National College of Ireland

A Machine Learning Pose Detection Framework to Identify Suspicious Activity

MSc Data Analytics
Research Project

Rajat Deepak Agrawal
Student ID: 21172030

School of Computing
National College of Ireland

Supervisor: Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Rajat Deepak Agrawal
Student ID:	21172030
Programme:	Research Project
Year:	2023
Module:	MSc Data Analytics
Supervisor:	Paul Stynes
Submission Due Date:	14/08/2023
Project Title:	A Machine Learning Pose Detection Framework to Identify Suspicious Activity
Word Count:	6049
Page Count:	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rajat Deepak Agrawal
Date:	14th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine Learning Pose Detection Framework to Identify Suspicious Activity

Rajat Deepak Agrawal
21172030

Abstract

Suspicious activity is a type of behaviour that can be classified as unusual and may indicate illicit intent. Detecting these types of activities using conventional methods is a big-time challenge as these methods suffer from computational complexity which limits their real time applicability. This research proposes a machine learning pose detection framework for identifying suspicious activity. Suspicious activities involve unusual movements of the body which might be an indication of a potential threat. By combining pose detection model and classification algorithms this framework tries to enhance the surveillance systems in real time. The proposed framework combines a pose estimation model, an activity classification model, an activity recognition model, and an alert mechanism. The framework proposes to use MediaPipe BlazePose pose estimation model to extract body posture features in the form of x, y, and z coordinates. These coordinates are later used to create customised geometric features and train the model. The framework proposes to create two separate models where the primary model is used to classify if the activity is suspicious or not and the second model is used to recognize the activity if predicted as suspicious by the first model. A large-scale dataset of 1900 real world surveillance videos which consists of 13 different classes is used for this study. Once the features were extracted from this data four algorithms XgBoost, Random Forest, LightGBM, and Deep Neural Network (DNN) were evaluated for both primary and secondary models. Random Forest was determined to be the optimal primary model with an accuracy of 93% and XgBoost was found to be more feasible as a secondary model with an accuracy of 70% in predicting activities. The potential users who can be benefited from this framework include security companies, private businesses, public safety organizations, etc. .

1 Introduction

Detecting suspicious activities plays an important role in ensuring the safety and security of society and contributing to create a stable environment. These types of activities pose a significant threat to society and should be prevented as soon as possible. Detecting suspicious activities can help in the prevention of criminal activities by early intervention and stop the situation from being escalated. To detect these activities surveillance cameras are usually used but the amount of data these surveillance cameras generate is huge and the traditional methods usually rely on human surveillance. As it is very difficult to

analyse this huge amount of data by humans continuously with respect to manual labour and time, the whole system becomes error-prone.

The motivation behind creating a Machine learning pose detection framework to identify suspicious activity emerges from the limitations of state-of-the-art methods. A few applications of suspicious activity detection use the Convolutional Neural Networks(CNN) model to extract features from the video frames and predict if the activity is suspicious or not Jain and Vishwakarma (2020)Affonso et al. (2021). These CNN models fail in a few cases such as complex backgrounds or spatial context that is the unusual body movements which was not covered in train data etc. Human pose estimation models are designed to determine human body joints and their coordinates. Most of the existing systems use pose estimation models in yoga pose detection, sign recognition, and gesture recognition. Not much of the study is done to use pose estimation models in suspicious activity detection. These pose models can be combined with machine learning or deep learning models to create a framework for suspicious activity detection. These models can capture spatial features which include finding out relationships between body parts and posture which helps in differentiating between suspicious and non-suspicious activities. There are many pose detection algorithms available publicly such as OpenPose, AlphaPose, PoseNet, etc which are used in various applications. Even though these algorithms provide good accuracy they are computationally expensive and very complex to implement.

Hence, the main goal of this study is to investigate to what extent a pose estimation model with machine learning can identify suspicious activities. In this study, a framework has been created to detect suspicious activities based on pose estimation features.

This study develops two robust machine learning models in conjunction with Mediapipe's BlazePose pose estimation model. The first model is created to detect if the activity is suspicious or not and the second model is created to identify the type of suspicious activity if the activity is detected as suspicious by first model. Mediapipe BlazePose pose estimation model can accurately locate pose landmarks and as it is lightweight it can be integrated with any edge devices which demand low latency and high accuracy. The model gives 33 landmark keypoints in the form of x, y, and z coordinates. These coordinates are later converted into customised features which include shoulder distance, elbow distance, shoulder angle, height-to-elbow ratio, etc. These features help in capturing the change in the body movements which can be used to determine the type of activity. The study utilizes a publicly available large video dataset which consists of 13 types of anomalous activities Sultani et al. (2018). These anomalous activities contain activities such as robbery, assault, vandalism, etc. The research objectives derived to address the research question were as follows: -

1) Broadly investigate and implement the state-of-the-art framework developed by Amrutha et al. (2020) which uses VGG16 and Long Short Term Memory (LSTM) for classifying suspicious and non-suspicious activities.

2) Develop a novel machine learning pose detection model to detect if the activity is suspicious or not.

3) Implement a machine learning model to detect the type of activity if detected as suspicious.

4) Evaluate and discuss the obtained results from the implemented framework and compare its results with state-of-the-art results by using different matrices.

A major contribution of this research is the development of a novel machine learning

pose detection framework to detect suspicious activities. This research tries to help in the advancement of suspicious activity detection systems which will help in the enhancement of surveillance systems in the future. This study has wide applications in the area of security, law enforcement, smart environments, etc.

The paper is structured as follows. Section 2 contains Related Work about various suspicious activity detection methods. The methodology followed in this study is described in section 3. Section 4 of the paper describes design specifications while section 5 describes the implementation steps. Section 6 consists of the results and evaluation of all the experiments conducted regarding the research. Section 7 does a thorough discussion of the findings of the study. In section 8, the conclusion and future work is discussed.

2 Related Work

In today's modern world, the detection of suspicious activities accurately and more efficiently through human activity recognition has allowed more and more advancement in creating automated frameworks using machine and deep learning. Even before deep learning technologies emerged there were efforts to create automated surveillance systems.Elhamod and Levine (2012) uses blob matching technique to extract 3d object-level information from videos while Ouivirach et al. (2013) uses the same blob-matching technique to create an incremental learning statistical model. Even though both approaches gave good results at that time the blob-matching techniques have some limitations with respect to rapid movements, object deformation, or in the case of complex backgrounds. The researchers Ben-Musa et al. (2014) and Hegde et al. (2016) presented approaches for suspicious activity detection through surveillance videos. The first approach detects suspicious activities going on in the exam hall with the help of Viola-Jones face detection, tracking, and classification algorithms while the second focuses on detecting unusual activities using the Mean Feature Point Method(MFPM). Both of these approaches use the SURF algorithm for feature extraction. Although these methods gave good accuracy, the SURF algorithm has a lot of limitations when it comes to complex 3D interactions between people, and lighting conditions and it is also not suitable for real-time scenarios where timely detection is necessary, as its relatively slow than other modern techniques.

OpenPose is one of the most powerful pose estimation models which is known for its accuracy in detecting multiple key points on the human body. As it allows multi-person pose estimation it is well suited for pose estimation of people in crowded places. Moyo et al. (2023) proposes a framework to detect suspicious activities in exam halls with the help of OpenPose. The researchers here tried to predict 4 types of activities such as passing pieces of paper, passing answer sheets, using reference material like phones, or exchanging objects. The researchers used skeleton features given by OpenPose and then converted them into desired geometric features such as joint angles. While this method seemed to perform well on the test data, it was noted that the researchers only used 10000 images for training which is very less. This might impact models' performance in real-world scenarios while detecting complex or unseen scenarios. Riaz et al. (2021) also proposed a new approach to detect suspicious activities in the exam hall with OpenPose and a cascade of deep convolutional neural network models. In this approach, researchers apply the OpenPose model to extract patches of the human body from the data using key points and later this data is used for training of DenseNet model. The researchers obtained an accuracy of 95.88% and precision, sensitivity, and F1 Score of 96% each while

detecting if the activity is normal or abnormal. The data for this research was collected and labeled manually by researchers which may cause unintentional bias in the data and there is a possibility that data might not be well generalized which may affect the model's performance in detecting new or evolving abnormal behaviors.

Rathod et al. (2020) proposes a novel approach for the purpose of smart surveillance. The proposed framework uses drone videos for the research. In this approach, the videos are sent to the server where the videos are converted into a form of GIF and then the OpenPose estimation model is applied to them for activity recognition. The researchers compared multiple models for training such as SVM, CNN, and RCN Inception models. From the results, it was found that the RCN inception model gives a decent amount of maP(mean absolute Precision) of 83.8% as compared to SVM and CNN for suspicious activity detection. Even though the novelty of the research is the use of GIFs, it might make the resolution of train data frames lower as compared to original videos which might impact OpenPose's feature extraction capability. Complex activities might involve some abnormal activity patterns lasting a few seconds, but this might get compressed or distorted which might affect model performance in the future. Emanuel et al. (2021) proposed a design for action recognition. The researchers used a dataset with 2132 images with 11 different activities. The pose detection model was used to detect 25 key points and the input was sent to 4 different classifiers to classify the 11 activities. The classifiers used were Neural Network, Random Forest, KNN, and SVM. After the performance evaluation, Random Forest was found to be having highest accuracy of 77.52%. The research showed that Random Forest works best with data in the form of key point coordinates.

Even though in most of cases the OpenPose model gives higher accuracy it demands high computational resources which makes them not suitable for lightweight applications. To overcome this limitation Mediapipe's BlazePose model can be used. Varghese et al. (2023) proposed a real framework for fitness activity recognition using BlazePose algorithm. This framework provides real time recognition and also provides real time feedback. The authors of this research framework have made use of angular features constructed using vector geometry. The features constructed using BlazePose were used to train 3 models namely, GAN's, Convo-LSTM, and LSTM-RNN. Out of these models, LSTM-RNN got the highest accuracy of 97% followed up by Convo-LSTM with 95% accuracy. It was noted by the researchers that the GAN gave an accuracy of 92% but struggled in real time because of its computational complexity. Mohammed et al. (2022) proposed a robust and lightweight framework for yoga asana recognition. During the implementation, a public dataset of 6 yoga asanas was used. The features in this research were extracted using BlazePose model and fed to a deep learning model which was a combination of CNN and LSTM. The researchers here got an accuracy of 95.29 on train data and 98.65 on test data. In real time scenario researchers tested the performance of features extraction from both BlazePose and OpenPose where it was found that features extracted by OpenPose gave an accuracy of 99.04% but OpenPose was able to work only at the rate of 0.4 FPS while BlazePose worked at 30FPS and gave the accuracy of 98.65%. It was concluded from the research that BlazePose was the optimal model for their framework in real time. Similarly, Sunney et al. (2023) proposed a home-based yoga system that uses machine learning and BlazePose model. The researchers here used a publicly available dataset with 5 yoga poses. In this approach, researchers used skeleton images given by BlazePose model and converted them into CSV format so that machine learning algorithms can process them. The researchers of this research used 6 models

for comparison namely Random Forest, XgBoost, SVM, Decision Tree, LSTM and CNN etc. From the results of the research, it was found that XgBoost performed best out of all the models with an accuracy of 95.14% and a latency of 8ms, and size of 513kb. The second-best model as per the accuracy was Random Forest and CNN, both with an accuracy of 94.70 but in this case, Random Forest was considered second best as it has lower latency and size as compared to CNN.

Nguyen et al. (2022) proposed a yoga pose grading tool that uses BlazePose model for feature creation. The researchers here created their own data and annotated it with the help of a yoga teacher. The research also highlighted the use of angular features while training the models. Four pre-trained models were used for the training and analysis of this data. These models were VGG-16, VGG-19, ResNET-101 and MobileNet. Out of these models, VGG 19 performed well with an accuracy of 95.70. Similarly,Setiyadi et al. (2022) proposed an activity detection system that used 2D function of BlazePose model and LSTM. The proposed system is used to recognize 5 types of activities such as walking, falling, sleeping on the back, sleeping on the right side, and sleeping on the left side. The researchers compare the results of three BlazePose models which are Full, Lite, and Heavy. The researchers observed that the Full type BlazePose model gives the highest accuracy of 80% with a frame rate of 25+ for all activities using LSTM.

From this section, it can be concluded that BlazePose model works more efficiently and effectively than any other model for lightweight applications. At the same time, this section identifies the importance of angular heuristics which were calculated from BlazePose model's output. One of the limitations of BlazePose was discovered in this section. This limitation states that the BlazePose model doesn't support multi-person pose estimation. To overcome this limitation 'YOLO'(YOU ONLY LOOK ONCE) which is a real time object detection algorithm was used in this research Redmon et al. (2016).

3 Methodology

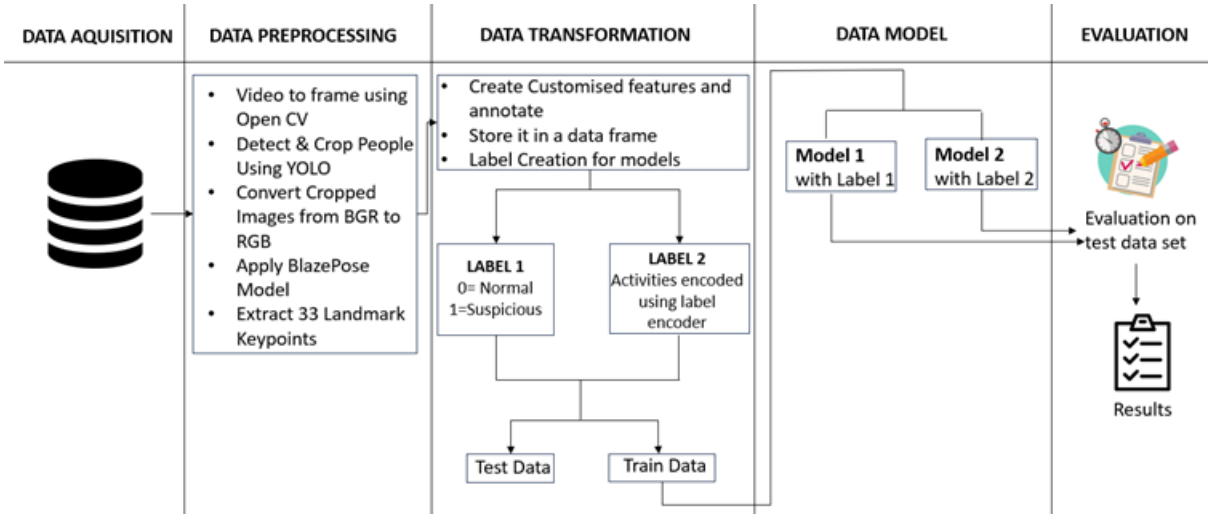

Figure 1: Methodology - Suspicious Activity Detection Framework

To achieve the aim of the research a methodology of five steps was followed. These five steps were Data Gathering, Data Preprocessing, Data Transformation, Data Modelling,

and Evaluation.

In the first step of Data gathering, a dataset consisting of different types of suspicious activities was acquired for this research. The dataset was specifically created for the task of real-world anomaly detection using deep learningSultani et al. (2018). This dataset consists of surveillance videos of 13 different types of anomalies such as Fighting, Arrest, robbery, etc which were collected from YouTube and different new websites. As the data size was huge, it was stored on the open-access Dropbox link from the creator. This data was copied to the Dropbox of the author of this research for further use. To access the data into the research script a scoped application was created on the 'Dropbox Developer' website and a secured access token was generated to access the data. After the data access token was generated, the data was downloaded onto google collab using Python's 'Dropbox' library.

The data was in the form of five zip folders where four zip folders consisted of anomaly videos and the fifth folder consisted of videos of 'Normal' events. Every folder in anomaly folders consisted of subfolders with the name of the activity and its videos inside it. All these five main folders were then extracted and combined into one master folder consisting of subfolders of all the 13 anomaly activities and 1 normal activity video. To achieve the goal of the research which is to detect suspicious activities using the pose detection model it was necessary to target human-centric activities. This led to the removal of two types of anomaly videos which were 'Road Accidents' and 'Explosion' as these videos didn't contain any human body movements which made them irrelevant to the research.

For the process of transformation, it's necessary to extract the body coordinates of each of the individuals present in every video frame. To identify the people in every frame of the video an object detection model called 'YOLO' was used Redmon et al. (2016). This model creates a red bounding box around every person in the frame and crops it. After initializing an instance of this algorithm, all the videos were looped through this model. Every nth frame was extracted from the video and passed through the YOLO model. After the YOLO model crops the people from every frame these frames are sent to the pose estimation model. To extract the body coordinates from the frame a pose estimation model called 'BlazePose' provided by 'Mediapipe' was used Bazarevsky et al. (2020). The output from the YOLO model is in BGR format but the BlazePose model requires input in RGB format. To provide the input to the pose estimation model all the cropped images were converted into RGB format using OpenCV library. The BlazePose model gives 33 different types of landmarks. These landmarks include coordinates of different body parts such as right and left eyes, both shoulders, elbows, wrists, hips etc. To make these coordinates more interpretable and useful for detecting specific human actions some additional features were calculated. These features mainly included features such as shoulder distance, elbow distance, shoulder angle and arm-to-height ratio. The formulas for these features are as follows: -

1.Shoulder Distance: -

$$shoulder\_distance = ||keypoints[LeftShoulder]–keypoints[RightShoulder]|| \quad (1)$$

2.Elbow Distance: -

$$elbow\_distance = ||keypoints[LeftElbow]–keypoints[RightElbow]|| \quad (2)$$

3.Shoulder_angle: -

$$Shoulder\_angle = arctan2(keypoints[LeftShoulder, 1]–keypoints[RightShoulder, 1], \tag{3}$$

$$keypoints[LeftShoulder, 0]–keypoints[RightShoulder, 0]) \tag{4}$$

4. Arm_to_Height Ratio: -

$$Arm\_to\_height\_ratio = elbow\_distance/keypoints[LeftHip, 1]–keypoints[RightHip, 1] \tag{5}$$

Where,

Keypoints: - coordinates of landmarks represented in arrays.
Left Shoulder, Right Shoulder, Left Elbow, Right Elbow, Left Hip, Right Hip: - Pose landmarks
$\|.\|$: −Euclidean distance between two points (Euclidean Norm)
Arctan2(y, x):-Arctangent function used to calculate angles and directions based on coordinate values

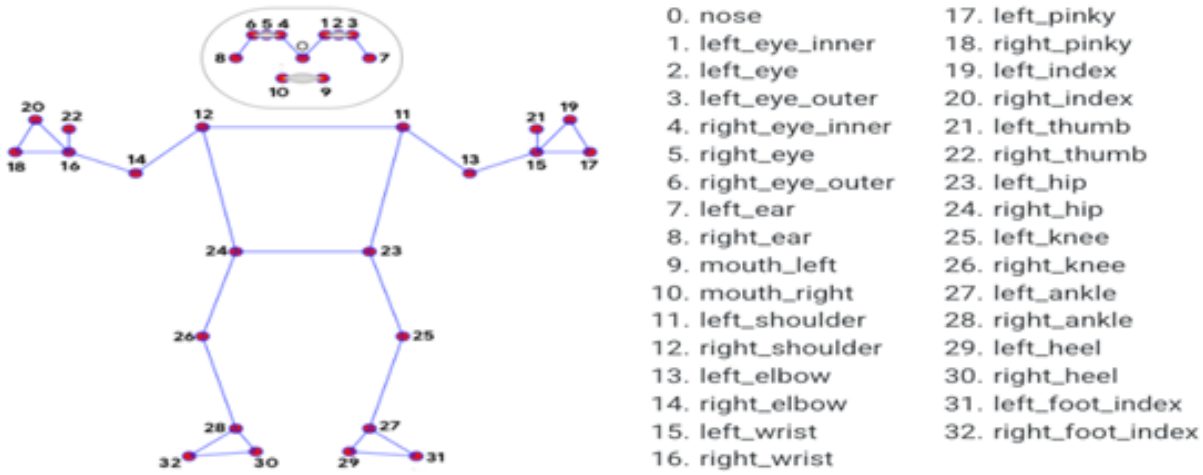

Figure 2: BlazePose Model - Topology Bazarevsky et al. (2020)

These extracted features convert complex coordinate data into meaningful numeric data which is later used by different machine learning and deep learning models for finding the pattern for suspicious behaviour. After transforming the landmark key points into meaningful features they are stored into a data frame. As these features were in different scales, Min-Max scaling was applied to these features to eliminate the possibility of potential bias due to the original scale. The Min-Max scaling helped in bringing down these features into a specific range of 0 to 1 which helped in preventing a single feature dominating the model's learning performance. After the scaling, one more set of labels was constructed where all the labels except 'Normal' were replaced with 'Suspicious' activity. Later the data was separated into dependent and independent variables and then was divided into two parts training and testing. 80% of the data was considered as training data and the remaining 20% of the data was considered as test data.

In the data modeling step, two separate models were created. The primary model contained features and the newly created labels 'Normal' and 'Suspicious'. If the output

of this model is 'suspicious', then the result is fed to the secondary model as an input to recognize the suspicious activity type. This approach was followed by the researcher of this research to reduce the complexity of the models and to maintain interpretability. In these two separate model approaches both the models used were different but were provided with similar feature extraction and the same preprocessing steps to reduce computational costs. Another advantage of this two-model approach is to enhance interpretability where insights from the interpretation of one model can be navigated to the other model. This approach included the use of 3 machine learning models Xg-Boost, Random Forest, LightGBM, and one deep learning model called Deep Neural Network. Hyperparameter Optimization and cross-validation were applied to select the best parameters and get better results for machine learning models. For DNN models' multiple dense layers were added with 'ReLU' as an activation function. The first layer input was given based on the shape of the input features provided. After every dense layer, one dropout layer was added to prevent the model from overfitting. For the primary model, the output layer was added with a number of classes to predict as 2, and for the secondary model number of classes was added as 13. Both the models were trained using the activation function as 'Softmax'. In the end, the models were compiled with 'adam' optimizer and loss function as 'binary_crossentropy' and 'sparse_categorical_crossentropy' for primary and secondary models respectively. The number of layers and neurons in these models was decided based on the evaluation matrices.

In the final step, all the models were evaluated based on different evaluation matrices such as Accuracy, Precision, Recall, etc. Based on the experimentations done, it was observed that Random Forest and XgBoost models performed best as compared to other models. In the case of the primary model Random Forest model was able to predict if the activity is suspicious or not while for the secondary model XgBoost was able to predict the suspicious activity type effectively based on BlazePose features. In the end, a combined framework was developed to detect suspicious activities using the pose estimation model BlazePose, ML models and an alert mechanism.

4 Design

The main purpose of this research is to improve state-of-the-art suspicious activity detection methods by introducing a new approach of using a pose detection model with classification algorithms. The architecture of the framework consists of Data Capture and preprocessing module, Feature Creation Using the BlazePose module, Classification Module, and Alarm System module.

The data Capture and Preprocessing module is responsible to capture live video from security cameras and preprocess them for the purpose of feature extraction through BlazePose model. The OpenCV's VideoCapture() method is used to simulate live video scenarios in this research. This method can be used to integrate with existing surveillance systems or with the user's webcam to capture the live video stream. The function creates a video capture object and all the frames given by this object are further processed to detect the number of people in the frames using pretrained 'YOLO5' model. These detected people in each frame are cropped and converted from BGR to RGB format and then sent to the BlazePose model for the detection of pose keypoints.

BlazePose Model detects 33 pose keypoints from every person's image provided. After the keypoints are extracted these keypoints are converted into features by calculating

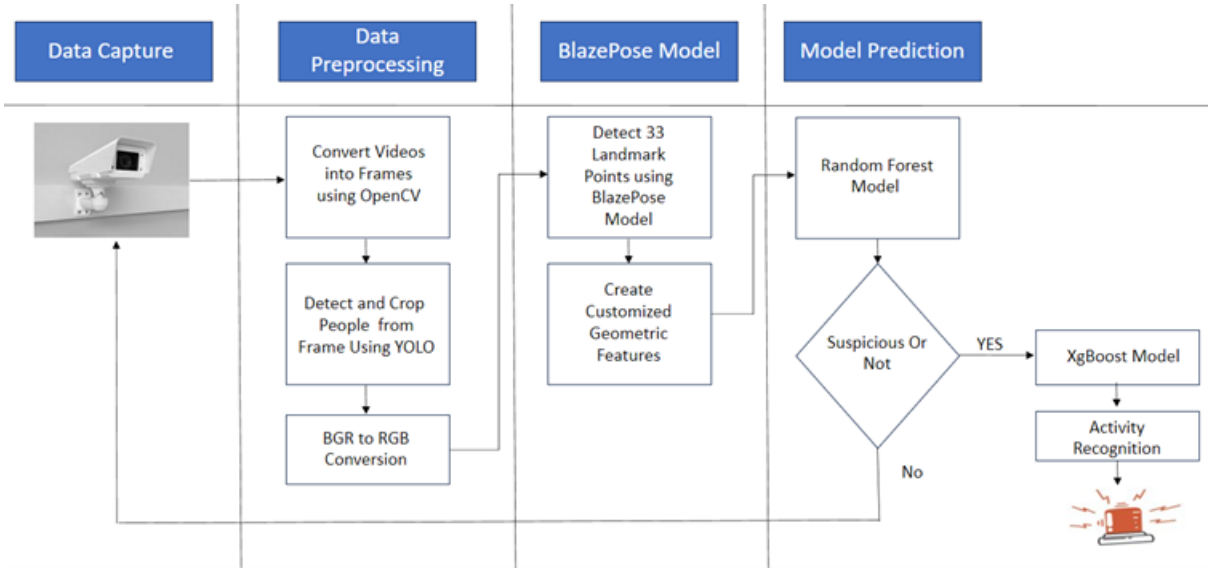

Figure 3: System Architecture

shoulder distance, elbow distance, shoulder angle, and arm-to-height ratio. These features are calculated using Euclidean distance from the output provided by the BlazePose model. These types of features are calculated to help the classification model's decision-making process be more accurate and transparent. Features like distances and angles are easier to understand by the models and can help them to capture patterns to identify different activities.

Figure 4: Original Image

In the classification module, two machine-learning models were used. The primary model was used to identify if the activity is suspicious or normal and the secondary model was used to identify the type of activity if recognized as suspicious. The model predictions were followed up by an alarm system. This alarm system gets triggered if 'K' number of frames are predicted as suspicious. The value 'K' is dynamic and can be changed as per convenience. Once the alarm system is triggered it sends an 'SMS' to the user.

Figure 5: Cropped and BlazePose Landmarkes

5 Implementation

A suspicious activity detection framework using a pose detection model was implemented using machine learning algorithms and Mediapipe's BlazePose algorithm. This framework was implemented on google colab pro with Python 3.10.12. The GPU used was an A100 GPU. For implementation purposes, few main libraries were used such as OpenCV, Mediapipe, Torch, Pandas, NumPy, TensorFlow-Keras, Sklearn, and Dropbox. While implementing the framework 'YOLO 5' model was cloned from GitHub which was later used in preprocessing. The data is present in an open Dropbox link provided by the creator of the dataset. As the data size was approximately 100 GB the data was copied from the Dropbox of the creator to the Dropbox of the author of this research. To obtain the data on colab from Dropbox one scoped application was created on dropbox developer website. This app provides an access token to the dropbox contents which is later used by Python's dropbox library to download the data on colab. Later the data was extracted using 'zipfile' library and then moved to a single folder that contained subfolders of the names of the suspicious activities. Due to the availability of limited infrastructure and the size of the data, a sampling technique was used for the video processing. Only 10 videos from each subfolder were taken which resulted in the processing of 140 videos.

After the data was ready to be processed 'YOLO 5' model and BlazePose pose estimation model instances were created. All the videos were looped for the preprocessing and feature creation. For every video 'n' number of frames were extracted where 'n' was dynamic. Once the frame is extracted it is passed to the 'YOLO 5' model instance to determine the number of people in the frame. The 'YOLO 5' model crops these people from the original image and later BlazePose model instance is applied to these cropped images to extract the 33 pose keypoints in the form of x, y, and z coordinates. These coordinates are later sent to create multiple features such as Shoulder Distance, Elbow Distance, Shoulder Angle Arm to Height Ratio, etc. These features were calculated for every frame extracted for every video and later stored in a data frame.

Four different machine learning and deep learning model performances were analysed for this study. The machine learning models such as XgBoost, Random Forest, and LightGBM were implemented and evaluated using Python's Sklearn library while the deep learning model DNN was implemented and evaluated using Keras with TensorFlow

backend. To predict the suspicious activity and its type two separate models were created. These two separate models used two different machine learning algorithms. The evaluation of the models was done based on accuracy, precision, recall, F1 Score. Based on the results from these matrices Random Forest model was found to be the most feasible for the first model while XgBoost was found to be most feasible for the second model.

The alarm system was created using 'Twilio'. Twilio is a platform that allows users to integrate communication systems such as SMS, Emails, Voice, and Video calls, etc with their applications using Python APIs. To integrate Twilio it is necessary to sign up on its website after which a phone number is assigned for free to send the SMS through Python APIs.

6 Results and Evaluation

Different types of experiments were conducted during this study. The aim of the study was to create a framework to detect suspicious activities using pose estimation model and machine learning. This section explains these experiments in detail.

6.1 Experiment 1

This experiment was conducted to replicate the state-of-the-art method used in detecting suspicious activities Amrutha et al. (2020). The authors of this research used pretrained CNN and LSTM models to detect suspicious activities from the videos. The author of this research used a combination of datasets such as KTH and CAVIAR. The researchers also integrated a few of the videos taken from YouTube and a few videos from the campus they were working on. Due to the use of multiple data sources which were used and some of them being private, only KTH dataset was used while replicating the state-of-the-art. The KTH dataset consists of videos of six types of actions which include walking, jogging, running, boxing, hand waving, and hand clapping. As per the researchers of this study, the activities such as running, and boxing were considered suspicious.

As a first step to replicate the experiment, the video data was converted into frames. Then this data was divided into an 80-20 ratio, and it was sent to pre trained VGG16 model with an image size of 224*224. As the number of LSTM layers and max pooling layers were not mentioned 3 LSTM layers of 128,128 and 64 units were used. These layers were followed by Dropout layers of 0.5 to avoid overfitting. The model was trained for 10 epochs where it was observed that it started overfitting. The validation accuracy achieved after the 10th epoch was 70.02% with the constant increase in the loss of 4.79 which is very high.

6.2 Experiment 2

The experiment of creating AI based yoga trainer using media pipe was conducted to understand the effective use of the pose detection model.Agarwal et al. (2022) The media pipe framework uses BlazePose as a pose detection framework. The researchers here have not mentioned the dataset used so an open-source dataset was used to replicate the study. The dataset contained 5 different types of yoga poses which needed to be predicted. As mentioned in the study first the videos were converted into frames using the OpenCV library. These frames were converted from BGR to RGB format as the BlazePose requires the input in RGB format. After the conversion, these frames were

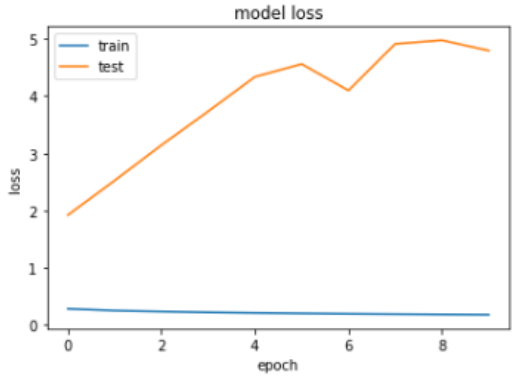

Figure 6: Training Loss Vs Validation Loss

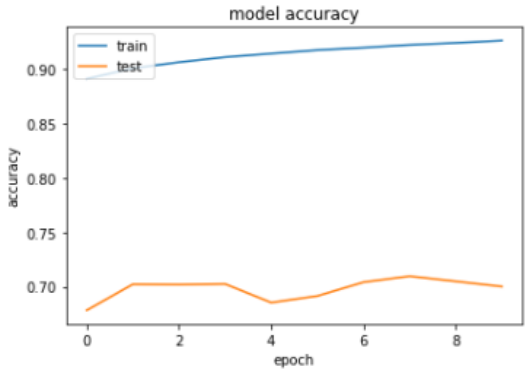

Figure 7: Training Accuracy Vs Validation Accuracy

passed as input to mediapipe BlazePose model. For each frame, BlazePose model gave a set of 33 keypoints which were later used to create angular features. The study states that three different model performances were studied during the research which were KNN, SVM, and Random Forest. The authors of this study used these models on their own dataset for training and testing. The results from the author concluded that the optimized SVM model works best with the accuracy of 94% and precision, recall, and F1 Score of 94%,92%, and 93% respectively. As the data used while replicating the study was different, the replicated results showed that the Random Forest model performed better than the SVM model with the accuracy of 86% and Precision, Recall, and F1 Score of 85%, 86% and 85% respectively as compared to the accuracy of SVM model with the accuracy of 80% and Precision, Recall and F1 Score of 79%, 80% and 78% respectively.

This study gives an idea about the use of angular heuristics for feature creation and helps in understanding how angular features derived from keypoints given by BlazePose can help in determining the anomalies in the deviation of body parts. The same technique is used in experiment 3 to create one part of the features.

Model Name	Accuracy	Precision	Recall	F1 Score
SVM	80%	79%	81%	78%
Random Forest	86%	85%	86%	85%

Table 1: Experiment 2 Model Results

6.3 Experiment 3

In this experiment, the performance of two types of models was evaluated. The primary model was created to identify if the activity is suspicious or not and the secondary model was created to identify the type of activity if predicted as suspicious by the primary model. Four types of models were tested to find out the best primary and secondary models. These models were XgBoost, Random Forest, LightGBM, and Deep Neural Network.

While evaluating the results it was found that Random Forest model performed best as compared to other models with the accuracy of 93% and precision, recall and F1 Score of 65%, 79% and 69% respectively. The performance of XgBoost and Random Forest

Model Name	Accuracy	Precision	Recall	F1 Score
XgBoost	92%	64%	79%	68%
Random Forest	93%	65%	79%	69%
LightGBM	92%	64%	80%	69%
Deep Neural Network	80%	79%	57%	57%

Table 2: Primary Model Results

models were almost similar, both with 92% of accuracy. The performance of the DNN model was the lowest as compared to other models where the accuracy was 80%.

Model Name	Accuracy	Precision	Recall	F1 Score
XgBoost	70%	65%	73%	68%
Random Forest	69%	65%	72%	68%
LightGBM	65%	53%	70%	58%
Deep Neural Network	24%	26%	39%	22%

Table 3: Secondary Model Results

While evaluating the results for the secondary model it was observed that XgBoost model performs really well with an accuracy of 70% and precision, recall, and F1 score of 65%,73%, and 68% respectively. The XgBoost is able to identify the types of activities more accurately than others. Again, the performances of XgBoost and Random Forest were almost similar and DNN was the worst performing algorithm amongst all.

7 Discussion

This section performs a detailed discussion of the findings obtained from all the experiments conducted during this research. The study utilized UCF-Crime dataset to create a suspicious activity detection framework using machine learning and pose detection. Mediapipe's BlazePose model was used to extract key point features which were later converted into customized pose geometric features to capture the relationship between key body joints and the joint angles. The framework consisted of two models, one model was aimed at detecting if the activity is suspicious or not, and the second model was aimed at recognizing the activity type. This study performed an evaluation of 4 different algorithms for both models. These algorithms were XgBoost, Random Forest, LightGBM, and DNN.

The first experiment in the study was conducted to understand current approaches to detect suspicious activities and understand their limitations. The second experiment was conducted to understand the working of Mediapipe's BlazePose pose estimation model. The third experiment consists of making a framework to detect suspicious activities and evaluating the performance of models on the generated data to detect the best model. All these experiments were conducted on Google Colab Pro with 83GB RAM and an A100 GPU hardware accelerator. These experiments can also be replicated on a machine with the configuration of 16 GB RAM and Nvidia Graphics Card of 8 GB but the machine might take some extra time for the execution of the experiments.

The results of the study indicated that Random Forest works best for the prediction of suspicious activities with an accuracy of 93% and XgBoost performs better as compared

to other models in the case of activity recognition with an accuracy of 70%. The findings of the research suggest that random forest is robust at handling complex datasets and XgBoost is effective at finding out complex relationships. This finding aligns well with the prior research in the literature review where it was discovered that tree ensemble methods such as XgBoost and Random Forest work better than other algorithms while handling complex classification tasks. The study also implies that the use of customised features can help in increasing the performance of the model.

Even though the results of the whole framework are average there are some limitations that need to be addressed. The data used in this research is a collection of surveillance videos. As these videos are captured from surveillance cameras the video quality of some of the videos is a little low. Even though the YOLO model used in this research has the ability to find out the people from blurred images and crop them into separate images, the BlazePose algorithm finds it difficult to process those images. The results from model one to detect if the activity is suspicious or not are good but the second model where the name of the suspicious activity is recognized are not good as compared to the first one. The model is getting confused in recognizing certain activities such as Fighting, Abuse, Arrest, etc. This might be happening due to the similar poses or body movements in these activities which is making the model struggle to differentiate between these poses for example, the model might get confused in scenarios where the person is changing its posture from normal posture to shooting posture and it might resemble as just raising an arm to the model. This shows that the results need to be analysed to find out where the model is getting confused, and the model needs to be fine tuned more for better performance.

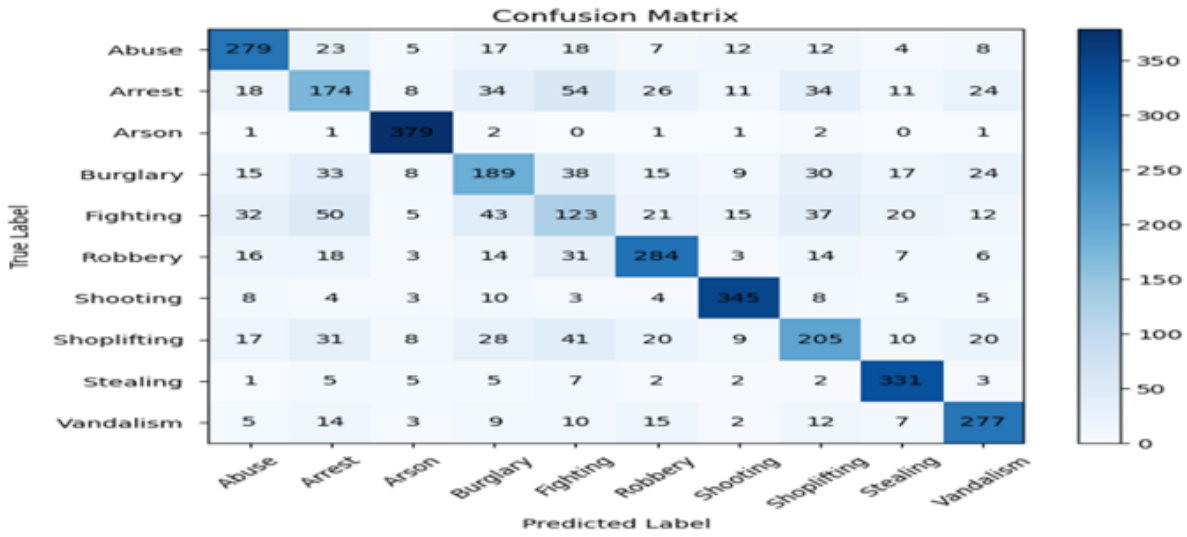

Figure 8: Confusion matrix for activity recognition model

8 Conclusion and Future Work

The aim of this research was to understand how well a lightweight pose estimation model in conjunction with machine learning can identify suspicious activity. This aim was achieved by using Mediapipe's BlazePose model and two machine learning algorithms.

The dataset used for the study was a publicly available dataset called UCF-Crime which consists of surveillance videos of different suspicious activities such as Abuse, Vandalism, Fighting, etc. To achieve the aim of the study, pose landmark key points were extracted and used to create customized geometric features. Later, two models were created where one model was used to predict if the activity is suspicious or not, and the second model was used to recognize the type of activity if it is predicted as suspicious by the first model. A comparison of 4 machine learning and deep learning models was performed in this research to find out the best model for classification. Out of these four models, the Random Forest model gave the highest accuracy of 93% in the detection of suspicious activity while XgBoost gave the highest accuracy of 70% in activity recognition. Based on the study findings, it was concluded that suspicious activity detection can be achieved using machine learning with a pose estimation model.

As the framework created for this study uses Blazepose model and machine learning algorithms, it is very lightweight which makes it easy to integrate into any surveillance system with some minor changes. Even though the Blazepose model doesn't support multi-person detection, this problem was solved with the integration of YOLO into the framework. The framework also uses geometric features for training and classification which allows suspicious activity detection and recognition in complex scenarios.

The results of the study showed that the models show good performance in suspicious activity detection, but their performance becomes moderate while recognizing the type of activity. This moderate performance allows further investigation to find advanced techniques to recognize these activities accurately. For future work, the research framework can be integrated with object detection techniques to identify if the person in the video frame is carrying a weapon or not which will help in making the framework more accurate. In the future, audio features from videos can be extracted as some unusual or suspicious scenarios are based on audio features like shouting, screaming, etc, and at the same time, contextual cues can also be provided to make the framework more robust and accurate. This research tries to set a stage for the advancement of security systems and ultimately contribute towards the creation of a safer society.

References

Affonso, G. A., De Menezes, A. L., Nunes, R. B. and Almonfrey, D. (2021). Using artificial intelligence for anomaly detection using security cameras, *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, IEEE, pp. 1–5.

Agarwal, V., Sharma, K. and Rajpoot, A. K. (2022). Ai based yoga trainer - simplifying home yoga using mediapipe and video streaming, *2022 3rd International Conference for Emerging Technology (INCET)*, pp. 1–5.

Amrutha, C., Jyotsna, C. and Amudha, J. (2020). Deep learning approach for suspicious activity detection from surveillance video, *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, IEEE, pp. 335–339.

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F. and Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking, *arXiv preprint arXiv:2006.10204* .

Ben-Musa, A. S., Singh, S. K. and Agrawal, P. (2014). Suspicious human activity recognition for video surveillance system, *International conference on control, instrumentation, communication and computational technologies (ICCICCT)*, pp. 214–218.

Elhamod, M. and Levine, M. D. (2012). Automated real-time detection of potentially suspicious behavior in public transport areas, *IEEE Transactions on Intelligent Transportation Systems* **14**(2): 688–699.

Emanuel, A. W., Mudjihartono, P. and Nugraha, J. A. (2021). Snapshot-based human action recognition using openpose and deep learning, *Snapshot-Based Human Action Recognition using OpenPose and Deep Learning* **48**(4): 2–8.

Hegde, C., Chundawat, S. S. and Divya, S. (2016). Unusual event detection using mean feature point matching algorithm, *International Journal of Electrical and Computer Engineering* **6**(4): 1595.

Jain, A. and Vishwakarma, D. K. (2020). State-of-the-arts violence detection using convnets, *2020 international conference on communication and signal processing (ICCSP)*, IEEE, pp. 0813–0817.

Mohammed, S. W., Garrapally, V., Manchala, S., Reddy, S. N. and Naligenti, S. K. (2022). Recognition of yoga asana from real-time videos using blaze-pose, *International Journal of Computing and Digital Systems* **12**(1): 1304–1295.

Moyo, R., Ndebvu, S., Zimba, M. and Mbelwa, J. (2023). A video-based detector for suspicious activity in examination with openpose, *arXiv preprint arXiv:2307.11413* .

Nguyen, T. N., Tran, T.-H. and Vu, H. (2022). An automatic tool for yoga pose grading using skeleton representation, *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*, IEEE, pp. 187–192.

Ouivirach, K., Gharti, S. and Dailey, M. N. (2013). Incremental behavior modeling and suspicious activity detection, *Pattern recognition* **46**(3): 671–680.

Rathod, V., Katragadda, R., Ghanekar, S., Raj, S., Kollipara, P., Anitha Rani, I. and Vadivel, A. (2020). Smart surveillance and real-time human action recognition using openpose, *ICDSMLA 2019: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications*, Springer, pp. 504–509.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016). You only look once: Unified, real-time object detection, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Riaz, H., Uzair, M., Ullah, H. and Ullah, M. (2021). Anomalous human action detection using a cascade of deep learning models, *2021 9th European Workshop on Visual Information Processing (EUVIP)*, IEEE, pp. 1–5.

Setiyadi, S., Mukhtar, H., Cahyadi, W. A., Lee, C.-C. and Hong, W.-T. (2022). Human activity detection employing full-type 2d blazepose estimation with lstm, *2022 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, IEEE, pp. 1–7.

Sultani, W., Chen, C. and Shah, M. (2018). Real-world anomaly detection in surveillance videos, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488.

Sunney, J., Jilani, M., Pathak, P. and Stynes, P. (2023). A real-time machine learning framework for smart home-based yoga teaching system, *2023 7th International Conference on Machine Vision and Information Technology (CMVIT)*, IEEE, pp. 107–114.

Varghese, M. M., Ramesh, S., Kadham, S., Dhruthi, V. and Kanwal, P. (2023). Real-time fitness activity recognition and correction using deep neural networks, *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, IEEE, pp. 1–6.