Efficient Monitoring of Illicit Activities: Identifying Smokers through Human Activity Recognition

by

Nafiun Al Amin
20201069
Ayan Haider
20201152
Tasmia Tarannum Naomi
20241013
Rafid Sadman Rahman
23141033
Nusaiba Zaman
20201104

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
July 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nafiun Al Amin

20201069

Ayan Haider

20201152

Tasmia Tarannum Naomi

20241013

Rafid Sadman Rahman

23141033

Nusaiba Zaman

20201104

Approval

The thesis/project titled "Efficient Monitoring of Illicit Activities: Identifying Smokers through Human Activity Recognition" submitted by

1. Nafiun Al Amin (20201069)
2. Ayan Haider (20201152)
3. Tasmia Tarannum Naomi (20241013)
4. Rafid Sadman Rahman (23141033)
5. Nusaiba Zaman (20201104)

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on July 14, 2024.

Examining Committee:

Supervisor:
(Member)

Md. Tawhid Anwar

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Md. Sabbir Ahmed

Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The use of security surveillance systems is becoming more prevalent, causing them to become an excellent tool for major crime prevention. However, niche crimes such as the use of cigarettes and other forms of smoking devices in non-smoking environments go unnoticed due to the range of motions required to perform smoking being very minimal. The problem is not only smoking but also the littering that follows afterward. Hence, our research implemented several models that can recognize smoking-related aspects. With these models, it would be possible to detect smokers through surveillance systems and prevent such niche crimes without requiring extensive manpower. In this research, we have used customized visual data to train our models, consisting of videos of smoking from different angles, lighting conditions, and population densities. Object detection models such as YOLOv5, YOLOv8, YOLOv9 and YOLOv10 were used to detect objects related to smoking and classify people as smokers, with YOLOv9 achieving the best results at 96.2% Precision and 76.3% Recall scores. Video Classification was also performed using YOLOv8 to recognize the different features in frames that constitutes to a person being a smoker which achieved a Precision of 90.6% and Recall of 94.2%.

Keywords: Security Surveillance System; Cigarettes; Smoking; Non-Smoking Environment; Recognize; Human Activities; Object Detection; Video Classification; YOLO;

Acknowledgement

We would like to extend our highest gratitude to the Almighty Allah, for allowing us to pursue our bachelor's degree at BRAC University.

Our academic endeavors have been made possible due to the support of many people, and we are deeply grateful to them. Firstly, we thank our parents for their constant encouragement and resources throughout our bachelor's degree and thesis. Their unwavering support provided a strong foundation of motivation that sustained us throughout our work.

We express our sincerest gratitude to our Supervisor, Md Tawhid Anwar, and Co-supervisor, Md Sabbir Ahmed, who have been the backbone of our work. Their continuous guidance and the resources they provided have been instrumental in advancing our knowledge. Their valuable feedback has significantly polished our work, making completing our thesis possible.

We also thank our peers, friends, and family members who have actively assisted in our data collection process. Their contributions have been a significant help in our research.

Lastly, we thank our team members for their constant efforts throughout this thesis. Completing this thesis in due time would not have been possible without our collective efforts, effective communication, and coordination.

We dedicate this thesis to all our helpful well-wishers.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Smoking has been a persistent problem in society for decades, maintaining popularity after constant awareness campaigns and imposed regulations aiming to decrease the consumption of cigarettes or other smoking devices such as vape to prevent the side effects of smoking on oneself and society. Smoking has caused a rise in health problems such as asthma, diabetes, and fatal diseases such as lung diseases, heart diseases, strokes, and cancer. However, smoking is not only harmful to the smoker but also to the people in the surroundings, this is known as passive smoking.

According to a recent article, [1] talked about the dangers and effects of passive smoking in children. Children who are exposed to second and third-hand smoke have an increased risk of diseases and early death as they have smaller airways than adults. The article also talks about how exposure to second and third-hand smoke can affect a child's developing brain. Hence, smoking inside schools, daycare centers, or children's hospitals may be harmful and indirectly contribute to harming children and future generations. Moreover, the third highest cause of fires in Bangladesh has been from carelessly discarding cigarette butts according to [2]. Fires caused by cigarette butts have accounted for 14.78% of total fires in the year 2023 in Bangladesh, becoming a major and avoidable issue. Thus, smoking cigarettes or vapes in public places such as schools, hospitals, and the workplace amongst many others should be banned to eliminate the problems faced such as passive smoking,

accidental fires, and many more.

However, the current methods of detecting the act of smoking may not be as efficient in recent times and thus, prone to neglect or human error. To combat the issues caused by smoking, technological advancements in the field of Deep Learning will be used to implement an efficient smoke monitoring system. To implement this system, we will first employ object detection to identify smoking objects such as cigarettes, vapes, the smoke and people will be detected. Using these we will then classify the smokers. The proposed architectures, such as the different models of YOLO will be used to compare between them and check which model achieves the highest Precision and Recall scores. Video Classification can also be performed to check whether the YOLO architecture can understand the features that are present when someone is smoking.

1.1 Research Problem

With the help of computer vision and deep learning, there have been significant advancements in the detection of illicit activities, however, there is a lack of research on the detection of smokers through the lens of Closed-Circuit Television (CCTV). As of 2021, the number of smokers had reached a record high of 1.1 billion according to [3]. Furthermore, there has also been a rise in E-cigarette smokers as stated in [4]. Smoking has always been a public health concern and having the ability to detect smokers through CCTV would prove to be incredibly beneficial for hospitals, schools, parks, and anywhere where smoking is not allowed. That too without the need for increased manpower.

However, training a model for this specific purpose brings about certain difficulties. One of which has to do with the datasets to train a model. Regarding CCTV footage, not all the samples will be the same, there will be footage from different camera angles, lighting conditions, occlusions, and population densities. Addition-

ally, acquiring labeled data poses a challenge due to its scarcity. Hence, requiring a more dynamic range of solutions to traverse said challenges.

Moreover, it is important to understand which architecture would be the best and the optimizations required to detect smokers in real-time accurately. The optimizations would have to be based on the size and speed of the models. To address this, past research will have to be looked into to get a better idea.

According to [5], it is highly crucial to have digital security surveillance instead of the traditional, manual surveillance that we see all around us. His paper proposes the idea of detecting guns using YOLOv3 models. In another case, [6] proposes the concept of accuracy in detecting small objects using R-CNN. Again, another paper [7] suggests a way to detect small objects using the YOLOv4 algorithm. While there is much literature about object detection, none specifically addresses smoking devices, which we are trying to address in our paper.

On top of that, [8] emphasizes smoke detection using the YOLO model. Another study on the detection of smoke in early wildfires was conducted in [9]. Multiple studies deal with the detection of smoke which we can utilize. We plan to test different versions of YOLO algorithms, along with modifications suggested in the papers, to detect smoke coming out of a smoking device.

Apart from these, we also plan to detect the activity of a human. [10] proposed a technique to effectively extract information from videos to identify human actions, which can be used to detect and movements done by someone who is either trying to smoke or is smoking under our digitized security surveillance.

In our paper, we plan to detect multiple objects in a frame to classify a person as a smoker and also implement video classification techniques to detect smoking.

With this in mind, the question arises about how we can implement all three different functions efficiently, compare different models, determine which model is providing us with the highest accuracy and confidence scores, and whether any modifications serve us better with the results.

Therefore, in this research, the question that we plan to answer is:

Which of the versions of YOLO is the most effective to classify a person as a smoker from the customized dataset created for this research and can any modifications to them produce better results?

This research will try to answer all the questions mentioned above, find an effective way of implementing the models, and compare different models along with modifications to detect smoking most efficiently.

1.2 Research Objectives

The purpose of this research is to develop an effective system to detect smokers through CCTV using different models of YOLO. The detection has to be done in real-time, and the model has to distinguish between similar gestures and objects. The objectives of this research are:

1. To understand the problem regarding smoking.
2. To understand the scope of different deep learning models.
3. To create a dataset depicting real-life smoking scenerios imitating a CCTV camera.
4. Detect smokers through multiclass object detection(person, smoke, cigarette, and vape) along with a logical reasoning.
5. Detect smokers using video classification.
6. To recommend further improvements.

Chapter 2

Literature Review

Despite being a major health concern, smoking continues to be a consistent habit worldwide for a significant amount of people. In an article published by the Guardian, a worldwide statistics study on smoking displayed that the number of smokers had reached an all-time high of 1.1 billion in 2021 [3]. The same study showed that within 9 years the number of smokers had increased by 150 million new smokers.

To explain such a large increase it is easy to simply look at the increase in population growth with each year; according to the same study too, the most vulnerable demographic to becoming smoking addicts were young people with up to 89% of new smokers having become addicted within the age of 25. The exponential growth of the human population will only result in this phenomenon occurring more frequently and increasing the number of new smokers even more.

2.1 Adverse Health Effects of Smoking

Smoking as an activity causes multiple different foreign chemical reactions within the body. According to [11], these reactions generate more than 7000 different chemicals with the majority being toxic, and with a minimum of 69 being cancer-causing

carcinogens. Among the long-term effects on the body, some of the major ones include increasing the risk of cancer across all organs, causing permanent lung damage, weakening the body's immune system, increasing chances of negative mood swings, increasing the chances of birth defects, and many others.

2.2 Adverse Health Effects of Second-Hand Smoke

According to [1], second-hand smoke refers to both smoke from cigarette butts and exhalants from smokers. Both types contain harmful toxic chemicals that remain in the environment long after smoking has ceased and they are especially harmful to babies and children with their smaller airways and weak immune systems. Such harmful effects include but are not limited to increasing the risk of sudden infant death syndrome (SIDS), increasing the likelihood of brain defects, long-term lung diseases like asthma or bronchitis, and increasing the likelihood that the child becomes a smoker later in life.

2.3 Related Works

Narejo et al. (2021) [5] highlighted in their paper the importance of having automated security surveillance instead of the traditional, manual surveillance that we see all around. In this paper, they have come up with a computer-friendly fully automated system that detects objects like guns, or rifles. In their procedure, they have used YOLOv3 trained on their customized dataset. Firstly, their approach is divided into three parts - object detection, analysis, and action. Firstly, it'll take input from the video which passes through frame conversion and then the image is processed. After that, the object is detected and later identified. The detected frame

is passed through trained data and classification, which detects if the object is a gun. If a gun is found, its location is found and it alerts the database. For this approach, they manually collected a large number of photos from publicly available sources on Google. It should be noted that the images they have taken were converted to "jpg" format and then resized to 416x416. In this paper, experiments were also done with YOLOv2, YOLOv3, and traditional CNN, and found that YOLOv3 tops them all, with an accuracy of 98.89%. In their research, they have also successfully created a database in the backend, which stores the latitude, longitude, incident time, and the location of the gun detected. Their proposed model with YOLOv3 is also compared with CNN VGG-16, Faster RCNN, Alexnet + SVM and found out that their proposed approach got 98.89% accuracy, topping the chart above all.

A method was stated by Fang and Shi (2018) [6] to improve the accuracy of detecting small objects in images. Their method used a more flexible content information fusion approach which is based on Faster R-CNN while also in tandem with Feature Pyramid Networks (FPN). They used the COCO dataset from which they classified the small objects in the dataset. Their approach achieved an mAP of 23.21 points which compared to the base model had a significant increase of 6.97 points while AP increased to 12.3 points for 3.25 points. Furthermore, their approach achieved 28.9 points for AR on small objects which is an increase from 13.0 of the base model. Therefore, using the flexible context information fusion method they were able to improve the accuracy in detecting small objects.

In a study, Zhu et al. (2021) [7] proposed a deep-learning algorithm to identify small objects in the streets during extreme weather conditions such as fog and rain. The algorithm they used is the improved YOLOv4-Tiny network which contains CSPDarknet53-Tiny, FPN, and YOLO Head. For clear images, they used the VOC dataset and the COCO dataset, while for the images with fog and rain, they used the RTTS dataset and images extracted from the web. The algorithm had to cor-

rectly identify between Car, Truck, Bicycle, Motorbike, and Person. The model used by the paper which was improved by YOLOv4-Tiny increased AP values of Bicycle, Motorbike, and Person by 5.05%, 5.64%, and 1.99% respectively from the non-improved best model while the overall mAP value increased to 68.88% from 66.72%. Hence, the enhanced YOLOv4-Tiny algorithm notably boosted the performance of small object recognition.

Cai et al. (2020) [8] focused on applying a deep learning algorithm that is faster and has high accuracy in the detection of smoke in complicated environments. The deep learning algorithm they used was the YOLO-SMOKE model which is based on YOLOv3. The model was trained by sorting out public smoke datasets combined with MSCOCO datasets and tested on pictures from the public datasets. The YOLOv3 model was trained on MSCOCO. The mAP achieved using their improved YOLOv3 model during initial testing was 86.86%, which was 4.91% better in performance than the original model. The model was further tested on smoke video datasets created by Bilkent University in Turkey and was able to achieve an average accuracy of 98.75% in identifying smoke and 98.89% in identifying non-smokes. With their proposed model, they were able to achieve higher accuracy, more robustness, and efficiency in the detection of smoke in complicated environments compared to previous algorithms used on the same datasets.

A research on the detection of smoke in early wildfires was conducted by Al-Smadi et al. (2023) [9] compares the YOLOv3, YOLOv5, and YOLOv7 with previous models such as Fast R-CNN and Faster R-CNN, to find the best model for detecting early smoke in wildfires. The original dataset used was found from kaggle.com and alongside the original dataset, new images were formed by using data augmentation methods, increasing the training sample size. The YOLOv5x model had the highest precision at 0.96027 and an mAP at IoU of 50% at 96.8%, however, it had the slowest rate of detection compared to other previous models. The YOLOv7 had an mAP

at IoU of 50% accuracy of 95.08% and YOLOv3 had 94.8% while Fast R-CNN and Faster R-CNN had mAP at IoU of 50% accuracy at 68.3% and 70.6% respectively.

A technique was put forward by Fu et al. (2018) [10] to effectively extract information from videos to identify human actions. To achieve this goal they used a combination of algorithms like K-means, LSTM auto-encoder, and 3D-CNN which was improved using the recursive feature elimination (RFE) algorithm. The dataset they used was UCF101, which was classified using SVM. Their approach achieved an overall accuracy of 75.4% but for some specific actions, it achieved 100% accuracy. However, their accuracy is lower than a few of the other methods as their approach had a 1% lower accuracy compared to the best approach. The reason for their lower accuracy could be because their method only trained on gray pixels and their 3D-CNN was not trained on their dataset but used parameters from pre-trained data. Overall, even though they achieved a slightly lower accuracy, they did however simplify the process of human activity recognition.

In their paper, Jiang et al. (2020) [12] focused on using a YOLOv4-tiny-based real-time object detection algorithm for embedded devices. The paper comes up with an efficient way of using YOLO-v4-tiny. Instead of two CSPBlock modules in YOLOv4-tiny, this method uses two ResBlock-D modules in the ResNet-D network. Several tests were conducted in this research to compare the suggested method with YOLOv4-tiny and YOLOv3-tiny. They've used MS COCO for training and testing to accomplish this (p. 6). They have found their proposed method giving 294 fps, which is the largest, followed by YOLOv3-tiny, with 277 fps, and YOLOv3-tiny, which has 270 fps. Additionally, they researched mAP, which revealed that YOLov4-tiny gives the highest mAP percentage at 38.1%, followed closely by their proposed method, at 38.0% (p. 7). Furthermore, they have also conducted tests on randomly selected images from the MS COCO test set and compared the object detection results of their proposed method with YOLOv4-tiny. In their tests, both

seemed to detect the same objects with some differences in the confidence scores. For most cases, they found out that the confidence scores obtained by their method were higher than YOLOv4-tiny, in fact, in only 5 of the 22 test scores, YOLOv4-tiny had a higher confidence score than their proposed method. They showed that their proposed method is faster and more accurate than YOLOv4-tiny and YOLOv3-tiny, making it more suitable for object detection.

In the research work by Saric et al. (2020) [13] deep learning algorithms were applied. These algorithms are remarkable at automatically identifying patterns of interests and distinct features of objects, for correctly classifying objects. The algorithms used were Faster R-CNN and YOLO, which were first trained on ImageNet and COCO datasets and then tested on the GRAZ-02 dataset. The algorithms had to identify between bikes, cars, and pedestrians. The accuracies achieved by Faster R-CNN were 93%, 95%, and 91% respectively, while for YOLO the results were 96%, 94%, and 98% respectively. They showed that YOLO had achieved a greater overall accuracy over Faster R-CNN while also having significantly lower runtime during prediction.

Ullah (2020) [14] developed an object-detecting approach that doesn't require a GPU. In his approach, YOLO, along with OpenCV, is optimized in a manner where real-time object detection is possible even in CPU-based computers. In this paper, they have used COCO for datasets and YOLOv3 which is loaded by OpenCV framework. In the proposed approach, although higher than the DarkFlow, but still very low. Thus, to increase the FPS, the Blob size of the input video was reduced. The paper found out that in 90x90, detection was fairly good with the fps above 10. In 64x64, the fps was even better but the results were slightly wrong for tinier objects. In this paper, this approach was tested on several computers to obtain the best blob size for the experiment. The results show that AMD Ryzen 3 2200G with a CPU of 3.50GHz, with a RAM of 8GB had an mAP of 31,05% and FPS of 16. This result, in

actuality, is good enough for low-configured computers based on CPU. The author also recommends future works using customized dataset training for better fps and accuracy with this model.

Wang and Wang (2023) [15] in their research designed a model to improve the effectiveness of Feature Pyramid Networks (FPN) for object detection. Their proposed method involves lessening the loss and misalignment of semantic information in FPNs during multi-scale feature fusion by using an improved bottom-up scaling approach with multi-scale residual aggregation (MSRA). The MSRA-FPN method was tested on PASCAL VOC and MS COCO datasets, as well as the Thangka Figure dataset which was constructed for this research. For the PASCAL VOC dataset, the RetinaNet object detection model coupled with MSRA-FPN showed improvements in the performance of 0.5-1.9% more than that of RetinaNet coupled with previous state-of-the-art FPN methods. Similarly, on the MS COCO dataset, RetinaNet and FCOS detection models displayed performance improvements of 0.8% and 0.6% respectively when coupled with MSRA-FPN instead of other FPNs. The final dataset Thangka Figure also experimentally demonstrated that the proposed method improves the performance baseline models by 2.9-4.7%, reaching a max mAP of 71.2%. With these results, they were able to show that their MSRA-FPN method has better performance than previously utilized FPNs.

The research by Han et al. (2019) [16] applied a deep learning model to accurately detect cigarettes in different image qualities and illumination. To achieve this they used the color segmentation algorithm alongside MTCNN and Faster R-CNN. They created their own dataset by sourcing images from the internet, smoking scenes from videos, and more videos that contain behaviors similar to smoking. They split the dataset into four different sets: A, B, C, and D. Testing on B, C, and D it was seen that their proposed model had lower false detections at 0%, 1%, and 0% for each set respectively compared to base Faster R-CNN. It also achieved higher mAP at

95.5%, 75.5%, and 95.0% respectively. Their model is also faster by being able to perform detection in images at 0.1 seconds less time and detection in videos at 53 seconds less time.

YOLO V5s with a spatial and channel attention mechanism alongside CARAFE upsampling was used by Yin et al. (2022) [17] for the detection of early-stage fires by detecting small and thin smoke. The authors created datasets that include a combination of smoke and smoke-free images in different scenarios, forming a complete dataset for smoke detection used in training, validation, and testing. The proposed model obtained a precision of 92.72%, the highest accuracy among other models such as SSD, Retinanet, YOLO V4, YOLO V5, and Efficientdet-D2. The proposed model was also compared to the other models for the small smoke dataset, obtaining a precision of 87.84% which was the highest. It was also tested for the non-smoke dataset, achieving 99.75% precision, which was the second highest. A detection speed of 69.00 fps was obtained which was fairly satisfactory and the third-highest speed compared to the other models. Overall, the proposed model was efficient in detecting smoke in real time.

In another paper, small targets were used by Zhao et al. (2022) [18] to detect fire and smoke. The proposed model uses the Fire-YOLO model and EfficientNet, to extract input image features. The first dataset used consisted of images of fire and smoke in different lighting conditions and the second consisted of flame and smoke with small targets, this dataset was self-made. These datasets were divided into training sets, validation sets, and test sets. The proposed Model had a precision of 0.915, an F1 value of 0.73, and an mAP value of 0.802, the highest among its comparison with YOLO-V3 and Faster R-CNN. The accuracy for detecting small targets was 75.48% and the efficiency for detection of fire and smoke-like targets under different natural lights was greater for the proposed model. Moreover, Fire-YOLO decreased computing costs, offered sustainable development, and saved resources.

A study conducted by Zhang and Wu (2018) [19] on real-time forest fire detection used popular object detection methods. The models studied are Faster R-CNN, different YOLO models, and SSD, the results obtained are compared to find the best model. The dataset used consisted of 1000 pictures of fire and smoke to test all three models. The Faster R-CNN model had an accuracy of 99.7% for the detection of fire and 79.7% for the detection of smoke, it was also found to operate at 7fps which was too slow for real-time detection. The YOLOv3 model outperformed the other YOLO models with an accuracy of 92.29% for the detection of fires and 65.53% for detecting smoke, however, it had bad performance when it came to detecting small fires. Lastly, the SSD model had an accuracy of 99.88% and 68.4% for the detection of fire and smoke respectively in real time.

A method presented by Abdusalomov et al. (2021) [20] would detect fire faster and in real time while focusing on the existing YOLOv3 algorithm. Their proposed method involves capturing live videos, resizing the images obtained from the input, applying data augmentation, and finally, using the YOLOv3 algorithm to evaluate the predictions. According to the authors, one of the biggest drawbacks in this task was obtaining the data and training sets. Thus, to increase the number of training sets, they gathered images from publicly available 'Google' as well as collected fire videos, from which images were extracted as frames using the OpenCV framework. Further, they increased the number of datasets by rotating the images at 90,180, and 270 degrees. In this paper, they tested several YOLO algorithms with 9200 fire images. In the test result, we can see that YOLOv3 had the highest accuracy of 82.4% in training and 77.8% in testing over 57 hours. Results obtained from YOLOv4 had similar outcomes, but it had higher processing time due to its larger weight. However, using YOLOv3 on their enhanced dataset, there seemed to be a rise in accuracy, 98.3% in training and 99.7% after further changes. While they successfully enhanced the data and test set required for the cause, it is also important

to note down their ways to work with the limitations they face while performing. For example, they focus on enhanced features for fire detection, instead of automatic color augmentations to refrain from detecting false positives, removing low-quality images, and filling the spaces by adjusting the brightness and contrast of the same images. This method also increases the number of test sets.

A large-scale benchmark smoke image dataset was introduced by Cheng et al. (2019) [21] to provide more effective training for data-driven smoke detection models. Their proposed dataset consists of 100k synthesized smoke images, corresponding smoke-free images, smoke masks, and bounding box positions. The smoke-free images were collected from the LabelMe dataset and NYU dataset, while base masks were created manually using Photoshop with 3 levels of density: low, middle, and high. To test the efficacy of their dataset, state-of-the-art CNN detectors MobileNetV2 and ResNet50 were trained using the dataset, showcasing adequate accuracy with Intersection over Union (IoU) of 84% and 85% respectively for low-density smoke, 79% and 81% respectively for medium density smoke, and 88% and 92% respectively for high-density smoke. Overall, the results showed that they were able to create a dataset capable of training a data-driven model for more accurate smoke detection.

A hybrid deep learning model of CNN with Random Forest and Bi-GRU was implemented by Ahmad et al. (2023) [22] to optimize the processing required for human activity recognition from visual data. The model processed data by extracting deep features using CNN, refining the deep features to obtain selected features using Random Forest, and learning the temporal motions of frames sequence with Bi-GRU. The proposed model was trained using both deep and selected features of realistic human activity datasets Youtube11, HMDB51, and UCF101, and the recognition accuracies for deep features were 92.05%, 67.53%, and 88.07% respectively while for selected features the accuracies of 93.38%, 71.89%, and 91.79% respectively, and also achieved the highest recognition accuracy of 71.8% for HMDB51 dataset. The

results proved the significance of the random forest features selection technique in improving the precision of the model. Additionally, the model achieved an average time complexity of 1.03 sec for training with Youtube11, improving upon previous models' 1.12 sec. Overall the proposed hybrid model displayed significant accuracy and computational complexity improvements over previous CNN and RNN models.

Jebur et al. (2023) [23] developed multiple deep learning (DL) methods being used to achieve high-performance anomaly detection (AD) in videos and provided a comprehensive overview. They analyzed different classifications of anomalies as well as DL methods and architectures, and their analysis was categorized into network type, architecture model, datasets, and performance metrics. Their classification of anomalies was Single-Point Anomaly and Group Anomalies where complexity and time requirements were greater in the latter, and their classification of models was Image-Based and Video-Based Detection where the latter achieves higher accuracy and flexibility. They also classified the different methods of detection that are applied and compiled multiple state-of-the-art DL architectures used within the last 3 years along with information regarding their use. For example, for the anomaly "smoking" there was the YOLOv5 plus ConvLSTM architecture which achieved 90% for the Cigarette smoker dataset and also there was the YOLOv5 architecture which achieved 91% accuracy for the Private dataset. Lastly, the paper provided information regarding benchmark datasets that are used, the key performance metrics used to evaluate effectiveness, and finally applications and challenges of DL anomaly detection, thus achieving the goal of providing a wide overview of the latest DL techniques used in AD in videos.

In [24], Algamdi et al. (2020) designed a CapsNet architecture that performs multi-label classification, meaning it recognizes two or more simultaneous human actions from scenes captured by drone videos. The proposed model called DroneCaps uses features extracted by 3D-CNNS and a new set of features extracted by a novel

Binary-Volume Comparison (BVC) layer. The model's performance was evaluated by comparing it with six different methods of HAR for multi-label classification on the Okutama-Action dataset, which contains aerial view videos of multiple human actions. 3DCapsNet-EM had the second-best performance due to EM routing separating activation from the pose, which helped in the identification of objects from different angular and elevation viewpoints. Since the proposed DroneCaps model had a similar architecture to 3DCapsNet-EM with the addition of the BVC layer, compared to 3DCapsNet-EM, the DroneCaps model achieved 5.63% more in Total Accuracy, 0.013 less in Hamming Loss, 0.086 less in One Error, 0.05 more in Exact Match Ratio. With a Total Accuracy of 47.50%, the DroneCaps model outperformed all the other methods tested on the same dataset by up to 13.75%.

For human activity recognition Shuvo et al. (2020) [25] implemented a flexible hybrid approach. They divided the learning process into two stages for HAR using a gyroscope sensor and waist-mounted accelerometer. They classified activities into static or moving using a Random Forest (RF) binary classifier in the first stage. In the second stage, they used a Support Vector Machine (SVM) for the recognition of specific static activity, and a deep learning model based on 1D Convolutional Neural Network (CNN) for the recognition of specific moving activity. They have tested and trained on the UCI-HAR dataset that contains recordings. The proposed hybrid model achieved an overall accuracy of 96.8% for classifying static activities and 98.6% for classifying moving activities. The respective precisions achieved for sitting, lying, standing, walking, walking upstairs, and walking downstairs were 98.2%, 99.8%, 92.7%, 99%, 97.3%, and 98.6%. After combining static and moving activities, the overall accuracy achieved was 97.66%. To evaluate the model, the performance was compared with other methods. Their proposed hybrid model outperformed all the other models in the evaluation comparison and achieved 0.06% more accuracy than the second-highest-performing model. They recommended using their model for future works in low-power circuits.

A technique was shown by Naik et al(2021) [26], to effectively detect objects in real time using the YOLOv4 algorithm. The custom dataset used in this proposed model features nine different classes, and the proposed custom model was trained for 500 epochs. The proposed model involves taking the image input from the created dataset involving the Indian road traffic images taken in real time by CCTV footage, then feeding the images to convolutional downsampling and then to a series of layers of dense connection blocks. After that, the results are passed down to the spatial pyramid pooling layer and, finally, to the object detection block to detect the object. The proposed model involving YOLOv4 achieved a precision of 98.32%, which, according to the authors of this paper, proves to be capable enough to detect road traffic classes in traffic areas.

Liu et al. (2017) [27] conducted research on Faster R-CNN for object detection based on Deep Learning methods. This model combines the Fast R-CNN with the RPN network to achieve a better model for object detection from end to end. The following experiment used Caffe, a deep learning framework for training different networks and the PASCAL VOC2007 dataset was used for training and testing. The mean average precisions for different models of Faster R-CNN and different networks were 69.4% for the VGG16 network, 72.5% for the ResNet101 network, and lastly, 84.9% with the PVANET network.

From the above discussion, it is observed that most researchers who achieved higher accuracy have designed hybrid models. These hybrid models included versions of YOLO with modifications, which succeeded in giving better results in the real time detection of smoke and objects. Another combination using the deep learning model 1D CNN also achieved high accuracy for detecting human movements. Although using a hybrid model is ideal, it is important to uphold that the combination of the models used should not be too big or too slow. Otherwise, real-time recording

devices such as CCTV cannot handle them.

Chapter 3

Methodology

3.1 Data Collection

Initially, we tried collecting data from online sources and existing datasets but, we were faced with a number of challenges. First of all, there was a lack of existing datasets on smokers. Furthermore, the existing datasets and videos focus directly onto the smokers. Hence, the smokers were the main subjects of the videos. There was also a lack of complex videos so using those would not be sufficient for our research as our models would then mostly train on simple videos which would not depict real-life scenarios. Therefore, we had to create our own dataset for this research.

The dataset we acquired were videos captured using the lens of an iPhone camera. All the videos are at 30 frames per second and are captured at various orientations. To depict real-life scenarios, complex environments were used for most of the videos. There are multiple subjects performing various activities including drinking tea, conversing with one another, eating food, and smoking in the videos. To simulate CCTV footage, most of the videos were taken from higher altitudes relative to the subjects. An important criterion for the videos in the dataset is that most of the videos do not focus on the smokers, they focus on the environment as a whole. There are cases when no smokers are present for a period of time, in some videos

the smokers are hard to notice at a glance while in others they are more visible. This was done to bring more diversity to the dataset and to cover a wider range of situations.

3.2 Data Preprocessing

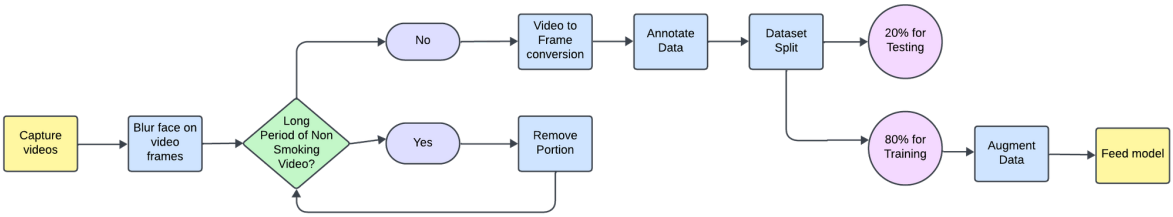


Figure 3.1: Flowchart for Preprocessing

Figure 3.1 visualizes the data preprocessing segment. After the collection of the dataset was completed, the faces of the subjects were blurred from each video to protect their privacy. The length of the videos were shortened to prevent unnecessary video data from being present in the dataset. Long periods of non-smoking and empty frames were removed from the videos to create a cleaner dataset. The cleaned videos were then converted to frames which provided 4013 frames in total.

Once all the frames were extracted, annotations were done on each frame using CVAT [28], an online annotations software. Annotations were done for five different classes namely Vape, Cigarette, Smoke, Smoker and Person. The criteria to annotate Smoker was that it had to be accompanied by either Vape, Smoke, or Cigarette. Therefore, if none of those were annotated we did not annotate the Smoker. This was done so that there are visible reasons why a person was being marked as a Smoker. And also to be a reference to see if the classification of the person being a smoker was accurate. This was also to take into account when someone stopped smoking as then they would not be performing that activity hence they should no longer be marked a smoker. Therefore, the main criterion was that a person had to

visibly perform the act of smoking.

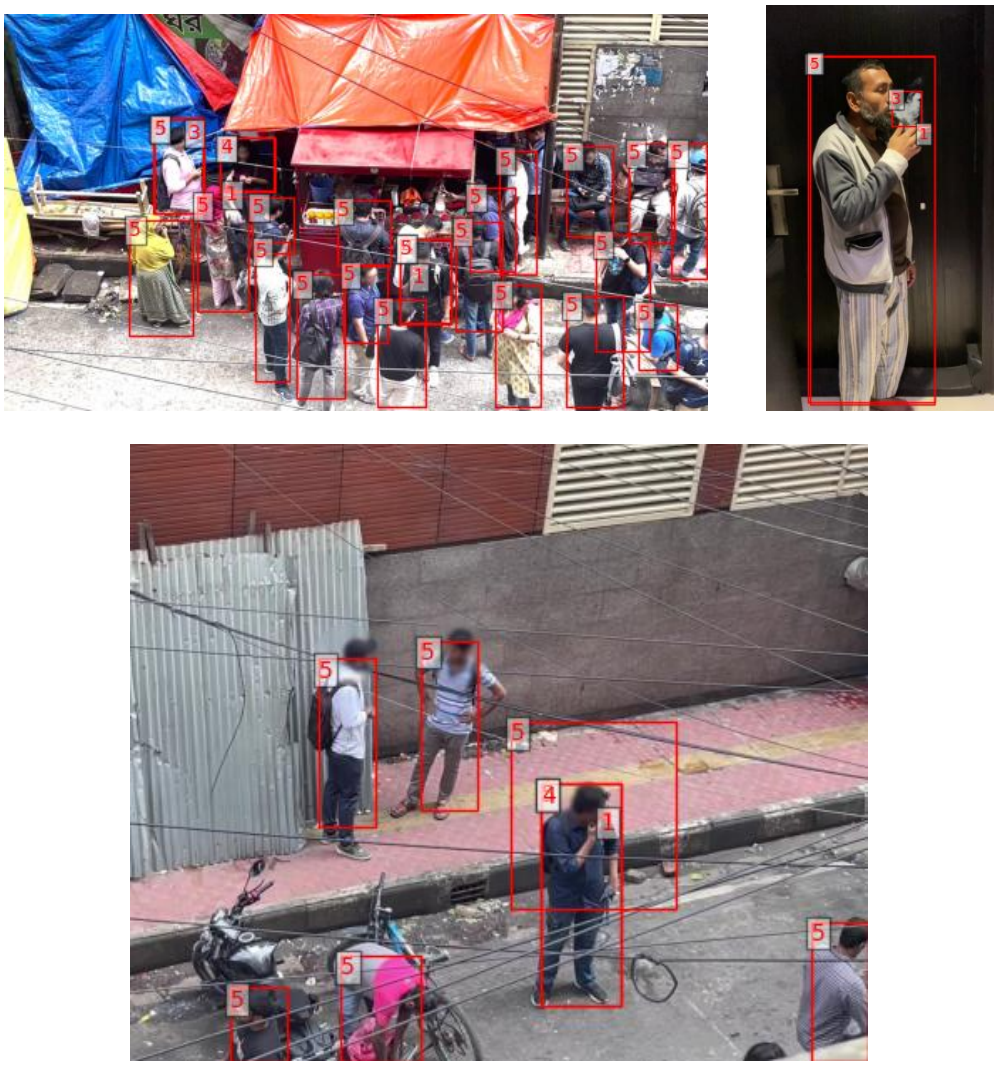


Figure 3.2: Annotated Frames

Lastly, the annotated frames were split into training and testing sets. 80% of the frames were used for training while 20% were used for testing. To prevent data leakage, the frames selected for testing were carefully checked to prevent any of the frames from being also part of the training set. Frame selection was done at random by taking small sequences from each video at different intervals. Then the frames of the training set underwent different augmentations to prevent overfitting. Flipping the images vertically and horizontally, changing the hue, saturation, and brightness values, and also mosaic. Once, all of these steps were concluded then the dataset was fed to the model for training.

Figure 3.3: Augmented Training Batch

For Video Classification, frames from each video were put into different folders. For each video, there were two folders, one for the sequences of smoking while the other was for sequences of non-smoking. Here, each video sequence underwent different augmentations which were the same as done before. The sequences were also split into 80% to 20% ratios for training and testing respectively. In the testing set there were a variety of sequences where each video did not always contain both smoking

and non-smoking sequences. This was done to bring more variety to the testing set and to not make it similar to the training set.

3.3 Model Description

- **YOLOv5s**

  According to [29], the components of YOLOv5 can be classified into 3 parts- backbone, neck and output. Data preprocessing and data augmentation tasks like mosaic are performed by the input terminal of the YOLOv5. YOLOv5 uses adaptive anchor free calculations for adapting better in different datasets. The backbone of YOLOv5 is a CNN network that forms features at different sets. It also uses a CSP network and spatial pyramid pooling for extracting feature maps of distinctive sizes. Moreover, to reduce the amount of calculations required, the bottleneck CSP architecture is used. Furthermore, to combine the features to make the prediction, the neural network is used. The detection capability is enhanced when YOLOv5 uses PAN structure and FPN structure jointly- which strengthens the features extracted from different layers. YOLOv5 has several versions and in this research, YOLOv5s is used, which is a small version fit for areas with limited computation resources.

- **YOLOv8n**

  A very popular architecture for object detection in the field of computer vision is YOLO (You Only Look Once). Its latest iteration YOLOv8, developed by Ultralytics [30], builds upon that architecture. At the backbone, it uses a modified version of the CSPDarknet53 architecture that has 53 convolutional layers [31]. This modified version uses the C2f module which combines the results of all the bottleneck modules instead of only the last module. A key feature of the YOLOv8 is its anchor-free detection mechanism which predicts the center of an object which speeds up the training phase as fewer bounding

boxes are predicted overall [32]. It also employs a feature pyramid network to detect objects at different scales, this is particularly useful for the detection of small objects which is a requirement for the task of this paper. Lastly, YOLOv8 contains mosaic which is an image augmentation method used during the training phase where the four images are combined into one image which drastically reduces overfitting.

- **YOLOv9t**

  YOLOv9's contribution over its predecessors came with a newer approach to tackle information loss challenges in deep neural networks [33]. It enhances its learning capability and retains important information during the detection process by using PGI and versatile GELAN architecture. YOLOv9 focuses on tackling the information bottleneck principle - challenges faced in deep learning, which states that as data goes through the layers, the chances of information loss increases. To tackle this problem, the YOLOv9 uses PGI, programmable gradient information, which helps in preserving data across the depth of the network. Moreover, the model also takes the help of the reversible functions to retain important data during object detection and reduce the loss of information in the deeper layers of the network. Furthermore, the GELAN architecture helps the model for greater utilization of parameters and better computational efficiency without the need of compromising in accuracy. In this research, the tiny version, YOLOv9t, has been used to ensure a lesser number of parameter usage and computational requirements.

- **YOLOv10-N**

  The latest YOLO model used for object detection is the YOLOv10 model [34]. The YOLOv10-N version was used for this research, which addresses issues within older models such as architectural inefficiencies and reliance on NMS. Inference latency is reduced through the elimination of NMS by utiliz-

ing consistent dual assignments. The efficiency and accuracy are optimized through using a rank-guided block design, lightweight classification heads and a down sampling method which separates spatial and channel information. Additionally, to improve performance without increasing computational cost, partial self-attention modules and large kernel convolutions are used. The backbone uses an enhanced CSPNet version for feature extraction, improving gradient flow and decreasing extra computation. The neck uses features from different scales to combine them and send to the head, which includes PAN layers to carry out multi-scale feature fusion. During training, multiple predictions for each object are made to improve learning accuracy and a single prediction is made for each object during inference, eliminating the need for NMS.

3.4 Prediction Criteria

The paper was worked on in two different segments. The first segment was about understanding the problem at hand and doing the research required to understand the models that could be used. This has been visualized in Figure 3.4.

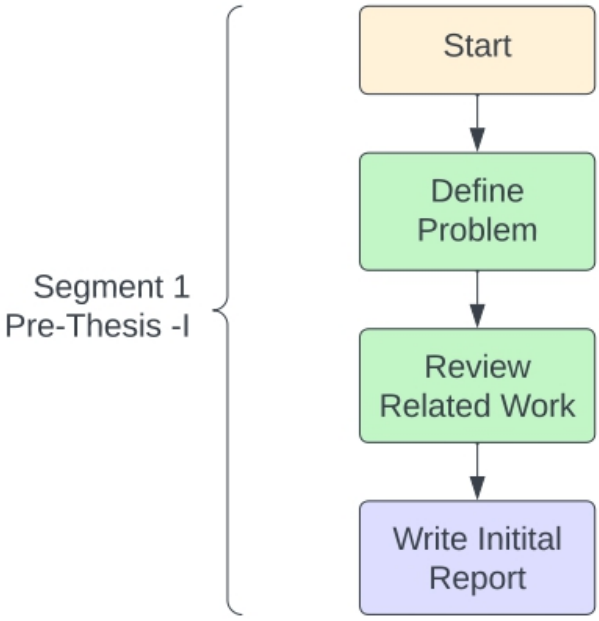


Figure 3.4: Work Flow for First Segment

The second phase was about data collection and creating a complete dataset. It also included deploying the selected models on the dataset and analyzing the predictions.

For object detection, we used a hybrid detection system where all the predictions work in tandem. The criteria for the predictions was that it predicted all the classes of objects present in each image. Then it took the Person class as the base and checked whether any of the bounding boxes of the other classes predicted overlapped. If that criterion was fulfilled, then the bounding box of the Person was classified as a Smoker and sent to calculate the IoU with the ground truth bounding boxes of Smoker. An IoU over 0.5 was then considered an accurate prediction. All the works of phase 2 can be visualized in Figure 3.6.

Figure 3.5: Smoker (top) and Non-Smoker (Bottom)

For Video Classification, the predicted sequences were checked with the ground truths, the top-5 accuracy was taken and if the labels had values over 0.5 then the labels were accepted. From there false positive, true positive and false negative predictions were calculated.

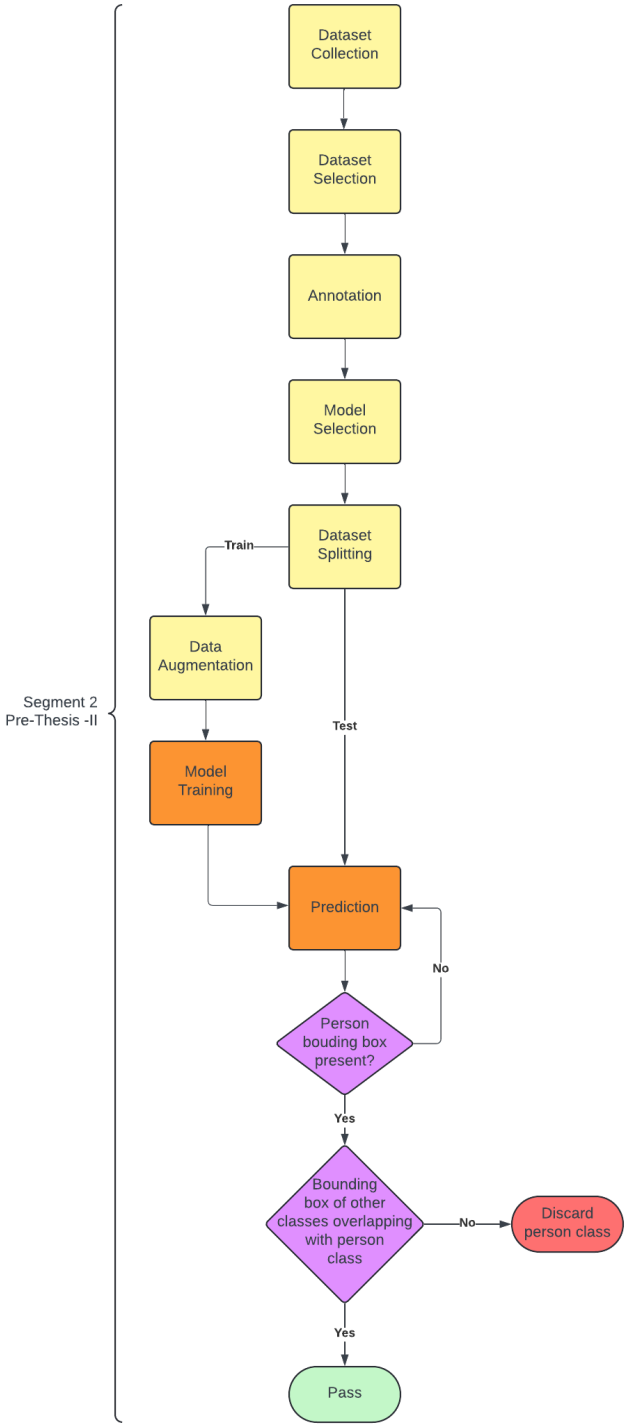


Figure 3.6: Flow Chart of Second Segment

Chapter 4

Results and Analysis

4.1 Results and Discussion

With the completion of our testing phase on the models for object detection and video classification that were trained for 50 epochs each and only choosing those predictions whose confidence scores were above 50%, we evaluated our models based on three metrics: F-1 Score, Precision, and Recall. The table below provides the results for our models.

Model	F-1 Score (%)	Precision (%)	Recall (%)
YOLOv5	77.6	90.3	68.1
YOLOv8	81.0	93.6	71.4
YOLOv9	85.1	96.2	76.3
YOLOv10	81.4	92.7	72.6

Table 4.1: Experimental Results for Object Detection

Model	F-1 Score (%)	Precision (%)	Recall (%)
YOLOv8	92.4	90.6	94.3

Table 4.2: Experimental Results for Video Classification

To achieve these results we trained the models based on the following hyper-parameters which can be found in Table 4.3.

Parameters	Values
Optimizer	AdamW
Learning Rate	0.001111
Weight Decay	0.0005
Momentum	0.9

Table 4.3: Hyper-parameter Settings

From Table 4.1 it can be seen that we have achieved overall decent results from all three metrics. For YOLOv5, YOLOv8, YOLOv9 and YOLOv10 we achieved F-1 Scores of 77.6%, 81.0%, 85.1% and 81.4% respectively. However, despite YOLOv10 being the newest YOLO model it performed worse compared to YOLOv9, which had the best results across all the metrics. The F-1 Scores show that all the models had well balanced Precision and Recall scores.

To get a better understanding of the results we can look at the Precision scores we can see that all the models had achieved scores of over 90%. With YOLOv9 achieving the highest Precision score at 96.2% followed by 93.6% of YOLOv8 and YOLOv10 trailing with 92.17%. The lowest score was achieved by YOLOv5 at 90.3% which is expected considering it is the oldest model among them. The improvements in the backbones of the newer models have significantly allowed the newer models to make lower false positives. Furthermore, it should also be stated that all the models made very little false predictions causing the very high Precision scores. A very high precision score is important for this problem as it is more important to not falsely predict a non-smoker as a smoker. This is due to the criteria that was used where only when a smoking device or smoke is being detected only then the person will be classified as a smoker. The criteria also helped to reduce the false positives by not classifying people as smokers when they have stopped smoking as none of the other

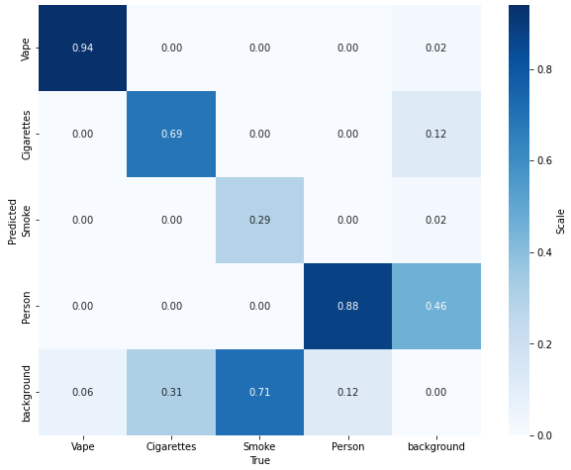
objects were detected. However, a reason why the precision scores of all the models are not higher is because in some of the cases the supporting classes were detected and they were located inside the bounding boxes of a person class that was not a smoker which can lead to lower Precision scores.

Moving on to the Recall scores we can see that YOLOv9 also achieved the highest score at 76.3% with YOLOv10, YOLOv8 and YOLOv5 trailing with 72.6%, 71.4% and 68.1% respectively. These show that when it comes to predicting all the instances of smokers, the models are extremely effective. Even though the scores are relatively high for all the models, there are cases where a lot of the smokers will go undetected. This can also be due to the criteria with which people are classified as smokers. Because, without the detection of smoking objects or smoke the models will not classify a person as a smoker even if that person is smoking.

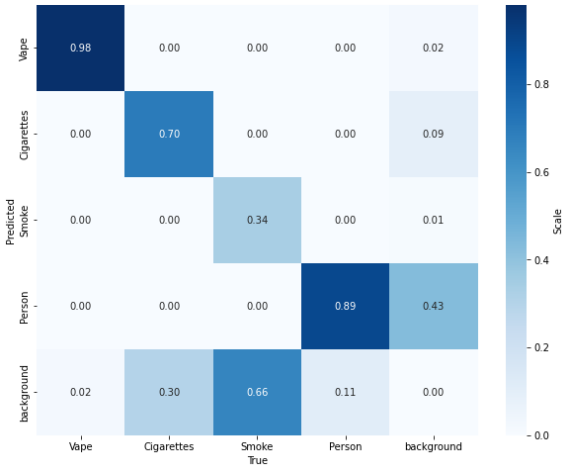
The above results can be evaluated using the confusion matrices in 4.1. Examining the matrices and the results we can see that they are consistent. From the confusion matrix of YOLOv9 we can see that it has the highest true positives for the cigarette class which has resulted in its recall being high as the instances of cigarettes far exceed the instances of vape and smoke. Furthermore, looking at the correct predictions of the person class we can see that all the models had high correct detection of the class in the range 88%-89%. This can also mean that even if a smoker was present, because they were not detected as a person, they were not classified as smokers even if the supporting objects were detected.

Moreover, the confusion matrix of YOLOv5 does give evidence that its architecture has become slightly outdated, considering it has the lowest accurate predictions of all the smaller classes which leads to it discarding a lot of the person class as non-smokers even if in reality they were smokers.

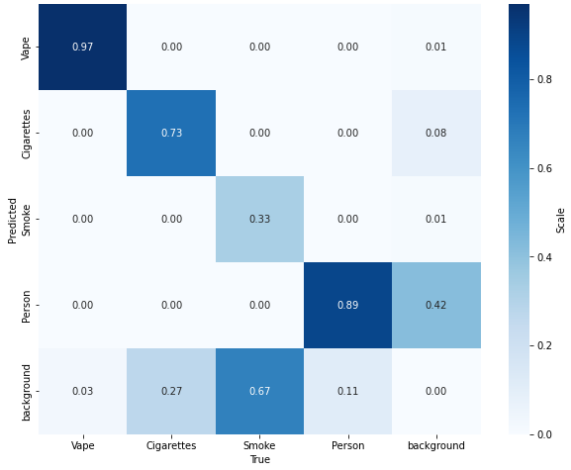
Therefore, the confusion matrices does provide reasoning why YOLOv9 is the best for this problem as its detection capabilities of the smaller objects especially exceed that of the newer YOLOv10 model.



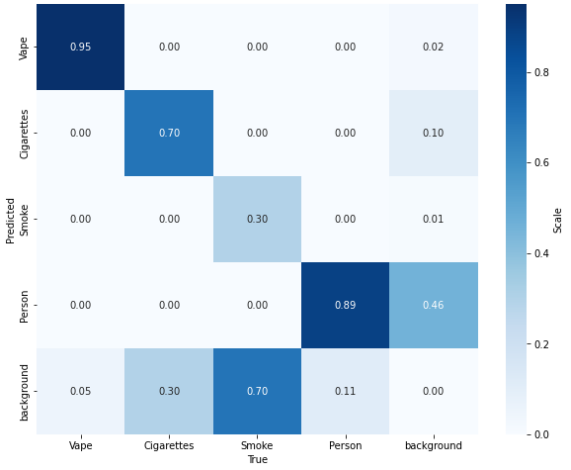
(a) YOLOv5



(b) YOLOv8



(c) YOLOv9



(d) YOLOv10

Figure 4.1: Confusion Matrices

From 4.2 we can see that using the latest YOLO model which has classification capabilities, we achieve very high precision, recall and F-1 scores at 90.6%, 94.3% and 92.4% respectively. These show that when it comes to video classification YOLOv8 performs very well at correctly classifying a sequence of frames. To examine further into the results, a reason for why the precision score is lower than the recall score is because of the nature of the videos. Because from each video there are sequences of both smoking and non-smoking, the model may not have been able to understand fully understand when the features of the sequence constitutes to smoking or non-smoking. Furthermore, as we are classifying each sequence as smoking or non-smoking based on the visibility of cigarette, vape and smoke, if the model does not understand the features of these small objects it will lead to it making mistakes correctly classifying the sequence.

With the aim to further evaluate and validate the performance of our model, we utilized Class Activation Mapping tools to obtain visual explanations of which features of an image the models focus on during classification. easy-explain [35] technique was used, which is a gradient-free CAM technique that identifies the features that are preserved after passing through the Convolutional Networks to create a class activation map without the need for model modification.

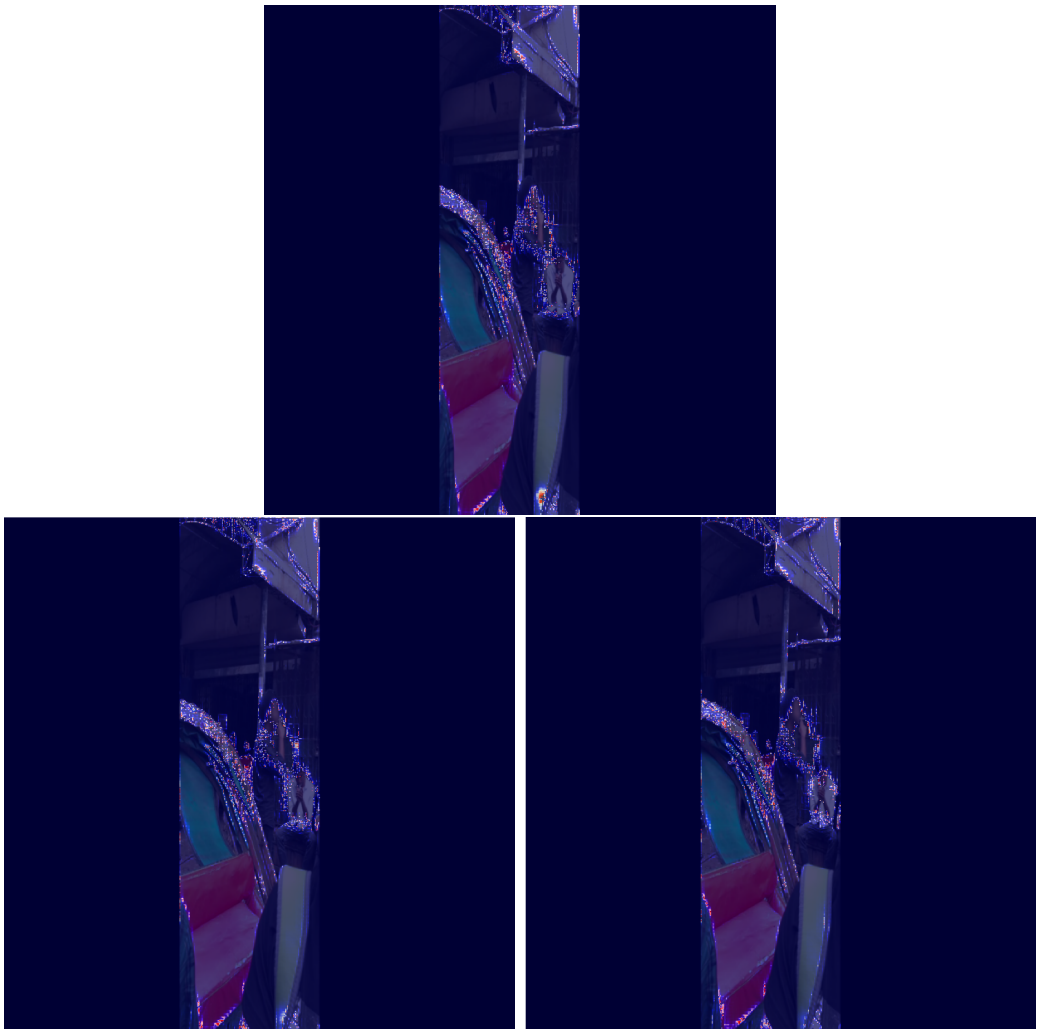


Figure 4.2: CAM Images

As seen in 4.2 for the sequence of the three frames we can see that YOLOv8 is basing its decision from mostly the correct features. From the images we can see that for all the three frames it is extracting features concentrated around the two people who are smoking. While it does take information from other sources as well such as towards the bottom of the image, the important features also play an important role in determining if the sequence is of smoking or not. As, the two people were indeed smoking it may have given the model enough information for it classify the sequence as smoking.

However, it is also important to see that it did not specifically focus on the cigarette itself. This shows that it did not specifically focus on the smoking objects present

but the gesture of the person as a whole as the range of motions performed by the person when smoking is very unique.

Therefore, the CAM images have given us in-depth understanding that for video classification the model can extract correct features when classifying the sequence but it sometimes also is not also able to take in all the important features especially when those features are located in a very small region of the frame.

4.2 Research Challenges

Nonetheless, we had to face significant challenges throughout the inception of our research. A major issue occurred during the creation of the dataset. As the dataset had to depict real-life scenarios and contained videos that were taken from a higher vantage point with the subjects further in the background, objects of the cigarette class took up very little space in each frame of the videos. These resulted in the models having very little information about the regions occupied by cigarettes. Hence, during the training phase, the models were unable to train properly on the cigarette class resulting in its lack of detection during the testing phase. Moreover, it was seen that the models had a high detection rate of the person class, but due to the lower detection of the other classes some of the people who were actually smokers never got marked as smokers which resulted in lower metric scores.

Moreover, we believe we could have achieved better results if we had a more balanced dataset. The instances of smoke and vape were low compared to cigarettes. This caused the models to not be fully trained in those 2 classes. A more balanced dataset could have improved our models further and allowed us to achieve better scores altogether.

The same is true for the dataset used in video classification. As frames from the

same dataset were used, due to the size of the the important smaller objects, the model may have been unable to fully locate the important features which is why the metric scores are not nearer to 100%.

4.3 Future Improvements

Be that as it may, to improve the results and the quality of the models, a larger and more diverse set of videos can be used to train the models so that they are able to fully learn about all the different situations. Higher quality videos could also be introduced into the dataset to help the models better train on small objects such as Cigarettes and Vape. Due to the limited time and conditions, they resulted in the lower instances of vape. As for how it stands now, the models we created have performed the required tasks well but for further improvements, they have to be trained under more diverse and clearer videos. Further models need to be tested to see which one works the best for this problem and ensemble learning methods need to be implemented to increase the scope of the solutions.

Chapter 5

Conclusion

The dangers of smoking cigarettes and vapes are known to most people and yet they decide to smoke, thus to prevent other non-smokers from experiencing these damaging effects, a surveillance system is important. The available research papers attempt to monitor small objects in real time or detect smoke in real time. The following research paper combines these functions to create a digitized security surveillance system by using different object detection models to find the most accurate model along with using a video classification technique to find smokers. Therefore, this research serves as an attempt to monitor illicit activities such as smoking without the additional use of manpower through CCTV, to monitor and maintain the ban of smoking in public areas.

Bibliography

[1] Second-hand and third-hand smoke and vapour: Effects on children. (2023). *Raising Children Network*. https://raisingchildren.net.au/babies/health-daily-care/health-concerns/second-hand-smoke?fbclid=IwAR1HZAQlaswxAIJTuQFjJII5PS-6Rlfb0Bio1wFJBeRhgrHvtjHxb38dwi0

[2] Islam, R. (2022). Carelessly discarded cigarette butts 3rd top cause of fires in 2021: Fire service. *The Daily Star*. https://www.thedailystar.net/news/bangladesh/accidents-fires/news/carelessly-discarded-cigarette-butts-3rd-top-cause-fires-2021-fire-service-2957361?amp&fbclid=IwAR2RGR64-mRsHjme3dRWoTGauEm7sIN-B34uKexTBXk_ptWeykdbEOZkssA

[3] Ahmed, K. (2021). Number of smokers has reached all-time high of 1.1 billion, study finds. *The Guardian*. https://www.theguardian.com/society/2021/may/27/number-of-smokers-has-reached-all-time-high-of-11-billion-study-finds

[4] Christensen, J. (2023). Us cigarette smoking rate falls to historic low, but e-cigarette use keeps climbing. *CNN*. https://edition.cnn.com/2023/04/27/health/cigarette-smoking-decline/index.html

[5] Narejo, S., Pandey, B., Vargas, D. E., Rodriguez, C., & Anjum, M. R. (2021). Weapon detection using yolo v3 for smart surveillance system. *Mathematical Problems in Engineering*, *2021*, 1–9. https://doi.org/10.1155/2021/9975700

[6] Fang, P., & Shi, Y. (2018). Small object detection using context information fusion in faster r-cnn. *2018 IEEE 4th International Conference on Computer and*

Communications (ICCC), 1537–1540. https://doi.org/10.1109/compcomm.2018.8780579

[7] Zhu, D., Xu, G., Zhou, J., Di, E., & Li, M. (2021). Object detection in complex road scenarios: Improved yolov4-tiny algorithm. *2021 2nd Information Communication Technologies Conference (ICTC)*, 75–80. https://doi.org/10.1109/ictc51749.2021.9441643

[8] Cai, W., Wang, C., Huang, H., & Wang, T. (2020). A real-time smoke detection model based on yolo-smoke algorithm. *2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC)*, 1–3. https://doi.org/10.1109/csrswtc50769.2020.9372453

[9] Al-Smadi, Y., Alauthman, M., Al-Qerem, A., Aldweesh, A., Quaddoura, R., Aburub, F., Mansour, K., & Alhmiedat, T. (2023). Early wildfire smoke detection using different yolo models. *Machines*, *11*(2). https://doi.org/10.3390/machines11020246

[10] Fu, Q., Ma, S., Liu, L., & Liu, J. (2018). Human action recognition based on sparse lstm auto-encoder and improved 3d cnn. *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 197–201. https://doi.org/10.1109/fskd.2018.8686921

[11] Heather, H. (2023). The effects of smoking on the body. healthline. *healthline*. https://www.healthline.com/health/smoking/effects-on-body#central-nervous-system

[12] Jiang, Z., Zhao, L., Li, S., & Jia, Y. (2020). Real-time object detection method based on improved yolov4-tiny. *Journal of Network Intelligence, Volume 7, Number 1*. https://doi.org/10.48550/arXiv.2011.04244

[13] Sarić, R., Ulbricht, M., Krstić, M., Kevrić, J., & Jokić, D. (2020). Recognition of objects in the urban environment using r-cnn and yolo deep learning algorithms. *2020 9th Mediterranean Conference on Embedded Computing (MECO)*, 1–4. https://doi.org/10.1109/MECO49872.2020.9134080

[14] Ullah, M. B. (2020). Cpu based yolo: A real time object detection algorithm. *2020 IEEE Region 10 Symposium (TENSYMP)*, 552–555. https://doi.org/10.1109/TENSYMP50017.2020.9230778

[15] Wang, H., & Wang, T. (2023). Multi-scale residual aggregation feature pyramid network for object detection. *Electronics*, *12*(1). https://doi.org/10.3390/electronics12010093

[16] Han, G., Li, Q., Zhou, Y., & Duan, J. (2019). Rapid cigarette detection based on faster r-cnn. *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2759–2765. https://doi.org/10.1109/SSCI44817.2019.9003034

[17] Yin, H., Chen, M., Fan, W., Jin, Y., Hassan, S. G., & Liu, S. (2022). Efficient smoke detection based on yolo v5s. *Mathematics*, *10*(19). https://doi.org/10.3390/math10193493

[18] Zhao, L., Zhi, L., Zhao, C., & Zheng, W. (2022). Fire-yolo: A small target object detection method for fire inspection. *Sustainability*, *14*(9). https://doi.org/10.3390/su14094930

[19] Wu, S., & Zhang, L. (2018). Using popular object detection methods for real time forest fire detection. *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, *01*, 280–284. https://doi.org/10.1109/ISCID.2018.00070

[20] Abdusalomov, A., Baratov, N., Kutlimuratov, A., & Whangbo, T. K. (2021). An improvement of the fire detection and classification method using yolov3 for surveillance systems. *Sensors*, *21*(19). https://doi.org/10.3390/s21196519

[21] Cheng, H.-Y., Yin, J.-L., Chen, B.-H., & Yu, Z.-M. (2019). Smoke 100k: A database for smoke detection. *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, 596–597. https://doi.org/10.1109/GCCE46687.2019.9015309

[22] Ahmad, T., Wu, J., Alwageed, H. S., Khan, F., Khan, J., & Lee, Y. (2023). Human activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features

selection. *IEEE Access*, *11*, 33148–33159. https://doi.org/10.1109/ACCESS.2023.3263155

[23] Jebur, S. A., Hussein, K. A., Hoomod, H. K., Alzubaidi, L., & Santamaría, J. (2023). Review on deep learning approaches for anomaly event detection in video surveillance. *Electronics*, *12*(1). https://doi.org/10.3390/electronics12010029

[24] Algamdi, A. M., Sanchez, V., & Li, C.-T. (2020). Dronecaps: Recognition of human actions in drone videos using capsule networks with binary volume comparisons. *2020 IEEE International Conference on Image Processing (ICIP)*, 3174–3178. https://doi.org/10.1109/ICIP40778.2020.9190864

[25] Hossain Shuvo, M. M., Ahmed, N., Nouduri, K., & Palaniappan, K. (2020). A hybrid approach for human activity recognition with support vector machine and 1d convolutional neural network. *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–5. https://doi.org/10.1109/AIPR50011.2020.9425332

[26] Naik, U., Rajesh, V., R, R., & ., M. (2021). Implementation of yolov4 algorithm for multiple object detection in image and video dataset using deep learning and artificial intelligence for urban traffic video surveillance application, 1–6. https://doi.org/10.1109/ICECCT52121.2021.9616625

[27] Liu, B., Zhao, W., & Sun, Q. (2017). Study of object detection based on faster r-cnn, 6233–6236. https://doi.org/10.1109/CAC.2017.8243900

[28] CVAT.ai Corporation. (2023, November). *Computer Vision Annotation Tool (CVAT)* (Version 2.8.2). https://github.com/opencv/cvat

[29] Horvat, M., Jelečević, L., & Gledec, G. (2022). A comparative study of yolov5 models performance for image localization and classification. *2*. https://www.researchgate.net/publication/363824867_A_comparative_study_of_YOLOv5_models_performance_for_image_localization_and_classification

[30] Jocher, G., Chaurasia, A., & Qiu, J. (2023, January). *Ultralytics YOLO* (Version 8.0.0). https://github.com/ultralytics/ultralytics

[31] Mehra, A. (2023). Understanding yolov8 architecture, applications & features. https://www.labellerr.com/blog/understanding-yolov8-architecture-applications-features/

[32] Boesch, G. (n.d.). A guide to yolov8 in 2024. https://viso.ai/deep-learning/yolov8-guide/

[33] Wang, C.-Y., & Liao, H.-Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information.

[34] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*.

[35] Theocharis, S. (2024, May). *Easyexplain* (Version 0.5.0). https://github.com/stavrostheocharis/easy_explain