned adapted decision-level protocols to establish the proposed fusion combined with artefact weight accuracy outcomes. Emphasising the notion of fusion combined with weight score embedding

| KEY | | Action Unknown | AU | | | | |
|-----|--------------|------------------|------------------|-------------------------------|--------------|--------|--------|
| | | No Detection | ND | | | | |
| # | Fusion Score | Main Artefact | Injury Artefacts | Weight Value Protocol Applies | | | |
| | | | | Outcome | Computation | | |
| 1 | Positive | Positive (Exist) | Positive (Exist) | Positive | Fusion Score | +0.03% | +0.01% |
| 2 | Positive | Positive (Exist) | Negative/ND | Positive | Fusion Score | +0.03% | |
| 3 | Positive | Negative/ND | Negative/ND | Negative | Fusion Score | | |
| 4 | Positive | Negative/ND | Positive (Exist) | Positive | Fusion Score | +0.01% | |
| 5 | Negative/AU | Negative/ND | Negative/ND | Negative/AU | Fusion Score | | |
| 6 | Negative/AU | Positive (Exist) | Positive (Exist) | Negative/AU | Fusion Score | | |
| 7 | Negative/AU | Negative/ND | Positive (Exist) | Negative/AU | Fusion Score | | |
| 8 | Negative/AU | Positive (Exist) | Negative/ND | Negative/AU | Fusion Score | | |

Table 6.5: Artefact Decision Level Fusion.

representing existing artefacts incorporated the following conditions discussed in Table 6.5, which enhanced the overall accuracy. The constraints provided through programming configurations initiated the procedures described in Figure 6.4’s illustration. The conditions are as follows.

Adapted Decision Level Protocols Incorporating Artefact Support:

1. If fusion-one overall accuracy is **positive** (between 50-100% ratio) and the **main**

and **injury artefacts** are **positive (artefacts detected)**, decision protocols apply a summation of 3% to the overall accuracy outcome for **main artefacts** present in the actions and 1% for **injury artefacts** for a total of 4% as a positive outcome. If the overall accuracy outcome exceeds 100%, decision protocols subtract the weights accordingly as a regulatory procedure to specify the significance of the artefacts and preserve the maximum outcome of 100%. The decision protocols ignore low overall accuracy scores below the 50% ratio, implying insignificant classification operations.

2. If fusion-one overall accuracy is **positive** (between 50-100% ratio), **main artefacts** are **positive (artefacts detected)**, and **injury artefacts** are **negative (No Detection)**, the case is positive. The protocols subtract 3% simultaneously, returning 3% to the overall accuracy to represent **main artefacts** and 0% for **injury artefacts**. If the overall accuracy outcome exceeds 100%, decision protocols subtract the weight as a regulatory procedure to indicate artefact presence and preserve the maximum outcome of 100%.
3. If fusion-one overall accuracy is **positive** (between 50-100% ratio), the **main** and **injury artefacts** dispense **negative (No Detection)**, and the outcome is **negative**. However, the decision protocol's process ignores Fusion-One's overall accuracy enhancements to suggest the absence of artefacts.
4. If fusion-one overall accuracy is **positive** (between 50-100% ratio), **main artefacts** are **negative (No Detection)**, and **injury artefacts** are **positive (artefacts detected)**, the prediction is **positive**. However, the decision level's programming applies a summation of 1% to fusion-one overall accuracy to insinuate **injury artefacts'** presence. If the overall accuracy is 100%, decision protocols subtract the existing weight value as a regulatory procedure to indicate artefact presence and preserve the maximum outcome of 100%. The anomaly occurs if

violence moves outside the field of view of the on-site camera sensors.

5. If fusion-one overall accuracy dispensed is **Negative/Action Unknown** (action unknown/ score below 50%), the **main** and **injury artefacts** display **negative (No Detection)**, and the outcome is **negative**. The artefact decision level protocol's programming applies no weight value enhancement to fusion-one's overall accuracy outcome.
6. If fusion-one overall accuracy dispensed **Negative/Action Unknown** (action unknown/ score below 50%), **main** and **injury artefacts** show **positive (artefacts detected)**, the outcome is **Negative/Action Unknown**. The artefact decision level protocol's programming applies no weight value enhancement to fusion-one overall accuracy outcome. The anomaly occurs because the individual models' processing detects artefact features. Although Fusion-one identified the artefact's presence in the individual activity recognition models, the decision protocols consider Fusion-1's outcome regulating its decision to coincide with such.
7. If fusion-one overall accuracy dispensed **Negative/Action Unknown** (action unknown/ score below 50%), **main artefacts** are **negative (No Detection)**, and **injury artefacts** are **positive (artefacts detected)**, the outcome is **Negative/Action Unknown**. Like protocol #6, The artefact decision level protocol's programming applies no weight value enhancement to the fusion-one overall accuracy outcome.
8. If fusion-one overall accuracy displays **Negative/Action Unknown** (action unknown/ score below 50%), **main artefacts** are **positive (artefacts detected)**, **injury artefacts** are **negative (No Detection)**, and the outcome is **Negative/Action Unknown**. Like process #6, The artefact decision level protocol's

programming applies no weight value enhancement to the fusion-one overall accuracy outcome.

6.3.3 Fusion Scheme-1 12-Frame Artefact Support (on 10-Test Samples):

Appreciating the artefact processing and its rationale above provided the results that validated its effectiveness in satisfying research question-6’s objectives in section 1.2.1. The analysis confirmed the dominance of the proposed fusion with artefact support with an increase in performance by 3% to suggest the presence of **main artefacts** over the original overall accuracy in Table 6.6. The main artefacts recorded 37 instances with no outcomes for **injury artefacts**, which signified the presence of the artefacts with no sustained injuries in the Fusion Scheme-1 scenario. The artefact findings demonstrated an enhanced original overall accuracy of 3DCNNsl at 43.83% at medium

| Dataset | Original Performance | | Artefact Enhancement | | | Confidence | Correctly_Classified_Instances | | | | Incorrectly_Classified_Instances | | | |
|--|----------------------|--------------------------------|----------------------|--------|--------|------------|--------------------------------|----------|------------------|-------|----------------------------------|----------|------------------|-------|
| | Model | Overall-Accuracy | Main | Injury | Score | Threshold | Fencing | Stabbing | Activity Unknown | Total | Fencing | Stabbing | Activity Unknown | Total |
| 10-Samples at 12-Frames Per Sample = 120 Frames | 3DCNNsl | 40.83% | 37 | 0 | 43.83% | Medium | 0 | 49 | 0 | 49 | 0 | 50 | 21 | 71 |
| | YOLOv5m | 33.33% | | | 36.33% | Low | 3 | 37 | 0 | 40 | 0 | 46 | 34 | 80 |
| | Artefact Fusion | 52.50% | | | 55.50% | Medium | 3 | 60 | 0 | 63 | 0 | 48 | 9 | 57 |
| | Led By | 11.67% Over 3DCNNsl | | | | | | | | | | | | |
| | | 19.17% Confidence Over YOLOv5m | | | | | | | | | | | | |

Table 6.6: Fusion Scheme-1 Artefact Results for 10-Samples at 12-Frames per Sample.

confidence compared to its original performance at 40.83%. 3DCNNsl operations recorded 49 correctly classified and 71 incorrect instances between the stabbing and activity unknown classes. YOLOv5m activity recognition artefact process dispensed 36.33% as the score enhancement result with low confidence compared to its original performance at 33.33%. YOLOv5m demonstrated 40 correctly classified instances with 80 incorrect instances amongst the sub-classes. The proposed artefact fusion recorded 55.50%, demonstrating its dominance over its original performance at 52.50% and other models. Fusion at this level produced 63 correctly classified instances and 57 incorrect categorisations at

this level. The proposed artefact fusion results substantiate its dominance over 3DCNNsl by 11.67% and YOLOv5m by 19.17% to validate its effectiveness.

6.3.4 Fusion Scheme-1 12-Frame Artefact Support (on 50-Test Samples):

The evaluation of surplus data in Table 6.7 demonstrated an increment in overall accuracy performance and the presence of the artefacts during inference at this level. The analysis showed the artefact’s effectiveness in performance by 3% with 204 main instances using 50 12-frames per test sample compared to section 6.3.3’s 10. The findings disclosed no support for the injury artefacts, signifying its absence. 3DCNNsl artefact support produced a high confidence outcome of 69.50% overall accuracy compared to its 66.50 % original score. The data proved that 3DCNNsl incurred 400 correct predictions between the stabbing and fencing sub-classes and 200 misclassifications for the activity unknown class. YOLOv5m artefact support dispensed a high confidence outcome of 73.16%, suggesting the presence of the main artefacts compared to its original score of 70.16%. At this stage, YOLOv5m correct classifications increased to 421 with a reduction in inaccuracy predictions at 179 over 3DCNNsl’s 200 cases. Fusion with artefact support recorded 86.83% at high confidence compared to its original performance at 83.83%. Fusion’s artefact

| Dataset | Original Performance | | Artefact Enhancement | | | Confidence | Correctly_Classified_Instances | | | | Incorrectly_Classified_Instances | | | |
|--|----------------------|--------------------------------|----------------------|--------|--------|------------|--------------------------------|----------|------------------|-------|----------------------------------|----------|------------------|-------|
| | Model | Overall-Accuracy | Main | Injury | Score | Threshold | Fencing | Stabbing | Activity Unknown | Total | Fencing | Stabbing | Activity Unknown | Total |
| 50-Samples at 12-Frames Per Sample= 600 Frames | 3DCNNsl | 66.50% | 204 | 0 | 69.50% | High | 203 | 197 | 0 | 400 | 0 | 0 | 200 | 200 |
| | YOLOv5m | 70.16% | | | 73.16% | High | 211 | 210 | 0 | 421 | 0 | 0 | 179 | 179 |
| | Artefact Fusion | 83.83% | | | 86.83% | High | 260 | 243 | 0 | 503 | 0 | 0 | 97 | 97 |
| | Led By | 17.33% Confidence Over 3DCNNsl | | | | | | | | | | | | |
| | | 13.67% Confidence Over YOLOv5m | | | | | | | | | | | | |

Table 6.7: Fusion Scheme-1 Artefact Results for 50-Samples at 12-Frames per Sample.

support incurred the highest correct subclass classifications at 503 instances, with 97 activity unknown misclassifications. Artefact support proved its effectiveness by dominating 3DCNNsl by 17.33% and YOLOv5m by 13.67%. The findings proved the effectiveness of the proposed fusion with artefact support, which satisfied research objective-6 towards using YOLOv5m’s object detection capabilities to enhance activity recognition’s status in 1.2.

6.3.5 Fusion Scheme-2 Whole Video Artefact Support (on 10-Test Samples):

Appreciating the evidence of the 10/50 12-frames per sample results in Table 6.6 and Table 6.7, the artefact support results using whole video samples proved its capability as a proof of concept utilising surplus data in Table 6.8. The findings projected the **main artefact** in 1059 cases, which triggered the 3% score increase with no predictions for the injury group. Like Table 6.6 and Table 6.7, the absence of the **injury artefacts** signified that no individual injuries occurred in the Fusion Scheme-2 scenario. 3DCNNsl’s artefact support results showed high confidence with an increase in performance of 85.43% over the original performance results at 82.43%, Table 6.6 at 43.83% and Table 6.7 at 69.50%. 3DCNNsl artefact processing produced 2473 correctly classified instances with 527 misclassifications at the subclass level. YOLOv5m artefact support results improved by 82.50% at high confidence compared to its original

| Dataset | Original Performance | | Artefact Enhancement | | | Confidence | Correctly_Classified_Instances | | | | Incorrectly_Classified_Instances | | | |
|--------------------------------------|-------------------------------|------------------|----------------------|--------|--------|------------|--------------------------------|----------|------------------|-------|----------------------------------|----------|------------------|-------|
| | Model | Overall-Accuracy | Main | Injury | Score | Threshold | Fencing | Stabbing | Activity Unknown | Total | Fencing | Stabbing | Activity Unknown | Total |
| 10-Whole Video Samples = 3000 Frames | 3DCNNsl | 82.43% | 1059 | 0 | 85.43% | High | 1275 | 1198 | 0 | 2473 | 0 | 1 | 526 | 527 |
| | YOLOv5m | 79.50% | | | 82.50% | High | 1105 | 1280 | 0 | 2385 | 102 | 0 | 513 | 615 |
| | Artefact Fusion Led By | 85.20% | | | 88.20% | High | 1228 | 1328 | 0 | 2556 | 102 | 0 | 342 | 444 |
| | 2.77% Confidence Over 3DCNNsl | | | | | | | | | | | | | |
| | 5.70% Confidence Over YOLOv5m | | | | | | | | | | | | | |

Table 6.8: Fusion Scheme-2 Artefact Results for 10-Samples in Whole Video Processing.

performance score of 79.50%, Table 6.6 at 36.33% and Table 6.7 at 73.16%. YOLOv5m artefact support produced 2385 correctly classified instances at this level, increasing misclassification by 615 compared to 3DCNNsl's 527 instances. Analysis accentuated the proposed fusion with artefact support superiority at 88.20% at high confidence compared to its original outcome at 85.20%, Table 6.6 at 55.50% and Table 6.7 at 86.83%. The proposed fusion generated the highest accurate prediction at 2556 instances with a decrease in misclassifications by 444 compared to 3DCNNsl's 527 and YOLOv5m's 615 instances. Fusion incorporating artefact support proved its effectiveness by demonstrating a performance lead of 2.77% over 3DCNNsl and 5.70% over YOLOv5msl. Fusion Scheme-2 operations results satisfied research objective-6 in 1.2 as a proof of concept by applying YOLOv5m activity recognition artefact power to enhance the status of its activity recognition.

6.3.6 Fusion Scheme-2 Whole Video Artefact Support (on 50-Test Samples):

Artefact support proved its effectiveness using surplus data for Fusion Scheme-2's operations by dispensing 4422 cases at a 3% score increase in Table 6.9. Like Table 6.6 to Table 6.8, the operations observed **no injury** to identify the possibility of increasing the original performance outcome further. Though the artefact support outcomes recorded high confidence thresholds, the operations decreased performance for all outcomes compared to Table 6.8's 3DCNNsl 85.43%, YOLOv5m 82.50% and Fusion artefact support 88.20%. The increase in surplus data caused performance issues linked to the complexity of processing more actions, which affected the models' ability to provide correct responses in such cases. 3DCNNsl artefact support recorded 72.05% compared to its original performance score of 69.05%. At this level, 3DCNNsl produced 11,394 correctly classified instances with 5,106 misclassifications for the subclass categories. The evidence exposed a decrease in performance of 0.19% for YOLOv5m at 71.86% compared to 3DCNNsl's 72.05%. YOLOv5m artefact support dispensed 11,198 correctly classified subclass instances and 5,302

misclassifications

| Dataset | Original Performance | | Artefact Enhancement | | | Confidence | Correctly_Classified_Instances | | | | Incorrectly_Classified_Instances | | | |
|---|----------------------|--------------------------------|----------------------|--------|--------|------------|--------------------------------|----------|------------------|-------|----------------------------------|----------|------------------|-------|
| | Model | Overall-Accuracy | Main | Injury | Score | Threshold | Fencing | Stabbing | Activity Unknown | Total | Fencing | Stabbing | Activity Unknown | Total |
| 50-Whole Video Samples =16,500 Frames | 3DCNNsl | 69.05% | 4422 | 0 | 72.05% | High | 5503 | 5891 | 0 | 11394 | 0 | 0 | 5106 | 5106 |
| | YOLOv5m | 67.86% | | | 70.86% | High | 5211 | 5987 | 0 | 11198 | 1348 | 0 | 3954 | 5302 |
| | Artefact Fusion | 84.75% | | | 87.75% | High | 6853 | 7131 | 0 | 13984 | 1082 | 0 | 1434 | 2516 |
| | Led By | 15.70% Confidence Over 3DCNNsl | | | | | | | | | | | | |
| | | 16.89% Confidence Over YOLOv5m | | | | | | | | | | | | |

Table 6.9: Fusion Scheme-2 Artefact Support for 50-Samples in Whole Video Processing.

compared to previous discussions on 3DCNNsl’s outcomes. Finally, fusion artefact support in Table 6.9 reduced performance by 0.45% at 87.75% compared to Table 6.8 88.20% fusion artefact support score. However, the model demonstrated its performance enhancement effectiveness by dominating the original performance score at 84.75%. Fusion with artefact support demonstrated capability, leading 3DCNNsl by 15.70% and YOLOv5m by 16.89% at this level. The evidence satisfied research objective-6 in section 1.2 as a proof of concept based on evaluating the results towards employing the power of object detection to enhance activity recognition’s status.

6.4 Artefact Operational Discussions

The results above proved the effectiveness of fusion fortified by artefact enhancements instead of Chapter 5’s fusion in a self-operating approach. Fusion Scheme-1 and 2 evaluated the models’ aptitude in conditions containing surplus and less data. The notion established fusion’s ability to produce consistent results, validating the idea as a proof of concept from an informed decision perspective. At this level, fusion containing artefact support generated score improvements for its confidence threshold regardless of the action similarity complexity within Fusion Scheme-1 and 2’s data. Investigating alternative weights established the scope of their impact towards achieving optimal results.

Although integrating fusion score regulators facilitated the processing, alternative weights increased the regulator’s classification bias beyond the overall accuracy threshold of 100% (101%-104%). Abandoning the idea of alternative weigh values proved helpful as it generated insignificant results that negatively impacted the entire operation.

6.4.1 Fusion Scheme-1 12-frame Discussions on Artefact Processing:

The operations experienced output anomalies reflecting the absence of artefacts during inference because YOLOv5m activity recognition artefact processing produced inconsequential outcomes. Insufficient data in experiments utilising 12 frames reflecting violence impacted its classification, producing fewer misclassifications and lower overall score outcomes. Providing scenarios replicating complex unidentified activity unknown classifications indicated in stabbing24.avi with no detections emphasises Fusion Scheme-1’s processing. The findings suggest the artefacts’ absence in the sample, as demonstrated in Figure 6.6. Figure 6.7 solidified YOLOv5m activity recognition artefact effectiveness on

| Fr | Correct_Class | YoloPreds | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | Blood |
|----|----------------|------------------|-----------|--------------|----------------|------|--------|-------|
| 12 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 13 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 14 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 15 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 16 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 17 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 18 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 19 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 20 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 21 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 22 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 23 | stabbing24.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |

The model produced no artefact outcomes on stabbing24.avi due to its weak prediction

Figure 6.6: Fusion Scheme-1 12-Frame Stabbing24.avi Activity Unknown Cases.

stabbing37.avi, showing 5-artefact instances. The model depicted 7-activity unknown misclassifications with several no-detection cases for the victim and blood artefacts. The results insinuated that the model experienced challenges discerning the action’s state re-

lating to its motion complexity. Because the presence of blood artefact is improbable in the pre-stages of violence, no detection outcomes proved as accurate as anticipated. However, the role of blood artefacts is to identify injuries sustained and individuals requiring urgent medical attention. The findings proved artefact support effectiveness on stabbing37.avi, which satisfied research question-6 classification enhancement objectives in section 1.2.1.

| Fr | Correct_Class | YoloPreds | Aggressor | Knife-Weapon | Knife-Deployed | Hand | Victim | Blood |
|----|----------------|------------------|-----------|--------------|----------------|------|--------|-------|
| 24 | stabbing37.avi | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | 1 | ND |
| 25 | stabbing37.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 26 | stabbing37.avi | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | 2 | ND |
| 27 | stabbing37.avi | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | 2 3 | ND |
| 28 | stabbing37.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 29 | stabbing37.avi | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | 4 | ND |
| 30 | stabbing37.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 31 | stabbing37.avi | Stabbing | Aggressor | Knife-Weapon | Knife-Deployed | Hand | 5 | ND |
| 32 | stabbing37.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 33 | stabbing37.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 34 | stabbing37.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |
| 35 | stabbing37.avi | Activity Unknown | ND | ND | ND | ND | ND | ND |

Figure 6.7: Fusion Scheme-1 12-Frame Stabbing37.avi Artefacts Special Cases

6.4.2 Fusion Scheme-2 Whole Video Discussions on Artefact Processing:

The evidence dispensed by the models at this level solidified the impact of incorporating more data to achieve optimal performance. The positive impact of the surplus data demonstrated a coherent increment in performance at the whole video processing level as opposed to the 12 frames per sample exercise of Fusion Scheme-1. Analysis of the whole video experiments disclosed vivid score differences because of the complexity of the action similarity conditions and the impact of surplus data. Observations showed an innate relation between applying surplus data and realistic outcomes. Fusion-Scheme-2's operations validated the relationship of applying surplus data to achieve realistic outcomes. Although artefact processing recorded high overall accuracy scores in Fusion Scheme-1 and 2, the score's relationship differed as expected among the video experiments utilising 10 and 50 samples because of the surplus data. The relationship between performance and surplus

data revealed a strong misclassification link when the sample size increased. As anticipated, the impact of the data/performance relationship relative to surplus data contained higher levels of complex actions, further intensifying the operation’s classification challenge on more action similarity conditions.

6.5 Conclusion

Chapter 6 investigated the prospects of Chapter 5’s fusion amalgamating YOLOv5m activity recognition artefact processing in 2-fusion schemes to achieve optimal results. The evidence demonstrated the power of fusion with artefact support by designing specific weights to reflect the presence of artefact categories, thus boosting the classification outcome within violent scenarios. Table 6.10 artefact enhancements solidified its dominance over fusion without support, 3DCNNsl, and YOLOv5m activity recognition operations. Like discussions on state-of-the-art comparisons in Section 4.6 and Section 5.7, it proved impractical to compare other state-of-the-art solutions because of the aim and objectives coupled with the dataset pre-stage structure. The analysis accentuated Fusion Scheme-2’s artefact processing dominance of 88.20% and 87%.75% at high confid-

| Dataset | Original Performance | | Artefact Enhancement | | | Confidence |
|-----------------------------------|----------------------|------------------|----------------------|----------|---------------|---------------|
| | Model | Overall Accuracy | Main | Injury | Score | Threshold |
| 10-Samples 12-Frames | 3DCNNsl | 40.83% | 37 | 0 | 43.83% | Medium |
| | YOLOv5m | 33.33% | | | 36.33% | Medium |
| | Fusion | 52.50% | | | 55.50% | Medium |
| 50-Samples 12-Frames | 3DCNNsl | 66.50% | 204 | 0 | 69.50% | High |
| | YOLOv5m | 70.16% | | | 73.16% | High |
| | Fusion | 83.83% | | | 86.83% | High |
| 10-Whole Video Samples | 3DCNNsl | 82.43% | 1059 | 0 | 85.43% | High |
| | YOLOv5m | 79.50% | | | 82.50% | High |
| | Fusion | 85.20% | | | 88.20% | High |
| 50-Whole Video Samples | 3DCNNsl | 69.05% | 4422 | 0 | 72.05% | High |
| | YOLOv5m | 67.86% | | | 71.86% | High |
| | Fusion | 84.75% | | | 87.75% | High |

Table 6.10: Fusion Scheme-1 & 2 Artefact Processing Dominance.

-ence thresholds over Fusion Scheme-1 overall accuracy scores at 52.50% and 87%.75%. The evidence proved fusion’s artefact enhancement superiority over 3DCNNsl’s 40.83%, 82.43%, 69.05% and YOLOv5m activity recognition’s 33.33%, 70.16%, 79.50%, 67.86% as individual operations. Figure 6.8 graphical projection visually validated the proposed fusion processing superiority emphasised in Table 6.10 over 3DCNNsl and YOLOv5m. The findings substantiated Fusion Scheme-2 robustness, which fulfilled research question-6 in section 1.2.1 and section 1.2.2 to conclude the investigations.

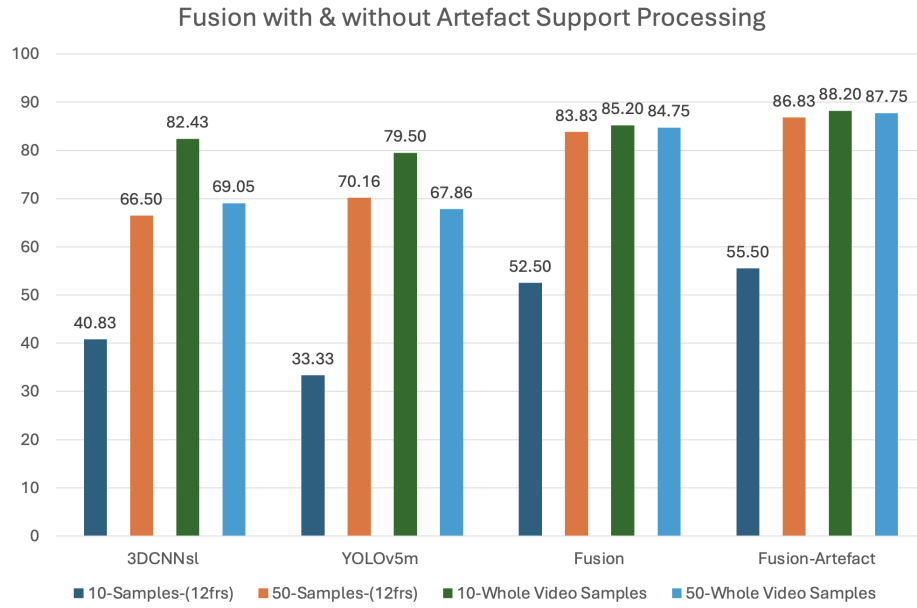


Figure 6.8: Fusion’s Dominance Over Individual Processing

Chapter 7

Conclusion

The current research proposal presents violent activity recognition pre-empting techniques that considered the amalgamation of 2-state-of-the-art convolution models through fusion to remedy individual processing limitations regarding homogeneous and heterogeneous human actions. A recap of the 2-stage literature investigations disclosed 3-dimensional convolution neural networks (3DCNN) state-of-the-art activity recognition model and you-only-look-once (YOLO) object detection with architecture conforming possibilities, which fosters activity recognition utilising frame sequences during input stages. The proposal also addresses contributory decision-level fusion approaches through programming configurations aligned with quantitative experiments. This research reveals processing dominance by deliberately incorporating complex action similarity data to validate the model's effectiveness and achieve high classification outcomes. Achieving the solution incorporated the development of research objectives driven by the research motivation, which defined the investigation's structure and experimental approach within Chapter 1. The following are milestones organised into sections to summarise the main findings and contributory prospects regarding the proposal's objectives. Section 7.1 considers research question evaluations to establish the attainment of the objectives. Section 7.2 features the proposal's

contributory factors with the proposal's limitations' in Section 7.3. Section 7.4 projects future research initiatives and concerns, thus highlighting prospects for violent artefact activity recognition.

7.1 Research Questions' Assessment

Chapter 1: Accentuated the structure of the expected elements within the thesis, followed by a lucid summary of the research's aim and objectives driven by the motivation to solve the primitive stages of violence before its lethal impact.

Chapter 2: This chapter emphasises the general background on classification intricacies regarding several sequential processes, data handling, software/hardware requirements, and evaluation procedures, which promotes processing efficiency. The chapter simplifies artificial intelligence operations, linking the reader's understanding of essential components applied to achieve the proposal's outcome. The chapter also accentuates a 2-stage literature review evaluation for potential mediums leading to the discovery of 3DCNN/YOLOv5m. The investigation targeted research question-1 in 1.2 to establish avenues for achieving violent activity recognition with weapons.

Chapter 3: This chapter detailed action class descriptions of the dataset acquisition, which accumulated 3DCNNsl RWVAD1st dataset and YOLOv5m RWVAD2nd dataset. The subdivisions expound on procedures satisfying each model's processing standards concerning architectural differences. Specifying the protocols sets the structure for the experimental operations towards appraising 3DCNN/YOLOv5 individual processing limitations in Chapter 4. The context designed the experimental foundation to achieve research question-1 in 1.2 as the primary objective concerning recognising violent activity with weapons as artefacts.

Chapter 4: This chapter presents the experimental assessments which defined the feasibility of YOLOv5/3DCNN's processing and the fulfilment of research question-2 in 1.2 via the impact of data pre-processing concepts. Chapter 4 defined YOLOv5m activity recognition as identifying specific objects across multiple frames compared to its object detection static image approach. The section satisfied research objective-1 via in 1.2, which considered violence and weapon artefact predictions in CCTV videos. The chapter expressed YOLOv5 activity recognition challenges when identifying small weapon instruments emulating sporadic accelerated motion specific to its convolutional architecture. Although YOLOv5 activity recognition created moderate result ratios of 67-74% in Section 4.2.1, the sporadic motion of the small object, combined with complex homogeneous attributes amid violent and non-violent human actions, impacted its classification operations negatively. 3DCNN's investigations substantiated its classification limitations when processing complex action similarity experimental categories. Because of action complexity classification issues, the model's action recognition processing capabilities declined. At this level, 3DCNN state-of-the-art dispensed 62% for fencing and 75% for stabbing. The appraisal substantiated YOLOv5m activity recognition, as 3DCNNsl emerged from fashioning additional layer support using reconfiguration. The analysis fulfilled the action class classifications using dataset variations to satisfy research questions 3 and 4 in 1.2. Chapter 4 proposed the object combining technique towards enhancing the model's processing outcomes. At this level, the operations fulfilled research question-1 in section 1.2 as the primary objective, concluding YOLOv5m/3DCNNsl violent activity recognition before considering alternative measures.

Chapter 5: This section explored fusing 3DCNNsl/YOLOv5m activity recognition with decision-level support to establish the fulfilment of research question-5's processing superiority in section 1.2.1. The operations investigated several techniques (max

score) to conclude fusion’s feasibility in addition to 2-fusion schemes, which validated the performance. Chapter 5 projected the proposed fusion status by analysing 2-fusion schemes to achieve optimality. The procedure considered a 10/50-sample dataset where each sample represented 12 frames to simulate conditions lacking data as Fusion Scheme-1 and conditions with surplus data concerning whole video operations as Fusion Scheme-2. The idea established the difference in model effectiveness in conditions reflecting short videos with limited frames and videos with longer durations. Fusion-Scheme-2 dominated Fusion Scheme-1 to insinuate the need for more data to dispense more robust overall accuracy results. At this level, the evidence emphasised YOLOv5m activity recognition dominance on the fencing27.avi action similarity sample compared to 3DCNNsl. The findings substantiate the proposed fusion’s effectiveness, fulfilling research question-5 model superiority in section 1.2.1 by producing significant outcomes.

Chapter 6: With knowledge of fusion’s effectiveness in Chapter 5, Chapter 6 surveyed YOLOv5m activity recognition’s feasibility of exploiting its object detection power. Chapter 6 investigated weapon artefact classification resonance to enhance further the proposed fusion activity recognition status incorporating weight value embedding. The analysis corroborated YOLO artefact activity recognition’s effectiveness with high ratings using the same fusion scheme strategies in Chapter 5 at this level. By achieving the artefact classification milestone, Chapter 6 proposed an artefact weight value embedding operation that enhances the proposed fusion activity recognition status regardless of the violent conditions. Table 6.10 ’s results accentuated the proposed fusion’s robustness fortified by artefact embedding procedure on violence and conditions containing homogeneous action attributes. Task-3’s 88.20% and Task-4’s 87.75% overall accuracy at high confidence proved the effectiveness of fusion artefact enhancement processing. Chapter 6 artefact processing mitigated the

distinction of actions between action similarity containing strong characteristics of violence. The findings indicated that combining objects and data processes discussed in section 4.5.4 and section 4.5.5 strongly influenced the outcomes. Chapter 6 fulfilled the research question-6 objective in section 1.2, which illustrates investigating object detection's power towards enhancing the classification status of activity recognition.

7.2 Contributory Factors

The proposal's contributions disclosed in **Section 1.4** are emphasised as follows:

1. **Chapter-3 to 4** Conducted performance testing of two known machine learning techniques (YOLOv5m and 3DCNNsl) in independently recognising violent and non-violent activities in CCTV video footage. The findings demonstrated the model's processing capability to classify complex violence in multiple scenarios. The concept stimulated the need for the proposed fusion as an alternative robust approach.
2. **Chapter-5** Demonstrated violent activity recognition performance in such videos when both machine learning techniques operate in tandem through a decision-level fusion operation. The strategy combined the processing of the individual models utilising decision-level techniques to generate robust results regardless of the homogeneous action attributes.
3. **Chapter-6** Implemented performance enhancement by further incorporating threat object detection in the previous combined solution. The idea applied weight value embedding to suggest the presence of lethal weapon objects/artefacts, thus enhancing the outcome. The method incorporated classification redundancy to suggest certainty of action if weapons belonging to a specific class of violence exist within

a given scenario. The proposed approach demonstrated classification effectiveness, which satisfied research questions 1-6.

7.2.1 Introduction to publications:

The research efforts presented the first publication, "Suspicious Activity Detection for Defence Applications". Though it is the final review stage, the paper demonstrated the operations utilising 3DCNNsl as a tool for activity recognition. The following papers are due for publication in the coming future.

1. **Leveraging YOLOv5m for Detecting Violent Activities: A Novel Approach in Activity Recognition**
2. **Fusing YOLOv5m and 3DCNNsl Activity Recognition for Defence Application**
3. **Weighted Fusion of YOLOv5m and 3DCNNsl for Robust Violent Activity Classification in Defence Systems**

7.3 Limitations of the Study

Although the research question's assessment and contributory factors suggest high performance, acknowledging restrictions impacting the operations and the scope of the research efforts at this level provided context for further research.

Dataset Acquisition Limitations: Violent data proved challenging to obtain, containing the crux of the pre-start, middle and end attributes (primitive stages of violence) in publicly available datasets. Although some publicly available datasets contain violence relative to movies, the application of such data conveying violent movies and acting contributed to

flawed results. Because of this notion, a manual investigation into social media platforms acted as a buffer to acquire a significant sample size of raw data reflecting pre-start violent scenario elements. Following the acquisition task, the operations experienced an increase in the project’s operational timeline utilising data conforming, pre-processing, and blob analysis tools. Although automated software tools were acknowledged, those initiatives applied unwanted blob features as objects of interest during training. As the proposal’s nature involves the criticality of human lives, a manual blob analysis approach avoided object issues, ultimately affecting the overall classification and the proposal’s timeline.

Violent Class Evaluation Reduction: Underestimating the scope of analysis using an 8-class category weighed heavily on the time constraints to formulate a balance between violent/non-violent for 3DCNNsl/YOLOv5m activity recognition during development and the hypothesis phase. Because of the weight of the analysis, a 2-class action reduction comprising similarity proved effective towards evaluations between fencing and stabbing. The 8-category task negatively affected the proposal’s timeline because of the volume of assessments per class.

Field of View: Instances of erratic and sporadic violent actions occurred outside the camera sensor’s field of view, thus obscuring the activity recognition prediction processing.

Human Dependency in Some Cases: Intermittently, the model produced prediction anomalies, thus requiring manual operator assistance to establish the actual nature of the actions given the criticality of life in violent scenarios. The event creates human dependency to mitigate such circumstances and verify the action validation procedures.

7.4 Future Research and Recommendations

The thesis proposes the merging of violent artefact activity recognition with GPU support and the new drone technology to reduce memory and the field of view challenge to en-

courage absolute monitoring as prospects. The idea integrates an onboard camera sensor to generate the video data, performing identical classification tasks as motionless CCTV devices. Drones promote critical data capturing for processing using wireless connectivity, which accesses the device's framework and controls amassing data via a mainframe dispersed geographically, facilitating its activity recognition capacity. Although drone technology has disadvantages, its influence has no impact on the current proposal. Thus, its feasibility is possible via investigations to achieve a formidable strategy. The concept increases the possibility of violent activity recognition artefacts, extends its mobility, and expands the scope of the CCTV sensor device from aerial perspectives. Drone technology significantly reduces human dependency input because violent activity recognition artefact operations using drones expand its classification range. However, maintaining human support to enhance operations and save lives is encouraged. For proper prevention, the model predicts the pre-empted stages of violence and facilitates integration with an alarm system that alerts persons near the altercation, providing crucial time to evade lethal outcomes in preliminary stages.

A continuation of the data acquisition procedures plays an integral role in achieving a substantial volume of videos as a major element for violent activity recognition in its pre-stages. The proposed solution not only addresses the current challenges in security technology but also paves the way for a futuristic approach. The investigations experienced time constraint issues analysing eight classes balanced between violent and non-violent actions. However, future works entail exploring the concept of more actions to increase the complexity, further testing the model's robustness. Regardless of the data issues, the proposed solution facilitates a futuristic approach when combined with innovative facial recognition and robotics that only captures the data necessary to suggest the presence of violence and its perpetrators from a lethal perspective. The idea allows the restructuring of security/government institutions to incorporate a robust solution without intruding into the private lives of its citizens. With the deployment of such technology the scale of

employment vs achieving a secure environment is a factor to consider.

The algorithmic approaches present future interest for external researchers to explore human activity recognition and the flexibility of the proposed technique in several object detection/recognition tasks. For this task, the idea of fusion utilising the decision level facilitates flexibility in hardware at this level. The experimental evidence proved that as a future strategy, employing arithmetic processing utilising GPU may provide substantiating analysis to suggest its effectiveness if applied accurately. The deployment of the proposed fusion can positively impact society's global scourge of crime. Some countries detest the constant CCTV monitoring as a breach of human rights and privacy, which can affect data-generating devices.

Moreover, ethical concerns concerning European general data protection regulations may hinder CCTV surveillance. However, if configured appropriately, the proposed solution's processing through new-age robotics can facilitate the classification task with a reduced impact on human privacy. The current proposed software as a service tool facilitates only violence. As artificial intelligence becomes increasingly powerful, control issues (preventing AI from adverse outcomes) must undergo thoroughly examined before societal deployment.

Violent activity recognition is continually evolving as a direct link created by the advent of new artificial intelligence innovations in computing devices. With the growing rate of computing power, the availability of GPU and TPU devices will positively impact the processing speed of the application, thus increasing its potential exponentially. Amalgamating violent artefact activity recognition with drone technology increases its range towards solving society's security challenges relative to the pre-empting of criminal activities and the loss of lives via violence. Although the model explicitly remedies the complexity of pre-empting violence, because of the 2-state-of-the-art frameworks, the model facilitates any task involving object detection and activity recognition. The proposal identified the issue of the field of view challenges with a demand for higher bands of processing as disadvantages that

affected the effectiveness of the application's output. The proposal acknowledged integrating drone technology with GPU processing that motions high-performance classification efficiency to mitigate the previously mentioned challenges. The strategy involves altering CPU to GPU processing, with drone technology, to eradicate memory and the field of view issues as an advantage compared to the static camera sensor classification. Drone technology encourages classification efficiency in scenarios impeding the inference processing of static camera sensors. In the future, further evolution of applying violent artefact activity recognition to aid security and other institutions in pre-empting violence will become a global innovation. The fusion idea complements the judicial system by providing evidence to display altercations requiring specific details from an in-depth perspective. Because fusion artefact processing incorporates object detection and activity recognition profiling, it can facilitate disaster-prone possibilities on construction sites as a plot to enforce litigation, predict/prevent accidents and reduce incidents leading to death in public (prisons, schools, public bars) and domestic areas (fall detection/ action abnormality in hospitals). The advent of innovative artificial intelligence models and package upgrades pushes the boundaries of pre-empting violence beyond the scope of 15-20 seconds before the act occurs. The challenges identified will be mitigated in the future as technological companies jostle for market dominance, thus causing the cost of processing devices to become affordable. At this stage, drone technology disadvantages and the advent of new technology are unknown. Therefore, further investigations must consider the implications of drone technology and its classification impact using the violent artefact activity recognition application with GPU capabilities.

References

- [1] Bryn King. “Psychological theories of violence”. In: *Journal of human behavior in the social environment* 22.5 (2012), pp. 553–571.
- [2] John Monahan. “The causes of violence”. In: *FBI L. Enforcement Bull.* 63 (1994), p. 11.
- [3] IFSEC Global. *Revealed The UK’s Most-Watched Cities*. 2020. URL: <https://www.ifsecglobal.com/video-surveillance/revealed-the-uks-most-watched-cities/>.
- [4] Robert E. Emery and Laumann-Billings Lisa. “An overview of the nature, causes, and consequences of abusive family relationships: Toward differentiating maltreatment and violence.” In: *American psychologist* 53.2 (1998), p. 121.
- [5] Hall Dan, White Debbie, and Burrows Thomas. *The Sun, CRIME COUNT London killings 2019 – latest news on knife crime, attacks and statistics*. 2019. URL: <https://www.bbc.co.uk/news/uk-49923129>.
- [6] IFSEC Global. *Omdia reports global video surveillance market revenues to reach \$24bn by 2022*. 2021. URL: <https://www.ifsecglobal.com/video-surveillance/omdia-reports-global-video-surveillance-market-revenues-to-reach-24bn-by-2022/>.

- [7] Moore James. *Video Surveillance Report 2021 Integration, the cloud and AI: Expanding horizons for video surveillance technology*. 2013. URL: <https://www.ifsecglobal.com/downloads/the-video-surveillance-report-2021/>.
- [8] Surveillance Camera Commissioner. *Surveillance Camera Commissioner Annual Report January 2020 – March 2021 Presented to Parliament pursuant to Section 35(1)(b) of the Protection of Freedoms Act 2012*. 2022. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1037228/E02682343_SCC_ARA_2020-21_Accessible.pdf.
- [9] James Ritchey. *Number of CCTV cameras in the UK reaches 5.2 million*. 2020. URL: <https://counterterrorbusiness.com/news/19112020/number-cctv-cameras-uk-reaches-52-million>.
- [10] Butcher Ben, Corker Sarah, and Stephenson Wesley. *BBC News The places knife crime is rising fastest*. 2019. URL: <https://www.bbc.co.uk/news/uk-49923129>.
- [11] Stripe Nick. *Homicide in England and Wales: year ending March 2021*. 2022. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/homicideinenglandandwales/yearendingmarch2021>.
- [12] Office for National StatisticsONS. *released 19 October 2023, statistical bulletin, Crime in England and Wales, year ending June 2023*. 2023. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingjune2023>.
- [13] Office for National StatisticsONS. *released 24 July 2024, ONS website, statistical bulletin, Crime in England and Wales: year ending March 2024*. 2024. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingmarch2024>.
- [14] Ministry of Justice. *Knife and offensive weapon sentencing statistics: year ending March 2021*. 2021. URL: <https://www.gov.uk/government/statistics/knife->

- and-offensive-weapon-sentencing-statistics-year-ending-march-2021#sentencing.
- [15] Menendez Elisa and Williams Tom. *Metro, PC Wayne Couzens charged with kidnap and murder of Sarah Everard*. 2019. URL: <https://metro.co.uk/2021/03/12/pc-waynecouzens-charged-with-the-kidnap-and-murder-of-sarah-everard-14227853/>.
 - [16] Kottasová Ivanav. *Murder of young teacher makes women in London worry it could have been them*. 2021. URL: <https://edition.cnn.com/2021/09/25/europe/sabina-nessa-vigil-london-gbr-intl/index.html>.
 - [17] Nadeem Badshah. *Murder investigation after ‘horrific assault’ on woman in east London*. 2022. URL: <https://www.theguardian.com/uk-news/2022/jun/26/investigation-launched-after-horrific-assault-in-east-london>.
 - [18] West-Midlands Police. *Murder investigation launched after stab victim died*. 2022. URL: <https://www.west-midlands.police.uk/news/murder-investigation-launched-after-stab-victim-died>.
 - [19] Davidson Peter and Daily Record. *Tory MP David Amess dead after horrific stabbing during constituency meeting*. 2021. URL: <https://www.dailyrecord.co.uk/news/politics/tory-mp-david-amess-dead-25224685>.
 - [20] Singh Kanishka and Sandle Paul. *Reuters, Six people killed in mass shooting in Plymouth, England*. 2021. URL: <https://www.reuters.com/world/uk/emergency-services-deployed-at-scene-incident-city-plymouth-england-2021-08-12/>.
 - [21] BBC. *Greenford: Manhunt for mobility scooter murder suspect*. 2022. URL: <https://www.bbc.co.uk/news/uk-england-london-62575679>.
 - [22] Metropolitan Police. *Homicide victims in London in 2022*. 2022. URL: <https://www.murdermap.co.uk/statistics/murder-london-2022-latest-total/>.

- [23] James A Nichols, Hsien W Herbert Chan, and Matthew AB Baker. “Machine learning: applications of artificial intelligence to imaging and diagnosis”. In: *Biophysical reviews* 11 (2019), pp. 111–118.
- [24] Iqbal H Sarker. “Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions”. In: *SN computer science* 2.6 (2021), p. 420.
- [25] Hang Yuan et al. “Self-supervised learning for human activity recognition using 700,000 person-days of wearable data”. In: *NPJ digital medicine* 7.1 (2024), p. 91.
- [26] Mike Lakoju et al. “Unsupervised learning for product use activity recognition: An exploratory study of a “chatty device””. In: *Sensors* 21.15 (2021), p. 4991.
- [27] Bahareh Nikpour, Dimitrios Sinodinos, and Narges Armanfard. “Deep reinforcement learning in human activity recognition: A survey”. In: *Authorea Preprints* (2023).
- [28] Isna Alfi Bustoni et al. “Classification methods performance on human activity recognition”. In: *Journal of Physics: Conference Series*. Vol. 1456. 1. IOP Publishing. 2020, p. 012027.
- [29] Henry Friday Nweke et al. “Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges”. In: *Expert Systems with Applications* 105 (2018), pp. 233–261.
- [30] Ravpreet Kaur and Sarbjeet Singh. “A comprehensive review of object detection with deep learning”. In: *Digital Signal Processing* 132 (2023), p. 103812.
- [31] Shan Lu et al. “Blob analysis of the head and hands: A method for deception detection”. In: *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE. 2005, pp. 20c–20c.
- [32] SAS institute Incorporated. *Computer Vision What it is and why it matters*. 2022. URL: https://www.sas.com/en_us/insights/analytics/computer-vision.html.

- [33] Irhum Shafkat. *Intuitively Understanding Convolutions for Deep Learning*. 2018. URL: <https://towardsdatascience.com/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1>.
- [34] Cornelisse Daphne. *An intuitive guide to Convolutional Neural Networks*. 2018. URL: <https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050/>.
- [35] Hand-Break. *HandBrake: The open source video transcoder*. 2022. URL: <https://www.bbc.co.uk/news/uk-49923129>.
- [36] Xuguang Zhang, Honghai Liu, and Xiaoli Li. “Target tracking for mobile robot platforms via object matching and background anti-matching”. In: *Robotics and Autonomous Systems* 58.11 (2010), pp. 1197–1206.
- [37] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. “Example-based object detection in images by components”. In: *IEEE transactions on pattern analysis and machine intelligence* 23.4 (2001), pp. 349–361.
- [38] MathWorks. *Tracking and Motion Estimation*. 2020. URL: <https://uk.mathworks.com/help/vision/tracking-and-%20motion-estimation.html>.
- [39] Kaelon Lloyd et al. “Violent behaviour detection using local trajectory response”. In: *7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016)*. IET. 2016, pp. 1–6.
- [40] Dominik Endres et al. “Hooligan detection: the effects of saliency and expert knowledge”. In: (2011).
- [41] Yan Sun, Jonathon S Hare, and Mark S Nixon. “Detecting acceleration for gait and crime scene analysis”. In: *7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016)*. IET. 2016, pp. 1–6.
- [42] Geoffrey Mison. “Crime in England and Wales”. In: *Sociology* 3.3 (1969), pp. 441–443.
- [43] Sherry Hamby. “On defining violence, and why it matters.” In: (2017).

- [44] Home Office. *Guide on firearms licensing law*. 2013. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/518193/Guidance_on_Firearms_Licensing_Law_April_2016_v20.pdf.
- [45] Paul K Davis et al. *Using behavioral indicators to help detect potential violent acts: A review of the science base*. RAND, 2013.
- [46] Cem Direkoglu, Melike Sah, and Noel E O'Connor. "Abnormal crowd behavior detection using novel optical flow-based features". In: *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE. 2017, pp. 1–6.
- [47] Naeem Ahmad. "Modelling, optimization and design of visual sensor networks for sky surveillance". PhD thesis. Mid Sweden University, 2013.
- [48] Kyung Joo Cheoi. "Temporal saliency-based suspicious behavior pattern detection". In: *Applied Sciences* 10.3 (2020), p. 1020.
- [49] Shuiwang Ji et al. "3D convolutional neural networks for human action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 221–231.
- [50] Xinfeng Zhang et al. "Anomaly detection and localization in crowded scenes by motion-field shape description and similarity-based statistical learning". In: *arXiv preprint arXiv:1805.10620* (2018).
- [51] Ahlam Al-Dhamari, Rubita Sudirman, and Nasrul Humaimi Mahmood. "ABNORMAL BEHAVIOR DETECTION IN AUTOMATED SURVEILLANCE VIDEOS: A REVIEW." In: *Journal of Theoretical & Applied Information Technology* 95.19 (2017).
- [52] Eddins Steve. *Steve on Image Processing and MATLAB*. 2021. URL: <https://blogs.mathworks.com/steve/2014/04/25/color-threshold-app-in-r2014a/>.

- [53] Thanh Binh Nguyen and Sun Tae Chung. “An improved real-time blob detection for visual surveillance”. In: *2009 2nd International Congress on Image and Signal Processing*. IEEE. 2009, pp. 1–5.
- [54] Jie Yang, Jian Cheng, and Hanqing Lu. “Human activity recognition based on the blob features”. In: *2009 IEEE international conference on multimedia and expo*. IEEE. 2009, pp. 358–361.
- [55] K Kraus et al. “Hot-Spot Blob Merging for Real-Time Image Segmentation”. In: *International Journal of Electrical and Computer Engineering* 2.10 (2008), pp. 2167–2172.
- [56] Vladimir L Petrović and Jelena S Popović-Božović. “A method for real-time memory efficient implementation of blob detection in large images”. In: *Serbian Journal of Electrical Engineering* 14.1 (2017), pp. 67–84.
- [57] Per-Erik Forssén and Gösta Granlund. “Robust multi-scale extraction of blob features”. In: *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer. 2003, pp. 11–18.
- [58] Vision Adaptive. *Blob Analysis*. URL: https://docs.adaptive-vision.com/current/studio/machine_vision_guide/BlobAnalysis.html.
- [59] S Padmappriya and K Sumalatha. “Digital image processing real time applications”. In: *International Journal of Engineering Science Invention (IJESI)* 6.3 (2018), pp. 46–51.
- [60] MathWorks. *Computer Vision for Student Competitions: Object Detection using Blob Analysis*. 2020. URL: https://uk.mathworks.com/matlabcentral/fileexchange/53333-computer-vision-for-student-competitions-object-detection-using-blob-analysis?s_tid=srchtitle.
- [61] Pierre Soille. “On morphological operators based on rank filters”. In: *Pattern recognition* 35.2 (2002), pp. 527–535.

- [62] MathWorks. *Morphological Operations*. 2020. URL: <https://uk.mathworks.com/help/images/morphological-filtering.htm>.
- [63] MathWorks. *Vision.BlobAnalysis*. 2020. URL: <https://uk.mathworks.com/help/vision/ref/vision.blobanalysis-%20system-object.html>.
- [64] J Trein, A Th Schwarzbacher, and B Hoppe. “FPGA implementation of a single pass real-time blob analysis using run length encoding”. In: *MPC-Workshop, February*. Citeseer. 2008.
- [65] Ricardo Acevedo-Avila, Miguel Gonzalez-Mendoza, and Andres Garcia-Garcia. “A linked list-based algorithm for blob detection on embedded vision-based sensors”. In: *Sensors* 16.6 (2016), p. 782.
- [66] Stephen Bates, Trevor Hastie, and Robert Tibshirani. “Cross-validation: what does it estimate and how well does it do it?”. In: *Journal of the American Statistical Association* (2023), pp. 1–12.
- [67] Sitefanus Hulu, Poltak Sihombing, et al. “Analysis of performance cross validation method and K-Nearest neighbor in classification data”. In: *International Journal of Research and Review* 7.4 (2020), pp. 69–73.
- [68] Sanjay Yadav and Sanyam Shukla. “Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification”. In: *2016 IEEE 6th International conference on advanced computing (IACC)*. IEEE. 2016, pp. 78–83.
- [69] Indratmo Soekarno, Iwan K Hadihardaja, M Cahyono, et al. “A study of hold-out and k-fold cross validation for accuracy of groundwater modeling in tidal lowland reclamation using extreme learning machine”. In: *2014 2nd International Conference on Technology, Informatics, Management, Engineering & Environment*. IEEE. 2014, pp. 228–233.
- [70] Quang Hung Nguyen et al. “Influence of data splitting on performance of machine learning models in prediction of shear strength of soil”. In: *Mathematical Problems in Engineering* 2021.1 (2021), p. 4832864.

- [71] Teck Fu Thien and Wan Sieng Yeo. “A comparative study between PCR, PLSR, and LW-PLS on the predictive performance at different data splitting ratios”. In: *Chemical Engineering Communications* 209.11 (2022), pp. 1439–1456.
- [72] Li-Jen Weng and Chung-Ping Cheng. “Parallel analysis with unidimensional binary data”. In: *Educational and Psychological Measurement* 65.5 (2005), pp. 697–716.
- [73] V Roshan Joseph and Akhil Vakayil. “SPlit: An optimal method for data splitting”. In: *Technometrics* 64.2 (2022), pp. 166–176.
- [74] Borislava Vrigazova. “The proportion for splitting data into training and test set for the bootstrap in classification problems”. In: *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy* 12.1 (2021), pp. 228–242.
- [75] MathWorks. *What Is Artificial Intelligence (AI)? 3 Things You Need To Know*. 2019. URL: <https://uk.mathworks.com/discovery/artificial-intelligence.html>.
- [76] Welch Stephen. *Popular Python AI libraries*. 2020. URL: <https://www.udacity.com/blog/2020/05/popular-python-ai-libraries.html>.
- [77] Dunn Timothy. *Breakdown of the Convolution 1D and 2D*. 2021. URL: <https://www.coursera.org/lecture/machine-learning-duke/breakdown-of-the-convolution-1d-and-2d-peSoI>.
- [78] Zhong-Qiu Zhao et al. “Object detection with deep learning: A review”. In: *IEEE transactions on neural networks and learning systems* 30.11 (2019), pp. 3212–3232.
- [79] Aston Zhang et al. “Dive into deep learning”. In: *arXiv preprint arXiv:2106.11342* (2021).
- [80] Dingjun Yu et al. “Mixed pooling for convolutional neural networks”. In: *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9*. Springer. 2014, pp. 364–375.
- [81] Aston Zhang et al. *Dive into deep learning*. Cambridge University Press, 2023.

- [82] Fath U Min Ullah et al. “Violence detection using spatiotemporal features with 3D convolutional neural network”. In: *Sensors* 19.11 (2019), p. 2472.
- [83] Thushan Ganegedara. *Intuitive Guide to Convolution Neural Networks*. 2018. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/homicideinenglandandwales/yearendingmarch2021>.
- [84] J. Brownlee. *A Gentle Introduction to Padding and Stride for Convolutional Neural Networks*. 2019. URL: <https://machinelearningmastery.com/padding-and-stride-for-convolutional-neural-networks/>.
- [85] Juan Yopez and Seok-Bum Ko. “Stride 2 1-D, 2-D, and 3-D Winograd for convolutional neural networks”. In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 28.4 (2020), pp. 853–863.
- [86] Nan Yang et al. “Random Padding Data Augmentation”. In: *Australasian Conference on Data Science and Machine Learning*. Springer. 2023, pp. 3–18.
- [87] Chen’ ’Ting-Hao. *What is padding in Convolutional Neural Network?* 2017. URL: <https://medium.com/machine-learning-algorithms/what-is-padding-in-convolutional-neural-network-c120077469cc>.
- [88] Anton Hristov, Maria Nisheva, and Dimo Dimov. “Filters in convolutional neural networks as independent detectors of visual concepts”. In: *Proceedings of the 20th International Conference on Computer Systems and Technologies*. 2019, pp. 110–117.
- [89] Aditya Singh, Alessandro Bay, and Andrea Mirabile. “Assessing the importance of colours for cnns in object recognition”. In: *arXiv preprint arXiv:2012.06917* (2020).
- [90] J. Brownlee. *A Gentle Introduction to the Rectified Linear Unit ReLU*. 2019. URL: <https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/#:~:text=The%20rectified%20linear%20activation%20function,otherwise%2C%20it%20will%20output%20zero>.

- [91] Rachana Patel and Sanskruti Patel. “A comprehensive study of applying convolutional neural network for computer vision”. In: *International Journal of Advanced Science and Technology* 6.6 (2020), pp. 2161–2174.
- [92] Rikiya Yamashita et al. “Convolutional neural networks: an overview and application in radiology”. In: *Insights into imaging* 9 (2018), pp. 611–629.
- [93] Afia Zafar et al. “A comparison of pooling methods for convolutional neural networks”. In: *Applied Sciences* 12.17 (2022), p. 8643.
- [94] Laith Alzubaidi et al. “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions”. In: *Journal of big Data* 8 (2021), pp. 1–74.
- [95] Miriam Seoane Santos et al. “Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]”. In: *IEEE Computational Intelligence Magazine* 13.4 (2018), pp. 59–76.
- [96] Gandhi Rohith. *R-CNN, Fast R-CNN, Faster R-CNN, YOLO Object Detection Algorithms*. 2018. URL: <https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e>.
- [97] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.
- [98] Jasper RR Uijlings et al. “Selective search for object recognition”. In: *International journal of computer vision* 104 (2013), pp. 154–171.
- [99] Kunwar P Singh, Nikita Basant, and Shikha Gupta. “Support vector machines in water quality management”. In: *Analytica chimica acta* 703.2 (2011), pp. 152–162.
- [100] Rasmus Rothe, Matthieu Guillaumin, and Luc Van Gool. “Non-maximum suppression for object detection by passing messages between windows”. In: *Computer Vision—ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Sin-*

- gapore, November 1-5, 2014, *Revised Selected Papers, Part I* 12. Springer. 2015, pp. 290–306.
- [101] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. “Learning non-maximum suppression”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4507–4515.
 - [102] Alexander Neubeck and Luc Van Gool. “Efficient non-maximum suppression”. In: *18th international conference on pattern recognition (ICPR’06)*. Vol. 3. IEEE. 2006, pp. 850–855.
 - [103] Parthasarathy D. *A Brief History of CNNs in Image Segmentation: From R-CNN to Mask R-CNN*. 2017. URL: <https://blog.athelas.com/a-brief-history-of-cnns-in-image-segmentation-from-r-cnn-to-mask-r-cnn-34ea83205de4>.
 - [104] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.
 - [105] Yifan Liu et al. “Research on the use of YOLOv5 object detection algorithm in mask wearing recognition”. In: *World Sci. Res. J* 6.11 (2020), pp. 276–284.
 - [106] Aduen Benjumea et al. “YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles”. In: *arXiv preprint arXiv:2112.11798* (2021).
 - [107] Xingkui Zhu et al. “TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 2778–2788.
 - [108] Glenn Jocher et al. “ultralytics/yolov5: v3. 1-bug fixes and performance improvements”. In: *Zenodo* (2020).
 - [109] Hyun-Ki Jung and Gi-Sang Choi. “Improved yolov5: Efficient object detection using drone images under various conditions”. In: *Applied Sciences* 12.14 (2022), p. 7255.
 - [110] Ismat Saira Gillani et al. “Yolov5, yolo-x, yolo-r, yolov7 performance comparison: A survey”. In: *Artificial Intelligence and Fuzzy Logic System* (2022), pp. 17–28.

- [111] Chien-Yao Wang et al. “CSPNet: A new backbone that can enhance learning capability of CNN”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 390–391.
- [112] Upesh Nepal and Hossein Eslamiat. “Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs”. In: *Sensors* 22.2 (2022), p. 464.
- [113] Ziwei Liu, Ying Shi, and Mingjun Sun. “A pedestrian detection algorithm based on improved YOLOv2”. In: *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*. IEEE. 2018, pp. 488–492.
- [114] Dahang Wan et al. “Yolo-hr: Improved yolov5 for object detection in high-resolution optical remote sensing images”. In: *Remote Sensing* 15.3 (2023), p. 614.
- [115] Marios Antonakakis et al. “Real-Time Object Detection using an Ultra-High-Resolution Camera on Embedded Systems”. In: *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE. 2022, pp. 1–6.
- [116] Glenn Jocher et al. “ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation”. In: *Zenodo* (2022).
- [117] Roberta Vrskova et al. “A new deep-learning method for human activity recognition”. In: *Sensors* 23.5 (2023), p. 2816.
- [118] Liu Peng. “3DCNN-with-keras”. 2013. URL: <https://github.com/liupeng678/3DCNN-with-keras>.
- [119] D Manju, M Seetha, and P Sammulal. “Early action prediction using 3DCNN with LSTM and bidirectional LSTM”. In: *Turkish Journal of Computer and Mathematics Education* 12.6 (2021), pp. 2275–2281.
- [120] Roberta Vrskova et al. “Human activity classification using the 3DCNN architecture”. In: *Applied Sciences* 12.2 (2022), p. 931.

- [121] Ji Li et al. "Efficient violence detection using 3d convolutional neural networks". In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2019, pp. 1–8.
- [122] Salah Mostafa El Abyad, Mona M Soliman, and Khaled Mostafa El Sayed. "Deep Video Hashing Using 3DCNN with BERT." In: *International Journal of Intelligent Engineering & Systems* 15.5 (2022).
- [123] Md Atiqur Rahman Ahad et al. "Action recognition using kinematics posture feature on 3D skeleton joint locations". In: *Pattern Recognition Letters* 145 (2021), pp. 216–224.
- [124] Jonathan Huang et al. "Speed/accuracy trade-offs for modern convolutional object detectors". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7310–7311.
- [125] Ji Li et al. "Efficient violence detection using 3d convolutional neural networks". In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2019, pp. 1–8.
- [126] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [127] MN Favorskaya and VV Andreev. "The study of activation functions in deep learning for pedestrian detection and tracking". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42 (2019), pp. 53–59.
- [128] Jawad Nagi et al. "Max-pooling convolutional neural networks for vision-based hand gesture recognition". In: *2011 IEEE international conference on signal and image processing applications (ICSIPA)*. IEEE. 2011, pp. 342–347.
- [129] Paul Chris and Godambe Mandar. *Image Downsampling Upsampling*. 2000. URL: https://www.researchgate.net/publication/359772964_Image_Downsampling_Upsampling#fullTextFileContent.

- [130] Monalika Padma Reddy et al. “Human Activity Recognition using 3D CNN.” In: *Turkish Online Journal of Qualitative Inquiry* 12.7 (2021).
- [131] Tapani Raiko, Harri Valpola, and Yann LeCun. “Deep learning made easier by linear transformations in perceptrons”. In: *Artificial intelligence and statistics*. PMLR. 2012, pp. 924–932.
- [132] Ali Seydi Keceli and Aydin Kaya. “Violent activity classification with transferred deep features and 3d-Cnn”. In: *Signal, Image and Video Processing* 17.1 (2023), pp. 139–146.
- [133] Simone Accattoli et al. “Violence detection in videos by combining 3D convolutional neural networks and support vector machines”. In: *Applied Artificial Intelligence* 34.4 (2020), pp. 329–344.
- [134] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi. “Multi-modal convolutional neural networks for activity recognition”. In: *2015 IEEE International conference on systems, man, and cybernetics*. IEEE. 2015, pp. 3017–3022.
- [135] Wenchao Xu et al. “Human activity recognition based on convolutional neural network”. In: *2018 24th international conference on pattern recognition (ICPR)*. IEEE. 2018, pp. 165–170.
- [136] Parisa Fard Moshiri et al. “Using GAN to enhance the accuracy of indoor human activity recognition”. In: *arXiv preprint arXiv:2004.11228* (2020).
- [137] Usman Azmat and Ahmad Jalal. “Smartphone inertial sensors for human locomotion activity recognition based on template matching and codebook generation”. In: *2021 International Conference on Communication Technologies (ComTech)*. IEEE. 2021, pp. 109–114.
- [138] Massinissa Hamidi and Aomar Osmani. “Human activity recognition: A dynamic inductive bias selection perspective”. In: *Sensors* 21.21 (2021), p. 7278.
- [139] Jingsai Liang. “Confusion matrix: Machine learning”. In: *POGIL Activity Clearing-house* 3.4 (2022).

- [140] Xuan Liu, Xiaoguang Wang, and Stan Matwin. “Interpretable deep convolutional neural networks via meta-learning”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–9.
- [141] Serafeim Loukas. *Multi-class Classification: Extracting Performance Metrics From The Confusion Matrix*. 2020. URL: <https://towardsdatascience.com/multi-class-classification-extracting-performance-metrics-from-the-confusion-matrix-b379b427a872>.
- [142] Feras A Batarseh and Ruixin Yang. “Data democracy: at the nexus of artificial intelligence, software development, and knowledge engineering”. In: (2020).
- [143] Fatih Demir. “Deep autoencoder-based automated brain tumor detection from MRI data”. In: *Artificial Intelligence-Based Brain-Computer Interface*. Elsevier, 2022, pp. 317–351.
- [144] Ioannis Markoulidakis et al. “Multi-class confusion matrix reduction method and its application on net promoter score classification problem”. In: *Proceedings of the 14th Pervasive Technologies Related to Assistive Environments Conference*. 2021, pp. 412–419.
- [145] Bomi Jeong et al. “Comparison between statistical models and machine learning methods on classification for highly imbalanced multiclass kidney data”. In: *Diagnostics* 10.6 (2020), p. 415.
- [146] David Colquhoun. “An investigation of the false discovery rate and the misinterpretation of p-values”. In: *Royal Society open science* 1.3 (2014), p. 140216.
- [147] Karimi Zohreh. *Confusion Matrix*. 2021. URL: https://www.researchgate.net/publication/355096788_Confusion_Matrix,%20Date%20Accessed.
- [148] Megan Hollister Murray and Jeffrey D Blume. “False discovery rate computation: Illustrations and modifications”. In: *arXiv preprint arXiv:2010.04680* (2020).

- [149] Umair Ahmad et al. “Analysis of classification techniques for intrusion detection”. In: *2019 International conference on innovative computing (ICIC)*. IEEE. 2019, pp. 1–6.
- [150] Muhammad Hasnain et al. “Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking”. In: *IEEE Access* 8 (2020), pp. 90847–90861. DOI: 10.1109/ACCESS.2020.2994222.
- [151] Rafael Padilla et al. “A comparative analysis of object detection metrics with a companion open-source toolkit”. In: *Electronics* 10.3 (2021), p. 279.
- [152] Aqeel Anwar. “What is Average Precision in Object Detection & Localization Algorithms and how to calculate it”. In: *Towards Data Science* (2022).
- [153] Ambik Mitra et al. “Deep learning approach for object features detection”. In: *Advances in Communication, Devices and Networking: Proceedings of ICCDN 2020*. Springer. 2022, pp. 251–259.
- [154] Himanshu Singh. *Practical machine learning and image processing: for facial recognition, object detection, and pattern recognition using Python*. Springer, 2019.
- [155] Kashish Naqvi et al. “Employing real-time object detection for visually impaired people”. In: *Data Analytics and Management: Proceedings of ICDAM*. Springer. 2021, pp. 285–299.
- [156] Jun Liu et al. “Spatio-temporal lstm with trust gates for 3d human action recognition”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 816–833.
- [157] Sainbayar Sukhbaatar et al. “Training convolutional networks with noisy labels”. In: *arXiv preprint arXiv:1406.2080* (2014).
- [158] Swathi Pothuganti. “Review on over-fitting and under-fitting problems in Machine Learning and solutions”. In: *Int. J. Adv. Res. Electr. Electron. Instrum. Eng* 7 (2018), pp. 3692–3695.

- [159] Paul Choen and Jensen Donald David. *Overfitting Explained*. 2000. URL: https://www.researchgate.net/publication/2475394_Overfitting_Explained.
- [160] Arthur EW Venter, Marthinus W Theunissen, and Marelle H Davel. “Pre-interpolation loss behavior in neural networks”. In: *Southern African Conference for Artificial Intelligence Research*. Springer. 2020, pp. 296–309.
- [161] Yingxin Gu et al. “An optimal sample data usage strategy to minimize overfitting and underfitting effects in regression tree models based on remotely-sensed data”. In: *Remote sensing* 8.11 (2016), p. 943.
- [162] J. Brownlee. *How to use Learning Curves to Diagnose Machine Learning Model Performance*. 2019. URL: <https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/>.
- [163] Anindita Saha et al. “A survey of machine learning and meta-heuristics approaches for sensor-based human activity recognition systems”. In: *Journal of Ambient Intelligence and Humanized Computing* 15.1 (2024), pp. 29–56.
- [164] Kyle Dillon Feuz and Diane J. Cook. “Heterogeneous transfer learning for activity recognition using heuristic search techniques”. In: *International journal of pervasive computing and communications* 10.4 (2014), pp. 393–418.
- [165] Tayyip Ozcan and Alper Basturk. “Human action recognition with deep learning and structural optimization using a hybrid heuristic algorithm”. In: *Cluster Computing* 23.4 (2020), pp. 2847–2860.
- [166] Abdulaziz Alarifi and Ayed Alwadain. “Killer heuristic optimized convolution neural network-based fall detection with wearable IoT sensor devices”. In: *Measurement* 167 (2021), p. 108258.
- [167] Lin Fan, Zhongmin Wang, and Hai Wang. “Human activity recognition model based on decision tree”. In: *2013 International Conference on Advanced Cloud and Big Data*. IEEE. 2013, pp. 64–68.

- [168] Haiyong Zhao and Zhijing Liu. “Human action recognition based on non-linear SVM decision tree”. In: *Journal of Computational Information Systems* 7.7 (2011), pp. 2461–2468.
- [169] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. “A review of human activity recognition methods”. In: *Frontiers in Robotics and AI* 2 (2015), p. 28.
- [170] KG Manosha Chathuramali and Ranga Rodrigo. “Faster human activity recognition with SVM”. In: *International conference on advances in ICT for emerging regions (ICTer2012)*. IEEE. 2012, pp. 197–203.
- [171] Davide Anguita et al. “Human activity recognition on smartphones using a multi-class hardware-friendly support vector machine”. In: *Ambient Assisted Living and Home Care: 4th International Workshop, IWAAL 2012, Vitoria-Gasteiz, Spain, December 3-5, 2012. Proceedings 4*. Springer. 2012, pp. 216–223.
- [172] Sahak Kaghyan and Hakob Sarukhanyan. “Activity recognition using k-nearest neighbor algorithm on smartphone with tri-axial accelerometer”. In: *International Journal of Informatics Models and Analysis (IJIMA), ITHEA International Scientific Society, Bulgaria* 1 (2012), pp. 146–156.
- [173] Ferhat Attal et al. “Physical human activity recognition using wearable sensors”. In: *Sensors* 15.12 (2015), pp. 31314–31338.
- [174] Labiba Gillani Fahad, Syed Fahad Tahir, and Muttukrishnan Rajarajan. “Activity recognition in smart homes using clustering based classification”. In: *2014 22nd International conference on pattern recognition*. IEEE. 2014, pp. 1348–1353.
- [175] Xiaodong Yang and Ying Li Tian. “Eigenjoints-based action recognition using naive-bayes-nearest-neighbor”. In: *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE. 2012, pp. 14–19.

- [176] Mustafa Kose, Ozlem Durmaz Incel, and Cem Ersoy. “Online human activity recognition on smart phones”. In: *Workshop on mobile sensing: from smartphones and wearables to big data*. Vol. 16. 2012. 2012, pp. 11–15.
- [177] Antonio R Jiménez and Fernando Seco. “Multi-event Naive Bayes classifier for activity recognition in the UCAmI Cup”. In: *Proceedings*. Vol. 60. 1. MDPI. 2018.
- [178] Thi V Duong et al. “Activity recognition and abnormality detection with the switching hidden semi-markov model”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. IEEE. 2005, pp. 838–845.
- [179] M Humayun Kabir et al. “Two-layer hidden Markov model for human activity recognition in home environments”. In: *International Journal of Distributed Sensor Networks* 12.1 (2016), p. 4560365.
- [180] Dorra Trabelsi et al. “An unsupervised approach for automatic activity recognition based on hidden Markov model regression”. In: *IEEE Transactions on automation science and engineering* 10.3 (2013), pp. 829–835.
- [181] Felicity R Allen et al. “Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models”. In: *Physiological measurement* 27.10 (2006), p. 935.
- [182] Fatma Najar et al. “Unsupervised learning of finite full covariance multivariate generalized Gaussian mixture models for human activity recognition”. In: *Multimedia Tools and Applications* 78 (2019), pp. 18669–18691.
- [183] Jenny Margarito et al. “User-independent recognition of sports activities from a single wrist-worn accelerometer: A template-matching-based approach”. In: *IEEE Transactions on Biomedical Engineering* 63.4 (2015), pp. 788–796.
- [184] Long-Van Nguyen-Dinh et al. “Improving online gesture recognition with template matching methods in accelerometer data”. In: *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*. IEEE. 2012, pp. 831–836.

- [185] Abdulhamit Subasi et al. “Smartphone-based human activity recognition using bagging and boosting”. In: *Procedia Computer Science* 163 (2019), pp. 54–61.
- [186] Yuchuan Wu et al. “Recognizing activities of the elderly using wearable sensors: A comparison of ensemble algorithms based on boosting”. In: *Sensor Review* 39.6 (2019), pp. 743–751.
- [187] Chunyu Hu et al. “A novel random forests based class incremental learning method for activity recognition”. In: *Pattern Recognition* 78 (2018), pp. 277–290.
- [188] Niall Twomey et al. “A comprehensive study of activity recognition using accelerometers”. In: *Informatics*. Vol. 5. 2. MDPI. 2018, p. 27.
- [189] Wesllen Sousa et al. “A comparative analysis of the impact of features on human activity recognition with smartphone sensors”. In: *Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web*. 2017, pp. 397–404.
- [190] Stephen J Preece et al. “A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data”. In: *IEEE Transactions on Biomedical Engineering* 56.3 (2008), pp. 871–879.
- [191] Rui Yan et al. “HiGCIN: Hierarchical graph-based cross inference network for group activity recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 45.6 (2020), pp. 6955–6968.
- [192] Maja Stikic, Diane Larlus, and Bernt Schiele. “Multi-graph based semi-supervised learning for activity recognition”. In: *2009 international symposium on wearable computers*. IEEE. 2009, pp. 85–92.
- [193] Pei Wang et al. “Graph based skeleton motion representation and similarity measurement for action recognition”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer. 2016, pp. 370–385.
- [194] Yan Yan et al. “Egocentric daily activity recognition via multitask clustering”. In: *IEEE Transactions on Image Processing* 24.10 (2015), pp. 2984–2995.

- [195] Zhelong Wang et al. “An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors”. In: *IEEE Transactions on Information Technology in Biomedicine* 16.4 (2012), pp. 691–699.
- [196] Earnest Paul Ijjina and Krishna Mohan Chalavadi. “Human action recognition using genetic algorithms and convolutional neural networks”. In: *Pattern recognition* 59 (2016), pp. 199–212.
- [197] Saisakul Chernbumroong, Shuang Cang, and Hongnian Yu. “Genetic algorithm-based classifiers fusion for multisensor activity recognition of elderly people”. In: *IEEE journal of biomedical and health informatics* 19.1 (2014), pp. 282–289.
- [198] Carlos E Galván-Tejada et al. “An analysis of audio features to develop a human activity recognition model using genetic algorithms, random forests, and neural networks”. In: *Mobile Information Systems* 2016.1 (2016), p. 1784101.
- [199] Jindong Wang et al. “Deep learning for sensor-based activity recognition: A survey”. In: *Pattern recognition letters* 119 (2019), pp. 3–11.
- [200] Saedeh Abbaspour et al. “A comparative analysis of hybrid deep learning models for human activity recognition”. In: *Sensors* 20.19 (2020), p. 5707.
- [201] Shibo Zhang et al. “Deep learning in human activity recognition with wearable sensors: A review on advances”. In: *Sensors* 22.4 (2022), p. 1476.
- [202] Kaixuan Chen et al. “Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities”. In: *ACM Computing Surveys (CSUR)* 54.4 (2021), pp. 1–40.
- [203] Paul Viola and Michael Jones. “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Ieee. 2001, pp. I–I.
- [204] Michał Grega et al. “Automated detection of firearms and knives in a CCTV image”. In: *Sensors* 16.1 (2016), p. 47.

- [205] Ondrej Miksik and Krystian Mikolajczyk. “Evaluation of local detectors and descriptors for fast feature matching”. In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE. 2012, pp. 2681–2684.
- [206] Loris Nanni, Alessandra Lumini, and Sheryl Brahnam. “Survey on LBP based texture descriptors for image classification”. In: *Expert Systems with Applications* 39.3 (2012), pp. 3634–3641.
- [207] Rui Hu and John Collomosse. “A performance evaluation of gradient field hog descriptor for sketch based image retrieval”. In: *Computer Vision and Image Understanding* 117.7 (2013), pp. 790–806.
- [208] Alaa E Abdel-Hakim and Aly A Farag. “CSIFT: A SIFT descriptor with color invariant characteristics”. In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*. Vol. 2. Ieee. 2006, pp. 1978–1983.
- [209] Herbert Bay et al. “Speeded-up robust features (SURF)”. In: *Computer vision and image understanding* 110.3 (2008), pp. 346–359.
- [210] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. “BRISK: Binary robust invariant scalable keypoints”. In: *2011 International conference on computer vision*. Ieee. 2011, pp. 2548–2555.
- [211] Michael Calonder et al. “Brief: Binary robust independent elementary features”. In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV* 11. Springer. 2010, pp. 778–792.
- [212] Krystian Mikolajczyk and Cordelia Schmid. “A performance evaluation of local descriptors”. In: *IEEE transactions on pattern analysis and machine intelligence* 27.10 (2005), pp. 1615–1630.
- [213] Konstantinos G Derpanis. “The harris corner detector”. In: *York University* 2.1 (2004), p. 2.

- [214] Ethan Rublee et al. “ORB: An efficient alternative to SIFT or SURF”. In: *2011 International conference on computer vision*. Ieee. 2011, pp. 2564–2571.
- [215] Morteza Ghahremani, Yonghuai Liu, and Bernard Tiddeman. “Ffd: Fast feature detector”. In: *IEEE Transactions on Image Processing* 30 (2020), pp. 1153–1168.
- [216] Shaoqing Ren et al. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1137–1149.
- [217] Peng Tang et al. “Weakly supervised region proposal network and object detection”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 352–368.
- [218] Julian Müller, Andreas Fregin, and Klaus Dietmayer. “Disparity sliding window: object proposals from disparity images”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 5777–5784.
- [219] Ilker Buzcu and A Aydın Alatan. “Fisher-selective search for object detection”. In: *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2016, pp. 3633–3637.
- [220] Koen EA Van de Sande et al. “Segmentation as selective search for object recognition”. In: *2011 international conference on computer vision*. IEEE. 2011, pp. 1879–1886.
- [221] Marko Heikkilä, Matti Pietikäinen, and Janne Heikkilä. “A texture-based method for detecting moving objects.” In: *Bmvc*. Vol. 401. Citeseer. 2004, pp. 1–10.
- [222] Priyanto Hidayatullah and Miftahuddin Zuhdi. “Color-texture based object tracking using HSV color space and local binary pattern”. In: *International Journal on Electrical Engineering and Informatics* 7.2 (2015), p. 161.
- [223] Saka Kezia, I Santi Prabha, and Vijaya Kumar Vakulabharanam. “A color-texture based segmentation method to extract object from background”. In: *International Journal Of Image, Graphics And Signal Processing* 5.3 (2013), p. 19.

- [224] Ajmal S Mian, Mohammed Bennamoun, and Robyn Owens. “Three-dimensional model-based object recognition and segmentation in cluttered scenes”. In: *IEEE transactions on pattern analysis and machine intelligence* 28.10 (2006), pp. 1584–1601.
- [225] Arthur R Pope. “Model-based object recognition”. In: *A Survey of Recent Techniques, Technical Report* (1994).
- [226] Georgia Gkioxari et al. “R-cnns for pose estimation and action detection”. In: *arXiv preprint arXiv:1406.5212* (2014).
- [227] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. “Contextual action recognition with r* cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1080–1088.
- [228] Anuja Jana Naik and MT Gopalakrishna. “Deep-violence: individual person violent activity detection in video”. In: *Multimedia Tools and Applications* 80.12 (2021), pp. 18365–18380.
- [229] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.
- [230] Joseph Redmon et al. “You only look once: Unified, realtime object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.
- [231] Shubham Shinde, Ashwin Kothari, and Vikram Gupta. “YOLO based human action recognition and localization”. In: *Procedia computer science* 133 (2018), pp. 831–838.
- [232] Christian Wolf et al. “Evaluation of video activity localizations integrating quality and quantity measurements”. In: *Computer Vision and Image Understanding* 127 (2014), pp. 14–30.

- [233] Wan Emilyya Izzety Binti Wan Noor, Naimah Mat Isa, et al. “Object Detection: Harmful Weapons Detection using YOLOv4”. In: *2021 IEEE Symposium on Wireless Technology & Applications (ISWTA)*. IEEE. 2021, pp. 63–70.
- [234] Ujwalla Gawande Ujwalla Gawande, Kamal Hajari Kamal Hajari, and Yogesh Golhar. “Novel Pedestrian Detection and Suspicious Activity Recognition Using Enhanced YOLOv5 and Motion Feature Map”. In: (2023).
- [235] Suryanti Awang, Mohd Qhairel Rafiqi Rokei, and Junaida Sulaiman. “Suspicious Activity Trigger System using YOLOv6 Convolutional Neural Network”. In: *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*. IEEE. 2023, pp. 527–532.
- [236] G Kranthi Kumar et al. “Detection of Real Time Anomalous Activities using YOLOv7”. In: *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*. IEEE. 2023, pp. 974–979.
- [237] Roboflow Blog. *What is yolov8? the ultimate guide*. 2023.
- [238] K Ganagavalli and V Santhi. “YOLO-based anomaly activity detection system for human behavior analysis and crime mitigation”. In: *Signal, Image and Video Processing* (2024), pp. 1–11.
- [239] Gallagher James and Skalski Piotr. *How to Train YOLOv10 Model on a Custom Dataset*. 2024. URL: <https://blog.roboflow.com/yolov10-how-to-train/>.
- [240] Zhen Li et al. “An adaptive hidden Markov model for activity recognition based on a wearable multi-sensor device”. In: *Journal of medical systems* 39 (2015), pp. 1–10.
- [241] Mingjie Li et al. “An Action Recognition network for specific target based on rMC and RPN”. In: *Journal of Physics: Conference Series*. Vol. 1325. 1. IOP Publishing. 2019, p. 012073.
- [242] Sameh Neili Boualia and Najoua Essoukri Ben Amara. “3D CNN for human action recognition”. In: *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE. 2021, pp. 276–282.

- [243] Ming Cheng, Kunjing Cai, and Ming Li. “RWF-2000: an open large scale video database for violence detection”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 4183–4190.
- [244] Grand Canyon University. *Quantitative Research Design Methods for Writing Dissertations*. 2021. URL: <https://www.gcu.edu/blog/doctoral-journey/quantitative-research-design-methods-writing-dissertations>.
- [245] University of Central Florida. *Data Sets*. 2011. URL: <https://www.crcv.ucf.edu/data/UCF101.php>.
- [246] RoboFlow. *Everything you need to build and deploy computer vision models*. 2023. URL: <https://roboflow.com>.
- [247] PyTorch. *Introducing Accelerated PyTorch Training on Mac*. 2022. URL: <https://pytorch.org/blog/introducing-accelerated-pytorch-training-on-mac/>.
- [248] C Lawrence Zitnick and Piotr Dollár. “Edge boxes: Locating object proposals from edges”. In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 391–405.
- [249] Daniel Roggen, Gerhard Tröster, and Andreas Bulling. “Signal processing technologies for activity-aware smart textiles”. In: *Multidisciplinary know-how for smart-textiles developers*. Elsevier, 2013, pp. 329–365.
- [250] P Kuppusamy and VC Bharathi. “Human abnormal behavior detection using CNNs in crowded and uncrowded surveillance—A survey”. In: *Measurement: Sensors* 24 (2022), p. 100510.
- [251] Guoliang Zhang et al. “Weighted score-level feature fusion based on Dempster–Shafer evidence theory for action recognition”. In: *Journal of Electronic Imaging* 27.1 (2018), pp. 013021–013021.
- [252] Arun A Ross, Karthik Nandakumar, and Anil K Jain. *Handbook of multibiometrics*. Vol. 6. Springer Science & Business Media, 2006.

- [253] Brad Ulery et al. “Studies of biometric fusion”. In: *NIST Interagency Report 7346* (2006).
- [254] Ludmila I Kuncheva. “A theoretical study on six classifier fusion strategies”. In: *IEEE Transactions on pattern analysis and machine intelligence* 24.2 (2002), pp. 281–286.
- [255] Yufeng Zheng and Erik Blasch. “Score fusion and decision fusion for the performance improvement of face recognition”. In: *16th International Conference on Information Fusion*. 2013.
- [256] Qian Tao. “Face verification for mobile personal devices”. In: *University of Twente: Enschede, The Netherlands* (2009).
- [257] Sumegh Tharewal et al. “Score-level fusion of 3D face and 3D ear for multimodal biometric human recognition”. In: *Computational Intelligence and Neuroscience* 2022 (2022).
- [258] Kumari Priyanka Sinha and Prabhat Kumar. “Hybrid Classification with Score Level Fusion for Human Activity Recognition”. In: *Available at SSRN 4111254* ().
- [259] Firas S Assaad and Gursel Serpen. “Transformation based score fusion algorithm for multi-modal biometric user authentication through ensemble classification”. In: *Procedia Computer Science* 61 (2015), pp. 410–415.
- [260] Meziane Iftenea, Qingjie Liub, and Yunhong Wangc. “Very high resolution images classification by fusing deep convolutional neural networks”. In: *The 5th International Conference on Advanced Computer Science Applications and Technologies (ACSAT 2017)*. 2017, pp. 172–176.
- [261] Amjad Rehman et al. “Internet-of-things-based suspicious activity recognition using multimodalities of computer vision for smart city security”. In: *Security and communication Networks* 2022 (2022).

- [262] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. “Improving human action recognition using fusion of depth camera and inertial sensors”. In: *IEEE Transactions on Human-Machine Systems* 45.1 (2014), pp. 51–61.
- [263] Alina Roitberg et al. “A comparative analysis of decision-level fusion for multimodal driver behaviour understanding”. In: *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2022, pp. 1438–1444.
- [264] Wei Sun and Jun Yan. “A CNN based localization and activity recognition algorithm using multi-receiver CSI measurements and decision fusion”. In: *2022 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE. 2022, pp. 1–7.
- [265] Bihan Jiang et al. “Decision level fusion of domain specific regions for facial action recognition”. In: *2014 22nd international conference on pattern recognition*. IEEE. 2014, pp. 1776–1781.
- [266] Chia Chin Lip and Dzati Athiar Ramli. “Comparative study on feature, score and decision level fusion schemes for robust multibiometric systems”. In: *Frontiers in computer education* (2012), pp. 941–948.
- [267] Arijit Nandi et al. “Real-time multimodal emotion classification system in e-learning context”. In: *International Conference on Engineering Applications of Neural Networks*. Springer. 2021, pp. 423–435.
- [268] Rabab A Rasool. “Feature-level vs. score-level fusion in the human identification system”. In: *Applied Computational Intelligence and Soft Computing 2021* (2021), pp. 1–10.
- [269] Md Nazmuzzaman Khan and Sohail Anwar. “Paradox elimination in Dempster–Shafer combination rule with novel entropy function: Application in decision-level multi-sensor fusion”. In: *Sensors* 19.21 (2019), p. 4810.

- [270] Wen Jiang et al. “An improved method to rank generalized fuzzy numbers with different left heights and right heights”. In: *Journal of Intelligent & Fuzzy Systems* 28.5 (2015), pp. 2343–2355.
- [271] Lisa Ann Osadciw and Kalyan Veeramachaneni. *Fusion, Decision-Level*. 2009.
- [272] Debaditya Roy, Sarunas Girdzijauskas, and Serghei Socolovschi. “Confidence-calibrated human activity recognition”. In: *Sensors* 21.19 (2021), p. 6566.
- [273] Md Milon Islam et al. “Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects”. In: *Computers in Biology and Medicine* 149 (2022), p. 106060.
- [274] Md Milon Islam et al. “Human Activity Recognition Using Tools of Convolutional Neural Networks: A State of the Art Review, Data Sets, Challenges and Future Prospects”. In: *arXiv e-prints* (2022), arXiv–2202.
- [275] Flávia Alves et al. “Sensor Data for Human Activity Recognition: Feature Representation and Benchmarking”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–8.
- [276] Anthony J Alberg et al. “REVIEW The Use of “Overall Accuracy” to Evaluate the Validity of Screening or Diagnostic Tests.” In: *JGIM: Journal of General Internal Medicine* 19.5 (2004).
- [277] Suneth Ranasinghe, Fadi Al Machot, and Heinrich C Mayr. “A review on applications of activity recognition systems with regard to performance and evaluation”. In: *International Journal of Distributed Sensor Networks* 12.8 (2016), p. 1550147716665520.
- [278] Ebrahim Mortaz. “Imbalance accuracy metric for model selection in multi-class imbalance classification problems”. In: *Knowledge-Based Systems* 210 (2020), p. 106490.
- [279] Sushovan Chanda et al. “A deep audiovisual approach for human confidence classification”. In: *Frontiers in Computer Science* 3 (2021), p. 674533.

- [280] Yagna Gudipalli et al. “Deep Modelling Strategies for Human Confidence Classification using Audio-visual Data”. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2023, pp. 1–4.
- [281] WANG Changhai, ZHANG Jianzhong, XU Jingdong, et al. “Identifying the confidence level of activity recognition via HMM”. In: *Journal on Communication* 37.5 (2016), pp. 143–151.
- [282] Zahraa S Abdallah et al. “Activity recognition with evolving data streams: A review”. In: *ACM Computing Surveys (CSUR)* 51.4 (2018), pp. 1–36.
- [283] Elizabeth F Chua, Daniel L Schacter, and Reisa A Sperling. “Neural basis for recognition confidence in younger and older adults.” In: *Psychology and Aging* 24.1 (2009), p. 139.
- [284] Nathan Weber and Neil Brewer. “Confidence-accuracy calibration in absolute and relative face recognition judgments.” In: *Journal of Experimental Psychology: Applied* 10.3 (2004), p. 156.
- [285] Weiyao Lin et al. “Activity recognition using a combination of category components and local models for video surveillance”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 18.8 (2008), pp. 1128–1139.
- [286] Haibo He and Edwardo A Garcia. “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.
- [287] Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* ” O’Reilly Media, Inc.”, 2013.
- [288] Tanvi S Motwani and Raymond J Mooney. “Improving video activity recognition using object recognition and text mining”. In: *ECAI 2012*. IOS Press, 2012, pp. 600–605.
- [289] Alexander Diete and Heiner Stuckenschmidt. “Fusing object information and inertial data for activity recognition”. In: *Sensors* 19.19 (2019), p. 4119.

- [290] Bingbing Ni et al. “Multilevel depth and image fusion for human activity detection”. In: *IEEE transactions on cybernetics* 43.5 (2013), pp. 1383–1394.
- [291] Vishva Payghode et al. “Object detection and activity recognition in video surveillance using neural networks”. In: *International Journal of Web Information Systems* ahead-of-print (2023).
- [292] Anagha Deshpande and Krishna Warhade. “SADY: Student Activity Detection Using YOLO-based Deep Learning Approach.” In: *International Journal on Advanced Science, Engineering & Information Technology* 13.4 (2023).
- [293] M Sunil Suthar and PS Nithya Darisini. “Comparative analysis of Human Activity Recognition and object detection”. In: *Journal of Physics: Conference Series*. Vol. 1716. 1. IOP Publishing. 2020, p. 012054.
- [294] Gilder Lucy and Clarke Jennifer. *How many violent attacks and sexual assaults on women are there?* 2022. URL: <https://www.bbc.co.uk/news/explainers-56365412>.
- [295] Grahame Allen and Harding Megan. *Research Briefing, Knife crime statistics*. 2021. URL: <https://commonslibrary.parliament.uk/research-briefings/sn04304/>.
- [296] BBC. *Plymouth shooting: Jake Davison was licensed gun holder*. 2021. URL: <https://www.bbc.co.uk/news/uk-england-devon-58197414.amp>.
- [297] Shaw Danny. *Ten charts on the rise of knife crime in England and Wales*. 2019. URL: <https://www.bbc.co.uk/news/uk-42749089>.
- [298] Scarff Brent. *Understanding Backpropagation A visual derivation of the equations that allow neural networks to learn*. 2018. URL: <https://towardsdatascience.com/understanding-backpropagation-abcc509ca9d0>.
- [299] Hongqing Fang et al. “Human activity recognition based on feature selection in smart home using back-propagation algorithm”. In: *ISA transactions* 53.5 (2014), pp. 1629–1638.

- [300] Rajarshi Guha, David T Stanton, and Peter C Jurs. “Interpreting computational neural network quantitative structure- activity relationship models: A detailed interpretation of the weights and biases”. In: *Journal of chemical information and modeling* 45.4 (2005), pp. 1109–1121.
- [301] Le Dung and Makoto Mizukawa. “A pattern recognition neural network using many sets of weights and biases”. In: *2007 International Symposium on Computational Intelligence in Robotics and Automation*. IEEE. 2007, pp. 285–290.
- [302] Grzegorz Dudek. “Generating random weights and biases in feedforward neural networks with random hidden nodes”. In: *Information sciences* 481 (2019), pp. 33–56.
- [303] Nadia Oukrich et al. “Activity recognition using back-propagation algorithm and minimum redundancy feature selection method”. In: *2016 4th IEEE international colloquium on information science and technology (CiSt)*. IEEE. 2016, pp. 818–823.
- [304] Adna Sengto and Thurdsak Leauhatong. “Human falling detection algorithm using back propagation neural network”. In: *The 5th 2012 Biomedical Engineering International Conference*. IEEE. 2012, pp. 1–5.
- [305] Girma A. *Part-1: Convolutional Neural Network in a Nutshell*. 2019. URL: <https://abenezer-g.medium.com/part-1-convolutional-neural-network-in-a-nutshell-89f81a329ec3>.
- [306] Muhamad Yani, Irawan Budhi, and Setiningsih Casi. “Application of transfer learning using convolutional neural network method for early detection of terry’s nail”. In: *Journal of Physics: Conference Series*. Vol. 1201. 1. IOP Publishing. 2019, p. 012052.
- [307] Ahsen Tahir et al. “Hrnn4f: Hybrid deep random neural network for multi-channel fall activity detection”. In: *Probability in the Engineering and Informational Sciences* 35.1 (2021), pp. 37–50.
- [308] Ernest Jeczmioneek and Piotr A Kowalski. “Flattening layer pruning in convolutional neural networks”. In: *Symmetry* 13.7 (2021), p. 1147.

- [309] Letizia Gionfrida et al. “A 3dcnn-lstm multi-class temporal segmentation for hand gesture recognition”. In: *Electronics* 11.15 (2022), p. 2427.
- [310] VL Helen Josephine, AP Nirmala, and Vijaya Lakshmi Alluri. “Impact of hidden dense layers in convolutional neural network to enhance performance of classification model”. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 1131. 1. IOP Publishing. 2021, p. 012007.
- [311] Francisco M Castro et al. “Multimodal feature fusion for CNN-based gait recognition: an empirical comparison”. In: *Neural Computing and Applications* 32 (2020), pp. 14173–14193.
- [312] Qiuyu Zhu et al. “Improving classification performance of softmax loss function based on scalable batch-normalization”. In: *Applied Sciences* 10.8 (2020), p. 2950.
- [313] Shmueli Boaz. *Multi-Class Metrics Made Simple, Part I: Precision and Recall*. 2019. URL:] :%20https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2.
- [314] Margherita Grandini, Enrico Bagli, and Giorgio Visani. “Metrics for multi-class classification: an overview”. In: *arXiv preprint arXiv:2008.05756* (2020).
- [315] Xue Ying. “An overview of overfitting and its solutions”. In: *Journal of physics: Conference series*. Vol. 1168. IOP Publishing. 2019, p. 022022.
- [316] Amir Ghasemian, Homa Hosseinmardi, and Aaron Clauset. “Evaluating overfit and underfit in models of network community structure”. In: *IEEE Transactions on Knowledge and Data Engineering* 32.9 (2019), pp. 1722–1735.
- [317] H Jabbar and Rafiqul Zaman Khan. “Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)”. In: *Computer Science, Communication and Instrumentation Devices* 70.10.3850 (2015), pp. 978–981.
- [318] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. “Human activity recognition: A survey”. In: *Procedia Computer Science* 155 (2019), pp. 698–703.

- [319] Chris Ellis et al. “Exploring the trade-off between accuracy and observational latency in action recognition”. In: *International Journal of Computer Vision* 101 (2013), pp. 420–436.
- [320] Ce Li et al. “Memory attention networks for skeleton-based action recognition”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.9 (2021), pp. 4800–4814.
- [321] Valeria Andreieva and Nadiya Shvai. “Generalization of cross-entropy loss function for image classification”. In: (2020).
- [322] Pieter-Tjerk De Boer et al. “A tutorial on the cross-entropy method”. In: *Annals of operations research* 134 (2005), pp. 19–67.
- [323] Nan Cui. “Applying gradient descent in convolutional neural networks”. In: *Journal of Physics: Conference Series*. Vol. 1004. IOP Publishing. 2018, p. 012027.
- [324] Mohamed Mostafa Soliman et al. “Violence recognition from videos using deep learning techniques”. In: *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE. 2019, pp. 80–85.
- [325] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. “Violent flows: Real-time detection of violent crowd behavior”. In: *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE. 2012, pp. 1–6.
- [326] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.
- [327] Waqas Sultani, Chen Chen, and Mubarak Shah. “Real-world anomaly detection in surveillance videos”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6479–6488.
- [328] Peng Wu et al. “Not only look, but also listen: Learning multimodal violence detection under weak supervision”. In: *Computer Vision–ECCV 2020: 16th Euro-*

- pean Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer. 2020, pp. 322–339.
- [329] Yuxin Peng, Yunzhen Zhao, and Junchao Zhang. “Two-stream collaborative learning with spatial-temporal attention for video classification”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.3 (2018), pp. 773–786.
 - [330] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. “UCF101: A dataset of 101 human actions classes from videos in the wild”. In: *arXiv preprint arXiv:1212.0402* (2012).
 - [331] AppleInc. *iMovie, Turn your videos into movie magic*. 2023. URL: <https://www.apple.com/uk/imovie/>.
 - [332] Ji Li et al. “Efficient violence detection using 3d convolutional neural networks”. In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2019, pp. 1–8.

Appendix

1 Overview of the Demographic Influence of Violence

Further investigations determine whether the act of violence had a more significant impact on ethnic backgrounds, age, and the variations in the categories of attacks to pinpoint its severity. The author [11] presented current data as 69% of homicide victims circumventing specific ethnic groups in 2021, but 98% of the victims were of black descendants with fluctuations during the 2019 period. The analysis highlighted the Asian-India community at 8% to demonstrate the effects of violence and its 3% impact on other ethnic groups. The statistics proved that violence is not partial to one specific group. Although most homicide victims were white, the black ratings depicted an act of victimisation as its average rate was six times higher and four times higher for the other groups. The demographic influence held no significant relation to homicide, and the socioeconomic indicators across other ethnic groups showed no impact. The focus emphasises how violence perpetuates as the rate of homicide increases drastically across the demographic groups over the period March 2015–2021, regardless of age, in Figure 1.1 below. The author [11] above reports showed that violence is predominant in the age group between 16-64 for 74% of the victims when amalgamating the ethnic groups, thus validating the signs of violence as a scourge on society.

1.1 Appendix: Overview of Categories/Weapons Influence

Relating to the previous analysis on the impact of violence per demographic grouping, it was necessary to understand the categories of violence and the instruments used as the target objects of interest for the current research objective. Authors' [294], [11], and [295] reported rising homicides relative to the use of bladed objects and how the classes of instruments were specific to a category of violence. The rationale for presenting the analysis on such objects emphasises the rise in the use of bladed instruments and its category of violent attacks from March 2014-2015-2021 in 1 and 2 in Figure 1.2.

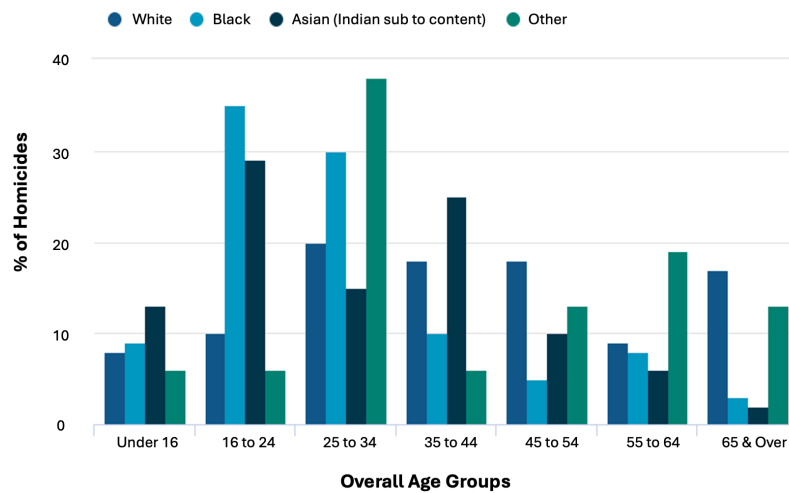


Figure 1.1: Appendix: Victims %: Age-Ethnicity, England/Wales, March 2011-2021.

The data substantiated the significance of such weapons within the classes of violence relative to the investigations. Using bladed objects or weapons adds value during the developmental stages to specify the duration and confirmation of the attack from its inception to its end. The previous notion proved the significance of its presence in violence and validated such objects as features of interest for the current research in Figure 1.2 graphs 1 and 2. Observations presented other statistics emphasising the impact of the loss of human life in [296] because of violence across the UK and the world. The data in [297]

proved that violence is drastically increasing, with the murder tolls significantly targeting women, children, senior citizens, and the innocent over the years with limited preventative mediums.

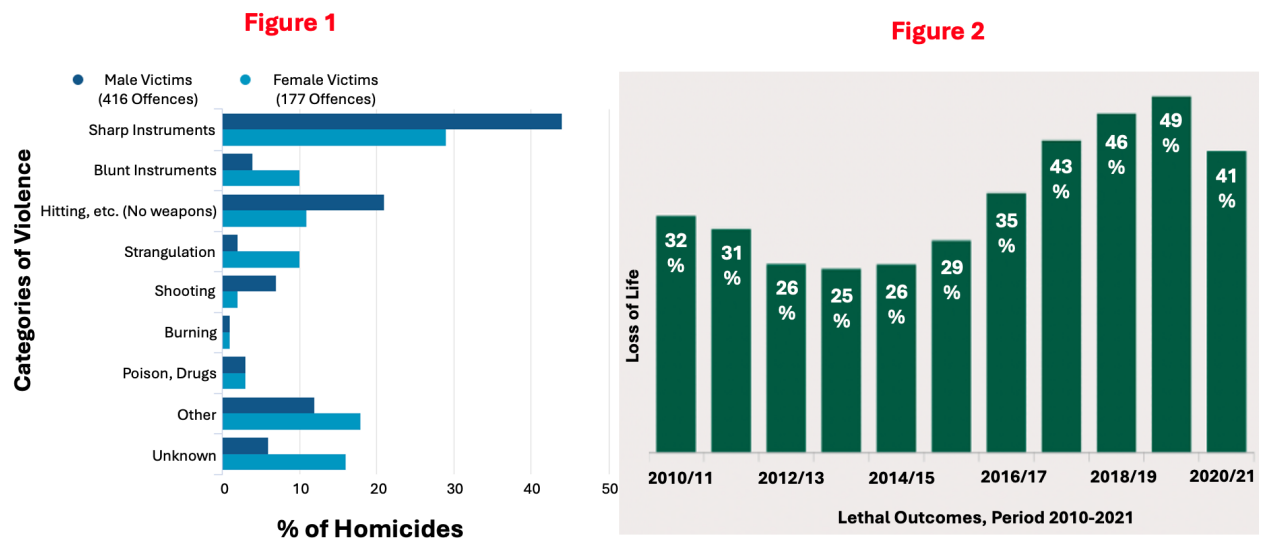


Figure 1.2: Appendix: Sharp Object Offences Escalates In The UK, March 2011-2021.

2 Overview of YOLOv5 Processing

With the fundamental understanding of YOLOv5 input-output processing operations, further analysis introduced violent data at the input stage in Figure 2.14 to begin the training operations. The data undergoes a series of feed-forward convolutional procedures towards the backbone's first and second bottleneck CSPL layers. These generate feature map representations of objects of interest from images of a 416x416 scale. Classification demonstrated in [33] occurs by inputting pre-processed image data into an artificial intelligence model in a feed-forward manner to generate values that indicate distinct features of violent action objects utilising a series of processing layers.

2.1 Appendix: Overview of YOLOv5 Activity Recognition Layers

Authors' [80] and [298] mentioned that YOLOv5's learning strategy employs multiple kernel filters that compute weights (modifying parameter values assigned to computing nodes) and biases to track and analyse the proficiency of the output values through back-propagation (backwards and forward passing of data) and parameter fine-tuning. The weighted values discussed in [299] are applied to regulate the input signal's accuracy path, which moves from one neuron to the next during the processing. This feature reflects values allocated to specific connections amidst the neurological nodes within the architecture [300]. The network's biases represent a surplus of computational values that formulate relationships emphasised in [301] amidst processing parameters. During the operations, the model generates an output identical to [302] regardless of the numerical denomination of the input value. The model's back-propagating operations accentuated in [303] helped solve output-weighted error values relating to regions of interest in the image scenery. According to [298], such outputs are processed backwards through the fully connected nodes to determine the influencing path as the output. Authors' [304] mentioned that identifying the latter allows the computation to influence the final output's strength or fine-tune the connections to achieve the highest prediction results.

2.2 Appendix: Overview of 3DCNN Processing Layers.

The breakdown of 3DCNN's operations commence by presenting a summary of operations from (A)-(F) as follows.

(A) 9 Convolutional (Conv3d) Layers: The layer enhances class feature associations of 3D objects across spatiotemporal boundaries relative to its optical flow in each image during its generalisation operations. Its layered processing conveyed 3-dimensional kernel filters that operate in an (x, y, and z) direction relative to a stride of one in a

three-dimensional sphere.

- (A) Its convolutional operations discussed in [120] produced a cube-like tensor, activation, and feature map that provides additional object details, which denote edges and demarcations for other layers.
- (B) **4 Activation layers (Act):** The RELU (Rectified Linear Unite) activation function discussed in [104] consists of a piece-wise linear operation that outputs the input directly if it is positive. Otherwise, its processing produces a zero output. Its processing detailed in [90] accelerated the training phases of the network by increasing non-linearity in the input data to remove all black attributes from the violent activity features and retain only those that represent positive values.
- (C) **4 Max-Pooling Layers:** The layer down-samples the resolutions of the feature map and tensor utilising a mathematical computation to control the dimensionality of the activation block (feature map block) at the feature level. Author's [305] confirmed its operational and efficacy enhancements during the training stages.
- (D) **3 Dropout Layers:** The layer regulates the violent activity data output to reduce over-fitting by randomly dropping computing neurons at this stage to force the operations to generate more robust results. The operations prevent interconnecting neurons from promoting computational dependency on high-computing neurons [306]. The approach in [307] suggested slight modifications with a dropout parameter that forces the model to learn violent human actions independently and encourages less processing complexity when handling unseen data.
- (E) **1 Flatten Layer:** The layer conforms multidimensional output from convolution into a 1-dimensional vector consisting of a string or row of values as a 1-dimension tensor. This computation emphasised in [308] and [309] is designated before the dense layers as it anticipates data in a 1-dimensional format and retains all values in the tensor relative to the violent object features for classification.

(F) Dense Fully Interconnected Layers: The layer further extracts high-level violent activity feature details and maps such attributes of the convolutional layers with probability scores into the correct labels with a Soft-Max function [104] and [310]. The purpose of the Soft-Max function discussed in [311] and [312] assigns score probabilities that equate to a 1 or 0 to the violent actions feature labels. That insinuates the ratio of its accuracy and the confidence of the model’s classification operations.

2.3 Appendix: Evaluating Binary/Multi-Class Classification Models

The evaluation concept demonstrates the approach taken to evaluate the potential of artificial intelligence models. The discussions in this section demonstrate the confusion matrix computations and how it applies to establish accuracy in a binary classification (two-class) model.

2.3.1 Appendix: Overview of Table 2.4.1 Metrics/Evaluation Measures:

At this level, examples of the evaluation metrics projected the viability of object detection and activity recognition towards violence. It is necessary to comprehend the available metrics that evaluate convolution operations. The ubiquitous metrics below produce results in various contexts. The evaluation metrics are as follows.

| | Estimators | Processing Description |
|-----------|---------------------------|---|
| TP | Recall/True Positive Rate | predictions that are Correct or True |
| TN | True Negative rate | predictions that are NOT Correct and it is True |
| FP | False Positive rate | predictions that are NOT True |
| FN | False Negative | predictions that are NOT Correct and it is NOT True |

Table 2.1: Appendix: Confusion Matrix Performance Evaluation Logic.

| | | Actual Values | |
|------------------|--------------|---------------|--------------|
| | | POSITIVE (1) | NEGATIVE (0) |
| Predicted Values | POSITIVE (1) | TP | FP |
| | NEGATIVE (0) | FN | TN |

Figure 1.3: Appendix: An Illustration of Confusion-Matrix (CM) Binary Operations

2.3.2 Appendix: Summary of Confusion Matrix Multi-Class Manual Process:

In this section, a projection of the confusion matrix operations emphasises the idea of its processing compared to binary classification. Each column of the confusion matrix depicts an instance of the predicted class, where each row highlights outputs as accurate predictions. A perfect classification operation generates values only on the diagonal plane to indicate that the model accurately categorised each class of the test data samples with a 100% proficiency rating. However, the diagonally projected values described in [141] represent the operation's actual performance, and the vertical values starting from the top and bottom of each positive centre value represent false positive instances. Each prediction on the horizontal plane indicates false negative values, and the vertical values indicate false positive output. The processing formulates a cross. Every instance external to this crossing represents actual negative values [313]. The multi-classification operations depicted in Figure 1.4 reflect an accuracy colour bar indicator. The bar outlined in [314] insinuates the threshold degree of accuracy where dark colours suggest high performance.

The illustration presented an overview of the logic circumventing its accuracy towards computing all instances. The confusion matrix demonstrates output that reflects the proper labels of the classes and the actual prediction of objects (knife, stick, ball, gun) from the model's perspective.

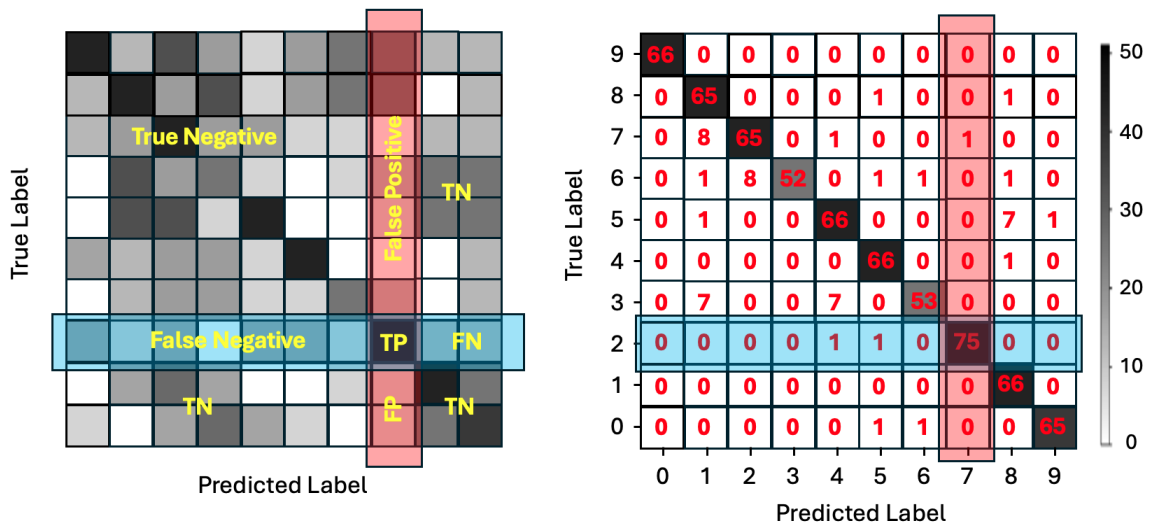


Figure 1.4: Appendix: Confusion Matrix (CM) Multi-Class Processing Source: [141].

2.3.3 Appendix: Overview of Confusion Matrix TP, TN, FP, FN Process:

Table 2.2 and Table 2.3 present a visual perspective to simplify the computation of overall performance manually to fortify the reader's perspective. At this stage, the data proved that the weapon gun attained the highest performance of 28, followed by ball 25 and stick 23. At the lower end of the spectrum is a knife at 19. Knife attained the highest TN at 79, followed by ball at 75, stick at 73 and gun at 70. In this instance, the model attained the highest FPs for the stick at 5 and 1 misclassification for the ball. The table also demonstrates computational deficiency as the stick FN prediction is at 5, along with the ball at one as a crucial systematic error in judgment. These predictions are **NOT Correct**, and it is **NOT TRUE**. The understanding provides insight into the criticality of the

proposed pre-empted logic in Chapters' 4, 5 and 6 when considering the misclassification of violence relative to human life.

| | Class | TP | TN | FP | FN |
|---|--------------|-----------|---------------------------|-----------|-----------|
| 1 | Knife: = | 19 | (23+0+0+0+25+0+2+1+28) 79 | (0+0+0) 0 | (3+0+0) 3 |
| 2 | Stick: = | 23 | (19+0+0+25+0+1+28+0+0) 73 | (3+0+2) 5 | (0+0+0) 0 |
| 3 | Ball: = | 25 | (19+3+0+23+28+0+0+0+2) 75 | (0+0+1) 1 | (0+0+0) 0 |
| 4 | Gun: = | 28 | (19+3+0+0+23+0+0+0+25) 70 | (0+0+0) 0 | (0+2+1) 3 |

Table 2.2: Appendix: Confusion Matrix Manual Computation.

| | Kinfe | Stick | Ball | Gun |
|-----------|--------------|--------------|-------------|------------|
| TP | 19 | 23 | 25 | 28 |
| TN | 79 | 73 | 75 | 70 |
| FP | 0 | 5 | 1 | 0 |
| FN | 3 | 0 | 0 | 3 |

Table 2.3: Appendix: Overview of the CM Multi-Classification Results.

Table 2.3 Manual Overview of the confusion matrix Overall Accuracy (**OA**). Computation:

$$\text{Sum All Values} = 19 + 23 + 25 + 28 + 3 + 0 + 0 + 0 + 0 + 0 + 2 + 1 + 0 + 0 + 0 + 0 = \mathbf{101}$$

$$\text{Sum All Diagonal Values} = 19 + 23 + 25 + 28 = \mathbf{95}$$

$$\mathbf{OA} = \frac{95}{101} = \mathbf{0.94}$$

2.4 Appendix: Summary: Good Fit, Over-fitting/Under-fitting Issues:

To understand the evaluation processing, concerns linked to over-fitting and under-fitting as significant concerns are defined during the developmental stages to disclose processing effectiveness. Figure 1.5 below presented a case of over-fitting to demonstrate a scenario

where the model learns the features of the training data in a fashion that appeared efficient. The operation discussed in [315] and [316] specified that its processing proved superb to that point it cannot infer substantial predictions results on new data, thus increasing generalisation errors. The illustration depicts the initial start of the training processing in blue at a high loss, gradually settling closer to zero. However, due to the data's complexity, the values gradually ascend as the epoch iterations increase.

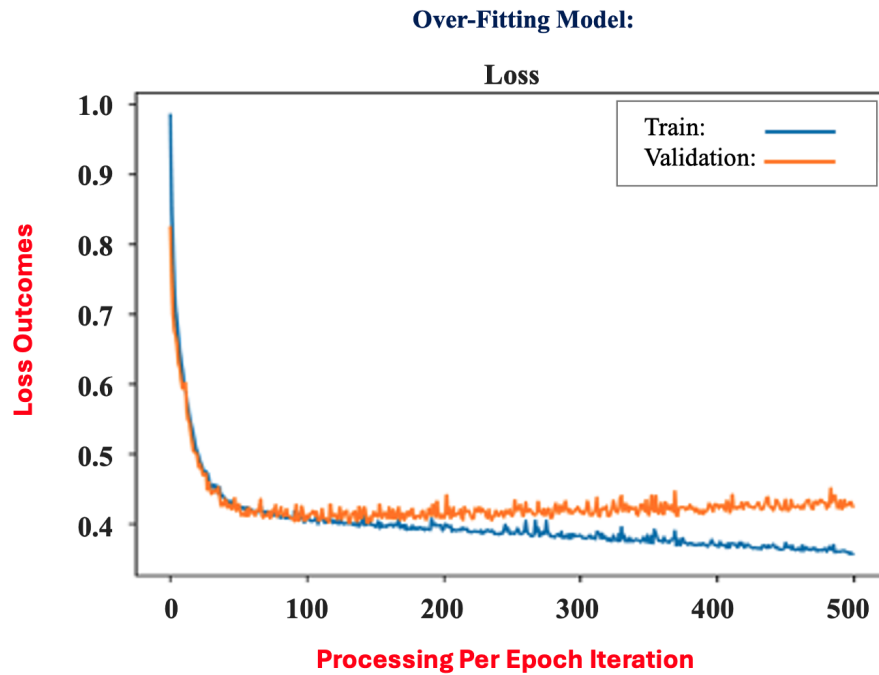


Figure 1.5: Appendix: Graphical Representation of Over-fitting Classification.

2.4.1 Appendix: Summary of Classification Issues: (Under-fitting):

With an understanding of over-fitting in [317], a representation of an under-fitting scenario outlined its context in Figure 1.1 in models (a)-(b). The illustration portrays the framework processing deficiencies, primarily when integrating small dataset volumes during training. Instances reflecting distinct separations of the training in blue and validation curve in

orange suggest abnormal processing. As a result of the previously mentioned issues, the operation produced a small number of low error values that impacted the gradual curvature of the graphical representations. Author's [318] interpreted such instance as an insignificant operation. The processing described in [161] demonstrates its starting point from the left highest value and ending at the right lowest value for every epoch iteration during training. Bridging the knowledge gap concerning iteration meant that the epoch iteration reflects the total repetitions of forward and backward passes of all training data in one cycle during training.

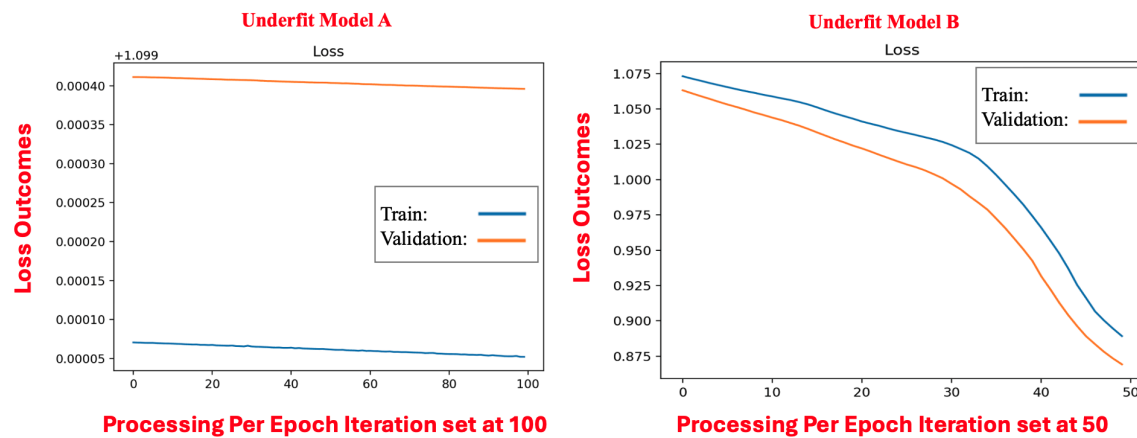


Figure 1.6: Appendix: Model (A)/(B) under-fitting Classification Operations.

2.4.2 Appendix: Summary of Convolution Performance Issues/a Good Fit:

A well-fitted operation demonstrates the model's loss and accuracy values as a smooth output per each iteration of the epoch cycle displayed in Figure 1.7. The loss values outlined in [160] represent the summations of error values that measure the overall performance of (how good or bad) the model's processing abilities. An insignificant result dispenses high error values, causing the loss value to have a similar effect to a fluctuation of the accuracy output. The latter issue described by [319] implies that the operations did not attain optimality with a low classification result. Understanding how the loss values impact the

overall performance is critical. The idea fortified by [320] insinuates that the lower the loss values, the better the performance result relative to the previously mentioned cross entropy approach for classification and regression application. The loss function performance in

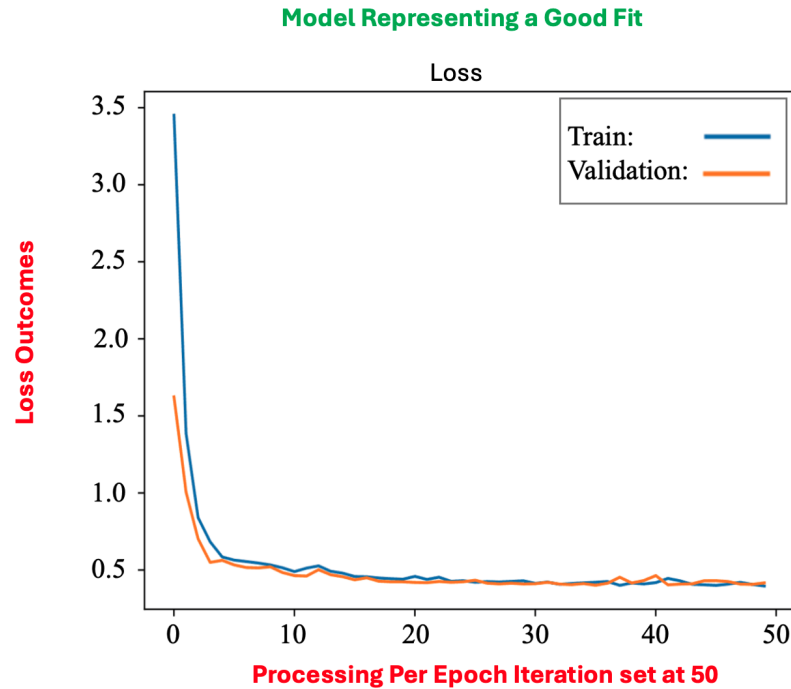


Figure 1.7: Appendix: Graphical View of Good Data Fitting Operations.

Figure 1.7 encourages learning as per [321] by employing a gradient descent technique to evaluate the loss results details. The strategy outlined in [322] alters the architecture's processing parameters accordingly. Author [323] 's operations present a smooth descending graphical curve as a nuance of the training performance blue in colour, which is always relatively higher than the validation result in orange. The outcome occurs because the operation utilises data already processed during the training stages. With an understanding of processing issues and their ability to impact operations negatively, further investigations into the literature provided available approaches to promote efficiency.

3 The Methodology

Further analysis of the datasets, such as Real-life violent situations with 2000 samples, provided a ratio balanced between violent and non-violent human actions. Its contents in [324] reflect samples conveying similarity in environments and poor resolution, with few files of significance regarding stabbing patterns. Violent Flows disclosed by [325] consists of 246 videos of crowd behaviour accumulated via YouTube. Their dataset incorporates excessively noisy background image sceneries with a few suitable samples of stabbing actions. Microsoft COCO dataset explored by [326] consists of 91 object classes and 2.5 million labelled instances in 328k images. Their dataset contains objects that bear no significance towards the current task.

Regarding [327], the UCF-Crime dataset repository comprises 128 hours of untrimmed real-world surveillance videos with 13 realistic actions, excluding stabbing altercations. LIRIS dataset for action recognition explored by [231] and [232] contains 55,298 images of non-violent activity data with 828 actions excluding violent actions. XD-Violence in [328] comprises 4754 untrimmed videos relevant to shooting and fighting, excluding stabbing actions. UCF101 dataset outlined in [329], [330], and [245] entails 13,320-video samples sourced from YouTube with 101-categories of human actions exceeding 27-hours. This dataset has a fixed frame rate of 25-FPS (frames per second) and a resolution of 320×240 with a few stabbing samples. All sources contained actual samples from a neutral action perspective; however, they lack relevant violent actions representing essential primitive stages (pre-start) of stabbing scenarios.

3.1 Appendix: YOLOv5 Action Recognition Data Grouping Summary

Conforming the data into frame sequences to meet the architectural specifications of YOLOv5's design is astronomically essential. If this aspect is not satisfied, the new raw data increases the complexity of the model's processing capabilities and harms the pro-

-cessing. The Hand-Break tool [35] batch processed RWVAD1st data simultaneously to refine its contents to a specific size ratio. The operation transforms large volumes of image codecs and resolutions to an acceptable ratio, removing unwanted elements in the data in batches. iMovie software processing tool conforms and enhances RWVAD1st attributes via cropping, resizing, flipping, and deleting unwanted frames. The method employed [246]’s blob analysis augmentation tools, which generated surplus data by altering its orientations boasting RWVAD1st’s dataset volume. Keeping in line with YOLOv5’s by-frame training operations, a manual data cleaning approach extracted 3662 frames from 160 videos previously mentioned at 30 fps (frames per second) comprising 1831-violent and 1831-non-violent samples to facilitate training and the proposed objectives in research questions 2 and 3 in 1.2. Aligning RWVAD1st concepts with research questions 1, 2, and 3 in 1.2 ensures the operations consider the activity’s life cycle, artefacts, start, middle, and end duration for efficiency. The data orientation strategy applied evaluated YOLOv5’s processing effectiveness. The concept illustrated in Figure 3.1 depict an example of actions portrayed in RWVAD1st for impact experiments in Section 4.

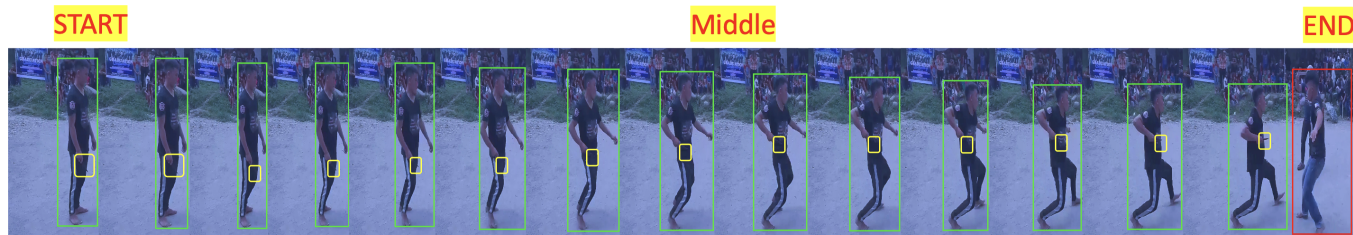


Figure 3.1: Appendix: Stabbing Sequences from its Start, Middle and End.

3.2 Appendix: Overview of 3DCNN Dataset Grouping for Activity Recognition

The rationale behind acquiring a new volume of data was to train the model on new variations of violence to enhance its operations and encourage robust results. Further investigations into social media sources accumulated 100-stabbing samples conveying action similarities. Our neutral samples highlighted in Table 3.1 derived from [245] at a 15-second

| # | Class Label | 3DCNN Class Label Description |
|-----------------------------|------------------------|---|
| non-violent(neutral) | | Generic Category: Indication of normal human actions |
| 0 | Cutting-in-Kitchen (C) | Indications of food preparation in a kitchen as a neutral class |
| 1 | Nun-chucks (N) | Indications of individual/s using nun-chuck alone in a non-violent manner |
| 2 | Fencing (Fe) | Actions relative to the fencing sport/ For Action Similarity experiments |
| 3 | Sumo-wrestling (Su) | Similarity relative to the wrestling sport |
| 4 | Walk-with-dog (W) | Actions relative to person/s walking dog/s |
| 5 | Knitting (K) | Actions relative to person/s sitting/ knitting |
| Violent Classes | | Generic Category: 1 vs 1, many vs 1, 1 vs many, group violence |
| 6 | Fighting (Fi) | Striking with arms/legs to cause harm / For Action Similarity Experiments |
| 7 | Beating (B) | Striking with object to cause bodily harm |
| 8 | Shooting (Sh) | Utilising projectile weapon/s to cause human endangerment |
| 9 | Stabbing (St) | Utilising bladed/sharpened instrument/s to cause bodily harm |

Table 3.1: 3DCNN Activity Recognition Subclass Description.

maximum duration with a frame rate of 25fps (frames per second) and an image dimensionality of 320x240. In UCF, each class folder contained an imbalance of samples. A manual selection comprising 40 samples per class, including the new raw data, facilitated 3DCNN’s processing standards. Following YOLOv5’s pre-processing initiatives, it maintained data consistency utilising [35] and [331]. That operation standardises the raw samples to a dimensionality of 320x240, formulating the new RWVAD2nd dataset of 560 samples for the 3DCNN’s processing. 3DCNN classification

differs by representing violent class labels instead of blob objects, which generate probability scores as a final output. Therefore, applying blob analysis to determine regions of interest in the image scenery is not viable for activity recognition at this stage. During development, omitting data samples conveying unrealistic violent activity relative to acting and movie scenes reduced the high risks of false positives during 3DCNN's processing.

| # | Criteria Required | YOLOv5 Dataset Enhancement Description |
|----|--|---|
| 1 | Images per class greater than 1500 | Number of images per class recommended |
| 2 | Instances per class greater than 10000 | Labelled objects per class recommended |
| 3 | Image variety | Different: times of day, seasons, weather, lighting, angles, sources |
| 4 | Label consistency | All instances/classes in images must be labelled |
| 5 | Label accuracy | Limited spacing between object and its bounding box. No objects should be missing a label. |
| 6 | Background image support | Images with no ROI objects: Reduces False Positives (FP). The author recommends about 0-10%. No labels are required for background images |
| 7 | Epochs | 300 epochs for training |
| 8 | Image size | Native resolution of image size = 414 x 414 & 640 x 640 |
| 9 | Batch size | Largest batch-size hardware allows |
| 10 | Hyper-parameters | Default options embedded in script file hyp.scratch.yaml |

Table 3.2: YOLOv5 Activity Recognition Enhancement Standards.

3.3 Appendix: Overview of YOLOv5 Operational Standards to Promote Robust Processing

YOLOv5's architecture required explicit standards and dataset reinforcements to encourage robust results. Table 3.2's enhancement standards aligned RWVAD1st's dataset with [108]'s precepts to encourage efficiency. Following the

dataset enhancements, it was necessary to incorporate [246]’s blob analysis tools to conform RWVAD1st’s file structure criterion to each image’s desired object output format. Its output structure entails (.txt) files containing the identification of the subclass object value (the object’s centre reflecting x, y, width, and height) as coordinates for each object’s location in the scenery of every image in the root processing folder. For example, one violent image processed via Robo-Flow produces several objects of interest. This operation generates a single (.txt) file with details representing the coordinates of each blob per object and image. After blob analysis, each image contains several labelled sub-classes to signify the identity, presence and location of blob objects. Those coordinates constitute the **object-class, x, y, width(w), height(h)** in Table 3.3 as a nuance of a Stabbing01.jpg

| Subclass ID | X | Y | W | H |
|-------------|----------|----------|----------|----------|
| 7 | 0.716797 | 0.295833 | 0.316406 | 0.847222 |
| 3 | 0.487109 | 0.579167 | 0.355469 | 0.858333 |
| 9 | 0.568709 | 0.769832 | 0.767667 | 0.219496 |
| 14 | 0.659803 | 0.353554 | 0.878912 | 0.989821 |
| 2 | 0.320312 | 0.340625 | 0.695833 | 0.966667 |
| 6 | 0.516797 | 0.495833 | 0.316406 | 0.847222 |
| 8 | 0.887109 | 0.679167 | 0.355469 | 0.958333 |
| 10 | 0.212111 | 0.532211 | 0.455667 | 0.598223 |
| 4 | 0.120312 | 0.395833 | 0.340625 | 0.766667 |
| 1 | 0.316797 | 0.295833 | 0.316406 | 0.747222 |
| 11 | 0.206090 | 0.436787 | 0.214309 | 0.902121 |
| 13 | 0.100834 | 0.766761 | 0.673443 | 0.455359 |
| 0 | 0.287109 | 0.779167 | 0.355469 | 0.858333 |
| 15 | 0.870415 | 0.888892 | 0.457671 | 0.655586 |
| 5 | 0.720312 | 0.795833 | 0.340625 | 0.766667 |

Table 3.3: 3DCNN Activity Recognition Subclass Description.

image after blob analysis operations. YOLOv5’s root folder directory path facilitates all program scripts constituting its object detection framework. The collaborative processing effort of the scripts executed vital instructions from a hierarchy directory tree path

illustrated in Figure 1.2. Without the processing order, the significance of the model's core operations is futile. YOLOv5 root folder relies on library packages to interconnect its processing to assist the architecture in generating desired results. Table 3.4 defined the contents of each folder and provided a means of identifying scripting and package integration error prompts during development. Exploring multiple versions of YOLOv5 frameworks

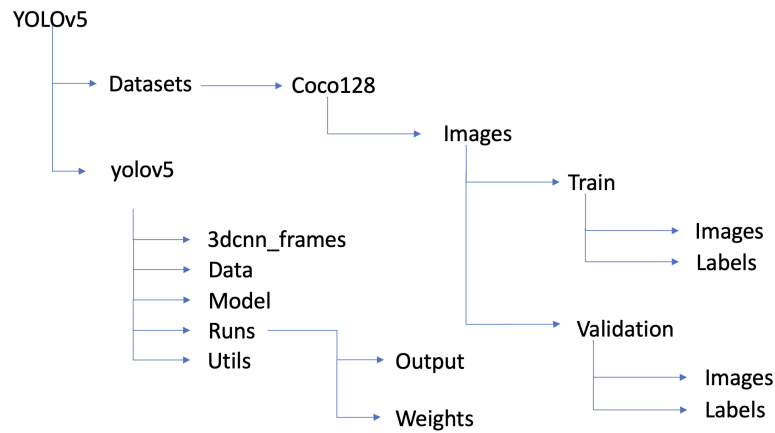


Figure 1.2: Appendix: Overview of YOLOv5's Directory Tree-Path.

via transfer learning initiatives utilising Microsoft benchmark COCO dataset provided the means to satisfy the research objectives. A redirection of YOLOv5 processing weight parameters from COCO's dataset towards RWVAD1st promoted processing efficacy. Each file specified by [108] contained 16 adjacent frames representing the class of activity and duration specifications, which endured a resizing operation to facilitate the input height and width conditions at 416x416x3x16. Regulating performance incorporated the finetuning of YOLOv5's hyper-parameter options reflecting a batch size of 16 and 32, an epoch of 32 and 300 to observe model superiority relative to research question-5 in Section 1.2.1. Retaining YOLOv5's default transfer learning on Microsoft COCOs running a requirements.txt script from the repository folder via the hardware terminal improved the library updates and model training. The operations investigated data containing pre-processing and without

pre-processing with and without background image enhancements to satisfy research questions 1 to 3 and 5 in Section 1.2.1.

| # | Folder | YOLOv5 Root Folder Contents Description |
|----|-------------|---|
| 1 | YOLOv5 | Main directory folder containing all files |
| 2 | Datasets | All training and validation data |
| 3 | Coco128 | kept the naming convention and swapped its data with the VRWA dataset |
| 4 | Images | Specified to differentiate between the validation and test data |
| 5 | Train | The data files used for training |
| 6 | Images | All images sectioned for the training operations e.g. (Stab01.jpg...n/ Fencing01.jpg...n) |
| 7 | Labels | All blob labels specified in the training images e.g. (Stab01.txt...n/ Fencing01.txt...n) |
| 8 | Validation | Data used for cross validation to guide the learning/ inference (Stab05.jpg...n/ Fencing03.jpg...n) |
| 9 | Images | All images sectioned for the validation operations (Stab05.jpg...n/ Fencing03.jpg...n) |
| 10 | Labels | All blob labels specified in the validation images (Stab05.txt...n/ Fencing03.txt...n) |
| 11 | yolov5 | All model configuration/processing files (.PY, .YAML, .TXT, .CSV) |
| 12 | 3dcnnframes | Stores the same 3dcnn inference image data for inference processing with yolov5 |
| 13 | Data | Additional .YAML files for finetuning options |
| 14 | Model | All yolov5 architectures (nano, small, medium, large & extra-large) |
| 15 | Runs | Training and inference results dispensed from processing |
| 16 | Output | Yolov5 output results for the proposal fusion concept |
| 17 | Weights | The pretrained weights for the models specified in #14 |
| 18 | Utils | Metric logging files and data loading/ processing scripts |

Table 3.4: YOLOv5 Folder Contents Definition.

3.3.1 Appendix: Overview of Yolov5 Activity Recognition Operations

At this level, the concept of YOLOv5 as object detection using static unrelated action images compared to YOLOv5m as activity recognition utilising a sequence of frames as video input provides a clear distinction between models. Ad-

hering to configuration standards specified in Table 3.2 onwards aids in promoting option consistency and processing efficiency. Object detection and Activity recognition follow the same stages. The difference between the detection models is that activity recognition utilised a series of frames during input, which conveys the target actions and object detection utilised static unrelated images at the input level. Activity recognition operations entail feeding a sequence of image frames demonstrated in Figure 1.3 and Figure 1.4 at the input stage to determine the classification results of the data. Contrarily, the concept of object detection depicted in Figure 1.5 and Figure 1.6 utilised the same convolutional operations. However, the data at the input stage convey no interrelations or relationship in velocity, acceleration or trajectory when establishing the object of interest in the scenery of each static image.



Figure 1.3: Appendix: YOLOv5 Activity Recognition Processing.

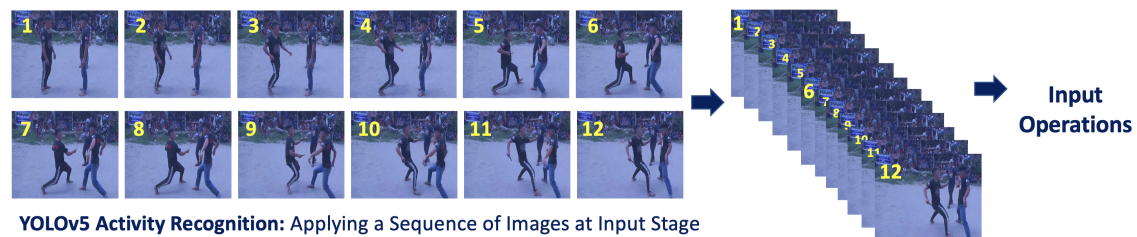


Figure 1.4: Appendix: YOLOv5 as Activity Recognition Input.

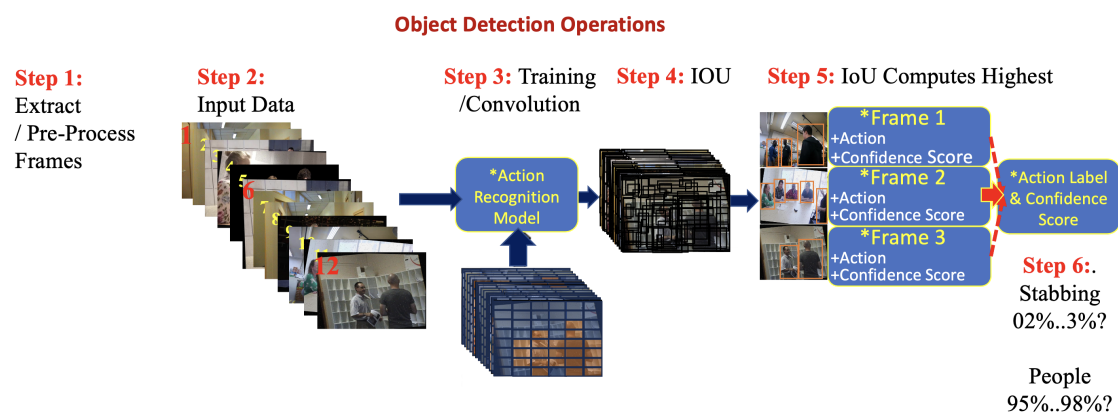


Figure 1.5: Appendix: YOLOv5 as Object Detection.

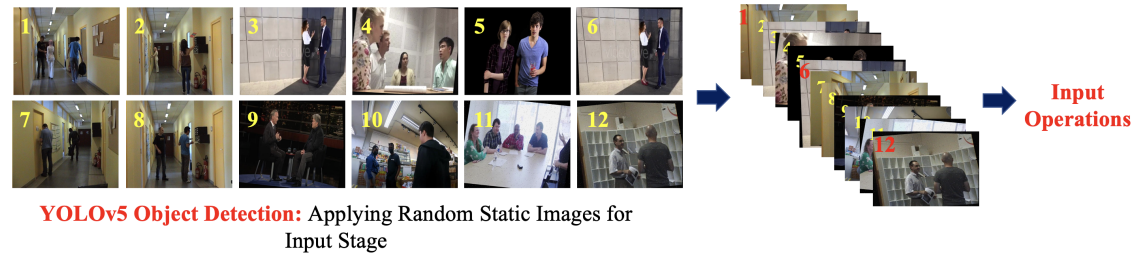


Figure 1.6: Appendix: YOLOv5 as Object Detection Input.

3.4 Appendix: Overview of 3DCNN Standards for Robust Processing

3DCNN requires a root directory path defined in Figure 1.7 to facilitate processing efficiency, which executes its computations utilising various library packages. Like YOLOv5, each folder in the 3DCNN file root path has a distinct function. Table 3.5 projects the contents of the 3DCNN root folder to fortify the notion of its purpose during processing. 3DCNN utilising transfer learning initiatives detailed in Section 2.6 in the report elaborates on the UCF dataset during the training procedures. A redirection of 3DCNN processing weight parameters from UCF towards RWVAD2nd video data, sampled 12-adjacent frames representing the CoAT (Class of Activity Template) aided satisfying research questions 1 and 4 in Section 1.2. 3DCNN's image height and width input conditions encompass 320x240x12x3. Its hyper-parameter performance options reflect a batch size of 130, an epoch of 32 and 90, and a depth of 12 to regulate the size of its tensor block output, which encourages optimal performance. Further investigations into data containing pre-processing, without pre-processing, with and without background image enhancements dis-

-closed model superiority towards satisfying research questions 2, 3, and 5 in Section 1.2.

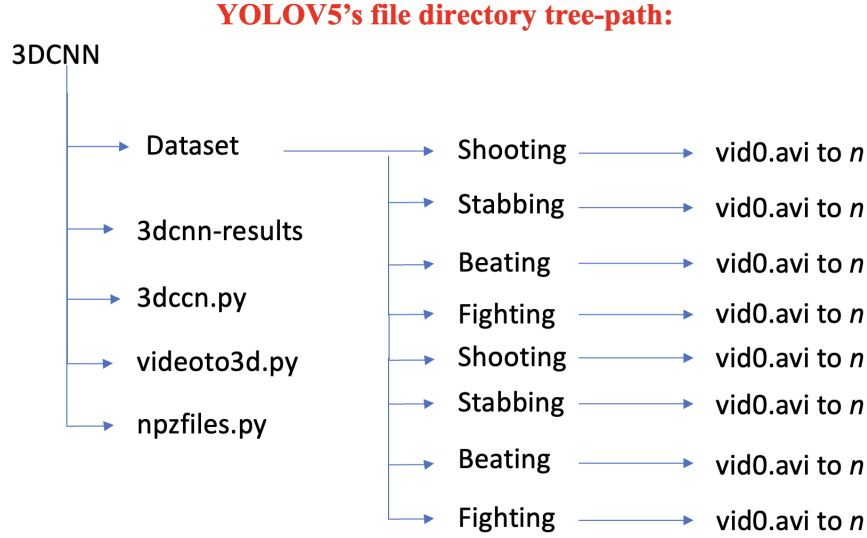


Figure 1.7: Appendix: YOLOv5 as Object Detection Input.

| # | Item | 3DCNN Root Folder Contents Description |
|---|---------------|--|
| 1 | 3DCNN | Root configuration folders, stores 3DCNN's script and data files |
| 2 | Dataset | Contains all violent and non-violent sectioned folders of data |
| 3 | 3dcnn-results | Stores the output results from the operations |
| 4 | 3dccn.py | Consist of the 3DCNN's architecture configurations |
| 5 | Video3d.py | Consist of video handling and frame extraction scripts |
| 6 | Npzfiles.py | Loads/saves data to a .NPZ format for accessibility operations |

Table 3.5: 3DCNN Activity Recognition Folder Contents Description.

4 Appraising YOLOv5 and 3DCNN

A model distinction between object detection and activity recognition operations projected the need for processing feasibility to enhance one's understanding further.

4.1 Appendix: Activity Recognition/Object Detection Input Summary

Figure 4.1 synopsis illustrates object detection utilising 12-static images in **simulation(A)** compared to applying sequences of frames across time for activity recognition in **simulation(B)**. Simulation(A) employed 12 random static images at the input stage to demon-

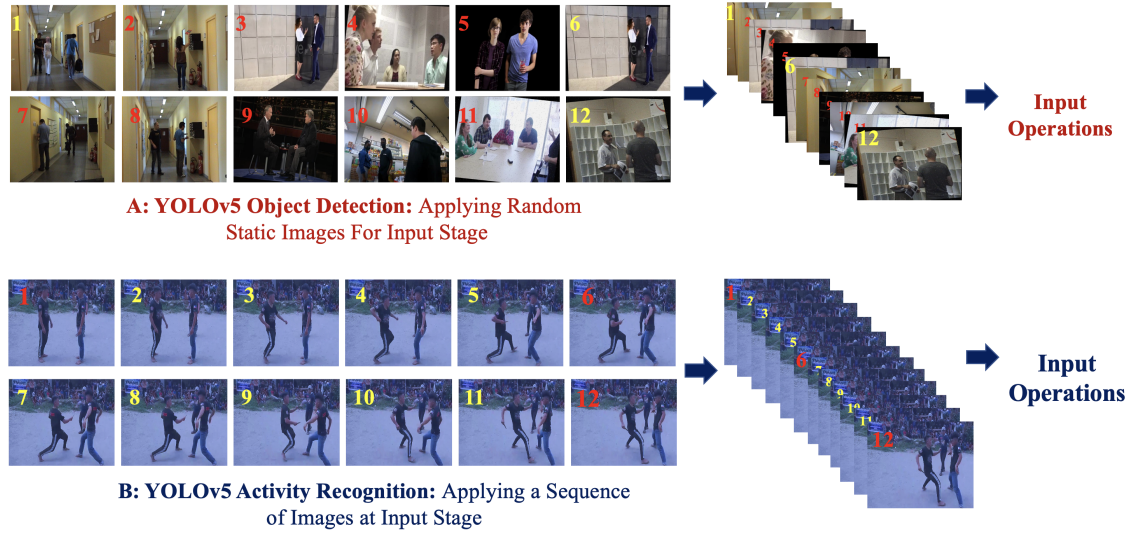


Figure 4.1: Appendix: Rational for Avoiding YOLOv5 Object Detection Simulation(A).

strate the action restrictions from that perspective. Simulation(B) applied a sequence of 12 portraits of actions from video data representing the primitive (pre-empted start) stages of a stabbing motion. Figure (B)'s processing proved viable for satisfying the first research question in Section 1.2.

4.2 Appendix: YOLOv5 Activity Recognition Limitation Summary

The approach entails training YOLOv5 on a section of two classes of pre-processed data from the RWVAD1st dataset. A deliberate application of classes representing action similarity (actions depicting close similarity in characteristics) posed a deliberate plot to chal-

lenge the model's true potential to establish its effectiveness. The class selection consists of the stabbing class projected as violent and a neutral class portrayed by the fencing sport, with 80 videos. Hyper-parameter options utilising 30 epochs and a batch size of 32 maintained option consistency to regulate classification performance. YOLOv5 operations commenced with manually extracting action sequences from the 80 samples for training procedures and configuring the program scripts to process those specified frames. The main configuration script from the terminal command line initialised the operations. The evidence in Image-1 eradicated its classification limitations by evaluating the stabbing class. Observations disclosed an anticipated prediction on the stabbing class with two scores, one at 67% as the highest accuracy. Figure 4.1 Image-2 generated three scores with two critical misclassifications. The analysis proved that the processing conceded crucial errors by misidentifying the victim as a stabbing category at 19% and the aggressor's lower extremities at 12% accuracy. Image 2 contained one correct prediction of the aggressor at 74% accuracy. Image-3's prediction peaked just above an average attempt of 57% with no victim.

Finally, Image 4 displayed an abstract cartoon to challenge the model's capability. The evidence disclosed 2-misrepresented predictions proving the model's partial status during processing at 20% and 13%. The assessment aligned the deficiency in processing with the action similarity of the neutral fencing class in Image-5. Image-6's operations dispensed no results that fortified the fusion concept regarding YOLOv5's processing deficiencies. The operation of 49% on Image-7 validated the model's consistent challenges with the complexity of actions conveying attribute similarity. Image-8 produced no results, thus solidifying the notion of the proposed fusion due to YOLOv5's sporadic interpretations between violent and non-violent actions in conditions conveying complex gait patterns. The analysis projected another scenario with 2-misrepresented predictions validating the model's partial operational state at 20% and 1%. The analysis linked the deficiency in processing to action similarity issues as it struggled to differentiate the neutral fencing class

in Image-5. The crucial misclassification instance of stabbing at 39% and 20% in a fencing event demonstrated the critical need for a robust mechanism that efficiently anticipates lethal actions compared to any other human activity. Though some cases inferred that the fencing prediction was accurate relative to the stabbing activity’s gesturing and lunging, the reality of the class categories is that they convey two separate natures. The overall processing proved detrimental if the model’s operations consistently misidentified authentic objects of interest in the scenery. Other instances confirmed YOLOv5’s limitations as it presented no scores for legitimate actions, thus substantiating the proposed fusion concepts.

4.3 Appendix: Overview of 3DCNN Ideal Confusion Matrix Results

To commence the individual (3DCNN without fusion) simulations, the application of 16-test videos of a balanced class ratio reflecting stabbing and fencing to facilitate the aim and objectives. However, due to 3DCNN’s construct (utilising entire videos), applying eight additional test video samples aided in demonstrating proper classification ratios during the confusion matrix analysis. Adhering to specific 3DCNN standards adds configuration consistency and eradicates biased notions to promote the actual outcome of the model. The strategic approach bears no impact on training exercises for YOLOv5 processes. Table 4.1 present an example of an ideal confusion matrix classification performance to bridge the gap in understanding the misclassification rate of 3DCNN actual processing. The example displayed a perfect classification state of 16-video samples relative to the confusion matrix true positive and accurate negative predictions. Table 4.1 classification performance insinuated that all classes were correctly classified instances. Out of the 16 videos applied for testing, eight stabbing and eight fencing were true positives as accurate outcomes in this scenario. The operations highlight eight videos that are not fencing and eight that are not stabbing, also as an accurate response. No misclassification rates reflected false positives and false negatives as a perfect processing operation. The operation measured overall accuracy to evaluate the classification state of the entire operation. The idea considers the

accuracy performance of the entire process compared to individual classes. The notion projects the actual performance of the entire operation. Considering Table 4.1, the highlights on 3DCNN deficiency in Table 4.2 nuance further indicated that violent generic categories and stabbing sub-classes occurred, which partially predicted the accuracy of scenario. False negatives convey a crucial state of the operations. Its impact directly affects the prevention of lethal scenarios and the loss of human life. In the case of the false positives, the evidence showed five instances: three misrepresentations for fencing and two for stabbing. These instances signified that the model predicted actions that

| Rating | Fencing | Stabbing | Confusion Matrix Description |
|--------------------|---------|----------|---|
| True Positives TP | 8 | 8 | Predictions are Actually True |
| True Negatives TN | 8 | 8 | Predictions NOT True |
| False Positives FP | 0 | 0 | Predictions NOT True but Actually Predicted to be True (It is ok if this is misclassified) |
| False Negatives FN | 0 | 0 | Predictions True but, Actually Predicted to be False (Not ok if this is misclassified) |

Table 4.1: Appendix Example of an Ideal Confusion Matrix Classification State

| Rating | Fencing | Stabbing | Confusion Matrix Description |
|--------------------|---------|----------|--|
| True Positives TP | 6 | 5 | Predictions are Actually True |
| True Negatives TN | 5 | 6 | Predictions NOT True |
| False Positives FP | 3 | 2 | Predictions NOT True but Actually Predicted to be True (It is ok if this is misclassified) |
| False Negatives FN | 2 | 3 | Predictions True but, Actually Predicted to be False (Not ok if this is misclassified) |

Table 4.2: Appendix: Rational for Fusion 3DCNN's Misclassification Break Down.

were not violent activity or the stabbing subclass. Nevertheless, the model predicts the action to be violent. False positive instances are less severe than false negatives in the case of violent activity recognition. The model may create false alerts, but in this instance, it is safer to apply additional support to validate the circumstance of an action to pre-empt and prevent lethal scenarios. Table 4.2 ratings insinuated that the predicted actions were not a violent generic class or a subclass stabbing, but the model identified this as violent. The results validated 3DCNN's unstable performance in scenarios containing complex action similarity focusing on violence. The deficiency in its processing validated the proposed fusion technique as a supporting mechanism that produces robust results regardless of the model's misclassification state and the complexity of violent human gaits.

4.4 Appendix: Validating 3DCNN Limitations Manually

A manual computation accentuated the true summations of diagonal elements relative to the confusion matrix from the left top corner to the bottom right, divided by the accuracy to validate the overall performance in Table 4.2 and Figure 4.2. The idea encompasses a summation of all scores to determine the accuracy generated with a breakdown of the misclassification rates and a final accuracy score to conclude this section. Figure 4.2 fluctuating scores compared to Table 4.1 emphasises the importance of attaining the correct result as its impact can adversely affect the overall predictions, leading to erroneous outcomes exclusively in lethal scenarios. The confusion matrix output projected the model's actual non-violent and violent human gaits processing with high posturing similarities at 68% overall accuracy at this stage. The manual accuracy and overall accuracy operations fortify the concept of the confusion matrix processing, which is as follows.

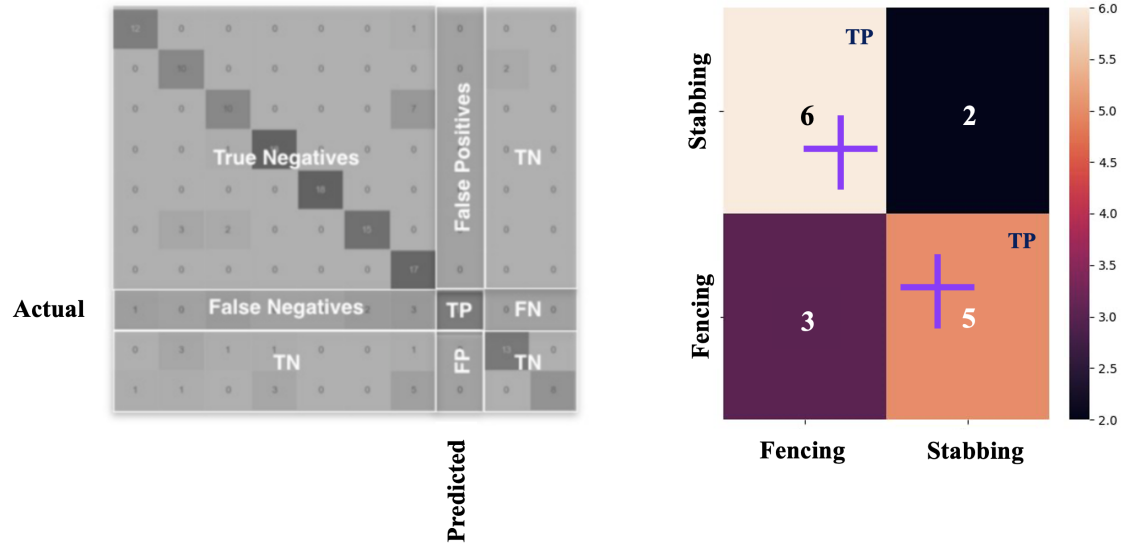


Figure 4.2: Appendix: Unstable 3DCNN Action Similarity Processing Source: [141].

Accuracy Generated = 6 + 5 + 3 + 2 = 16(No. of Test Videos Utilised for Operations))

$$\text{Overall Accuracy} = 6 + \frac{5}{16} = \frac{11}{16}$$

$$\text{3DCNN Overall Accuracy} = 0.6875$$

4.5 Appendix: Overview of Experimental Conditions

The strategic experiments satisfied research questions 1-5 in section 1.2.1 to finally justify the integration of the proposed fusion strategy accentuated in Chapter 5. The experimental conditions are emphasised in two phases. The notion determined the advantages and

disadvantages of YOLOv5 in phase-1 and 3DCNN’s classification state in phase-2.

| # | Conditions | Dataset Size | Impact Experiment Definition |
|---|--|--|---|
| 1 | No pre-processing and NoBackground Images | Original Dataset 160 videos 2284 Images | Experiment contains no data pre-processing or enhancements or background image support. |
| 2 | No pre-processing and WithBackground Images | 5944 Images | Experiment contains no data pre-processing or enhancements but, contains background image support. |
| 3 | With pre-processing and NoBackground Images | 5944 Images | Experiment contains data pre-processing enhancements but, no background image support. |
| 4 | With pre-processing and With Background Images | 6204 Images | Experiment contains data pre-processing enhancements and has background image support. |

Table 4.3: Appendix: Summary of Activity Recognition Impact Experiment Conditions.

4.5.1 Appendix: Phase-1 YOLOv5 Experiment Conditions

YOLOv5 operations consider two forms. The first is the From-Scratch technique, which trains the model using random weight (designated computational values) and option values (fine-tuning values). The second approach denotes a pre-trained transfer learning method, which entails training the YOLOv5 model with parameters attaining benchmark performance utilising the Microsoft COCO dataset. The rationale behind this evaluation approach disclosed model superiority to supplement the decision-level strategy during the proposed fusion’s developmental stages. The notion entail four categories of RWVAD1st reflecting action similarity, no pre-processing, and pre-processing enhancements as specified in Table 4.3. RWVAD1st pre-processing contains cropping, resolution altering 5%, blurring 5%, Gaussian noise (5% salt/pepper), and flipping (right/left).

4.5.2 Appendix: Phase-1 YOLOv5 Framework Version Selection

YOLOv5 investigations revealed p5 and p6 models that facilitated the research questions' intricacies during the developmental stages. Author [108] presented p5 and p6 solutions to handle challenges regarding objects with high acceleration, velocity, and sporadic trajectory. The difference between p5 and p6 is that p5 maintains a kernel processing stride of 32, while p6 employs a stride of 64 in the output layer. The additional p6 output layer increases the accuracy of image scales by facilitating significant object predictions from 414x414, 640x640, and 1280x1280 size ranges. The idea focuses on enlarging small objects of interest during the upscaling of the image size due to the p6's design. During development, up-scaling image sizes create more values for processes, and p6 models increase the demand for memory resources. The smallest image size reduced processing issues by bypassing p6 operations as a measure for future processing endeavours. Further investigations of the YOLOv5 p5 framework disclosed multiple designs that facilitate robust processing. YOLOv5 nano(n) design facilitates mobile solutions. YOLOv5 small(s) design superseded the nano version with 7.2 million processing parameters accommodating CPU inference. YOLOv5 medium(m) design employs 21.2 million parameters and is ideal for complex datasets. YOLOv5's large(l) design consists of 46.5 million parameters, which is superb for detecting smaller objects. However, it encouraged a rapid depletion of memory resources. Finally, the YOLOv5 extra-large (xl) design applies 86.7 million parameters that generated the highest accuracy with the slowest processing compared to all other models. A projection of YOLOv5's design range demonstrated the scope of its processing capabilities and highlighted the magnitude of the investigations to establish a suitable concept to facilitate the proposed fusion approach.

4.5.3 Appendix: Phase One Summary of YOLOv5m Experimental Conditions

Because the larger versions of YOLO facilitate high performance and exceed the hardware's computational limits, the analysis considered a range of versions to gather insight towards

reducing the high risks of generating critical memory latency issues. The task incorporated 12 pre-trained and 12 from-scratch experiments to investigate the research initiatives to evaluate research questions 1 and 5 in section 1.2.1. The idea disclosed processing feasibility and superiority by utilising YOLOv5’s nano, small and medium designs. A fine-tuning procedure maintained option consistency at this level, utilising hyper-parameters values to reflect a standard batch size of 32, an epoch of 30, classes in Table 3.1, and an image orientation set at 414x414 relative to [108]’s specifications. Appraising additional models with 300 epochs maintained a suitable approach for operation consistency. However, the approach peaked between 30 to 35-epochs. Integrating 32 epochs for 160 video samples with a split ratio of 80% training, 10% validation and 10% testing fostered high performance during development. A strict duration (5-15 seconds) through pre-processing increased the overall processing speed. The previously mentioned performance allowed the context of violence in two strategies concerning the primitive stages (pre-start) of each scenario.

Strategy-1 separate regions of interest in blob analysis: Separating regions of interest is crucial during blob analysis and training. This separation adversely affected performance because of the complexity of discerning multiple blobs of interest from unspecified background regions, which conveyed high acceleration, velocity, and trajectory in the image scenery. The challenge surrounds the classification of minute objects with rapid trajectories.

Strategy-2 Combined blobs in blob analysis: Separate ROIs were combined during blob analysis to suggest the pre-empting of violent activity. The idea ensured the integration of small objects to suggest heterogeneous activity. The approach reduced the computational resources required for output and the complexity of learning with fewer important objects for training and classification. Strategy two’s selection reduced the complexity of generalisation amidst layers, promoting high outcomes.

4.5.4 Appendix: Phase-1 3DCNN Framework Version Selection

To define the setup, investigations into 3DCNN multi-level and single-level frameworks proved necessary to fulfil research questions 4, 5 and 6 in section 1.2.1 utilising the RWVAD2nd dataset. The concept employs 2-levels of 3DCNN consisting of 46 layers fused with hidden layers to attain robust processing and encourage high accuracy scores. During development, multi-level processing added high risks of extreme flexibility, causing the layers to exceed the capacity required. Simultaneously, the approach drastically increased the computational parameters required for processing. The idea increased the rate of over-fitting discussed in [126] with a high increase in biased results. The analysis strongly suggested avoiding the multi-level approach because of its high over-fitting risks and its negative impact relative to research question 4 in section 1.2.1. The multi-level architectural limitations solidified 3DCNN single-level as a feasible processing option for activity recognition. The operations favoured the 3DCNN single-level (3DCNNsl) technique towards classifying the generic status and subclass categories of violent actions. 3DCNNsl operations achieved this by incorporating multiple feature reusing and element-wise backward propagation operations, identifying violent patterns reflecting anthropometric anomalies amidst its layers. 3DCNNsl solidified its status further because of its backward propagation operations to evaluate errors. The procedure entails applying processing backwards from its output nodes to the input nodes to improve the model's prediction accuracy. 3DCNNsl spatiotemporal feature extraction technique emphasised in [332] operates by integrating a 3-dimensional convolution kernel processing objects of interest relative to their height, width, image channels, and depth during activity recognition.

4.5.5 Appendix: Phase Two Summary of 3DCNNsl Experimental Conditions

Like YOLOv5m Phase-1, incorporating the identical strategy for 3DCNNsl using the RW-VAD2nd dataset with additional real-life videos proved necessary to maintain experiment consistency. The idea deliberately challenges 3DCNNsl regarding research questions 2, 3 and 4 in section 1.2.1. Each category contained 40-balance samples per folder. Each category contained 40-balance samples per folder. A class strategy of one violent class vs one neutral class of 80 samples, 2-by-2 of 160 samples, and 4-by-4 of 320 samples assists in regulating each experiment scenario. Evaluating the multi-class and generic classification objectives, focusing on the 2-by-2 classes containing 160 samples proved necessary to establish effectiveness. The analysis recorded additional results utilising the 1-by-1 80 samples set and the 4-by-4 of 320 samples set. Investigating those objectives risked exceeding the research developmental life cycle. Observations of 1-by-1 and 4-by-4 objectives satisfied experiments reflecting the impact of increasing data sample sizes for research questions 2 and 3 in section 1.2.1. In this instance, the result's importance is a model fine-tuning guide towards the proposed fusion concepts. The operations achieved further efficiency by implementing crucial framework adjustments to adapt 3DCNNsl to recognise sporadic violent patterns via transfer learning and boost robust classification for the proposed fusion concepts. Modifying the original 3DCNN structure to accommodate 23 layers with a tensor block depth of 12 allowed the model sufficient time to accumulate crucial high-level intricate feature details during processing. A set 50% regularisation in the drop-out layers reduced processing complications between adjacent layers and promoted high accuracy. Like YOLOv5m's setup, utilising a similar cross-validation split procedure (80%-10%-10%) aided with over-fitting. Regarding 3DCNN's construct, a notion to increase the hyper-parameter options to 90 epochs, a batch size of 130, a tensor block depth of 12 for training and 32 videos for testing generated significant operations for analysis. Additional fine-tuning of the hyperparameters during development aided in validating the processing of the previously mentioned option values. Following those alterations and fine-

tuning, YOLOv5m train.py and 3DCNNsl train.py scripts initialised the training, starting the inference stages utilising their detect.py scripts via the hardware's command terminal.

4.5.6 Appendix: Phase-2 Summary of Two 3DCNNsl Processing Methods

Method-1 Dataset Categorisation: The section involved re-configuring the original 3dcnn.py script to identify generic violent or non-violent folder categories/sub-classes. The idea integrated class labelling from the dataset level during training. The concept in the approach increased 3DCNNsl's demand for application memory. The notion forces the model to inflate its simultaneous computational load when analysing statuses and sub-classes in multiple sub-data folders within sub-folders during input data-loading operations. The approach reduced data real-time processing speeds, increasing the ambiguous results during training.

Method-2 Script 3DCNNsl Generic Status Output: Method-2 entailed restructuring the RWVAD2nd dataset to represent one data-store folder and its various sub-classes compared to specifying 2-generic category folders and their contents in sub-folders like Method-1. Modifying the main 3dcnn.py script specified probability scores for each subclass label in only one dataset folder. Applying the generic status and subclass programming at the output level provided processing efficiency. During development, Method 2 substantiated its efficiency as it required fewer data processing stages and ignored complex data categorisation procedures at the dataset level. Simultaneously, this improved the 3DCNNsl operational real-time speed with less dependency on computational resources. Subconsciously, 3DCNNsl lacks the knowledge of reality regarding complex, violent human actions. However, feeding significant volumes of violent action videos reflecting such anthropometric patterns, 3DCNNsl learns complex interlacing gaits across spatiotemporal boundaries in a sequence of frames. The technique proved efficient when discerning actions conveying violence. An overview of the following metrics emphasises the approach considered

to evaluate the effectiveness of the model's performance.

mAP subscript threshold of 0.5: mAP subscript threshold of 0.5 evaluates ground truth vs predicted bounding boxes. Higher scores insinuate a more efficient performance. The point of 0.5 denotes predictions above 50%, representing positive classification scores, and low scores insinuate the opposite.

Precision: It measured the frequency of YOLOv5m's ability to accurately recognise positive class occurrences considering all other instances it predicted to be positive. High scores signify low false positive classification discrepancies.

Recall: Estimated YOLOv5m's ability to accurately classify positive occurrences regarding all ground truth instances of the dataset. High scores insinuate the model's low classification state containing fewer false negatives, and low scores denote insignificant performance.

Accuracy: Estimated global accuracy of 3DCNN's classification operations and capability to predict objects.

4.6 Appendix: Overview of YOLOv5m Result, Analysis

At this point, analysis of the YOLOv5m results demonstrated the fulfilment of the research objectives in chronological order. The analysis is as follows.

A. Fulfilling Research Question-1 with YOLOv5m: In all simulations, YOLOv5m demonstrated its ability to recognise violent activity/weapons in CCTV videos above a threshold of 50%. In Table 4.4, 1 through 8 From-Scratch and pre-trained processing accentuated all mAP-0.5 scores above the previously mentioned thresholds. Research question 1's fulfilment came by achieving the highest performance from #8 pre-trained experiment 21 with a precision score of 0.85, recall of 0.82 and an mAP-0.5 of 0.85. The results proved that experiment 21 maintained superiority over all other experiments.

B. Fulfilling Research Question-2/3 via YOLOv5m pre-trained Operations:

In #3 experiment 6 From-Scratch, an implementation of a dataset size of 2284 images reflected data modifications with no background image enhancements to satisfy research questions-2/3 in section 1.2.1. The operations dispensed a precision score of 0.64, a recall of 0.68 and a mAP-0.5 of 0.67. Observations into experiment #1 in Table 4.4 utilising the same dataset with background image support increased the data to 5944 images. Those operations generated a precision score of 0.66, a recall of 0.51 and a mAP-0.5 of 0.53. The results proved that data increments positively affected the model's classification state. Contrasting #3 experiment-6 and #1 experiment 18 in the From-Scratch context experienced challenges with background image increments, negatively impacting the model's performance. Though #1 experiment 18 precision dispensed high results, the model's recall displayed signs of high false negative errors, where violent predictions were incorrect, and it was true. Experiment #1 results substantiated that data enhancements positively impacted the model's classification state via From-Scratch investigations. Moreover, analysis of #2 experiment 15, compared to experiments #3 and #1, disclosed a depreciation in performance, which insinuated that the data size increments were not the attribute affecting the performance. Experiment #2 generated a precision of 0.61, a recall of 0.59 and a mAP-0.5 of 0.63. Compared to experiment #1, experiment #2 utilised a larger dataset of 6204 images with pre-processing enhancements in From-Scratch operations. The evidence proved that the data was not the issue. However, the complexity of pre-empting violence utilising random weights in From-Scratch procedures proved volatile. Analysis of #4 From-Scratch experiment insinuated that pre-processing enhancements played a significant role during training to promote robust classification results. The analysis disclosed the highest results for From-Scratch operations on a dataset size of 5944 images containing pre-processing without background image support. The operations dispensed a precision score of 0.67, recall of 0.70 and a

mAP-0.5 of 0.72 that superseded all other From-Scratch operations. The experimental investigation projected the performance impact when utilising pre-processed data and no pre-processing (data without modification) via Table 4.4. The assessment commenced by analysing from scratch first and then from a pre-trained perspective to emphasise the contrast between both methods via performance. Analysis of From-Scratch experiments produced inferior results compared to the pre-trained initiatives. The evidence of Pre-trained operations for #5 experiment, 12 and #8 experiment 21 on a dataset containing 5944 images with no data enhancements confirmed the fulfilment of research questions 2 and 3. The #5 experiment 12 task generated a precision score of 0.77, recall of 0.72 and a mAP-0.5 of 0.75, which superseded all From-Scratch operations on every level. In #8 experiment 21, the findings disclosed an increase in performance in operations utilising an increment in data ratios from 5944 to 6204 images containing pre-processing and background image support. Observations showed fluctuations in the processing performance with and without pre-processing and background images between #6 and #7. The results proved the ability to classify violence regardless of real-world conditions. The previously mentioned results verified that pre-trained operations can efficiently recognise violent actions with minimum misclassification challenges by utilising large datasets. The evidence proved that pre-training methods superseded all from-scratch approaches and validated the completion and fulfilment of research questions 2 and 3's initiatives in 1.2.

4.6.1 Appendix: Phase-1 YOLOv5m From-Scratch Confusion Matrix Exp.1/2

At this level, YOLOv5m confusion matrix results provides analysis in Table 4.4 for individual class processing performance. Each experiment is aligned with the confusion matrix results to emphasise the fulfilment of the research questions and accentuate the need for the proposed fusion operations due to the misclassification rates. The diagonal values and bar colour coding (darker colours insinuate high performance) per experiment to mea-

sure the actual performance. Number #1 experiment 18 Figure 4.3 displayed a decline in predictions exceeding the 50-percentile range and stabbing class scores of 0.62 like #2 experiment 15 via From-Scratch operations. Discussion_woboard represented the highest prediction between violent and neutral classes at 0.88.

| Transfer Learning on YOLOv5m Trained From-Scratch Operations (random weights) | | | | | |
|---|--|-----------|-----------|--------|---------|
| # | Experiment | Model | Precision | Recall | mAP_0.5 |
| 1 | Exp18 No pre-processing/With Background Dataset Size 5944 images | | | | |
| | Exp-18 | Medium-FS | 0.66 | 0.51 | 0.53 |
| | | | | | |
| 2 | Exp15 pre-processing/With Background Dataset Size 6204 images | | | | |
| | Exp-15 | Medium-FS | 0.61 | 0.59 | 0.63 |
| | | | | | |
| 3 | Exp6 No pre-processing/No Background Dataset Size 2284 Images | | | | |
| | Exp-6 | Medium-FS | 0.64 | 0.68 | 0.67 |
| | | | | | |
| 4 | Exp3 pre-processing/No Background Dataset Size 5944 Images | | | | |
| | Exp-3 | Medium-FS | 0.67 | 0.70 | 0.72 |
| | | | | | |
| Transfer Learning on YOLOv5m Fine-tuned by COCO Dataset pre-trained Operations | | | | | |
| 5 | Exp12 No pre-processing/No Background Dataset Size 5944 images | | | | |
| | Exp-12 | Medium-PT | 0.77 | 0.72 | 0.75 |
| | | | | | |
| 6 | Exp24 No pre-processing/With Background Images Dataset Size 5944 images | | | | |
| | Exp-24 | Medium-PT | 0.84 | 0.74 | 0.79 |
| | | | | | |
| 7 | Exp9 pre-processing/No Background Dataset Size 5944 images | | | | |
| | Exp-9 | Medium-PT | 0.83 | 0.81 | 0.83 |
| | | | | | |
| 8 | Exp21 pre-processing/With Background Dataset Size 6204 images | | | | |
| | Exp-21 | Medium-PT | 0.85 | 0.82 | 0.85 |

Table 4.4: Appendix: YOLOv5m Activity Recognition Impact Results.

Following was discussions_ppl at 0.87, discussion_wgi at 0.77, and person at 0.62. In this instance, the From-Scratch models experienced challenges during the generalisation stages relative to low misclassification ratios in #2 experiment 15 compared to #1 experiment 18. The individual analysis proved the superiority of #2 experiment 15 over #1 experiment 18 by increasing the sample size and evaluating pre-processing versus no image enhancements. In #2 experiment 15, six classifications exceeded the 50th percentile ratio. The results on

neutral classes, discussion_wgi at 0.60, discussion_woboard at 0.88 like #2 experiment 15, discussions_ppl at 0.74, person at 0.55 with two violent types, knife_deployed at 0.67 and stabbing at 0.65. Both #1 experiment 18 and #2 fifteen generated high background image misclassification ratings. From-scratch approaches misinterpreted the true nature of 30 classes in #1 experiment 18 and 21 in #2 experiment 15. Though #1 experiment 18 precision was higher than #2 experiment 15 in Table 4.4, the confusion matrix validated the misclassification individually to fulfil research questions 1, 2 and 3 in 1.2.

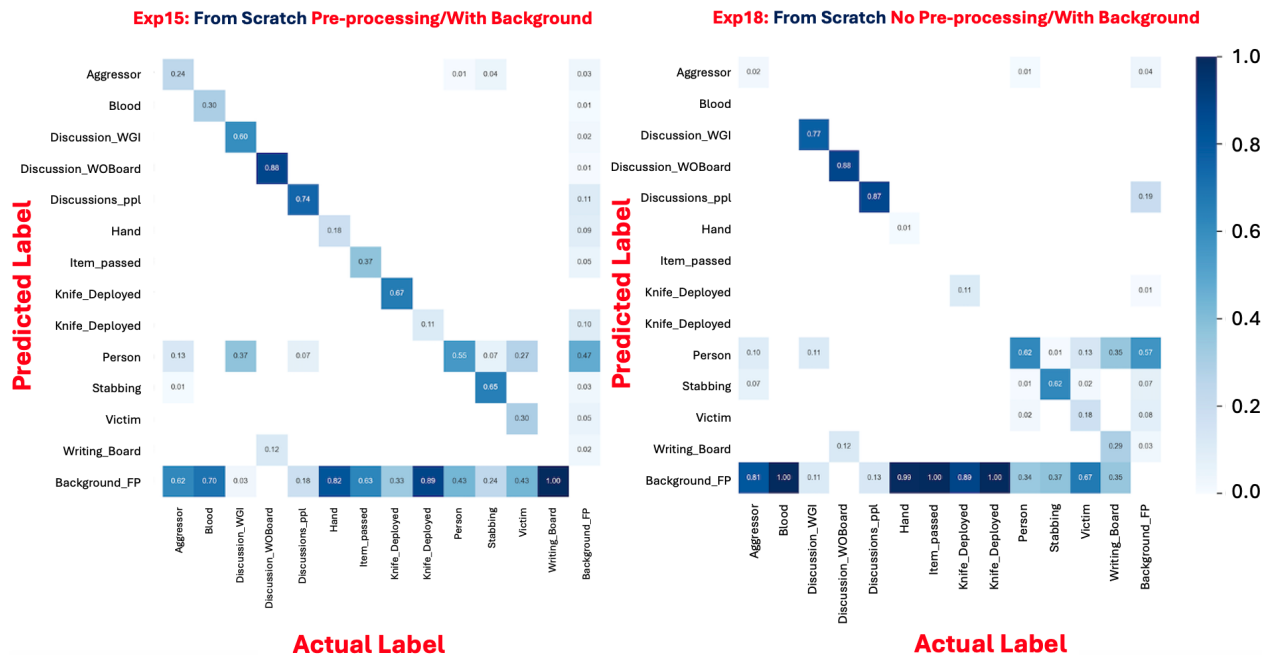


Figure 4.3: Appendix: YOLOv5m Confusion Matrix From-Scratch Exp. 1-2).

4.6.2 Appendix: YOLOv5m Exp.#1 From-Scratch Metric Results

Figure 4.4 accentuated performance for #1 experiment 18 via Table 4.4. The illustration emphasised the fluctuating precision metric at 0.66 and the recall value at 0.51 and an accuracy 0.53. The results validated Phase-1 analysis of From-Scratch misclassification in experiment 1.

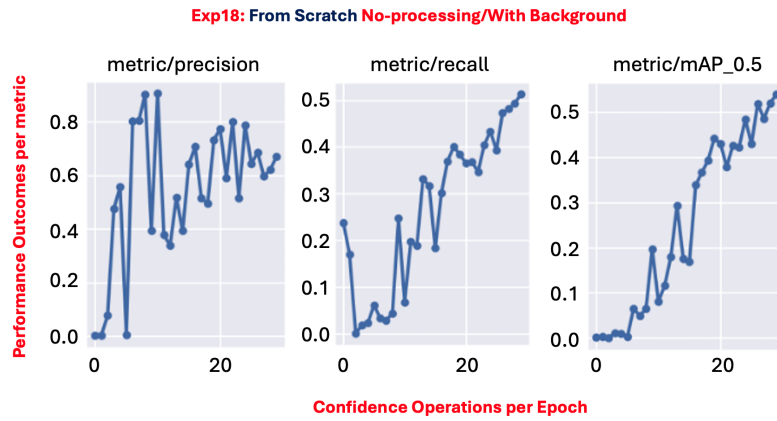


Figure 4.4: Appendix: YOLOv5m Impact Results for Exp. 18 From-Scratch.

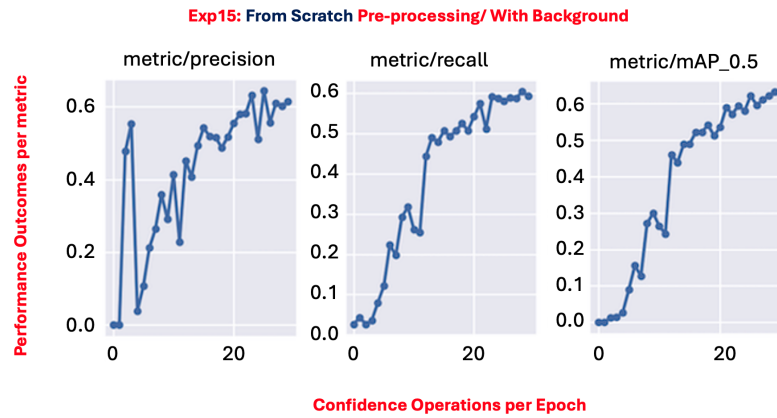


Figure 4.5: Appendix: YOLOv5m Impact Results for Experiment 15 From-Scratch.

4.6.3 Appendix: YOLOv5m Exp. #2 From-Scratch Metric Results

Figure 4.5 accentuated #2 experiment 15 performance via Table 4.4. The results displayed instability via the precision metric at 0.61, recall values at 0.59 and an accuracy at 0.63. The results validated Phase-1 analysis of From-Scratch confusion matrix in experiment 2.

4.6.4 Appendix: YOLOv5m Exp. #3 From-Scratch Metric Results

Figure 4.6 outlined #6 experiment performances via Table 4.4. The results displayed unstable precision at 0.64 and recall at 0.68, with an accuracy of 0.67, which validated Phase-1 From-Scratch confusion matrix in experiment 3.

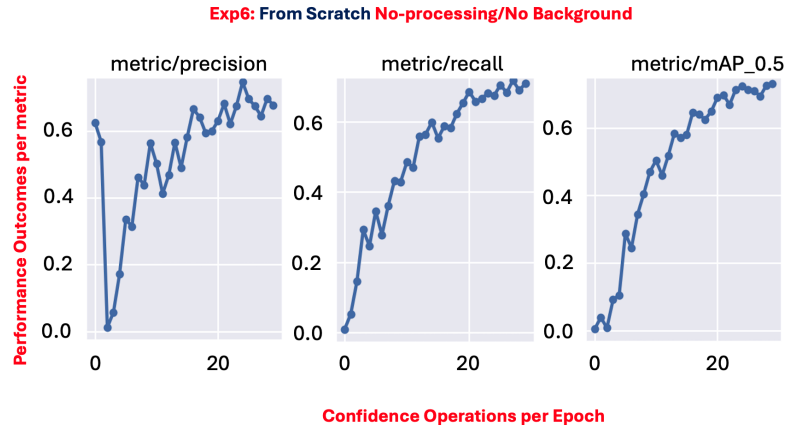


Figure 4.6: Appendix: YOLOv5m Impact Results for Experiment 6 From-Scratch.

4.6.5 Appendix: YOLOv5m Exp #4 Pre-Processing Metric Results

Figure 4.7 outlined #4 experiment 3 performance via Table 4.4. The previous results presented an overview of the precision metric at 0.67, recall at 0.70, with an accuracy of 0.72. The results validated Phase-1 From-Scratch classification analysis in experiment 3.

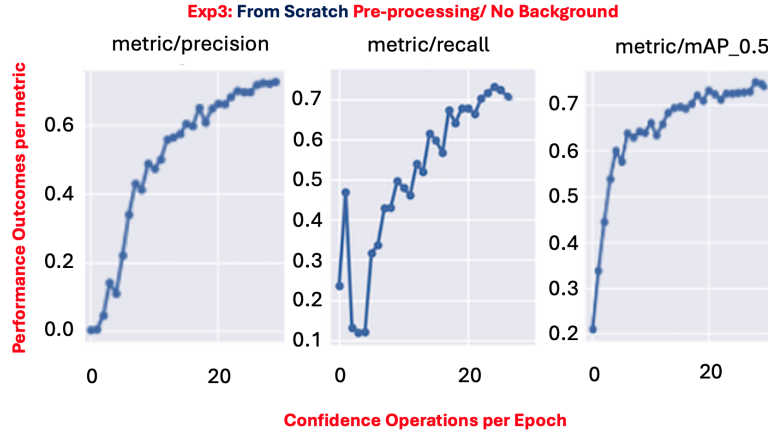


Figure 4.7: Appendix: YOLOv5m Impact Results for Experiment 4 From-Scratch.

4.6.6 Appendix: Phase-1 From-Scratch Analysis Exp. 3/4

The results demonstrated significant progress detailed in experiments #3 and #4 as the highest outcome dispensed in From-Scratch procedures, surpassing all other models. The model displayed a reduction in false negative and false positive output predictions, which exceeded the 60% accuracy threshold range with an overall accuracy of 0.72%. In #3 experiment 6 Table 4.4, analysis projected seven classifications above the 50% ratio with fewer background image misclassification ratings in Figure 4.8. Compared to experiments 1, 2 and 3, 4 ranked highest in performance with no image enhancements. The operations generated discussion_wgi at 0.97, discussion_woboard at 100%, discussions_ppl at 0.87, item_passed at 0.71, person at 0.72 and 1-violent class, stabbing at 0.72. In this instance, the model appeared unstable with signs of over-fitting utilising smaller datasets when evaluating #2 experiment 15. The evidence proved that increasing the sample size impacted the performance positively relative to analysis on experiment #4. Investigations on #4 revealed the highest From-Scratch ratings with pre-processing enhancements. The previous experimental approach validated pre-processing to foster performance improvements and fulfilled research questions-1-3, and, 5 in section 1.2.1. Analysis on experiment

#4 emphasized the fragility of From-Scratch approaches, and those operations misrepresented 28-classes compared to 21 in #3. Though the misclassification processing recorded higher ratings, #4 ranked higher in performance by generating eight accurate individual classifications above the 50th percentile. The operations predicted discussion_wgi 0.82, discussion_woboard 58%, discussions_ppl 0.82, item_passed 0.61, person 0.68 and 3-violent classes, blood at 0.54, knife_deployed 0.71 and stabbing at 0.56.

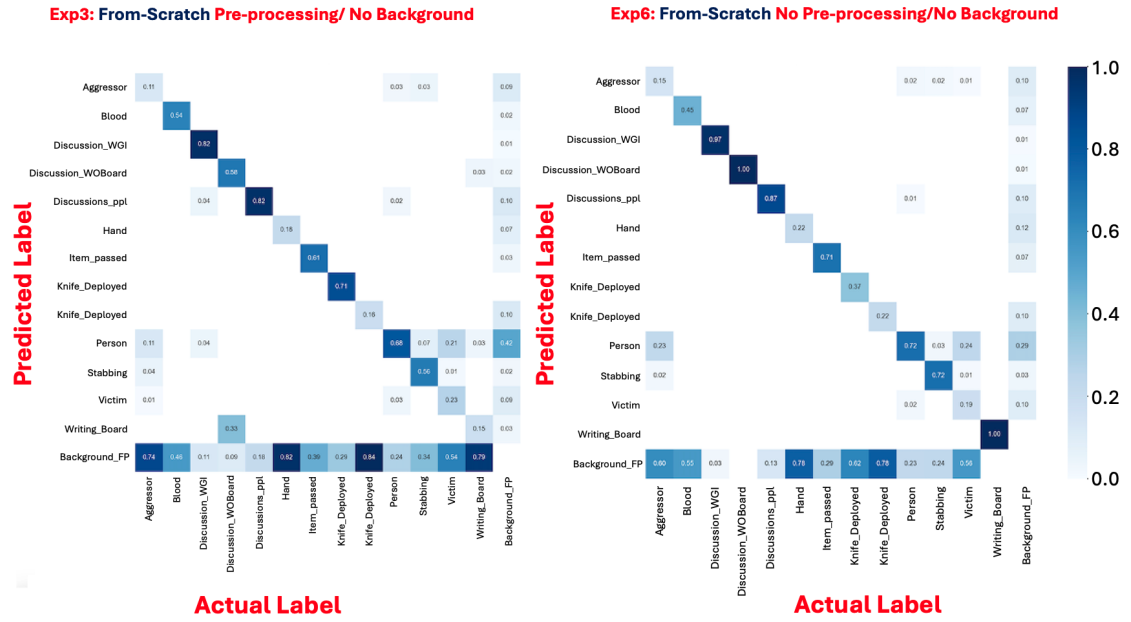


Figure 4.8: Appendix: YOLOv5m Confusion Matrix From-Scratch Exp. 3-4

4.6.7 Appendix: YOLOv5 Phase-1 Pre-trained Results Exp. 5(12)/6(24)

The model's performance improved significantly, thus exceeding the 70th percentile threshold range. Table 4.4 results suggested the pre-trained model's superiority over From-Scratch operations. Nevertheless, the findings projected substantial performance of the classification rates with stable precision and recall values. The analysis on #5 Exp. 12 projected 7-accurate predictions superseding all From-Scratch experiments with fewer back-

ground image misclassifications. In this instance, #5 Exp. 12 generated discussion_wgi 0.71, discussion_woboard 88%, discussions_ppl 0.80, person 0.70 as neutral classes and 2-violent classes concerning knife_deployed 100% and stabbing 0.64 in Figure 4.9. The operations confirmed research question-1 and 2, as they demonstrated positive results when no pre-processing and data increments were applied. The investigations validated impact using a significant volume of data relative to 2284 samples in #3 Exp. 6, compared to 5944 samples in #5 Exp. 12. Although the operations produced significant results, the pre-trained model demonstrated a deficiency by generating 28-class misrepresentations. The evaluations on #6 Exp. 24 disclosed an increment in performance using a similar dataset volume with 8-accurate predictions over #5 Exp. 12, as

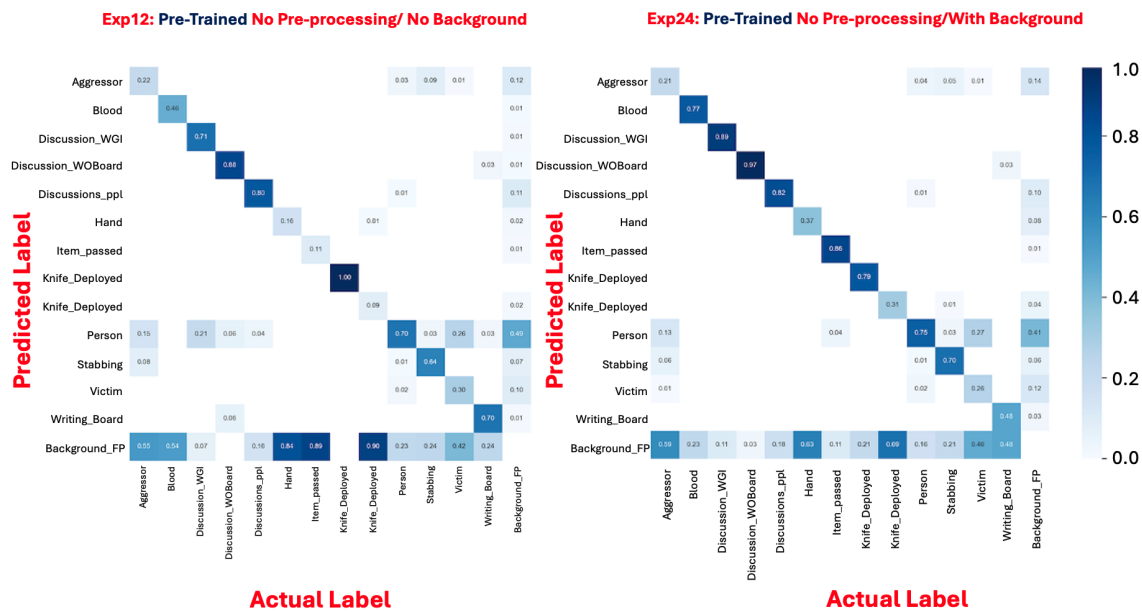


Figure 4.9: Appendix: YOLOv5m Confusion Matrix Pre-trained Exp. #5(12)/#6(24).

all predictions exceeded the 60th percentile with 27 misclassifications. The results disclosed classes relative to discussion_wgi 0.89, discussion_woboard 0.97, discussions_ppl 0.82, item_passed 0.86, the person 0.75. The evidence showed 3-violent classes, with an increase

in the stabbing class at 0.70, like #5 Exp. 12 at 0.64, blood0.77 and knife_deployed 0.79. The enhancement idea fulfilled research question-2 in section 1.2.1 as it positively affected the performance classifications state as demonstrated by lighter colour shades. The results signified that pre-trained experienced a positive enhancement by integrating background image support compared to From-Scratch approaches in #1 Exp. 18 and #2 Exp. 15 Table 4.4. The results proved the superiority of pre-trained models over all from-scratch methods, with a decrease in background false negative and false positive predictions indicative of the lighter-shaded blue.

4.6.8 Appendix: YOLOv5m No. 5 Exp.12 Pre-trained Metric Results

Figure 4.10 highlights ratings for #5 Exp. 12 concerning its performance via Table 4.4. The results showed the precision metric at 0.77 and the recall at 0.72, with an accuracy of 0.75. The results projected processing challenges, but provided a smoother graphical representation of the mAP and metrics.

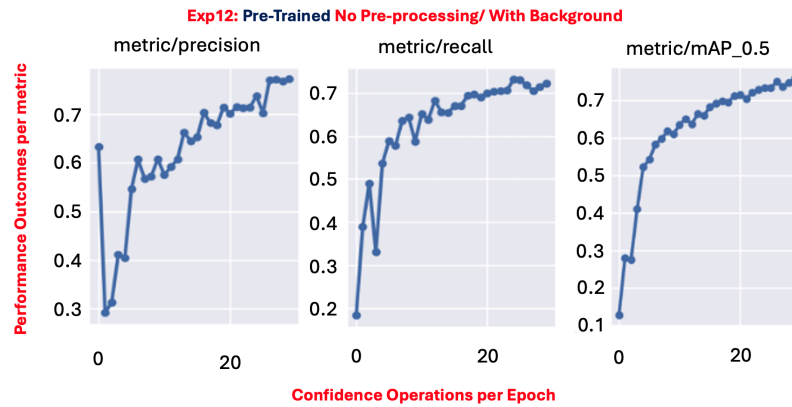


Figure 4.10: Appendix: YOLOv5m Impact Results for 5# Exp. 12 Pre-Trained.

4.6.9 Appendix: YOLOv5m No. 6 Exp. 24 Pre-trained Metric Results

Figure 4.11 projected ratings for #6 concerning its performance via Table 4.4. The evidence displayed the precision metric at 0.84 and the recall at 0.74, with an accuracy of 0.79. Like Experiment #5's processing, the operations showed similar characteristics with heavy false positive classification issues via precision.

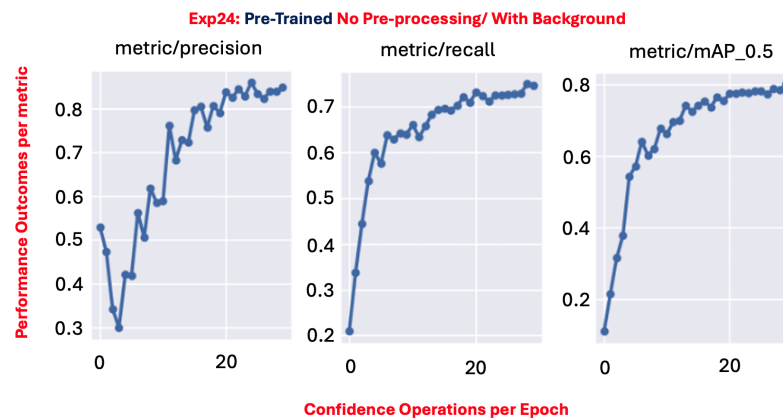


Figure 4.11: Appendix: YOLOv5m Impact Results for #6 Exp. 24 Pre-Trained.

4.6.10 Appendix: Phase-1 pre-trained Matrix Results Exp 7(9)/8(21)

The analysis projected the highest overall accuracy performance, superseding all pre-trained and From-Scratch experiments. The results validated YOLOv5m efficiency by utilising pre-trained operations in conditions containing pre-processing and background image supports. At this stage, pre-trained operations #7 and #8 via Figure 4.12 attained the highest performance ratings overall. Though the analysis on #7 disclosed high-performance ratings validating the metrics in Table 4.4, the confusion matrix projected signs of processing deficiencies. The findings proved 25-class misrepresentations with five accurate classifications exceeding the 60th percentile range in #7. The evidence fulfilled research question-2 in section 1.2.1 as stabbing attained the highest individual rating at

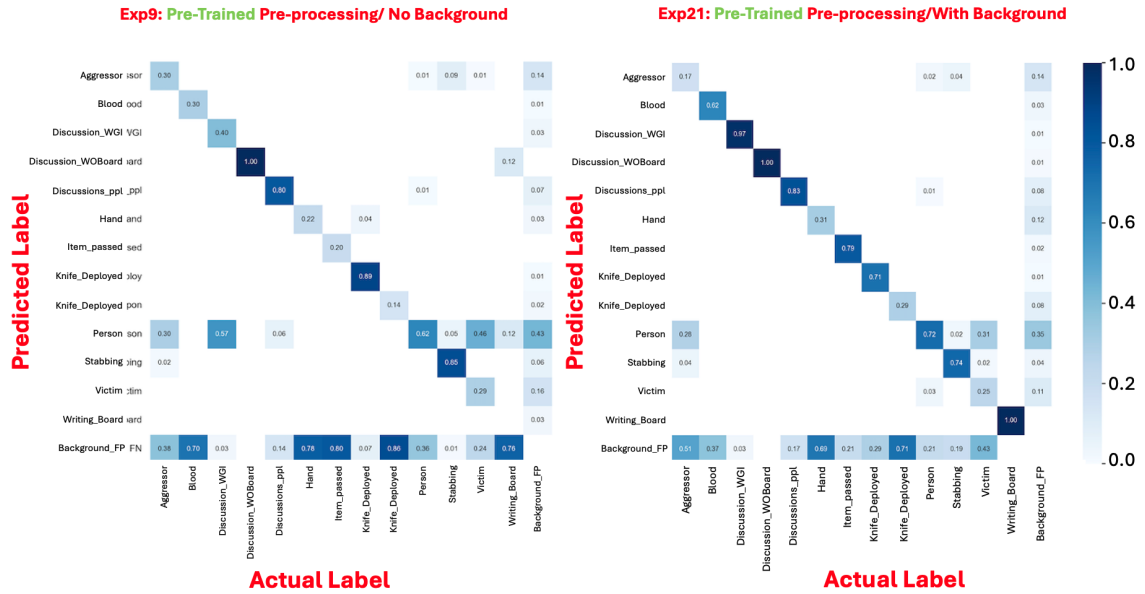


Figure 4.12: Appendix: YOLOv5m Confusion Matrix pre-trained Exp. #7(9)/#8(21).

0.85 overall. The application of background image enhancements impacted #7 negatively compared to #6 in Table 4.4. The predictions fulfilled the notions of research question-2 in section 1.2.1, proving that image enhancement is a critical component moving forward. The outcomes fortified the idea with discussion_woboard at 100%, discussions_ppl at 0.80, and person class 0.62%. The evidence showed that 2-violent categories, knife_deployed at 0.89 and stabbing at 0.85, as the highest rating overall, as a performance decline. The operations experienced a decline in performance with a reduction in accurate classifications. Because of fluctuating performances, the evidence signified that the model struggled to discern the true nature of violence. The analysis on #8 via Figure 4.12 disclosed the highest processing above the 60th percentile overall. The operations projected nine accurate classifications with reduced background false positives, false negatives and 20 misrepresentations. The outcomes displayed discussion_wgi at 0.97, discussion_woboard at 100%, discussions_ppl 0.83, item_passed 0.79, person 0.72. The 3-violent classes depicted stabbing decreasing to 0.74 compared to #7 0.85, blood 0.62 and knife_deployed 0.71. The evaluation proved that

YOLOv5m experienced stability issues processing stabbing in real-world conditions.

4.6.11 Appendix: YOLOv5m No. 7 Exp. 9 Metrics Results

Figure 4.13 represents #7's performance via Table 4.4. The results showed the precision metric at 0.83, recall at 0.81, an accuracy of 0.83, and more fluctuations compared to #6, proving its deficiency.

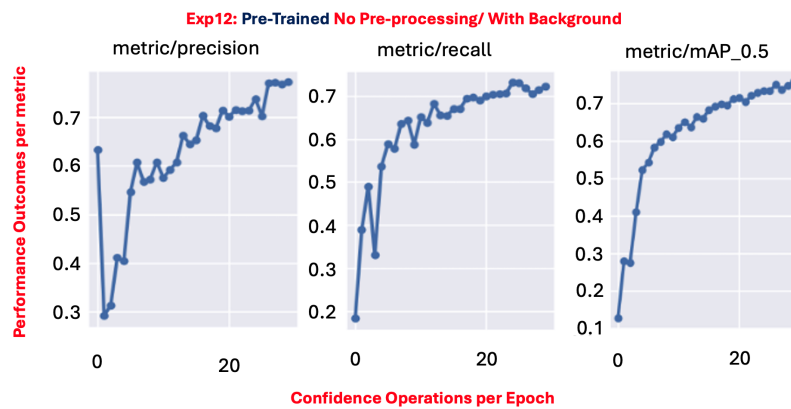


Figure 4.13: Appendix: YOLOv5m Impact Results for No. 7 Exp. 9 Pre-Trained.

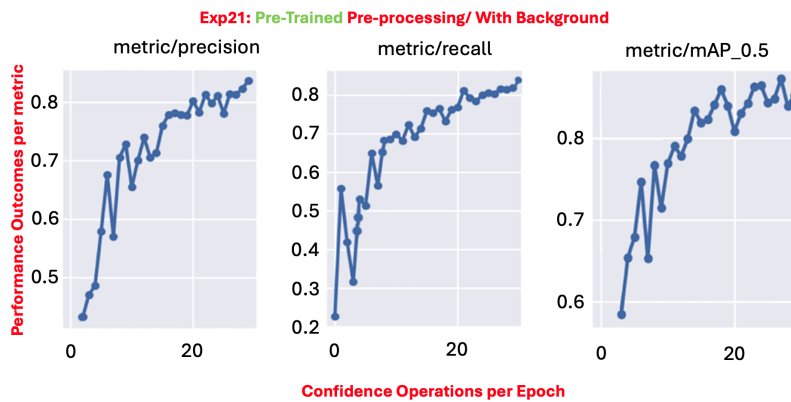


Figure 4.14: Appendix: YOLOv5m Impact Results for No. #8 Exp. 21 Pre-Trained.

4.6.12 Appendix: YOLOv5m No. #8 Exp. 21 Pre-trained Metric Results

Finally, Figure 4.14 showed performance for Exp.#8 via Table 4.4. The findings displayed the precision metric at 0.85 and recall at 0.82, with an accuracy of 0.85. The results projected a smoother curve, validating the classification notions specified in Phase-1's analysis of pre-trained results in #8 Exp. 21.

4.7 Appendix: Overview of 3DCNNsl Result, Analysis

A. Fulfilling Research Question-1 via 3DCNNsl Containing No pre-processing

The evidence proved that the model's performance fluctuations appeared predominantly in scenarios that intensified the complexity of violence, regardless of applying similar neutral classes during development. 3DCNNsl's results substantiated the classification proficiency on stabbing against other actions from an overall perspective. Fluctuating performances via precision and recall insinuated that the model experienced misclassification issues resulting in excessive false negatives and false positives. High individual accuracy scores for experiments 1-6 in Table 4.5 positively influenced the overall accuracy. 3DCNNsl's results are as follows.

4.7.1 Appendix: (A) Phase-2 Fulfilling Research Question 1 via 3DCNNsl No Pre-processing

In this instance, a projection of experiments 1-6 provides the context of no-processing data on 32 test samples to fortify the investigations.

4.7.2 Appendix: No. 1 Experiment 17 Results Overview on No pre-processing

Table 4.5 #1 represented an individual accuracy of 0.94, a precision of 0.75 and a recall at 0.38 for the Fighting(Fi) class. For Shooting(Sh), the individual category recorded 0.81, a precision of 0.57 and a recall of 0.50. The results projected 2-violent classes, which fulfilled research question-1's initiative in 1.2. 3DCNNsl dispensed an individual accuracy of Knit-

-ting(K) as 0.81, a precision of 0.67 and a recall of 0.75. Walk-with-Dog(W) recorded 0.69 with a precision of 0.58 and a recall value of 0.88. Though the individual accuracies were above the 60th percentile to the 90th range, the model's classification issues occurred because of fluctuating precision just above the 50th percentile with a recall under the 40th percentile range. Observations on 3DCNNsl's overall accuracy confirmed the false negative and false positive classification challenges by dispensing a score of 0.63. In this instance, the processing fulfilled research question-1 initiatives in section 1.2 with evidence of high individual scores concerning fighting and shooting.

| | | | | | | | | | | | | | |
|-----------|-----------|------------------|---------------|-----------|------------------|---------------|----------|------------------|---------------|----------|------------------|---------------|------------|
| #1Exp17 | Fi | Precision | Recall | Sh | Precision | Recall | K | Precision | Recall | W | Precision | Recall | ACC |
| | 0.94 | 0.75 | 0.38 | 0.81 | 0.57 | 0.50 | 0.81 | 0.67 | 0.75 | 0.69 | 0.58 | 0.88 | 0.63 |
| #2 Exp12 | B | | | Fi | | | K | | | W | | | |
| | 0.75 | 0.56 | 0.62 | 0.81 | 0.57 | 0.50 | 0.88 | 0.78 | 0.88 | 0.88 | 0.71 | 0.62 | 0.66 |
| #3 Exp26 | St | | | Fi | | | K | | | W | | | |
| | 0.69 | 0.38 | 0.38 | 0.94 | 0.75 | 0.38 | 0.75 | 0.67 | 100 | 0.94 | 0.88 | 0.88 | 0.66 |
| # 4 Exp14 | B | | | St | | | K | | | W | | | |
| | 0.81 | 0.73 | 100 | 0.88 | 0.33 | 0.12 | 0.81 | 0.67 | 0.75 | 0.94 | 0.89 | 100 | 0.72 |
| # 5 Exp23 | Sh | | | St | | | K | | | W | | | |
| | 0.81 | 0.62 | 0.62 | 0.94 | 0.75 | 0.38 | 0.88 | 0.78 | 0.88 | 0.81 | 0.73 | 100 | 0.72 |
| # 6 Exp20 | Sh | | | B | | | K | | | W | | | |
| | 0.94 | 0.80 | 0.50 | 0.94 | 0.89 | 100 | 0.88 | 0.67 | 0.50 | 0.75 | 0.67 | 100 | 0.75 |

Table 4.5: Appendix: (A) 3DCNNsl Impact Experiment with No Pre-processing.

4.7.3 Appendix: No.1 Exp.17 Projection of Accuracy No pre-processing

In this section, the confusion matrix operations in Figure 4.15 illustrates the overall performance in Table 4.6 and the actual classifications dispensing 12-critical false negative and 12-false positives. Figure 4.16 graph expressed processing challenges encountered via experiment #1. In this scenario, the expectation of the graphical curve should demonstrate a smooth representation of the output values to reflect processing stability. Nevertheless, the fluctuating

values validated the presence of misclassification during processing.

4.7.4 Appendix: No. 1 Experiment 17 Results Overview on No pre-processing

At this junction, Fighting(Fi) dispensed 0.81 as the highest individual score for violence with a precision of 0.57 and recall at 0.50. Beating(B) ranked second at 0.75 with a precision metric of 0.56 and a recall of 0.62. Knitting(K) and Walk-with-Dog(W) as neutral categories recorded 0.88 with fluctuating precision and recall metrics above the 60th percentile. The operations generated a higher accuracy of 0.66 because of the neutral class outputs and the violent classes in #1 experiment. Seventeen produced higher ratings proving that raw data with distinct categories and no pre-processing challenged the model’s discerning capability concatenating the fluctuating precision and recall scores.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Fighting | 3 | 23 | 1 | 5 | <ul style="list-style-type: none"> • TP Predictions Are Actually True |
| Shooting | 4 | 21 | 3 | 4 | <ul style="list-style-type: none"> • TN Preds NOT True |
| Knitting | 6 | 17 | 3 | 2 | <ul style="list-style-type: none"> • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 7 | 19 | 5 | 1 | <ul style="list-style-type: none"> • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.6: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 1 Exp.17.

#1 Exp17 No Pre-processing 32 test samples, Epoch 90, Batch Size 130

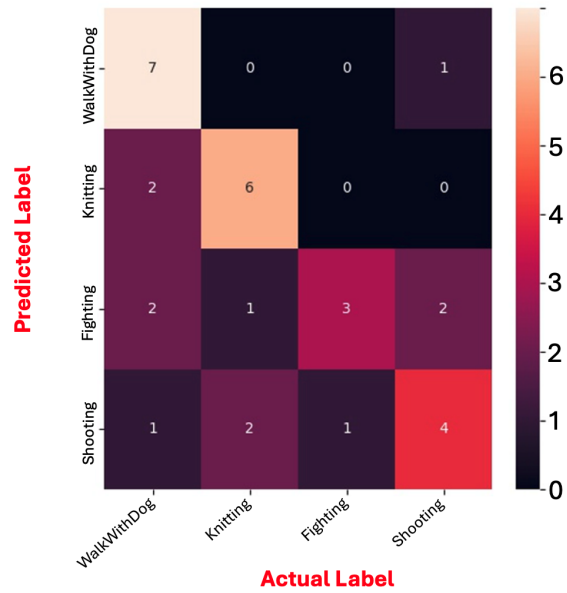


Figure 4.15: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 1 Exp.17.

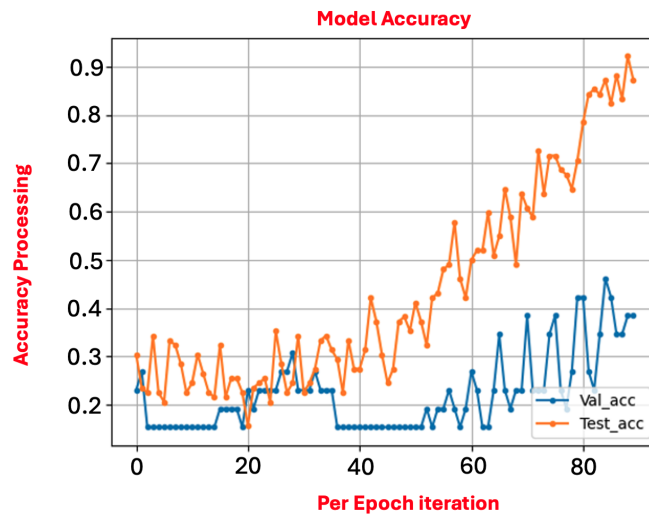


Figure 4.16: Appendix: No. 1 Experiment 17 Graphical View of No Pre-processing.

4.7.5 Appendix: No. 2 Exp.12 Confusion Matrix No Pre-processing Summary

Figure 4.17 confusion matrix illustrates the overall performance and the true classifications status per action in Table 4.7. The operations produced 11-false positive and 11-critical false negative predictions.

4.7.6 Appendix: No. 2 Exp.12 Projection of Accuracy No Pre-processing

Figure 4.18 graph expressed processing challenges via experiment #2 accuracy output. In this case, the expectation of the accuracy curve should be a smooth representation of the output values reflecting processing stability.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Beating | 5 | 20 | 4 | 3 | • TP Predictions Are Actually True |
| Fighting | 4 | 21 | 3 | 4 | • TN Preds NOT True |
| Knitting | 7 | 22 | 2 | 1 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 5 | 22 | 2 | 3 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.7: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 2 Exp.12.

4.7.7 Appendix: No. 2 Experiment 12 Results Overview on No pre-processing

Experiment #3 dispensed Fighting(Fi) at 0.94 as the highest individual accuracy score for violence with a precision of 0.75 and a lower recall at 0.38 compared to #2. Stabbing(St) produced 0.69 with a precision value similar to the fighting class and a low recall at 0.38. The neutral class Walk-with-Dog(W) produced 0.94 with a precision and recall of 0.88. Knitting(K) 's processing generated 0.75 with a precision of 0.67 and the highest recall at 100% compared to #1 and #2. The model's precision/recall fluctuations resulted in a similar overall accuracy of 0.66 compared to #2.

#2 Exp:12 No-Pre-processing 32 test samples, Epoch 90, Batch Size 130

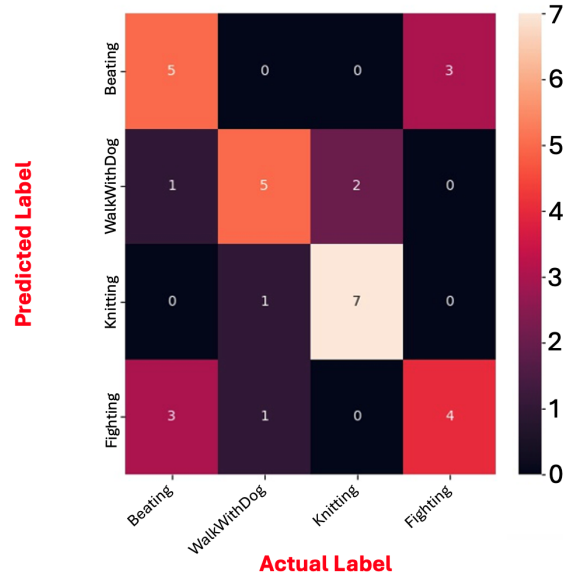


Figure 4.17: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 2 Exp.12.

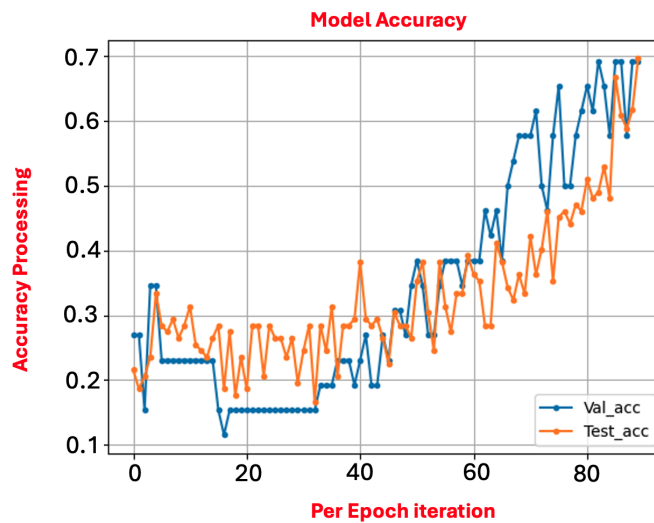


Figure 4.18: Appendix: No. 2 Experiment 12 Graphical Projection of No Pre-processing.

As anticipated, the overall accuracy category was negatively affected because the precision

reflected high false positives and recall values with multiple false negative classification outcomes. Nevertheless, individual performance recorded scores above the 60th percentile overall.

4.7.8 Appendix: No. 3 Exp.26 Confusion Matrix No Pre-processing Summary

In this section, Table 4.8 project the confusion matrix processing to illustrate the overall performance and the actual classification status per action. The operations validated the classification deficiency, producing 11 false positive and 11 critical false negative predictions.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Stabbing | 3 | 19 | 5 | 5 | • TP Predictions Are Actually True |
| Fighting | 3 | 23 | 1 | 5 | • TN Preds NOT True |
| Knitting | 8 | 20 | 4 | 0 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 7 | 23 | 1 | 1 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.8: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 3 Exp.26.

4.7.9 Appendix: No. 3 Exp.26 Projection of Accuracy No Pre-processing

Figure 4.19 graph demonstrated several processing challenges via experiment #3 accuracy output. In this case, the expectation of the accuracy curve should be a smooth representation of the output values; however, the blue line colour representing the validation operation appeared closer to the accuracy output. Though the graphical output at this stage demonstrated classification deficiencies, the results insinuated that the model is gradually improving its classification state.

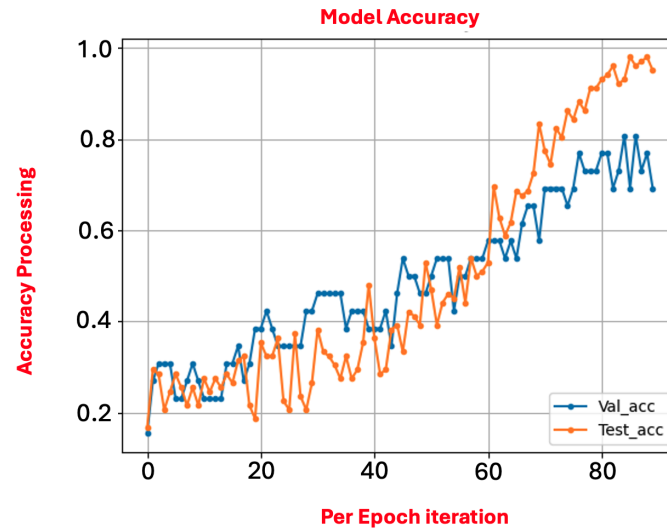


Figure 4.19: Appendix: No. 3 Exp.26 Graphical Projection of No Pre-processing.

4.7.10 Appendix: No. 4 Exp.14 Results on No Pre-processing

The individual classes recorded scores above the 80th percentile range, insinuating that category variations increased the model's processing complexity via #3 Table 4.5. Beating and stabbing recorded distinct attributes with higher ratings. The evidence disclosed Stabbing(St) at 0.88 as the highest output at this stage, with a low precision of 0.33 and recall of 0.12 compared to the previous experiments. Beating(B) generated an individual score of 0.81, a precision of 0.73 and a recall at 100 as evidence of a performance improvement. Regarding neutral classes, Walk-with-Dog(W) produced 0.94 with a precision score of 0.89 and a recall of 100 over knitting(K). Knitting(K) improved at 0.81 with a precision of 0.67 and a lower recall at 0.75 compared to #3 Exp. 26 recall at 100. The operations produced a higher overall accuracy above the 70th percentile at this level. Nevertheless, the operations demonstrated signs of

classification issues when evaluating the stability of the precision and the recall values.

4.7.11 Appendix: No. 4 Exp.14 Matrix No Pre-processing Summary

At this junction, the confusion matrix via Figure 4.20 emphasised the overall performance and classification status of Table 4.9 ’s classes. The operations showed signs of improvement. However, the misclassification results via stabbing’s output at one and beating eight confirmed the model’s limitations. 3DCNNsl produced 11 false positive and 11 critical false negative predictions overall.

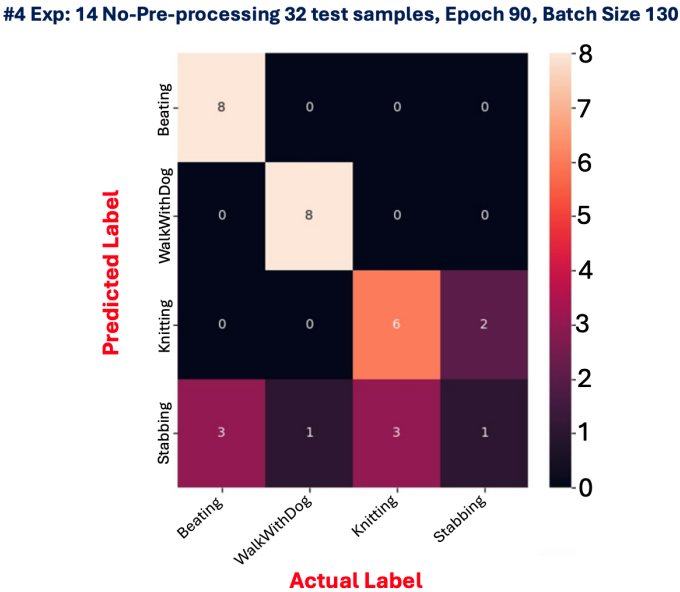


Figure 4.20: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 4 Exp.14.

4.7.12 Appendix: No. 4 Exp.14 Projection of Accuracy No Pre-processing

Figure 4.21 graph expressed 3DCNNsl's misclassification limitations via experiment #4 with fluctuating accuracy outputs. In this case, the expectation of the accuracy curve should be a smooth representation of the output values; however, the validation operation in blue is closer to the accuracy as it gradually tries to project the desired result. The graphical output insinuated that the model is steadily improving its classification state.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Beating | 8 | 21 | 3 | 0 | <ul style="list-style-type: none"> • TP Predictions Are Actually True |
| Stabbing | 1 | 22 | 2 | 7 | <ul style="list-style-type: none"> • TN Preds NOT True |
| Knitting | 6 | 21 | 3 | 2 | <ul style="list-style-type: none"> • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 8 | 23 | 1 | 0 | <ul style="list-style-type: none"> • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.9: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 4 Exp.14.

4.7.13 Appendix: No. 5 Exp.23 Results on No Pre-processing

At this level, individual accuracy recorded scores above the 80th percentile threshold. Those ratings proved that violence containing sections of fighting and beating can increase 3DCNNsl's learning complexity. The findings project Stabbing(St) at 0.94 as the highest output, with a higher precision of 0.75 and recall at 0.38 compared to #4 Table 4.5. Shooting(Sh) generated an individual score of 0.81 with a lower precision metric and recall of 0.62 compared to #4. The neutral Knitting(K) class increased to 0.88 with a higher precision at 0.78 and a recall at 0.88. Though the

knitting score superseded Walk-with-Dog(W), the classification showed stability issues with unstable precision and recall values. The Walk-with-Dog(W) neutral class produced an individual score of 0.81 with a precision of 0.73 and a recall of 100. The operations generated a similar result compared to #4 Table 4.4 overall accuracy score.

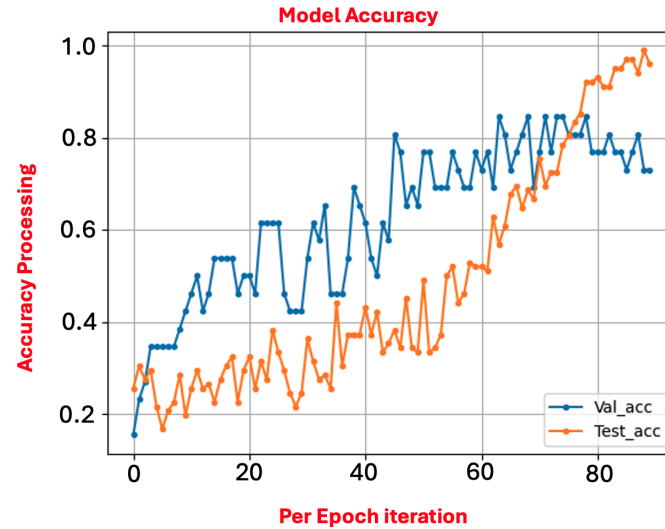


Figure 4.21: Appendix: No. 4 Exp.14 Graphical Projection of No Pre-processing.

4.7.14 Appendix: No. 5 Exp.23 Matrix No Pre-processing Summary

In this section, Figure 4.22 confusion matrix show the overall performance and the actual classification status of actions in Table 4.10. The operations demonstrated classification deficiency concerning shooting five and stabbing's output at 3. In this instance, 3DCNNsl produced nine false positive and nine critical false negative outcomes.

#5 Exp: 23 No-Pre-processing 32 test samples, Epoch 90, Batch Size 130

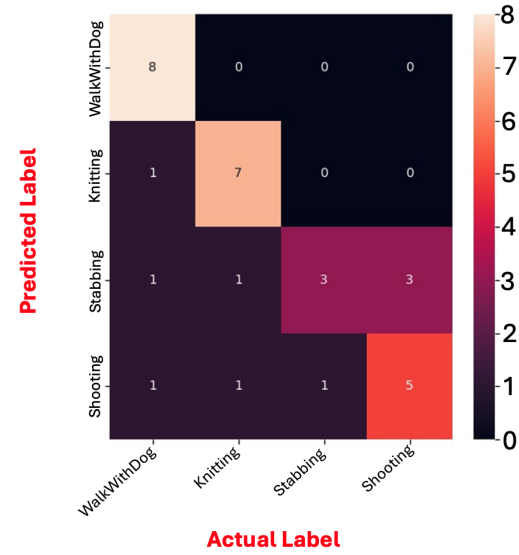


Figure 4.22: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 5 Exp.23.

4.7.15 Appendix: No. 5 Exp.23 Projection of Accuracy No Pre-processing

Figure 4.23 graph expressed 3DCNNsl's misclassification limitations via experiment #5 with fluctuating validation and accuracy scores. The expected graphical outline should reflect a gradual projection of the output values. The validation operation in blue is closer to the accuracy as it gradually tries to dispense the desired result. The graphical

output insinuated that the model is continuously improving its classification state.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Shooting | 5 | 21 | 3 | 5 | • TP Predictions Are Actually True |
| Stabbing | 3 | 21 | 1 | 5 | • TN Preds NOT True |
| Knitting | 7 | 22 | 2 | 1 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 8 | 21 | 3 | 0 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.10: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 5 Exp.23.

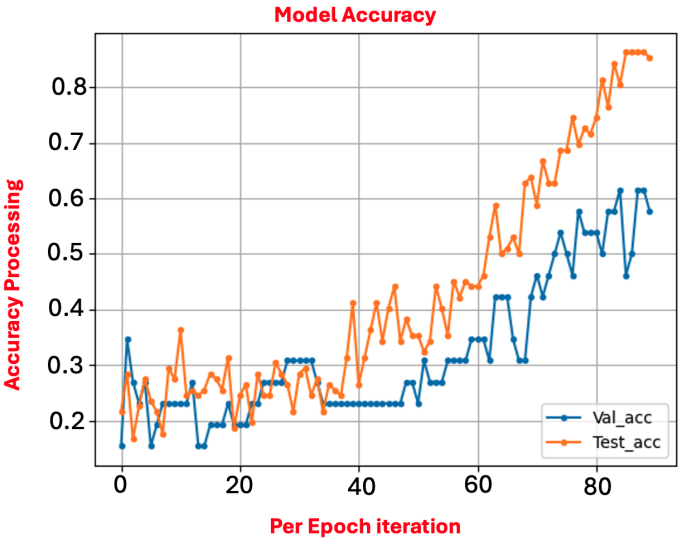


Figure 4.23: Appendix: No. 5 Exp.3 Graphical Projection of No Pre-processing.

4.7.16 Appendix: No. 6 Exp.20 Results Overview on No Pre-processing

The model generated the highest overall accuracy score of 0.75 for all experiments containing no pre-processing enhancements. The Shooting(Sh) individual score increased to 0.94 with a precision of 0.80 and a depreciated recall of 0.50. The data depicted similar Beating(B) scores with an increased precision score of 0.89 and a recall of 100 for violence. The neutral Knitting(K) class remained at 0.88, with a lower precision at 0.67 and recall at 0.50. Walk-with-dog(W) generated a lower score of 0.75 with a precision metric of 0.67 and a recall of 100.

4.7.17 Appendix: No. 6 Exp.20 Matrix No Pre-processing Summary

Figure 4.24 confusion matrix illustrated Table 4.11 overall performance and true classification status per action. The operations demonstrated classification deficiency relative to shooting 4 with an improvement in beating's output at eight. In this instance, the model displayed improvement as it produced eight false positive and eight critical false negative predictions overall.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Shooting | 4 | 23 | 1 | 4 | • TP Predictions Are Actually True |
| Beating | 3 | 21 | 1 | 5 | • TN Preds NOT True |
| Knitting | 4 | 22 | 2 | 4 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 8 | 20 | 4 | 0 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.11: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 6 Exp.20.

#6 Exp: 20 No-Pre-processing 32 test samples, Epoch 90, Batch Size 130

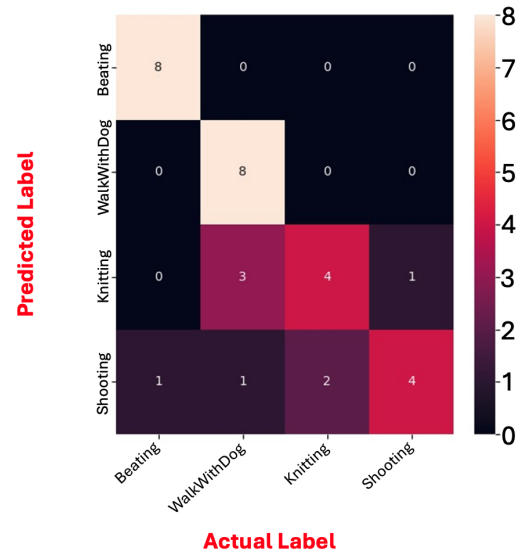


Figure 4.24: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 6 Exp.20.

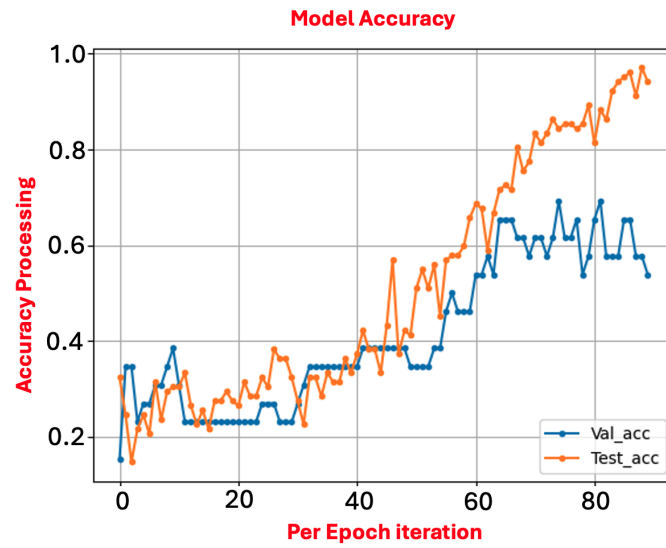


Figure 4.25: Appendix: No. 6 Exp.20 Graphical Projection of No Pre-processing.

4.7.18 Appendix: No. 6 Exp.20 Projection of Accuracy No Pre-processing

Figure 4.25 graph expressed 3DCNNsl’s misclassification limitation via experiment #6 with fluctuating validation and accuracy score outcomes. The expected accuracy curve outline should reflect a smooth representation of the output. The validation operation in blue is closer to the accuracy as it struggled to project the desired results. The graphical output insinuated that the model continuously improves its prediction at every stage.

| | Sh | Precision | Recall | St | Precision | Recall | K | Precision | Recall | W | Precision | Recall | ACC |
|-----------|-----------|-----------|--------|-----------|-----------|--------|----------|-----------|--------|----------|-----------|--------|------------|
| #1 Exp 56 | 0.75 | 0.60 | 0.75 | 0.81 | 0.62 | 0.62 | 0.94 | 0.86 | 0.75 | 0.94 | 0.86 | 0.75 | 0.72 |
| #2 Exp 59 | St | | | Fi | | | K | | | W | | | |
| | 0.81 | 0.70 | 0.88 | 0.88 | 0.60 | 0.38 | 0.94 | 0.88 | 0.88 | 0.81 | 0.67 | 0.75 | 0.72 |
| #3 Exp 50 | Fi | | | Sh | | | K | | | W | | | |
| | 100 | 100 | 0.25 | 0.94 | 0.88 | 0.88 | 0.94 | 0.89 | 100 | 0.69 | 0.62 | 100 | 0.78 |
| #4 Exp 53 | Sh | | | B | | | K | | | W | | | |
| | 0.88 | 0.71 | 0.62 | 0.88 | 0.75 | 0.67 | 0.88 | 0.80 | 100 | 0.94 | 0.88 | 0.88 | 0.79 |
| #5 Exp 47 | B | | | St | | | K | | | W | | | |
| | 0.81 | 0.70 | 0.88 | 0.88 | 0.71 | 0.62 | 100 | 100 | 0.88 | 0.94 | 0.88 | 0.88 | 0.81 |
| #6 Exp 44 | B | | | Fi | | | K | | | W | | | |
| | 100 | 100 | 0.88 | 0.94 | 0.86 | 0.75 | 0.94 | 0.88 | 0.88 | 0.88 | 0.80 | 100 | 0.88 |

Table 4.12: Appendix: (B) 3DCNNsl Experiment with Pre-processing Impact Results.

B. Fulfilling Research Question-2/3 Via 3DCNNsl Containing pre-processing: The results demonstrated

the contrast between operations including no pre-processing and processes that did. Table 4.12 **group B** disclosed performance superiority as a critical measure towards the proposed fusion developmental stages. The analysis projected via experiments 1-6 in the context of pre-processed 32-test samples fortified the research objectives. Operations at this level provided the opportunity to formulate comparisons using raw data and data containing pre-processing enhancements to foster robust results. The simulation is as follows.

4.7.19 Appendix: No. 1 Exp.56 Results Overview on Pre-processing

In this group, the findings projected Stabbing(St) individual score at 0.81 as the highest violence class with a precision and recall at 0.62 in #1 **Section-B Table 4.12**. Shooting(Sh) generated 0.75 with a precision metric of 0.60 and a recall of 0.75. In this instance, the model's performance in the neutral category superseded the violent classes. Knitting(K) and Walk-with-Dog(W) neutral (non-violent) classes generated scores of 0.94, with a precision of 0.86 and a recall of 0.75. Because of the high-performance ratings, the overall accuracy dispensed a higher score of 0.75 compared to #1, 0.63. The output insinuated that 3DCNNsl produced fewer misrepresentations even though the individual precision and recall fluctuated above the 60th percentile. The results projected evidence that satisfied research questions 1 and 2 when evaluating the overall accuracy score of #1.

4.7.20 Appendix: No. 1 Exp.56 Confusion Matrix Pre-processing Summary

Figure 4.26 confusion matrix illustrates the overall performance and the actual classifications in Table 4.13 per action. The operations showed processing issues for shooting six, improving with beating at 5. The model improved by nine false positive and nine false negative predictions overall. The scores exceeded 70% but produced lower ratings compared to # 5 no pre-processing in **Section B Table 4.12**.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Shooting | 6 | 20 | 4 | 2 | • TP Predictions Are Actually True |
| Stabbing | 5 | 21 | 3 | 3 | • TN Preds NOT True |
| Knitting | 6 | 12 | 1 | 2 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 6 | 23 | 1 | 2 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.13: Appendix: 3DCNNsl Confusion Matrix Pre-processing #1 Exp.56.

#1 Exp: 56 Pre-processing 32 test samples, Epoch 90, Batch Size 130

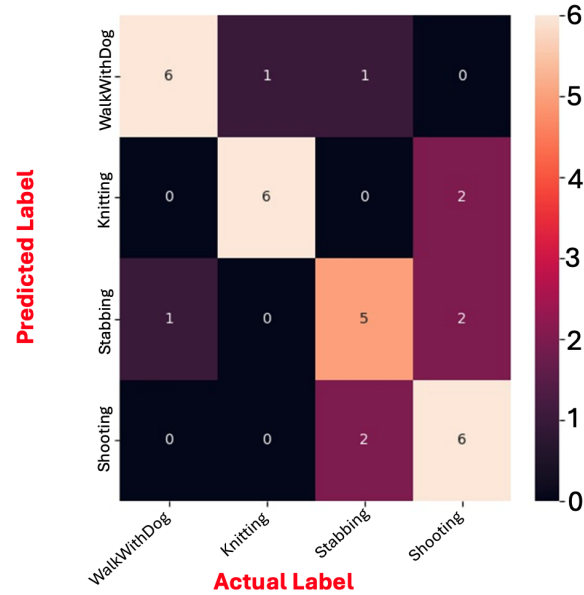


Figure 4.26: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 1 Exp.56.

4.7.21 Appendix: No. 1 Exp.56 Projection of Accuracy Pre-processing

Figure 4.27 graph expressed 3DCNNsl misclassification limitation via experiment #1 in **Section B Table 4.12** with fluctuating validation and accuracy scores. The expected accuracy curve outline must reflect a gradual descending representation of the output values. In this case, 3DCNNsl struggled to generate significant results. The graphical output insinuated that the model continuously improved its classification state at every stage but suffered from high misclassification ratings.

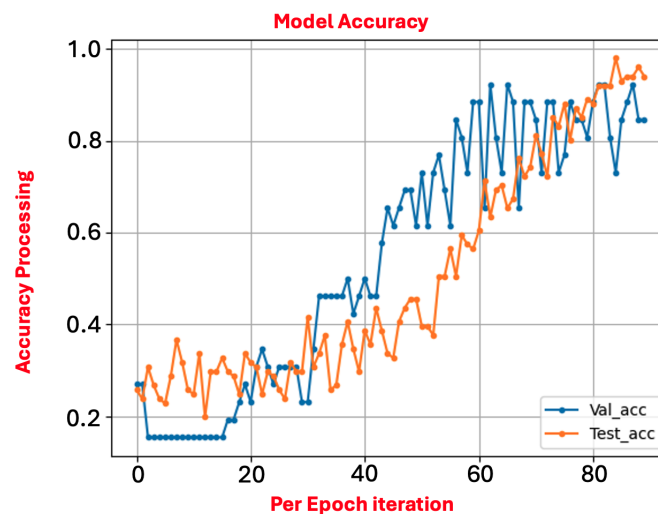


Figure 4.27: Appendix: No. 1 Experiment 56 Graphical Projection of Pre-processing.

4.7.22 Appendix: No. 2 Exp.59 Results Overview on Pre-processing

Scores on violence increased compared to #1 in **Section B Table 4.12** for stabbing and fighting. Stabbing(St) generated 0.81 with a precision of 0.70 and a recall of 0.88. Knitting generated 0.94 with higher precision and recall of 0.88 compared to #1. Walk-with-Dog(W) created a recall of 0.75 with a lower individual score of 0.81 and a precision of 0.67. The operation generated similarity in #1 via **Section B Table 4.12** performance. The outcome occurred due to similar characteristics between stabbing and fighting.

4.7.23 Appendix: No. 2 Exp.59 Confusion Matrix Pre-processing Summary

Figure 4.28 confusion matrix illustrates overall performance and the actual classifications in Table 4.14 results. The operations demonstrated classification deficiency concerning stabbing at 7 with a reduced fighting output at 3. 3DCNNsl displayed similar results compared to experiment #1 in **Section B Table 4.12**. That operation produced nine false positive and nine critical false negative predictions. The stabbing classifications showed no improvement due to precision and recall fluctuations. Analysis proved that the complexity of pre-processed approaches produced weaker outcomes for violent classes. The processing performance declined for the neutral classes with high misclassification similarities.

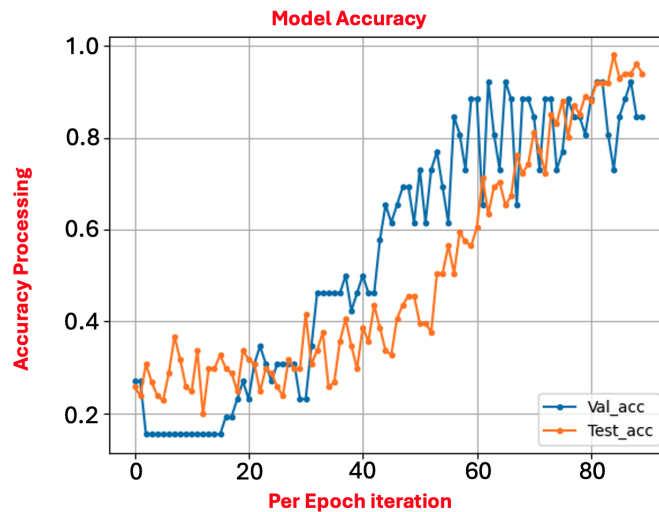


Figure 4.28: Appendix: No. 2 Exp.59 Graphical Projection of Pre-processing

4.7.24 Appendix: No. 2 Exp.59 Projection of Accuracy Pre-processing

Figure 4.29 expressed 3DCNNsl's graphical misclassification limitations via experiment #2 in **Section B Table 4.12**. Observation proved that 3DCNNsl struggled to generate significant representations of output values. The graphical output insinuated that the model suffered from high misclassification regarding fluctuating precision and recall scores.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Stabbing | 7 | 21 | 3 | 1 | • TP Predictions Are Actually True |
| Fighting | 3 | 22 | 2 | 5 | • TN Preds NOT True |
| Knitting | 7 | 23 | 1 | 1 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 6 | 21 | 3 | 2 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.14: Appendix: 3DCNNsl Confusion Matrix Pre-processing No. 2 Exp. 59

4.7.25 Appendix: No. 2 Exp.59 Results Overview on Pre-processing

Violence scores increased because of the distinction in pre-processing between Fighting(Fi) and Shooting(Sh). Fighting generated 100 with a precision of 100 and a recall of 0.25. Shooting scored higher with 0.94 compared to #1 in **Section B Table 4.12**, with higher precision and recall at 0.88. Knitting(K) produced a similar output of 0.94 compared to #1, with higher precision at 0.89 and recall at 100%. Walk-with-Dog(W) produced a lower score of 0.69 with a precision of 0.62 and a recall of 100. The results insinuated that pre-processing produced several false negative outputs linked to the fluctuating precision and recall. Nevertheless, the fluctuation increased, recording an overall accuracy score of 0.78, substantiating the pre-processing method's superiority over the no-processing approach.

#2 Exp: 59 Pre-processing 32 test samples, Epoch 90, Batch Size 130

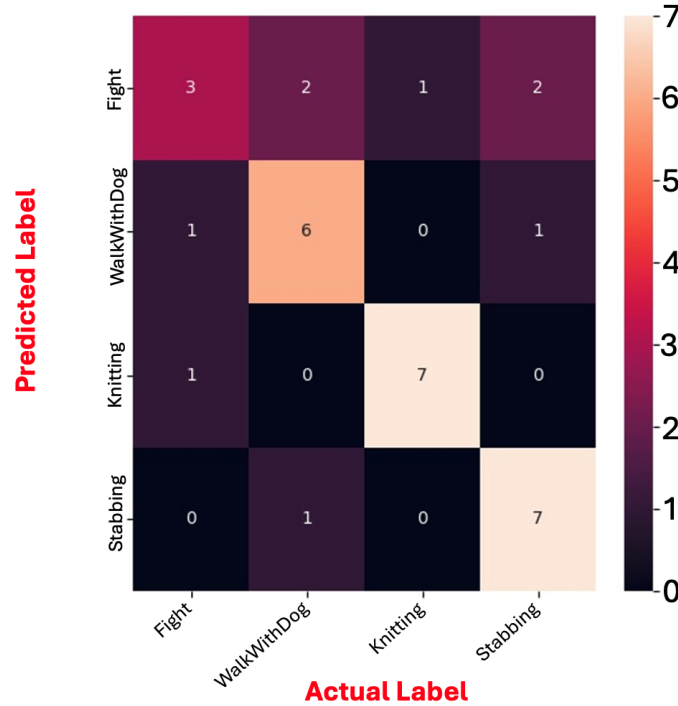


Figure 4.29: Appendix: 3DCNNsl Confusion Matrix No Pre-processing No. 2 Exp.59.

4.7.26 Appendix: No. 3 Exp.50 Confusion Matrix Pre-processing Summary

Figure 4.30 illustrates the confusion matrix performance and the actual classification status in Table 4.15 actions. The operations demonstrated classification improvement on fighting at 100% and shooting at 0.94. The model displayed higher performance compared to experiment # 2 in **Section B Table 4.12**. The operations produced seven false positive and seven critical false negative predictions overall. In this instance, the model improved its classification performance on the violent and neutral knitting classes. 3DCNNsl increased its misclassification for fighting via the recall and walk-with-dog; however, overall performance increased to 78% compared to 72% in #2.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Fighting | 2 | 24 | 0 | 6 | • TP Predictions Are Actually True |
| Shooting | 7 | 23 | 1 | 1 | • TN Preds NOT True |
| Knitting | 8 | 23 | 1 | 0 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 8 | 19 | 5 | 0 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.15: Appendix: 3DCNNsl Confusion Matrix Pre-processing No. 3 Exp.50.

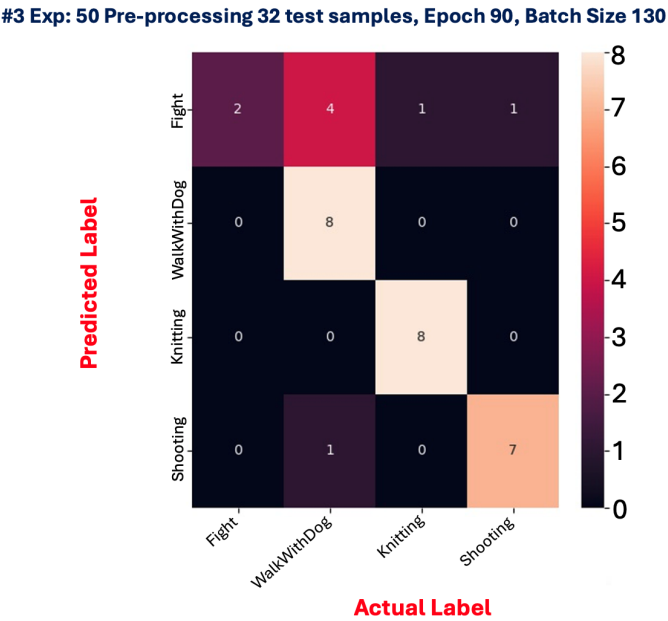


Figure 4.30: Appendix: 3DCNNsl Confusion Matrix Pre-processing No. 3 Exp.50.

4.7.27 Appendix: No. 3 Exp.50 Projection of Accuracy Pre-processing

Figure 4.31 expressed high misclassification limitation with intense validation and accuracy responses. The outcome occurred because of the dataset's size, negatively impacting 3DCNNsl operations. The model experienced limited computational outputs to generate high-performance results. Graphical analysis insinuated that the model suffered from high misclassification ratings because of the fluctuating precision and recall scores.

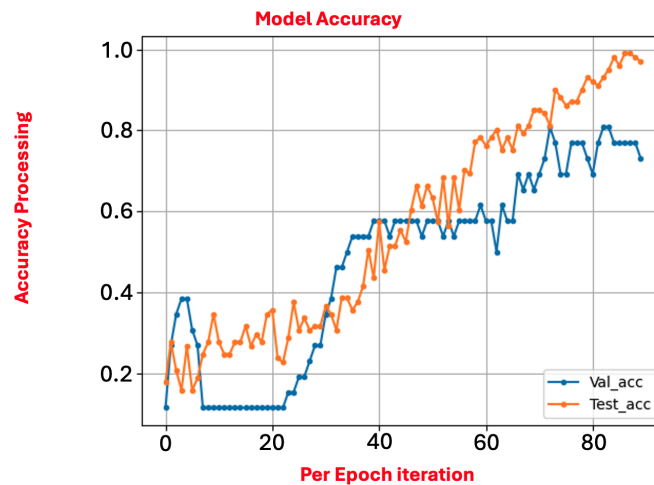


Figure 4.31: Appendix: No. 3 Exp.50 Graphical Projection of Pre-processing.

4.7.28 Appendix: No. 4 Exp.53 Results Overview on Pre-processing

3DCNNsl produced a higher overall accuracy score of 0.79 with a decrease in some individual scores but stabilised its precision and recall scores above the 60th percentile range. Shooting(Sh), Beating(B), Knitting(K) and Walk-with-Dog(W)'s precision and recall generated a similar individual score of 0.88. Precision recorded 0.71 with a recall of 0.62 for Stabbing(St). Walk-with-Dog(W) produced 0.94, Beating(B) generated a precision of 0.75 with a recall of 0.67 and Knitting(K) recorded a precision of 0.80 and a recall of 100%. Though individual scores were slightly lower compared to #3 in **Section B Table 4.12**,

analysis of precision and recall representations suggested that the 3DCNNsl classification state improved utilising pre-processing with distinct classes.

4.7.29 Appendix: No. 4 Exp.53 Confusion Matrix Pre-processing Summary

Figure 4.32 illustrates the confusion matrix overall performance in Table 4.16 and the classification status per class. The operations decreased its classification effectiveness to 0.88 for shooting, beating as violence and neutral class knitting. 3DCNNsl experienced a decline in performance compared to experiment # 3 in **Section B Table 4.12**. The occurrence produced seven false positive and six critical false negative predictions overall. Though 3DCNNsl depreciated in performance, the overall accuracy climaxed at 0.79. The unstable evaluation metrics proved that 3DCNNsl experienced a classification limitation in both action categories (non-violent/violent).

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Shooting | 5 | 23 | 2 | 2 | • TP Predictions Are Actually True |
| Shooting | 6 | 22 | 2 | 3 | • TN Preds NOT True |
| Knitting | 8 | 23 | 2 | 0 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 7 | 24 | 1 | 1 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.16: Appendix: 3DCNNsl Confusion Matrix Pre-processing No. 4 Exp. 53.

4.7.30 Appendix: No. 4 Exp.53 Projection of Accuracy Pre-processing

Figure 4.33 graphically expressed high misrepresentation because of the fluctuating validation and accuracy outcomes. The processing occurred because of the dataset size negatively impacting 3DCNNsl operations. The previously mentioned issues limited the computational output of values to reflect high-performance results. The graphical out-

-put insinuated that the model suffered high misclassification ratings due to fluctuating precision and recall scores.

#4 Exp: 53 Pre-processing 32 test samples, Epoch 90, Batch Size 130

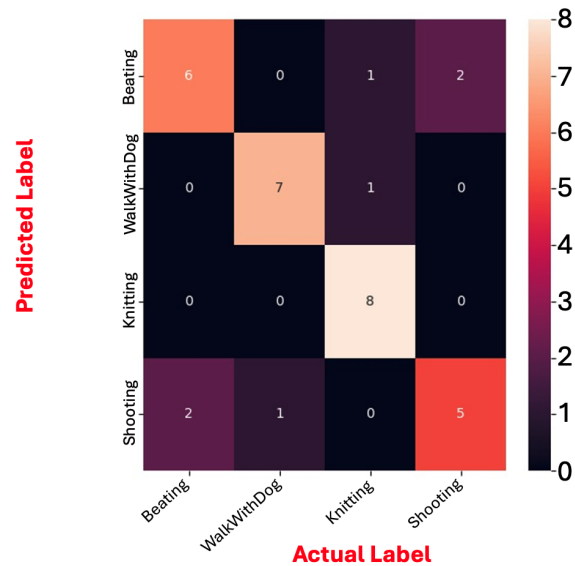


Figure 4.32: Appendix: No. 4 Exp.53 Graphical Projection of Pre-processing.

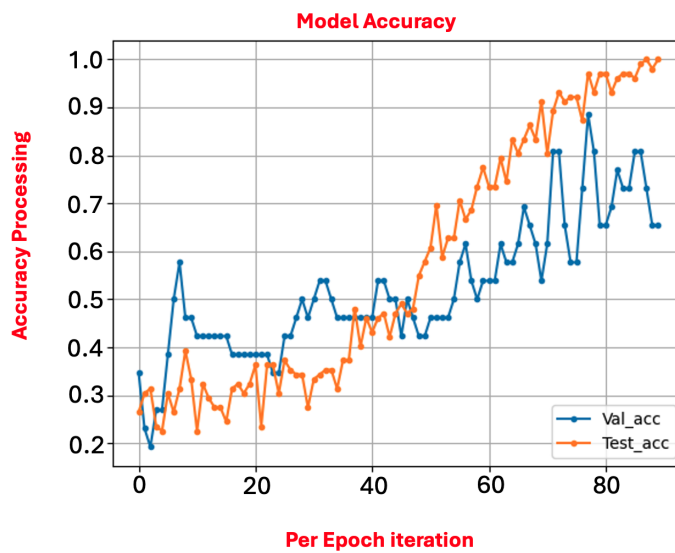


Figure 4.33: Appendix: No. 4 Exp. 53 Graphical Projection of Pre-processing.

4.7.31 Appendix: No. 5 Experiment 47 Results Overview on Pre-processing

3DCNNsl dispensed similar results compared to the individual scores projected in experiment #1 in **Section B Table 4.12**. Beating(B) produced an individual score of 0.81 with a precision of 0.70 and a recall of 0.88. The focus class, Stabbing(St), generated the highest individual score of 0.88 compared to all other experiments, with a precision of 0.71 and a recall of 0.62. The neutral class Knitting(K) generated individual and a precision score of 100% with a recall of 0.88. Walk-with-Dog(W) dispensed 0.94 with a precision and recall of 0.88. The evaluation proved that pre-processing improved the classification for the focus stabbing class and the overall accuracy score. The results increased to 0.81 compared to experiment #1.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|--|
| Shooting | 5 | 23 | 2 | 2 | • TP Predictions Are Actually True |
| Shooting | 6 | 22 | 2 | 3 | • TN Preds NOT True |
| Knitting | 8 | 23 | 2 | 0 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| WalkwithDog | 7 | 24 | 1 | 1 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.17: Appendix: 3DCNNsl Confusion Matrix Pre-processing No. 5 Exp. 47.

4.7.32 Appendix: No. 5 Exp.47 Confusion Matrix Pre-processing Summary

Figure 4.34 displays the confusion matrix overall performance in Table 4.17 and the actual classifications per action category. 3DCNNsl improved its classification with an overall accuracy of 0.81. The operations experienced fluctuations via the precision and recall ratings. Nevertheless, it generated six false positive and six critical false negative outcomes. The metrics appraisal proved stable; however, the volume of misrepresentations harms the objectives.

#5 Exp: 47 Pre-processing 32 test samples, Epoch 90, Batch Size 130

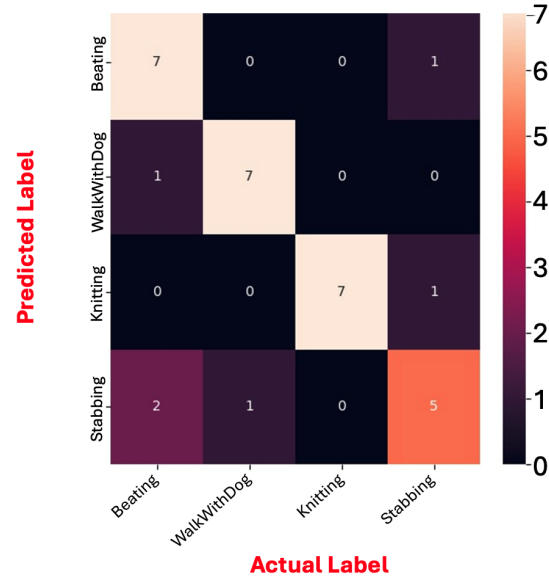


Figure 4.34: Appendix: 3DCNNsl Confusion Matrix Pre-processing No. 5 Exp. 47.

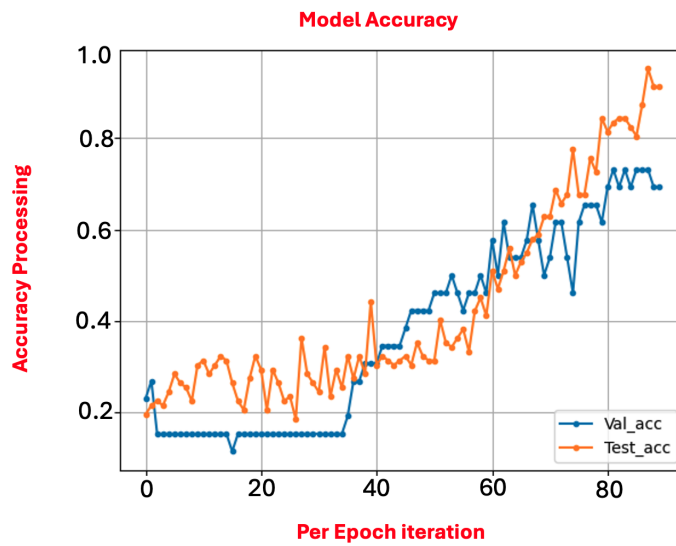


Figure 4.35: Appendix: No. 5 Exp.47 Graphical Projection of Pre-processing.

4.7.33 Appendix: No. 5 Exp.47 Projection of Accuracy Pre-processing

Figure 4.35 graphically expresses high misrepresentation, although the overall accuracy escalated. The operations proved that the dataset size and the complexity of the selected pre-processing techniques impacted the model's processing proficiency. The graphical output depicts high fluctuations in the precision and recall scores.

4.7.34 Appendix: No. 6 Exp.44 Results Overview on Pre-processing

At this pre-processing level, the overall accuracy dispensed the highest score at 0.88 regarding all experiments. Beating(B) produced 100% for the individual and precision scores with a recall of 0.88 in that order. Fighting(Fi) dispensed 0.94 with a precision of 0.86 and a recall of 0.75. Knitting(K) generated an individual score of 0.94 with a precision and recall of 0.88. Like knitting, Walk-with-Dog(W) produced 0.88 with a precision of 0.80 and a recall of 100%. Experiment #6 pre-processing concepts in **Section B Table 4.12** confirmed its superiority on the violent classes to satisfy research questions 1-5 in section 1.2.1.

4.7.35 Appendix: No. 6 Exp.44 Confusion Matrix Pre-processing Summary

Figure 4.36 displayed a confusion matrix overall performance in Table 4.18 and the actual classifications status per action. The data disclosed the highest score relative to the overall accuracy of 0.88 compared to all experiments. 3DCNNsl experienced reduced output fluctuations via the precision and recall ratings. The findings projected the lowest misclassification rate with four false positive and four critical false negative predictions overall. The individual evaluation metrics proved stable; however, the precision and recall metrics performed above the 75th percentile.

#6 Exp: 44 Pre-processing 32 test samples, Epoch 90, Batch Size 130

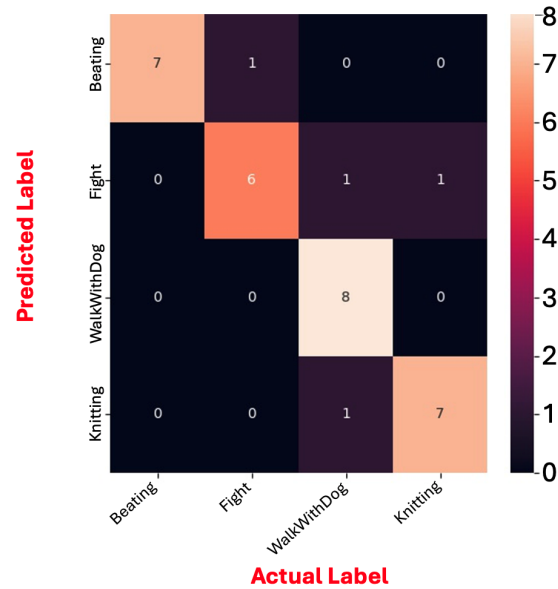


Figure 4.36: Appendix: 3DCNNsl Confusion Matrix Pre-processing No. 6 Exp.44.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-------------|----|----|----|----|---|
| Beating | 7 | 24 | 0 | 1 | <ul style="list-style-type: none">• TP Predictions Are Actually True• TN Preds NOT True• FP Preds NOT True But, Preds True (ok if misclassified)• FN Preds True But, Preds False (NOT ok if misclassified) |
| Fighting | 6 | 23 | 1 | 12 | |
| Knitting | 7 | 23 | 1 | 1 | |
| WalkwithDog | 8 | 21 | 2 | 0 | |

Table 4.18: Appendix: 3DCNNsl Confusion Matrix Pre-processing No. 6 Exp.44.

4.7.36 Appendix: No. 6 Exp. 43 Projection of Accuracy Pre-processing

Figure 4.37 graphically expressed high misrepresentation even though the overall accuracy escalated. Although 3DCNNsl attained the highest ratings for violent activity classification, the dataset size and the complexity of the selected pre-processing techniques impacted the model's effectiveness. The graphical output reflected high fluctuations in the precision and recall scores.

C. Fulfilling Research Question 2/3 via 3DCNNsl using Action Similarity:

At this level, experiments 1-6 in **Table 4.19-C** project the context of action similarity on 32-test samples to fortify the investigations. The emphasis on the results projects the theory of raw data without enhancements, data containing pre-processing enhancements and action similarity to observe optimal possibilities for the proposed fusion approach. The analysis commenced on experiment #1, which is as follows.

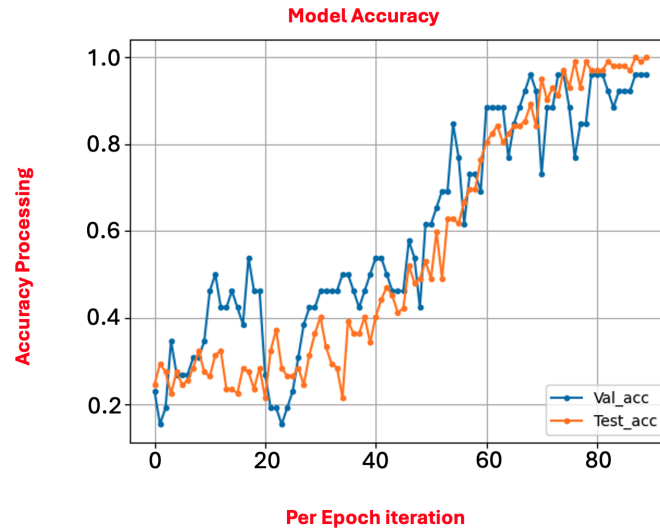


Figure 4.37: Appendix: No. 6 Exp. 44 Graphical Projection of Pre-processing.

| | | | | | | | | | | | | | |
|------------|-----------|-----------|--------|-----------|-----------|--------|-----------|-----------|--------|-----------|-----------|--------|------------|
| #1 (Ex 92) | St | Precision | Recall | Fi | Precision | Recall | C | Precision | Recall | Fe | Precision | Recall | ACC |
| | 0.63 | 0.40 | 0.50 | 0.81 | 0.50 | 0.38 | 0.88 | 0.78 | 0.88 | 0.94 | 0.86 | 0.75 | 0.63 |
| #2 (Ex 86) | Sh | | | B | | | N | | | Fe | | | |
| | 100 | 100 | 0.25 | 0.88 | 0.75 | 0.75 | 0.81 | 0.73 | 100 | 0.75 | 0.64 | 0.88 | 0.72 |
| #3 (Ex 77) | B | | | Fi | | | Su | | | Fe | | | |
| | 0.88 | 0.80 | 100 | 100 | 100 | 0.50 | 0.81 | 0.70 | 0.88 | 0.81 | 0.62 | 0.62 | 0.75 |
| #4 (Ex 83) | Fi | | | Sh | | | Su | | | N | | | |
| | 0.81 | 0.53 | 0.76 | 0.88 | 0.81 | 0.88 | 100 | 0.56 | 0.76 | 0.81 | 0.80 | 0.84 | 0.75 |
| #5 (Ex 89) | Sh | | | St | | | N | | | C | | | |
| | 0.81 | 0.57 | 0.50 | 0.69 | 0.50 | 0.62 | 100 | 100 | 0.88 | 100 | 100 | 100 | 0.75 |
| #6 (Ex 80) | B | | | St | | | Fe | | | C | | | |
| | 0.81 | 0.67 | 0.75 | 0.88 | 0.67 | 0.50 | 0.94 | 0.89 | 100 | 100 | 100 | 100 | 0.81 |

Table 4.19: **Appendix: (C)** 3DCNNsl Impact Experiment with Action Similarity.

4.7.37 Appendix: No. 1 Exp.92 Results Overview on Action Similarity

In Group C, the overall accuracy and stabbing class for action similarity operations generated a lower score of 0.63 with a precision of 0.40 and recall of 0.50. Fighting(F) dispensed an individual score of 0.81 with a precision of 0.50 and a recall of 0.38. Experiments #2 and #3 in **Section C Table 4.19** validated the notion that the model experienced processing challenges when discerning the true nature of violence in action similarity conditions. The neutral class Cutting-in-kitchen(C) produced an individual score of 0.88 with a higher precision of 0.78 and recall of 0.88. Fencing(Fe) dispensed an individual score of 0.94 with a precision of 0.86 with a recall of 0.75. The model demonstrated discerning abilities between stabbing and fencing above the 60th percentile to satisfy the research objectives. Though the performance proved promising, the model experienced severe misclassification linked to fluctuating precision and recall metrics.

4.7.38 Appendix: No. 1 Exp.92 Results Overview on Action Similarity

Figure 4.38 confusion matrix depicted the overall performance in Table 4.20 and the classification status of its classes. The findings disclosed a reduction in classification proficiency with an overall accuracy of 0.63. 3DCNNsl experienced intense output fluctuations with low precision and recall ratings. The evidence depicted increased misclassifications with 12

false positive and 12 critical false negative predictions overall. Because of the data complexity, the results confirmed that action similarity is a significant concern for 3DCNNsl at this stage. The evidence showed that stabbing produced the most misclassifications compared to fencing, cut-in-kitchen and fighting.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-----------------|----|----|----|----|---|
| Stabbing | 4 | 18 | 6 | 4 | <ul style="list-style-type: none"> • TP Predictions Are Actually True |
| Fighting | 3 | 23 | 3 | 5 | <ul style="list-style-type: none"> • TN Preds NOT True |
| Cutt-in-Kitchen | 7 | 22 | 2 | 1 | <ul style="list-style-type: none"> • FP Preds NOT True But, Preds True (ok if misclassified) |
| Fencing | 6 | 23 | 1 | 2 | <ul style="list-style-type: none"> • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.20: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 1 Exp.92.

4.7.39 Appendix: No. 1 Exp.92 Projection of Accuracy Action Similarity

Figure 4.39 graphically expressed high misrepresentation results as a direct link to the individual scores and overall accuracy. Though the model attained high scores for neutral activity classification on fencing, it dispensed stabbing as the lowest score for all experiments. The graphical output depicted reoccurring fluctuations via precision and recall scores for violent classes; however, its processing stabilised above the 75th percentile for the neutral classes. The operation made a significant attempt to generate a smooth representation of its output values, which validated the model's processing limitations.

#1 Exp: 92 Action Similarity 32 test samples, Epoch 90, Batch Size 130

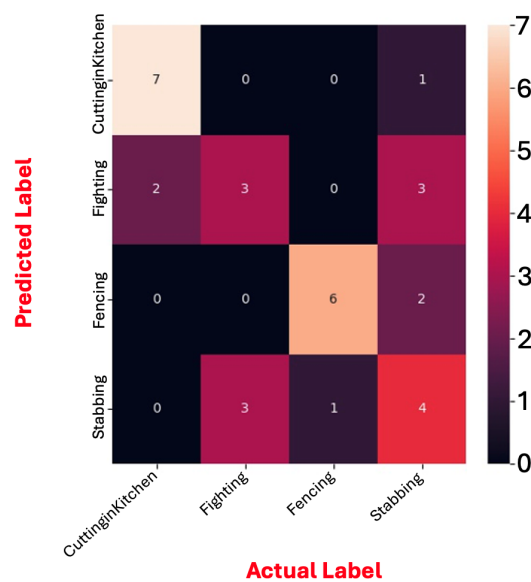


Figure 4.38: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 1 Exp.92.

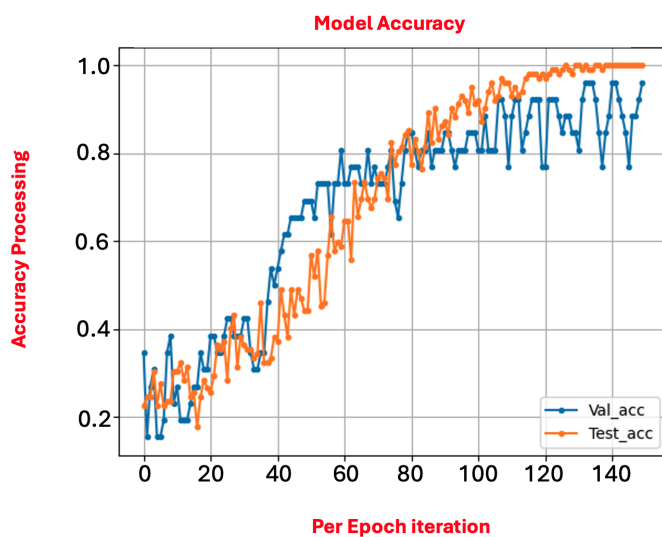


Figure 4.39: Appendix: No. 1 Exp.92 Graphical Projection of Action Similarity.

4.7.40 Appendix: No. 2 Exp.86 Results Overview on Action Similarity

3DCNNsl improved its classification with an overall accuracy of 0.72 and higher individual performance metrics. Shooting(Sh) via action similarity produced an individual and precision score of 100% each with a recall of 0.25. Beating(B) dispensed 0.88 with a precision and recall of 0.75 each. The violent actions were evaluated against the neutral Nun-chuck(N) action at 0.81 with a precision of 0.73 and a recall of 100%. Fencing(Fe) proved challenging to predict, depreciating from 0.94 in **Section C Table 4.19** experiment #1 to 0.75. 3DCNNsl demonstrated high performance; however, fluctuating precision and recall metrics confirmed that the generalisation challenges persist.

4.7.41 Appendix: No. 2 Exp.86 Confusion Matrix Action Similarity Summary

Figure 4.40 confusion matrix illustrates the overall performance in Table 4.21 and the actual classification status of the actions. The findings depicted an improvement in classification effectiveness with an overall accuracy of 0.72. 3DCNNsl experienced increased precision and recall, with the lowest recall rating for shooting at 0.25. The analysis projected a decrease in misclassifications with 6-false positives reflecting high precision scores and 12-critical false negative predictions directly linked to fluctuating recalls. The findings confirm the theory that actions similarity

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|------------|----|----|----|----|--|
| Shooting | 2 | 24 | 0 | 6 | • TP Predictions Are Actually True |
| Beating | 6 | 22 | 2 | 2 | • TN Preds NOT True |
| Nun-chucks | 8 | 21 | 0 | 3 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| Fencing | 7 | 20 | 4 | 1 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.21: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 2 Exp.86.

demonstrated little impact on the shooting against the nun-chucks class relative to violence but depreciated for the neutral samples.

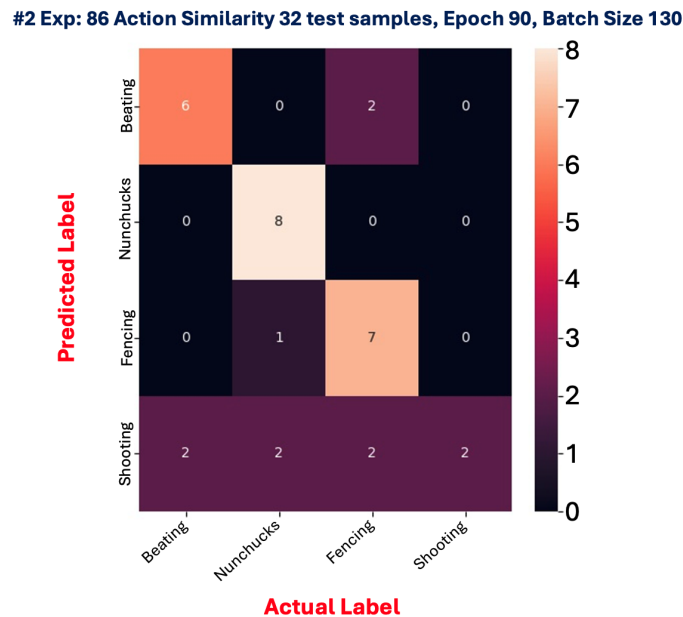


Figure 4.40: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 2 Exp.86.

4.7.42 Appendix: No. 2 Exp.86 Graphical Projection of Action Similarity

Figure 4.41 graphical representation expressed high fluctuations directly linked to the individual scores and overall accuracy values. The graphical results demonstrated a significant drop in validation performance because of the limited quantity and complexity of the data.

4.7.43 Appendix: No. 3 Exp.77 Results Overview on Action Similarity

The violent class Beating(B) projected no improvement as it generated the same score in **Section C Table 4.19** experiment #2 of 0.88 with a precision of 0.80 and a recall of 100%. Fighting(Fi) class also demonstrated improvement as it generated a score of 100% compared to #1, with a precision of 100% and a recall of 0.50. The neutral class Sumo-

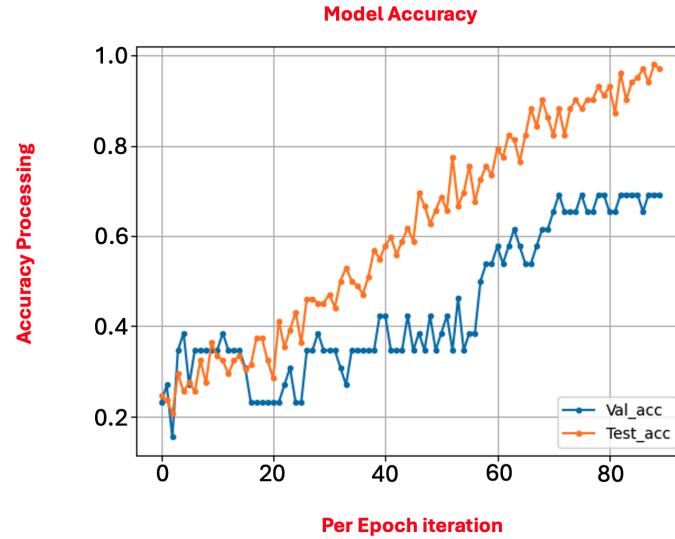


Figure 4.41: Appendix: No. 2 Exp.86 Graphical Projection of Action Similarity.

wrestling(Su) attained a classification score of 0.81 with a precision of 0.70 and a recall of 0.88. Like Sumo-wrestling(Su), Fencing(Fe) produced 0.81 with a lower precision and recall at 0.62 each. 3DCNNsl classification improved with an overall accuracy of 0.75 compared to experiments #1 and #2. Although the findings projected an overall classification improvement, 3DCNNsl displayed misclassifications for action similarity Group C.

4.7.44 Appendix: No. 3 Exp.77 Confusion Matrix Action Similarity Summary

Figure 4.42 depicts the confusion matrix's overall performance in Table 4.22 and the classification status per action. The result demonstrated an improvement in classification proficiency with an overall accuracy of 0.75. 3DCNNsl experienced increased precision and recall for violence but decreased for the non-violent neutral samples. The analysis reflected increased misclassifications with sixteen false positives and ten critical false negative predictions. The findings suggested prediction improvements, emphasising the distinction of the true nature of the actions with fewer false negative predictions compared to **Section C Table 4.19** experiment #2. Moreover, conceding high misclassification outcomes

suggested the insignificance of 3DCNNsl processing at this stage.

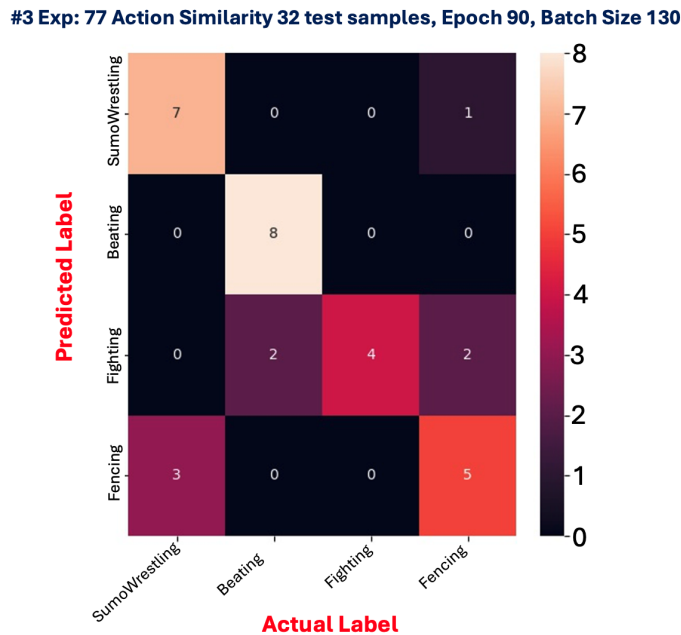


Figure 4.42: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 3 Exp.77.

4.7.45 Appendix: Exp.77 Graphical Projection Of Accuracy Action Similarity

Figure 4.43 graph validated the high fluctuations of the individual and overall accuracy processing relative to Figure 4.42. The graph expressed an increase in validation performance, emphasising processing challenges to distinguish the dissimilarity in the classes.

4.7.46 Appendix: No. 4 Exp.83 Results Overview on Action Similarity

3DCNNsl's findings displayed no improvement in performance as it regenerated an identical overall accuracy of 0.75 compared to **Section C Table 4.19** experiment #3. The operation recorded an individual score of 0.81 for Fighting(Fi) with a precision of 0.53 and a recall of 0.76. Shooting(Sh) dispensed an individual and recall score of 0.88, each with a precision

of 0.81. The neutral Sumo-wrestling(Su) class recorded a higher score of 100% with a lower precision of 0.56 and recall of 0.76. The Nun-chucks(N) category duplicated the scores once more compared to **Section C Table 4.19** experiment #2 with a precision of 0.80 and recall of 0.84. The findings reflected similar characteristics compared to experiment #4 processing and experiment #3.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|----------------|----|----|----|----|---|
| Beating | 8 | 22 | 2 | 0 | <ul style="list-style-type: none"> • TP Predictions Are Actually True |
| Fighting | 4 | 22 | 8 | 6 | <ul style="list-style-type: none"> • TN Preds NOT True |
| Sumo-Wrestling | 7 | 21 | 3 | 1 | <ul style="list-style-type: none"> • FP Preds NOT True But, Preds True (ok if misclassified) |
| Fencing | 5 | 21 | 3 | 3 | <ul style="list-style-type: none"> • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.22: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 3 Exp.77.

4.7.47 Appendix: No. 4 Exp.83 Confusion Matrix Action Similarity Summary

Figure 4.44 confusion matrix illustrates the overall performance in Table 4.23 and the classification status per class. The operations improved classification effectiveness with an identical overall accuracy score of 0.75, like **Section C Table 4.19** experiment #3. 3DCNNsl experienced intense fluctuations relative to precision and recall. The findings emphasised reduced misclassifications with eight false positives and eight critical false negative outcomes. The evidence substantiated the model’s discerning and prediction capability towards the true nature of the actions. Although 3DCNNsl proved effective, the high fluctuation indicated the need for auxiliary support to encourage more robust representations.

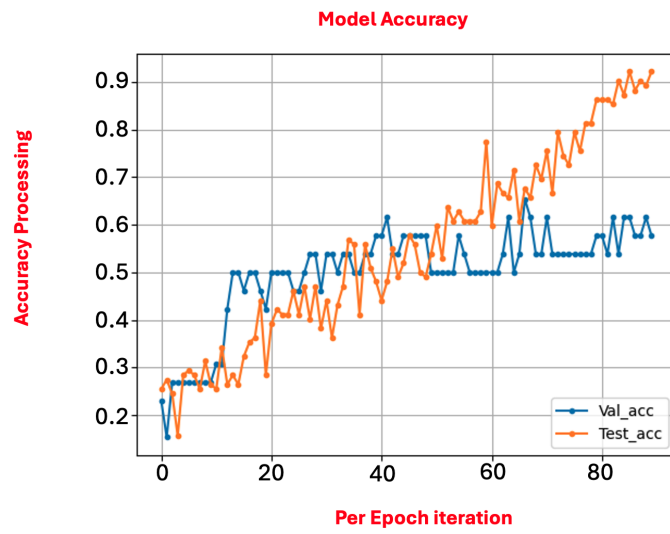


Figure 4.43: Appendix: No. 3 Exp. 77 Graphical Projection of Action Similarity.

#4 Exp: 83 Action Similarity 32 test samples, Epoch 90, Batch Size 130

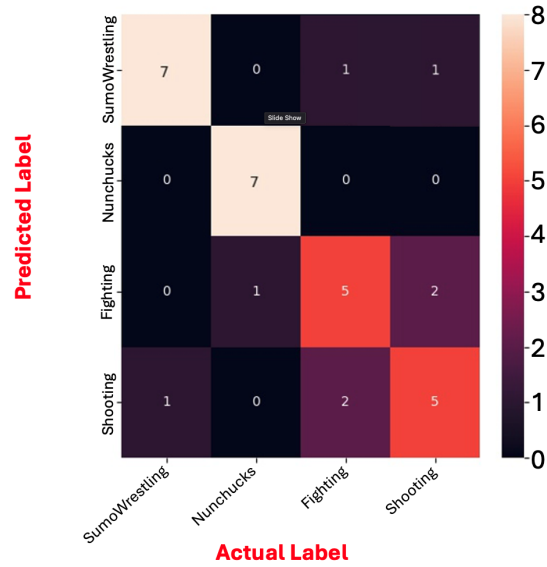


Figure 4.44: Appendix: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 4 Exp.83.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|----------------|----|----|----|----|--|
| Fighting | 5 | 21 | 3 | 3 | • TP Predictions Are Actually True |
| Shooting | 5 | 21 | 3 | 3 | • TN Preds NOT True |
| Sumo-Wrestling | 7 | 22 | 1 | 2 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| Nun-chucks | 7 | 24 | 1 | 0 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.23: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 4 Exp.83.

4.7.48 Appendix: No. 4 Exp.83 Projection of Accuracy Action Similarity

Figure 4.45 graph validated the high fluctuation linked to the individual scores and overall accuracy processing in Figure 4.44. The graph demonstrated the fluctuated output values as 3DCNNsl struggled to establish class dissimilarity. Although fighting and sumo-wrestling projected high prediction ratings, the graphical perspective insinuated the presence of over-fitting challenges in 3DCNNsl processing.

4.7.49 Appendix: Exp.89 Results Overview on Action Similarity

The neutral classes superseded the violent classes, demonstrating identical performance traits like **Section C Table 4.19** experiment #3 and #4. Nun-chuck(N) recorded an individual and precision score of 100% with a recall of 0.88. Cutting-in-kitchen(C) dispensed the highest performance at 100% for individual scores, precision and recall relative to the other actions. Shooting(Sh) dispensed 0.81 with a lower precision of 0.57 and recall at 0.50 versus **Section C Table 4.19** experiment #2. Stabbing(St) attained the second-lowest score at 0.69 with a precision of 0.50 and recalled at 0.62. At this level, 3DCNNsl recorded an overall accuracy score of 0.75, with lower ratings for the violent class and high ratings for the neutral class. The outcomes confirmed 3DCNNsl’s continued processing

limitations, displaying severe misclassifications from an individual standpoint.

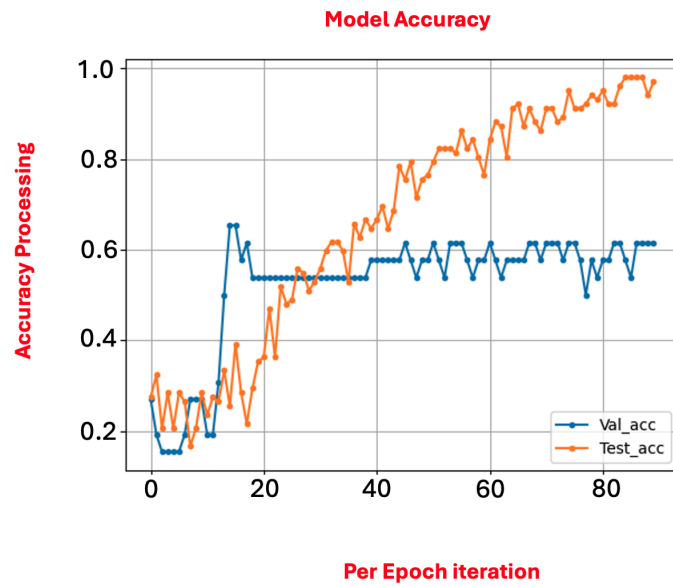


Figure 4.45: Appendix: No. 4 Exp.83 Graphical Projection of Action Similarity.

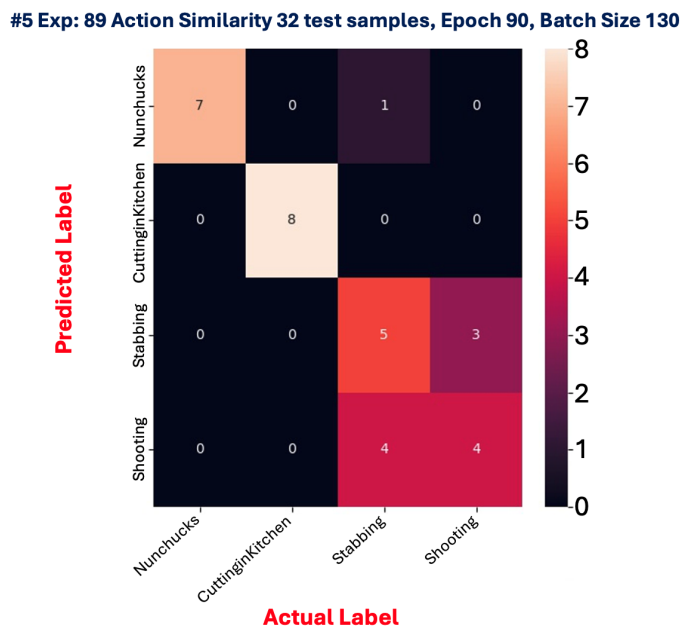


Figure 4.46: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 5 Exp.89.

4.7.50 Appendix: No. 5 Exp.89 Confusion Matrix Action Similarity Summary

Figure 4.46 confusion matrix illustrates the overall performance in Table 4.24 and the actual classification status per action. The results presented similarities in the overall accuracy ratings of 0.75 in **Section C Table 4.19** experiment #3 and experiment #4. The model experienced high fluctuation via the precision and recall metrics for the violent classes and higher performance for the non-violent neutral category. The results projected a decline in false positive misclassifications at five and eight critical false negative predictions. The analysis accentuated 3DCNNsl's processing improvements linked to predictions on the neutral classes from an individual perspective at 100% each. Nevertheless, the high fluctuations outcomes for the violent classes validated 3DCNNsl classification limitations towards dispensing more robust values. The complexity stabbing's score of 0.69 and shooting at 0.81 challenged the model as a direct link to the nature of the data and the deliberate pre-processing approach integrated to recreate real-world conditions.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-----------------|----|----|----|----|---|
| Shooting | 4 | 21 | 3 | 4 | • TP Predictions Are Actually True |
| Stabbing | 5 | 19 | 5 | 3 | • TN Preds NOT True |
| Nun-chucks | 7 | 24 | 1 | 0 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| Cutt-in-Kitchen | 8 | 24 | 0 | 0 | • FN Preds True But, Preds False (NOT ok) if misclassified) |

Table 4.24: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 5 Exp.89.

4.7.51 Appendix: No. 5 Exp.89 Projection of Accuracy Action Similarity

Figure 4.47 graphical representation validated the fluctuating results of individual scores for the violent classes and the similarity in overall accuracy. The graph expressed 3DCNNsl's insignificant processing attempt concerning

validation and accuracy towards discerning action similarity. 3DCNNsl's processing suggested the presence of over-fitted operations and a direct need for auxiliary classification support.

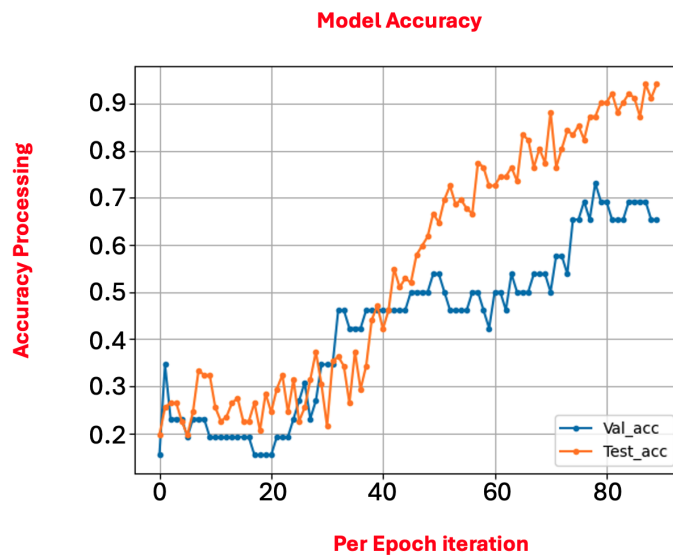


Figure 4.47: Appendix: No. 5 Exp.89 Graphical Projection of Action Similarity.

4.7.52 Appendix: No. 6 Exp.80 Results Overview on Action Similarity

The neutral class evidence on Cutting-in-kitchen(C) recorded the highest performance of 100% for the individual outcomes, precision and recall scores for **Section C Table 4.19** experiment #5. Fencing(Fe) dispensed an individual score of 0.94 with a precision of 0.89 and a recall of 100%. Stabbing(St) recorded the highest score at 0.88 for all experiments, with a precision of 0.67 and a recall of 0.50. Beating(B) dispensed 0.81 with a precision of 0.67 and a

recall of 0.75. Although the individual scores fluctuated, this model produced the second-highest overall accuracy for all experiments. The results insinuated that pre-processing with action similarity conditions can be facilitated utilising 3DCNNsl with high performance.

4.7.53 Appendix: No. 6 Exp.80 Confusion Matrix Action Similarity Summary

Figure 4.48 confusion matrix illustrates the overall performance in Table 4.25 and the actual classification status per action. The findings projected an identical overall accuracy rating of 0.81, like **Section C Table 4.19** experiment #5's outcome. Concerning all experiments, 3DCNNsl experienced high precision and recall fluctuations for the violent classes and higher performance for the neutral category above the 89th percentile. The data demonstrated a similar decrease in false positive misclassifications with an output of 5, like **Section C Table 4.19** experiment #4 and 7-critical false negative predictions. The results corroborated 3DCNNsl's prediction accuracy on the neutral classes, with decreased performance directly linked to the complexity of the violent classes. Moreover, analysis proved that deliberate pre-processing and action similarity conditions challenged 3DCNNsl's ability to generate robust ratings from an individual processing perspective.

| Rating | TP | TN | FP | FN | Confusion Matrix Description |
|-----------------|----|----|----|----|--|
| Beating | 6 | 21 | 3 | 2 | • TP Predictions Are Actually True |
| Stabbing | 4 | 22 | 2 | 4 | • TN Preds NOT True |
| Fencing | 8 | 23 | 0 | 1 | • FP Preds NOT True But, Preds True (ok if misclassified) |
| Cutt-in-Kitchen | 8 | 24 | 0 | 0 | • FN Preds True But, Preds False (NOT ok if misclassified) |

Table 4.25: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 6 Exp.80.

#6 Exp: 80 Action Similarity 32 test samples, Epoch 90, Batch Size 130

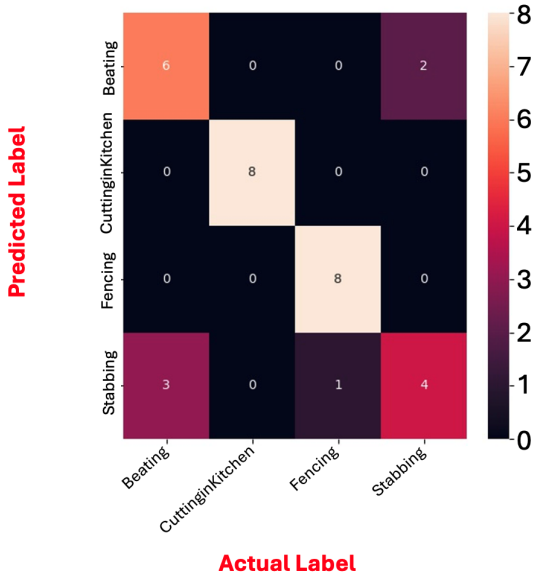


Figure 4.48: Appendix: 3DCNNsl Confusion Matrix Action Similarity No. 6 Exp.80.

4.7.54 Appendix: No. 6 Exp.80 Projection of Accuracy Action Similarity

Figure 4.49 graphical expression validated the rationale for conceiving fluctuations linked to the violent class individual scores and similarity in overall accuracy compared to **Section C Table 4.19** experiment #5. The data reflected 3DCNNsl insignificant attempt via validation and accuracy towards discerning action similarity. The previously mentioned results suggested that the model produced another severely over-fitted operation.

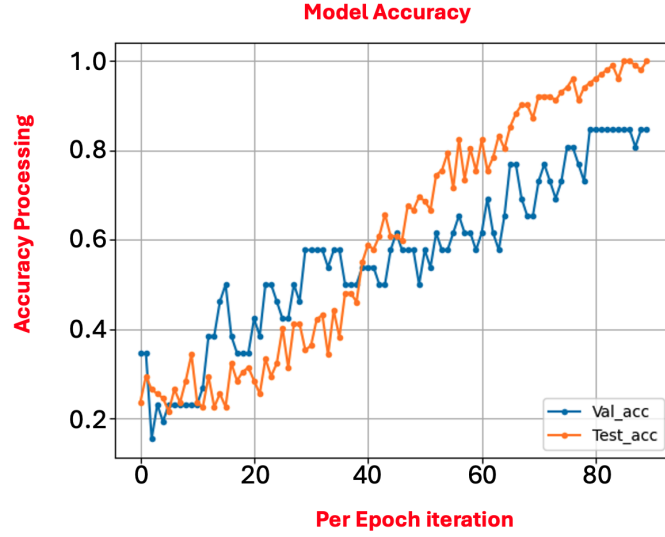


Figure 4.49: Appendix: No. 6 Exp.80 Graphical Projection of Action Similarity.

4.8 Appendix: 3DCNNsl (A) No Pre-processing/(B) Pre-processing/ (C) Action Similarity Summary

With insight into the operations and results dispensed from 3DCNNsl's simulations, it is necessary to outline model superiority to satisfy the research objectives. 3DCNNsl simulations displayed significant performance by applying the entire video as input to generate significant results. However, it was clear that pre-processing methods outperformed no pre-processing and action similarity. 3DCNNsl operations produced fewer misclassifications from an individual perspective with higher overall accuracy outcomes in pre-processing conditions. Analysing the results in this manner provided crucial insight into processing pre-empting violent activity recognition. Ranking the process emphasised 3DCNNsl's operating potential regarding real-world conditions, fortifying its application for the proposed fusion concepts. The ranking procedure relative to the results previously discussed in Appendix 4.7 is as follows.

Ranked 1st (B) Pre-processing with the least misclassification errors.

Ranked 2nd (C) Action Similarity compared to (A and B) results.

Ranked 3rd (A) No Pre-processing with the highest misclassification ratings.

4.9 Appendix: Discussions on YOLOv5m/3DCNNsl Performance

At this stage, discussions on both models projected their performance and the issues or anomalies encountered during development. The approach accentuated solutions to mitigate the issues discussed or minimise the negative impact experience concerning processing efficiency. The discussions are as follows.

4.9.1 Appendix: Discussions on YOLOv5m Confusion Matrix

Table 4.26 results emphasised pre-trained superiority in Appendix 4.6 but demonstrated high fluctuations from an individual class perspective. Appendix 4.6 and Table 4.26 provide an overall perspective of the challenge as an outcome. The findings validated YOLOv5m's stability and superiority in pre-trained experiment #8 regarding actual predictions, inaccurate predictions, and overall accuracy. The evidence satisfied research questions 1-3 in section 1.2.1. The analysis is as follows.

| From-Scratch | #1 Exp | #2 Exp | #3 Exp | #4 Exp |
|----------------------|---------------|---------------|---------------|---------------|
| Accurate Predictions | 4 | 6 | 7 | 8 |
| NOT Accurate | 30 | 21 | 21 | 28 |
| Overall Accuracy | 0.53 | 0.63 | 0.67 | 0.72 |
| Pre-Trained | #5 Exp | #6 Exp | #7 Exp | #8 Exp |
| Accurate Predictions | 7 | 8 | 5 | 9 |
| NOT Accurate | 28 | 27 | 25 | 20 |
| Overall Accuracy | 0.75 | 0.79 | 0.83 | 0.85 |

Table 4.26: Appendix: YOLOv5m From-Scratch/Pre-trained Errors

4.9.2 Appendix: Discussions on 3DCNNsl Confusion Matrix

Emphasis on 3DCNNsl processing provided further insight via the confusion matrix accentuating the superiority of YOLOv5m’s pre-processing operations. In previous discussions via Appendix 4.5’s conditions and the overview in Table 4.27, stabbing attained the lowest individual score compared to all other violent classes.

| (A) 3DCNNsl No Pre-processing (NP) | | | | | |
|---|-----------------|-----------------|-----------------|---------------------------|-----------------|
| # 6 Exp20 | Shooting | Beating | Knitting | WalkwithDog | Accuracy |
| | 0.94 | 0.94 | 0.88 | 0.75 | |
| Accurate Predictions | 4 | 8 | 4 | 8 | 0.75 |
| NOT Accurate | 5 | 1 | 6 | 4 | |
| (B) 3DCNNsl Pre-processing (PP) | | | | | |
| # 6 Exp44 | Beating | Fighting | Knitting | WalkwithDog | Accuracy |
| | 100 | 0.94 | 0.94 | 0.88 | |
| Accurate Predictions | 7 | 6 | 7 | 8 | 0.88 |
| NOT Accurate | 1 | 3 | 2 | 2 | |
| (C) 3DCNNsl Action Similarity (AS) | | | | | |
| #6 Exp80 | Beating | Stabbing | Fencing | Cutting-in-Kitchen | Accuracy |
| | 0.81 | 0.88 | 0.94 | 100 | |
| Accurate Predictions | 6 | 4 | 8 | 8 | 0.81 |
| NOT Accurate | 5 | 6 | 1 | 0 | |

Table 4.27: Appendix: Summary of 3DCNNsl NP, PP and AS Misclassification Rate.

4.9.3 Appendix: Discussions on 3DCNNsl Research Question 1 (A)

The analysis provided crucial insight into data requiring additional modifications to increase/decrease pre-processing, which facilitated further integration flexibility. The knowledge gain fosters robust performance during the proposed fusion’s development. The overall accuracy outcomes confirmed the model processing challenges when interpreting violent stabbings scenarios during classification. Nevertheless, 3DCNNsl’s evidence proved the effectiveness of individual classification on violent actions containing no pre-processing support with ratings exceeding the 75th percentile. The analysis highlighted stabbing at 0.94% as the highest performance rating overall to validate the interpretation of the 3DCNNsl results and satisfy research question 1 in section 1.2.1.

4.10 Appendix: YOLOv5m/3DCNNsl Fulfilling Research Question 1 (Can violence and weapons be recognised in CCTV data?)

The evaluation of new test data (data the model has not seen in any way) provided an opportunity to observe accurate perceptions of violence during inference. Regarding analysing the results of research question 1, the model confirmed the action representation with probability scores, utilising the Soft-Max function operations in the final processing layer. Its function produces probability score representations between 1 and 0 to insinuate its confidence in pending violent scenarios. The evidence emphasised the achievement of violent activity recognition by demonstrating YOLOv5m predictive capacity previously discussed in Appendix 4.4, Table 4.4 and 3DCNNsl Table 4.5 to Table 4.19 on each class. Because 3DCNNsl is the state-of-the-art for activity recognition, it was necessary to thoroughly evaluate its framework to determine the accurate classification of pre-empting violence given the criticality of human life and preventing these heinous outcomes.

4.11 Appendix: YOLOv5m and 3DCNNsl Fulfilling Research Question 2 (What’s the Impact of Data Modification?)

YOLOv5m with no pre-processing experienced lower performance ratings from an overall accuracy perspective in From-Scratch at 0.67 in Table 4.4 experiment #3 and pre-trained at 0.75. Contrarily, high fluctuating scores for individual performance confirmed the presence of misclassifications in

Appendix 4.4. YOLOv5m pre-processing operations demonstrated higher performance from pre-trained at 0.85% in Table 4.4 experiment #21 compared to From-Scratch methods at 0.72 #4. In addition to the 13% increase, the findings revealed high fluctuations with reduced misclassifications for the individual classes via pre-trained processing. 3DCNNsl via Table 4.5 to Table 4.19 demonstrated similar concepts of applying no pre-processing and pre-processing enhancements. Experiments containing action similarity acted as a deliberate approach to properly evaluate the state-of-the-art medium as a direct link to the criticality of human life. Overall accuracy recorded the highest score for no pre-processing operations via Table 4.5 experiment #6 at 0.75, which projected high individual class fluctuations in Table 4.19. Moreover, pre-processing in Table 4.12 experiment #6 at 0.88 increased by 15% with continuous fluctuations recorded for the individual classes. Action similarity operations generated comparisons between stabbing and fencing actions. The idea forced the model to pre-empt actions containing homogeneous attributes in the context of pre-processing towards satisfying research question-2 in section 1.2.1. Unlike pre-processing, action similarity operations maintained their performance above the 81st percentile for individual classification; nonetheless, its accuracy decreased by 7%.

4.12 Appendix: Fulfilling Research Question 3 (what’s the data impact if sample increase?)

The concept determines the significance of the model and its capacity to maintain efficiency if the volume of the data increases or not. The approach disclosed deficient areas requiring modifications to encourage high-performance results during the proposed fusion development stages. An increase in the volume of violent data in YOLOv5m operations in Table 4.4 from 2284 via experiment # 3, 5944 via experiment # 6 and 6204 via experiment # 8 assisted in achieving the notion. With insight into YOLOv5m’s increased performance at this stage, 3DCNNsl evaluations followed in a similar context with increased option values. Detailed in Appendix 4.11 to Appendix 4.12, an increment via the option parameters and the class samples from 1 violent class vs one neutral to four vs four classes acted as additional experiments to evaluate class performance rankings. The application of multiple experiments (one violent versus one neutral class) evaluated the processing complexity and 3DCNNsl’s ability to discern attributes from an individual and overall accuracy perspective to fulfil research question 3. It is crucial to indicate classes that were easier to discern

and those requiring pre-processing enhancements to promote robust classification results. The idea evaluated the class complexity impact in groups to simulated real-world conditions by incrementing the volume of samples and option parameter values. The investigations are as follows.

4.13 Appendix: 3DCNNsl Additional Experiments (4 violent classes vs 4 neutral classes) Summary

The understanding gained from the 1 x 1 experiments, particularly about epoch variations, was instrumental in projecting the impact of escalating sample sizes and exploring the effects of altering hyper-parameter values. This understanding was crucial in grasping the actual performance of the operations. The investigative approach validated performance impact, satisfying the research objectives. Analysis of one neutral class versus one violent class, two versus two, and four versus four proved essential for comparisons. Moreover, the rationale behind the additional experiments emphasised that various investigations were employed to validate the results and fortify the notions. Incrementing the class samples from one violent class discussed previously to determine the processing impact conditions proved essential to validate the impact of increasing the sample size. Maintaining configuration and experiment consistency alleviated bias results by integrating all attributes from the previous one versus one and two vs two experiments. The idea applied the insight to investigate four vs four from an individual class and overall accuracy standpoint. At this stage, four x four activity classes disclosed the impact status to satisfy research question-3 in section 1.2.1.

4.13.1 Appendix: 3DCNNsl via 90 Epochs in 4x4 (No pre-processing, pre-processing, Action Similarity)

The experiments reflected performance from an overall accuracy perspective followed by an individual perspective to satisfy research question-3 in section 1.2.1. Applying hyper-parameter values at 90, 150, and 180 epochs with a batch size of 120, 130, and 140 assisted in maintaining the previous context. Table 4.28 confirms the positive processing results when increasing the data samples from an overall accuracy perspective. Although the analysis projected fluctuations via individual classification, pre-processing proved superior in results. From an individual standpoint, pre-processed operations dispensed seven classes exceeding the 90th percentile, beating 0.78 in experiment #62

with no pre-processing and action similarity to validate the approach. With neutral activity classes between the 84th and 96th percentile, the violent classes proved challenging to discern, demonstrating fluctuating outcomes between the 65th and 96th percentile. Investigations of experiment #62 in action similarity determine the actual classification status of violence versus homogeneous neutral classes. From an individual performance perspective, all violent classes scored lower compared to the neutral non-violent classes. The findings on action similarity in Table 4.28 proved that 3DCNNsl’ experienced challenges processing violence utilising the 90-epoch parameter.

| No Pre-processing (NP) | | | | | | | | | |
|---------------------------------------|------|------|------|------|------|------|------|------|------|
| Fi, Sh, St, Be vs K, W, H, Br Exp 29 | St | Be | Fi | Sh | K | W | H | Br | ACC |
| | 0.84 | 0.90 | 0.96 | 0.65 | 0.84 | 0.96 | 0.84 | 0.96 | 0.50 |
| Pre-processing (PP) | | | | | | | | | |
| Fi, Sh, St, Be vs K, W, H, Br Exp 62 | St | Be | Fi | Sh | K | W | H | Br | |
| | 0.96 | 0.78 | 0.90 | 0.93 | 0.96 | 0.93 | 0.96 | 0.96 | 0.71 |
| Action Similarity (AS) | | | | | | | | | |
| Fi, Sh, St, Be vs C, Fe, Su, N Exp 62 | St | Be | Fi | Sh | C | Fe | Su | N | |
| | 0.78 | 0.84 | 0.90 | 0.87 | 0.96 | 0.96 | 0.93 | 0.93 | 0.60 |

Table 4.28: Appendix: 3DCNNsl 3 experiments, 4 vs 4 at 90-Epochs.

4.13.2 Appendix: 3DCNNsl via 150 Epochs in 4x4 (No Pre-processing, Pre-processing, Action Similarity)

Understanding the 90-epoch experiment, using eight classes consisting of a balanced ratio of violent and non-violent samples, the proficiency evaluation strategy encompassing an increment in the epoch iteration from 90 to 150, with a smaller batch size 120, proved necessary. The idea satisfied research question-3 in section 1.2.1 by evaluating 3DCNNsl’s confidence using suitable hyperparameters with a sample increase. Table 4.29’s results fulfilled research question-3 in section 1.2.1 by evaluating the overall accuracy in experiment #30 no pre-processing at 0.63, experiment #63 pre-processing at 0.81 and action similarity at 0.73. Previous results confirmed that applying more

violent samples and hyper-parameter tuning enhanced the outcome toward the objectives. The scores exceeded the 81st percentile in experiment #63 individual classification, thus projecting performance improvement. In this instance, pre-processing operations proved superior, generating the highest ratings for violence compared to no pre-processing and action similarity, with stabbing at 0.87. Pre-processing and action similarity concepts significantly improved, generating higher individual and overall accuracy scores. The results regarding the four versus four class using 150 epochs are as follows.

| No Pre-processing (NP) | | | | | | | | | |
|---------------------------------------|-------------------|-------------------|-------------------|-------------------|------------------|-------------------|------------------|-------------------|--------------------|
| Fi, Sh, St, Be vs K, W, H, Br Exp 30 | St 1 | Be 0.93 | Fi 0.81 | Sh 0.90 | K 0.94 | W 0.84 | H 0.84 | Br 0.97 | ACC 0.63 |
| Pre-processing (PP) | | | | | | | | | |
| Fi, Sh, St, Be vs K, W, H, Br Exp 63 | St 0.87 | Be 0.90 | Fi 1 | Sh 0.93 | K 0.97 | W 1 | H 0.97 | Br 0.97 | ACC 0.81 |
| Action Similarity (AS) | | | | | | | | | |
| Fi, Sh, St, Be vs C, Fe, Su, N Exp 63 | St 0.87 | Be 0.96 | Fi 1 | Sh 0.87 | C 1 | Fe 0.91 | Su 88 | N 0.97 | ACC 0.73 |

Table 4.29: Appendix: 3DCNNsl 3 experiments, 4 vs 4 at 150-Epochs.

4.13.3 Appendix: 3DCNNsl via 180 Epochs in 4x4 (No Pre-processing, Pre-processing, Action Similarity)

Analysis of the 150-epoch simulations corroborated the hypothesis that extending the epoch range to 180 satisfies performance gains regarding both individual and overall accuracy, as evidenced in Table 4.30. Like Table 4.29's results, the overall accuracy evaluation solidified the positive notion of increasing the samples. The operation extended training when comparing Table 4.28 to Table 4.30's results. At this level, pre-processing projected its superiority despite fluctuating individual scores. The analysis validated 3DCNNsl significant attempt at classifying the actions above the 84th percentile range. Table 4.31's results satisfied the fulfilment of research question-3 in section 1.2.1, which follows.

| No Pre-processing (NP) | | | | | | | | | |
|--|-----------|-----------|-----------|-----------|----------|-----------|-----------|-----------|------------|
| Fi, Sh, St, Be vs K, W, H, Br Exp 31 | St | Be | Fi | Sh | K | W | H | Br | ACC |
| | 90 | 0.93 | 0.84 | 0.87 | 0.96 | 0.93 | 0.84 | 1 | 0.65 |
| Pre-processing (PP) | | | | | | | | | |
| Fi, Sh, St, Be vs K, W, H, Br Exp 64 | St | Be | Fi | Sh | K | W | H | Br | |
| | 0.90 | 0.93 | 1 | 0.84 | 0.93 | 1 | 1 | 0.96 | 0.84 |
| Action Similarity (AS) | | | | | | | | | |
| FFi, Sh, St, Be vs C, Fe, Su, N Exp 64 | St | Be | Fi | Sh | C | Fe | Su | N | |
| | 0.87 | 0.90 | 93 | 0.90 | 1 | 1 | 1 | 0.93 | 0.75 |

Table 4.30: Appendix: 3DCNNsl 3 experiments, 4 vs 4 at 180-Epochs.

4.14 Appendix: 3DCNNsl via 180 Epochs in 4x4 (No Pre-processing, Pre-processing, Action Similarity)

Appreciating the results of 150 epochs in previous discussions, confirmed the research notion by increasing the epoch range to 180 to observe the performance impact from an individual and overall accuracy perspective via Table 4.31. Like Table 4.29’s results, the overall accuracy evaluation solidified the positive notion of increasing the samples; this extended training when comparing Table 4.28 to Table 4.30’s results. At this level, pre-processing confirmed its superiority despite fluctuating individual scores. The analysis proved that 3DCNNsl demonstrated a significant attempt at classifying the actions above the 84th percentile range. Table 4.31’s results validated the fulfilment of research question-3 in section 1.2.1, which follows.

4.15 Appendix: Fulfilling Research Question 4 (No Pre-processing, Pre-processing, Action Similarity)

Programming facilitated the classification status, leading to script techniques which specify the generic and subclass classification status for both models in the final processing layers. The discovery of programmable scripts for both YOLOv5m/3DCNNsl negatively impacted the real-time operations. With insight into the script development, the process incorporated only the generic and subclass classification features in 3DCNNsl. Ignoring identical script concepts for YOLOv5m improved the classification of real-time operations significantly, which satisfied research question-4

| No Pre-processing (NP) | | | | | | | | | |
|--|-----------|-----------|-----------|-----------|----------|-----------|-----------|-----------|------------|
| Fi, Sh, St, Be vs K, W, H, Br Exp 31 | St | Be | Fi | Sh | K | W | H | Br | ACC |
| | 90 | 0.93 | 0.84 | 0.87 | 0.96 | 0.93 | 0.84 | 1 | 0.65 |
| Pre-processing (PP) | | | | | | | | | |
| Fi, Sh, St, Be vs K, W, H, Br Exp 64 | St | Be | Fi | Sh | K | W | H | Br | |
| | 0.90 | 0.93 | 1 | 0.84 | 0.93 | 1 | 1 | 0.96 | 0.84 |
| Action Similarity (AS) | | | | | | | | | |
| FFi, Sh, St, Be vs C, Fe, Su, N Exp 64 | St | Be | Fi | Sh | C | Fe | Su | N | |
| | 0.87 | 0.90 | 93 | 0.90 | 1 | 1 | 1 | 0.93 | 0.75 |

Table 4.31: Appendix: 3DCNNsl 3 experiments, 4 vs 4 at 180-Epochs.

objectives in section 1.2.1.

4.16 Appendix: Fulfilling Research Question 5 (Can we determine model superiority between models?)

The artificial intelligence models’ true processing capabilities at this stage projected their classification results from a frame-by-frame processing level. Identifying model superiority proved challenging when one model (3DCNNsl) utilises the entirety of individual test videos to compute its inference and the other (YOLOv5m) applies frame-by-frame operations. To investigate this notion, a reconfiguring of 3DCNNsl inference evaluated action classes from a frame-by-frame level utilising the action similarity condition. Nevertheless, the idea establishes the actual processing power by implementing action similarity conditions to deliberately challenge the classification because of the focus on pre-empting violence and saving lives. The findings projected tabulated representations of the performance results to emphasise YOLOv5m and 3DCNNsl in scenarios conveying homogeneous and heterogeneous actions in real-world conditions.

4.16.1 Appendix: Overview of Superiority Regarding Processing

Testing the operation’s true processing capabilities by applying configurations to evaluate each classification from the frame-by-frame level compared to video-level processing examined superiority between YOLOv5m and 3DCNNsl. The approach provided crucial insight towards fulfilling

research question-5 in section 1.2.1. Initialising the processing scripts simultaneously on ten exclusive test videos proved crucial towards achieving the frame-by-frame outcomes. Adhering to consistency, integrating a frame extraction script to specifically select 12 frames in alignment with the class of activity template (CoAT) from each pre-processed video generated the outcomes for analysis. The method excluded beating, fighting, shooting, cutting-in-kitchen, nun-chucks, sumo-wrestling, walk-with-dog, and knitting to avoid the risks of exceeding the project’s life cycle. The idea allowed the investigations to focus on challenging actions (action similarity) utilising stabbing and fencing for frame-by-frame classification. Opting for the class reduction approach significantly reduced the analysis required to generate vital suggestive results. The actual performance discloses the operation’s pros and cons by implementing challenging actions from a frame-by-frame perspective. Moreover, implementing balanced ratios of violent and non-violent test videos to suggest 5-stabbing and 5-fencing categories reduced bias results. Those videos reflect stabbing8.avi, fencing11.avi, fencing12.avi, stabbing24.avi, fencing27.avi, fencing32.avi, stabbing37.avi, fencing38.avi, stabbing48.avi, and stabbing75.avi. The operations generated results for 119 frames, indicating both models’ processing limitations.

| Fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|---------------|----------|-----------|-----------|------------|---------------|
| 0 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.883333 | stabbing |
| 1 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.879167 | stabbing |
| 2 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.891667 | stabbing |
| 3 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.883333 | stabbing |
| 4 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.891667 | stabbing |
| 5 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.891667 | stabbing |
| 6 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.883333 | stabbing |
| 7 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.883333 | stabbing |
| 8 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.8875 | stabbing |
| 9 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.891667 | stabbing |
| 10 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.879167 | stabbing |
| 11 | stabbing8.avi | Stabbing | 0.9999702 | stabbing | 0.895833 | stabbing |

Table 4.32: Appendix: YOLOv5m-3DCNN Correctly Classified Stabbing8.avi.

4.16.2 Appendix: YOLOv5m3DCNNsl Accurate/ Partial Classifications Overview

With insight into the superiority evaluation previously discussed, the approach considered the processing of YOLOv5m and 3DCNNsl frame-by-frame action, classifying stabbing8.avi in Table 4.32,

stabbing48.avi in Table 4.33, and stabbing75.avi in Table 4.34. The analysis disclosed YOLOv5m efforts towards partially classifying fencing12.avi in Table 4.35 and stabbing37.avi in Table 4.36 with activity unknown outcomes. In contrast, 3DCNNsl correctly distinguished between fencing and stabbing; however, YOLOv5m’s misclassification via fencing11.avi considered violence only. Although YOLOv5m avoided misrepresentations by applying activity unknown to suggest that it cannot recognise the actions displayed, the model generated the correct response due to the homogeneous nature of fencing and stabbing via Table 4.35 and Table 4.36. The operations experienced YOLOv5m’s misclassification in stabbing24.avi Table 4.37, fencing32.avi Table 4.38, and fencing38.avi Table 4.39.

| Fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|----------------|----------|-----------|-----------|------------|---------------|
| 96 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.958333 | stabbing |
| 97 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.943056 | stabbing |
| 98 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.9375 | stabbing |
| 99 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.956944 | stabbing |
| 100 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.9375 | stabbing |
| 101 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.956944 | stabbing |
| 102 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.938889 | stabbing |
| 103 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.8125 | stabbing |
| 104 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.959722 | stabbing |
| 105 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.929167 | stabbing |
| 106 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.954167 | stabbing |
| 107 | stabbing48.avi | Stabbing | 0.9999722 | stabbing | 0.922222 | stabbing |

Table 4.33: Appendix: YOLOv5m-3DCNN Correctly Classified Stabbing48.avi.

4.17 Appendix: Summary of 3DCNNsl Misclassifications Perspective

Following the superiority evidence in Appendix 4.16, further evaluations on 3DCNNsl determines its effectiveness by aligning it with pros and cons. 3DCNNsl misinterpretations on fencing as stabbing activity in Table 4.40 proved interesting. Whilst 3DCNNsl appeared accurate as some fencing actions are homogeneous compared to stabbing, YOLOv5m correctly identified identical attributes with the label activity unknown. In some cases, the model accurately classified fencing’s actions regardless of the output of the 3DCNNsl model. YOLOv5m’s robust classification corroborated its superiority over the state-of-the-art 3DCNNsl; nevertheless, it projected multiple activity unknown scenarios. The analysis proved that the difference in classification between the models is minute.

| Fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|----------------|----------|------------|-----------|------------|---------------|
| 108 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.943056 | stabbing |
| 109 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.893056 | stabbing |
| 110 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.95 | stabbing |
| 111 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.929167 | stabbing |
| 112 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.825 | stabbing |
| 113 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.940278 | stabbing |
| 114 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.95 | stabbing |
| 115 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.954167 | stabbing |
| 116 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.911111 | stabbing |
| 117 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.893056 | stabbing |
| 118 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.908333 | stabbing |
| 119 | stabbing75.avi | Stabbing | 0.99994826 | stabbing | 0.922222 | stabbing |

Table 4.34: Appendix: YOLOv5m-3DCNN Correctly Classified Stabbing75.avi.

| Fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|---------------|---------|------------|------------------|------------|---------------|
| 24 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 25 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 26 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 27 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 28 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 29 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 30 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 31 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 32 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 33 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 34 | fencing12.avi | Fencing | 0.99565816 | Activity Unknown | 0 | fencing |
| 35 | fencing12.avi | Fencing | 0.99565816 | fencing | 0.708333 | fencing |

Table 4.35: Appendix: YOLOv5m Partially Classified Fencing12.avi.

Regarding the minute difference mentioned, in terms of life, this difference can be a measure between non-lethal scenarios and death. With this notion, the thesis proposed the fusion concepts as a robust tool towards mitigating misclassification challenges attained during action similarity scenarios regardless of the environmental conditions.

4.18 Appendix: Summary of YOLOv5m Operational Challenges

(A) YOLOv5m Steep Learning Curve Challenge Discussions: The investigations required substantial time to understand the framework components during development. The process

| fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|----------------|----------|-----------|------------------|------------|---------------|
| 72 | stabbing37.avi | Stabbing | 0.9999689 | stabbing | 0.5625 | stabbing |
| 73 | stabbing37.avi | Stabbing | 0.9999689 | Activity Unknown | 0 | stabbing |
| 74 | stabbing37.avi | Stabbing | 0.9999689 | stabbing | 0.5625 | stabbing |
| 75 | stabbing37.avi | Stabbing | 0.9999689 | stabbing | 0.558333 | stabbing |
| 76 | stabbing37.avi | Stabbing | 0.9999689 | Activity Unknown | 0 | stabbing |
| 77 | stabbing37.avi | Stabbing | 0.9999689 | stabbing | 0.558333 | stabbing |
| 78 | stabbing37.avi | Stabbing | 0.9999689 | Activity Unknown | 0 | stabbing |
| 79 | stabbing37.avi | Stabbing | 0.9999689 | stabbing | 0.470833 | stabbing |
| 80 | stabbing37.avi | Stabbing | 0.9999689 | Activity Unknown | 0 | stabbing |
| 81 | stabbing37.avi | Stabbing | 0.9999689 | Activity Unknown | 0 | stabbing |
| 82 | stabbing37.avi | Stabbing | 0.9999689 | Activity Unknown | 0 | stabbing |
| 83 | stabbing37.avi | Stabbing | 0.9999689 | Activity Unknown | 0 | stabbing |

Table 4.36: Appendix: YOLOv5m Partially Classified Stabbing37.avi.

| Fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|----------------|----------|------------|------------------|------------|---------------|
| 36 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 37 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 38 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 39 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 40 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 41 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 42 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 43 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 44 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 45 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 46 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |
| 47 | stabbing24.avi | Stabbing | 0.99986625 | Activity Unknown | 0 | stabbing |

Table 4.37: Appendix: YOLOv5m misclassified Stabbing24.avi.

considers interpreting behavioural patterns/results during/after processing and solving artificial intelligence script errors dispensed. YOLOv5m required significant amounts of time to understand the practical concepts of its application in the context of object detection and its feasibility towards current activity recognition endeavours.

Solution to (A), YOLOv5m Steep Challenges: The previous challenge in (A) demonstrated a lesser impact by enrolling on two additional artificial intelligence courses (scripting/development) to bridge the gaps in understanding and fortify the need for more knowledge regarding practical experience. Processing investigation strategies via Robo-Flow, Ghit-Hub, Discord and

| Fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|---------------|---------|-----------|------------------|------------|---------------|
| 60 | fencing32.avi | Fencing | 0.9997483 | fencing | 0.3 | fencing |
| 61 | fencing32.avi | Fencing | 0.9997483 | fencing | 0.3125 | fencing |
| 62 | fencing32.avi | Fencing | 0.9997483 | fencing | 0.283333 | fencing |
| 63 | fencing32.avi | Fencing | 0.9997483 | fencing | 0.3125 | fencing |
| 64 | fencing32.avi | Fencing | 0.9997483 | fencing | 0.258333 | fencing |
| 65 | fencing32.avi | Fencing | 0.9997483 | fencing | 0.3 | fencing |
| 66 | fencing32.avi | Fencing | 0.9997483 | Activity Unknown | 0 | fencing |
| 67 | fencing32.avi | Fencing | 0.9997483 | Activity Unknown | 0 | fencing |
| 68 | fencing32.avi | Fencing | 0.9997483 | Activity Unknown | 0 | fencing |
| 69 | fencing32.avi | Fencing | 0.9997483 | Activity Unknown | 0 | fencing |
| 70 | fencing32.avi | Fencing | 0.9997483 | Activity Unknown | 0 | fencing |
| 71 | fencing32.avi | Fencing | 0.9997483 | Activity Unknown | 0 | fencing |

Table 4.38: Appendix: YOLOv5m misclassified fencing32.avi.

| Fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|---------------|---------|-----------|------------------|------------|---------------|
| 84 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 85 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 86 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 87 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 88 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 89 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 90 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 91 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 92 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 93 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 94 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |
| 95 | fencing38.avi | Fencing | 0.9991573 | Activity Unknown | 0 | fencing |

Table 4.39: Appendix: YOLOv5m misclassified fencing38.avi.

PyTorch communities provided the means to acquire expertise on unknown systematic errors. Reducing development time depended exclusively on grasping the fundamental AI concepts and deploying such ideas to fortify the fusion concept.

(B) Discussions on YOLOv5m Data and Processing Issues: The context of violent data containing relevance of the pre-start, middle and end attributes in publicly available datasets proved challenging to obtain. Because applying fake data (movies, acting violent scenes) during training operations could contribute to erroneous results, a manual acquisition technique accumulated a significant sample size of raw data reflecting the pre-start attributes of

| Fr# | Video | 3DPreds | 3DScores | YoloPreds | YoloScores | Correct_Class |
|-----|---------------|----------|-----------|------------------|------------|---------------|
| 48 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |
| 49 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |
| 50 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |
| 51 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |
| 52 | fencing27.avi | Stabbing | 0.9006899 | fencing | 0.266667 | fencing |
| 53 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |
| 54 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |
| 55 | fencing27.avi | Stabbing | 0.9006899 | fencing | 0.583333 | fencing |
| 56 | fencing27.avi | Stabbing | 0.9006899 | fencing | 0.266667 | fencing |
| 57 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |
| 58 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |
| 59 | fencing27.avi | Stabbing | 0.9006899 | Activity Unknown | 0 | fencing |

Table 4.40: Appendix: 3DCNN misclassified fencing27.avi.

violent scenarios. The acquisition and data pre-processing took considerable time because of inadequate experience manipulating software for blob analysis and data cleaning. The issue impacted the project’s overall timeline. The notion stimulated the manual pre-processing of the data, as most automated options unintentionally introduced unwanted regions of interest for training. Investigations into the applicability of software tools that naturally defined bounding box regions circumventing the object’s edges proved futile. The previous approach proved fatal due to the algorithm’s unconscious ability to accumulate undesirable background elements as the target objects. The pre-processing duration to conform unwanted redundant frames and extensive file sizes affected research questions 2-3 in section 1.2.1. The issues surrounding processing encouraged high over-fitting risks, leading to unrealistic results if left unchecked. Annotating irregular dimensional scales during blob analysis intensified the processing challenges. The sporadic contours of the human gait during violence increased the challenge of specifying bounding box edges for each region of interest per class and per image frame to suggest the action category.

Solution to (B), YOLOv5m Data and Processing Issues: The data acquisition and pre-processing challenges were regulated by downloading the relevant raw data samples from YouTube, Facebook, and other Social Platforms and merging a few samples acquired from publicly available datasets. Once achieved, a data screen operation confirmed the specificity of the raw data towards meeting the framework’s standards and the pre-start attributes for investigations.

Understanding blob analysis software and manipulation is necessary to fashion the appropriate data context relative to experiments. The context of the manual pre-processing approach provided stability towards reducing the risk of adding unwanted attributes in the image scenery background during blob analysis. Although the processing proved lengthy, the operations avoided the risks of training the model with noisy data (containing undesirable details) that generates unrealistic bias results compared to the automatic blob analysis tools. The intention is to apply the object's mask during blob analysis as a future endeavour to mitigate the irregularity of the human gait issue.

(C) Discussions on YOLOv5m GPU Non-compliance: With a lack of sufficient GPU processing, YOLOv5m experienced a computational lag due to the incompatibility of the PyTorch platform utilities with the MacBook's processing hardware. The data increase introduced memory depletion challenges, which affected YOLOv5m's ability to converge its training in an acceptable period utilising CPU processing.

Solution to (C), YOLOv5m GPU Non-compliance: A duration assessment task reduced the lengthy raw video samples containing sporadic violent activity across several sequences from 2-10 minutes to 5-15 seconds per file at a rate of 30fps (frames per second). The original data extended the training sequence durations and negatively impacted the hardware's processing. By reducing the number of redundant and unwanted frames, the processing avoided the risks of immobilising the operation with inconclusive and insignificant results. Observation proved that the model's computational demand on the hardware exhausted its capacity when processing image dimensions exceeding ratios of 641x641 during training procedures. A dimensional assessment approach facilitated the deficiency issues by regulating the data specifications to 412x412. The approach positively impacted the training operations, reducing the convergence of the results from 2 weeks to 1.5 hours maximum. Simultaneously, the approach reduced the computational load, leading to predictions in seconds compared to 2-72 hours. Although the new hardware contains PyTorch metal performance GPU capability, some PyTorch packages lack full compatibility with the MacBook's new silicon processing chip models. Constant monitoring procedures target the PyTorch 3 support communities for platform and software upgrades to mitigate the onboard GPU challenge. Though free

GPU online services exist, their free processing aligns with costly stipulations that could harm the processing if the services cease during operations due to power failure and service provider disconnection disruptions relative to geographical boundaries. Other entities, such as Google-Collab, stipulated strict timelines for their integration, thus creating an annoyance with periodic operating system reassessments and configuration file management requirements. The issue resulted in the laborious task of commencing the entire scripting operation at intervals from the start. As a future endeavour, the primary intention is to target the integration of GPU processing support from online service providers relative to its interoperability, feasibility, and cost-effectiveness.

(D) YOLOv5m CCTV Sensor Field of View issue: In many instances, the operations experienced scenarios in which violence evaded the CCTV sensor’s field of view. By impairing the viewpoint of the CCTV device, the actual status of an attack remained unknown until the visuals of the actions returned within the device’s view. The challenge affected the severity of the classification and prolonged the actual status of the activities performed.

Solution to (D), YOLOv5m Field of View Challenge: The operations experienced a minimised impact of the field of view challenged by merging the proposed fusion idea with the new drone technology. Data generation utilising an onboard camera sensor performs the same prediction tasks as motionless CCTV devices. The idea allows the capture of critical data for processing through wireless connectivity that accesses the device’s framework controls and stores its data for processing via a mainframe from another geographical location within the drone’s operating capacity. A proposed strategy of disposing of unwanted data after convergence drastically reduces the drag on storage resources. The approach maintains functionality and effectiveness during operations. Although drone technology has disadvantages, its influence does not affect the current proposal. Thus, its feasibility will require investigation to obtain a suitable approach. The idea of applying innovative drones in this regard increases the application possibilities, extending its mobility and expanding the scope of the CCTV sensor device from aerial perspectives. Maximising the peripheral view and positioning of the sensors ensures that sporadic actions are within sight for classification at all intervals.

(E) Discussions on 2 Violent Classes Appose to 8: Because of the stipulated research timeline, the investigations on the number of violent and non-violent action classes proved critical to determining the impact of the results. At this stage, the complexity and depth of the experiments have proven challenging in thoroughly analysing class variations. The previous issue directly affected the time allocated to explore all specified violent and non-violent action classes to accentuate and interpret their significance.

Solution to (E), Discussions on 2 Violent Classes Appose to 8: As discussed in E, critical class input decisions reduce the focus on violent and non-violent classes from eight to two categories (stabbing and fencing). The rationale analyses the most complex conditions of action similarity to generate significant results in this domain. By reducing the focus classes, as previously mentioned, the scope of analysis decreased to allow effective investigations with contributory results.

(F) Discussions on YOLOv5m Human Dependency issue: Occasionally, YOLOv5m produced prediction anomalies (unable to discern the status of the action) that escalate system alerts. Simultaneously, the issue introduces irritations by generating excessive alerts due to misclassifications. Human support is required to assist with the classification anomaly distinction and provide the correct assessment measures.

Solution to (F), YOLOv5m Human Dependency issue: The system's configuration automatically escalates alerts to humans if the acceptable number of false positives/predictions alters as scenarios are life-dependent. In this regard, having a contingency plan incorporating additional human evaluation support to preserve life is promoted.

4.19 Summary of 3DCNNsl Operational Challenges

Following YOLOv5m operational discussions, the analysis emphasised additional challenges utilising 3DCNNsl during development with clear mitigation strategies. The approach is as follows.

(A) 3DCNN Language Barrier Issue Discussions: The first challenge emerged as a language

barrier whilst analysing configuration files during model acquisition operations. The configuration files written in [118]’s native language, Mandarin, proved tricky to interpret. Translating Mandarin during development proved extremely challenging to comprehend because of language barriers. The issue intensified the challenge of deciphering fundamental stepwise instructions to modify configuration parameters and reduce error prompts. Even though Google and MacBook translators were applied, some critical phrases required a more precise understanding as the instructions were not in the correct context. The issue significantly extends the timeline for comprehending its processing components thoroughly. An example of the language barrier issue affected fine-tuning features in the artificial intelligence framework, causing undecipherable errors.

Solution to (A), 3DCNN Language Barrier Issue: The first issue was mitigated by processing the Mandarin instructions using Google, open-sourced translators and the MacBook Pro Itranslate software to convert the core text to English. Correcting the misinterpretation required significant daily devotion to generating an understanding of 3DCNNsl standard requirements via practical exercises.

(B) Discussions on 3DCNNsl Padding Feature Issues: An outdated padding-valid script mismatch on two max pool layers emerged during development. The issue drastically reduced the image sizes beyond the capability of the convolution operations. The computation process generated a zero result due to an unfavourable border reduction within the image data. The issue reduced the image dimensions during generalisation and discarded padded border element options. As a result, padding-valid limited the model’s capability to discern the correlated labels, features, and coordinates of the objects of interest.

Solution to (B), 3DCNNsl Padding Feature Issues: Reconfigurations to [118]’s 3DCNN approach with additional level support reflecting 23 layers, including two additional max-pooling levels to encourage layer proficiency during training. Substituting the outdated processing (padding-valid) and feature limitation with (padding same) encouraged processing effectiveness. The modification reintroduced padded borders of zeros around the images during processing to encourage efficient convolution and generate robust results.

4.20 Summary of Observations and Lessons Learnt

Following the operational discussions for YOLOv5m and 3DCNNsl, this section outlined items concerning performance, development complexity, and processing constraints. Because YOLOv5m and 3DCNNsl relied on different operations, applying subcategories to maintain model distinction towards activity recognition proved effective. The categories itemised as (A) suggest lessons learnt from the perspective of YOLOv5m, and (B) for the context of 3DCNNsl demonstrate the processing distinctions. Subsequently, positive observations represent items that influenced the model's operations, and negative remarks describe the opposite.

(A) Positive Observations and Lessons Learnt Utilising YOLOv5m: Although formidable for object detection, YOLOv5m proved solid in its processing compared to 3DCNNs for activity recognition from a knowledge perspective. If more data is applied to enhance 3DCNNsl's learning, its framework will likely be less robust. 3DCNNsl projects its outcome considering the entirety of the video, as opposed to classification from a frame-by-frame basis like YOLOv5m activity recognition. Analysis favoured YOLOv5m processing over 3DCNNsl because it considers the frame and not the entire video duration to formulate probability scores. The thesis proposed achieving enhanced classification by combining activity and weapon artefacts to suggest the pre-empting of stabbing. With that notion, additional pre-processing to enhance inference is crucial towards fortifying performance. A method disclosed during development involves CPU boosting by introducing the right packages and platform, applying standard dimensionalities and pre-processing, and using significant sample sizes of real-world violent scenarios for training. The possibility of conforming YOLOv5m object detection to accurately classify the complexity of violence as activity recognition proved a valuable prospect. The idea reduces the configuration intricacy and encourages acceptable processing with 3DCNNsl.

Negative Observations on (A), Utilising YOLOv5m: Understanding the intricacies of the approaches for YOLO required time to accumulate practical experience. Because of its complexity, comprehending most error messages for the given objectives proved highly challenging. Rapid architecture upgrading counteracted processing challenges affecting older versions,

thus causing an implementation issue reverting to lower model versions. YOLOv5 faced limitations when classifying specific weapon objects, such as knives conveying high sporadic trajectory, acceleration, and velocity. During development, the larger versions of YOLOv5 produced higher results. However, when evaluating action similarity with real-world lethal scenarios, the model required a boost in robustness, thus hindering its classification state. Because of the high regard for human life, the application of fictitious data proved impractical as it escalated to over-fitting with unrealistic results. Separately specifying relevant objects during blob analysis with and without pre-processing support negatively impacted the inference, thus leading to further issues towards alternative support. Another challenge emerged when applying CPU processing compared to GPU on differing image dimensionalities. The impact reduced the performance significantly, and, in several cases, the relevance of its output proved futile. Selecting incompatible library package updates and support platforms (Anaconda, PyCharm, Spyder, MATLAB, Terminal) caused systematic challenges, resulting in unknown and insoluble errors. With an overview of the issues disclosed during development, the following discussions emphasised 3DCNNsl’s positive observations.

(B) Positive Observations and Lessons Learnt Utilising 3DCNNsl: Understanding the input-output processing between the adjacent layers to foster high-level feature learning proved highly insightful during convolution. Acquiring knowledge on correctly altering the number of layers in [118]’s frameworks to facilitate the research initiatives proved challenging. Analysis showed that 3DCNNsl engaged action similarity conditions with high robust accuracy ratings at the video level, utilising 23 layers compared to [118]’s 16.

Negative Observations on (B), Utilising 3DCNNsl: The magnitude of the research exceeded the expected analysis threshold. Because of time constraints, the scale of analysis was reduced from 8 classes to 2 to alleviate the risks of exceeding the project’s lifecycle. The current strategy entails continuing analysis as future endeavours and, at this stage, focuses on two categories for action similarity via stabbing and fencing. During development, observations disclosed a decline in CPU processing speed, primarily during the integration of the volume of samples. With insight into YOLOv5m’s processing, a GPU reverting issue emerged, negatively impacting the training and inference speeds. The endeavour encom-

passed exploring GPU integration to facilitate future initiatives as an alternative strategy. Although the operations yield acceptable outcomes using CPU processing, the persistence of the latency issue poses a high risk where human life is concerned during inference. Another factor became apparent when scripting generic status classification from the dataset level to encourage robust performance. However, that approach intensified the configuration complexity to project individual class analysis via the confusion matrix. During development, analysis showed that 3DCNN without modification lacked robustness in classification from a frame-by-frame viewpoint because of its design.