# CrimeEmoGuardNet : Dense-Custom Crime Detection with Emotion Recognition

**Abstract** In the ever-evolving field of video analytics, recent trends have led to more intelligent surveillance systems. This paper introduces an intelligence surveillance system by combining Crime Detection with Emotion Recognition. This system uses advanced computer vision and deep learning techniques to detect theft, abuse, and vandalism in public places such as malls. The increase in complexity of security challenges calls for a more nuanced approach and this paper proposes a system that integrates Crime detection and emotion recognition which adds an extra layer of insight into the situation in the realm of security applications. The Crime detection component of the Intelligent Surveillance System recognizes anomalies within a monitored environment like a shopping complex. Machine learning models are trained on diverse datasets, and the system learns to identify anomalies present. The emotion recognition component of the intelligent surveillance system adds a human-centric dimension to the surveillance system. The system analyzes body language for the anomalies before the actual Crime takes place. This paper presents an Intelligent surveillance system that embraces the recent trends in computer vision for video surveillance. The proposed system which integrates Crime detection and Emotion recognition addresses the limitations of traditional surveillance systems, providing a comprehensive and adaptive solution for enhanced security.

**Keywords** Crime Detection, Body Language Detection, Computer Vision, Emotion Recognition

## 1. Introduction

In recent years, there has been a surge in interest surrounding the utilization of CCTV footage for the advancement of surveillance systems focused on crime detection.By harnessing advanced computer vision algorithms and machine learning techniques, this interdisciplinary approach provides a comprehensive understanding of individuals' actions captured in video data, leading to enhanced crime detection capabilities.

However, amidst the pursuit of technological advancements, cities like Bhubaneswar in India are facing challenges that extend beyond the digital realm. As per a Fox Business

report, stores lost an estimated $86.6 billion to retail theft in 2022, with forecasts indicating that it might rise to $115 billion by 2025 after research conducted by Capital One Shopping [1]. The alarming statistics shed light on the pressing issue of retail theft and its detrimental effects on the economy. A report by Capital One Shopping delved into the root causes of this phenomenon, uncovering various factors contributing to the rise in retail theft. According to Egel, a spokesperson interviewed by Fox News, the COVID-19 pandemic played a pivotal role in exacerbating the problem. Prior to the pandemic, crime rates were on a downward trajectory. However, the widespread shutdowns and economic downturn triggered by the pandemic disrupted the social and economic fabric of America, leading to a surge in criminal activities. Egel emphasized that the pandemic-induced disruptions not only impacted the economy but also eroded common sense among individuals, resulting in a spike in retail theft incidents. As businesses grappled with the challenges posed by the pandemic, opportunistic individuals seized the opportunity to engage in illicit activities, further exacerbating the problem [1].

The repercussions of retail theft extend beyond financial losses for businesses, impacting consumers and communities at large. From higher prices for consumers to job losses and reduced investment in local communities, the ripple effects of retail theft are far-reaching and detrimental to societal well-being. In addition to retail theft, cities like Bhubaneswar are also grappling with the issue of vandalism [2]. Vandals have resorted to new methods, such as painting festival greetings on the surfaces of major roads, leading to concerns regarding both aesthetics and road safety. Road safety expert Syed Maqbool Ali emphasized that these acts not only detract from the city's aesthetic appeal but also pose a danger to road users. The use of bright white paint for these messages creates a jarring visual effect, increasing the risk of accidents as drivers and commuters become distracted.

Previously, vandalism was primarily confined to boundary walls and flyovers, but it has now escalated to more prominent and potentially hazardous targets, prompting concerns among city authorities and residents alike. Despite initiatives by the Bhubaneswar Municipal Corporation to decorate public spaces with thematic paintings as a deterrent, the problem persists, indicating a need for renewed efforts to protect the city's public spaces from defacement. Addressing the issues of retail theft and vandalism requires a multifaceted approach involving collaboration between businesses, law enforcement agencies, and policymakers. Enhanced security measures, investment in technology, public awareness

campaigns, and stricter measures to prevent vandalism are essential components of a comprehensive strategy to combat these challenges and safeguard businesses, communities, and public spaces against their adverse effects. In this context, the integration of Crime detection with emotion recognition presents a promising avenue for enhancing security and preserving the aesthetic and safety of urban environments. Through collaboration and innovation, cities can strive to create safer, more vibrant, and aesthetically pleasing environments for their residents and visitors alike.

To address the above challenges in public safety, this paper proposes an interdisciplinary approach that integrates Crime detection and emotion recognition. By utilizing Convolutional Neural Network(CNN) models, this approach aims to analyze individuals' actions and emotional states extracted from video data, thereby enhancing the effectiveness of crime detection measures.This paper covers the datasets employed, preprocessing procedures, model selection, training methodologies, and the validation process to ensure the reliability of the findings.

## 2. Literature Survey

Recent advancements in video analysis, driven by deep learning and large datasets, have propelled action recognition, Crime detection, and emotion recognition. This review synthesizes methodologies and outcomes from recent papers in these domains [3]. Human action recognition, Crime detection, and emotion recognition from videos are critical tasks with broad applications. Deep learning architectures have significantly improved accuracy and efficiency in these areas [4]. Innovations like HAANet (Hierarchical Attention-based Actor-Critic for Vehicle Navigation in Urban Environments) integrate gestures and emotions for improved action recognition, addressing challenges in differentiating actions with similar visual cues [5]. Sophisticated techniques such as double-flow Convolutional Long Short-Term Memory (ConvLSTM) variational autoencoders tackle unbalanced datasets and achieve robust Crime detection performance [6]. Works like "Happy Emotion Recognition From Unconstrained Videos Using 3D Hybrid Deep Features" emphasize facial expressions' significance, leveraging deep learning and multimodal datasets for accurate emotion classification [7]. Comparative analyses across papers highlight the effectiveness of various techniques in addressing specific challenges within each domain, with future directions

focusing on integration of multimodal cues and scalability [8]. Proposes a cross-epoch learning (XEL) strategy with a hard instance bank(HIB) to address the data imbalance issue and enhance the model's performance.It introduces additional information from previous training epochs and two new loss functions to achieve higher detection rates and lower false alarm rates for Crime events [9]. The pipeline for analyzing the behavior of students in online learning utilizes video facial processing techniques to predict engagement levels, individual emotions, and group-level affect in real-time. The pipeline, tested on datasets from EmotiW challenges, offers potential for enhancing online learning experiences [10]. ]The recognition of happy emotions from unconstrained videos is addressed using three dimensional (3D) hybrid deep features, leveraging convolutional neural networks and extreme learning machines. The proposed system demonstrates superior performance in challenging scenarios, emphasizing the impact of feature extraction methods and network sizes on recognition accuracy [11]. A two-stage classification approach outperforms a seven-class classifier for emotion recognition, leveraging CNNs for facial descriptors extraction and LSTMs for analyzing temporal dynamics of emotions in speech. Evaluation on the Multimodal Emotion Recognition dataset validates the efficiency of the proposed approach [12]. Face recognition based on videos is enhanced using convex hulls, reducing disk space requirements and speeding up testing processes. The proposed method utilizes support vector data description for essential sample extraction, offering efficiency benefits for large-scale face recognition systems [13]. Deep learning-based approaches for inappropriate content detection in YouTube videos leverage EfficientNet CNNs, Bidirectional Long-Term Short Term Memory (BiLSTM) networks, and attention mechanisms. While offering real-time processing capabilities and high accuracy, challenges include data scarcity and computational complexity [14]. The proposed framework for automatic detection of road accidents in surveillance videos utilizes stacked autoencoders to determine accident likelihood based on deep representation and reconstruction error. Incorporating vehicle trajectory intersection points minimizes false alarms, demonstrating effectiveness on real accident videos [15]. An ontology-based algorithm for event classification in basketball videos leverages global and collective motion patterns, extracting features using optical flow. The algorithm identifies effective video segments for classification, showing promise for enhancing event analysis in sports videos [16]. Hybrid Convolutional Variational Autoencoder (CNN-VAE) architecture is proposed for trajectory classification and Crime detection in videos, showing improved accuracy compared to traditional neural network classifiers. The system addresses violations in lane driving, sudden

speed variations, and abrupt vehicle terminations, demonstrating potential for enhancing traffic surveillance [17].

The main gap identified in these researches are the distinct lack of integrating crime detection and emotion recognition.The primary objective of this paper is to bridge this gap by combining these two approaches, thereby significantly enhancing the performance of surveillance systems which promises faster detection of incidents and deeper insights into the emotional states of subjects based on their body language cues.

## 3. Materials and Methods

This section encompasses the various stages involved in crime detection with emotion recognition. This includes the analysis of the dataset, preprocessing, model training, and evaluation. Each of these stages plays a crucial role in ensuring the integrity and reliability of the results obtained. In this section, a detailed overview of the steps undertaken to conduct the study effectively has been presented.

### 3.1 Datasets

To train and evaluate the proposed system, two datasets have been considered: University of Central Florida (UCF)-Crime Dataset [17], a benchmark for crime detection and Body Language Dataset(BoLD)[18], for analyzing body language. Both the datasets offer unique characteristics and features that contribute to the development and evaluations of machine learning models in their respective domains.

#### 3.1.1 UCF- Crime Dataset

The UCF-Crime dataset is a collection of real-world surveillance videos of criminal activities and suspicious behaviors. This dataset consists of a diverse range of criminal incidents such as theft , vandalism , assault and loitering recorded across various environments such as parking lots, streets and commercial establishments such as mall.The duration of each video footage varies , ranging from 30 seconds to almost one minute 30 seconds.

The dataset description is presented in Table 1 and one sample frame corresponding to the three classes are presented in Figure 1.

**Table 1** : Dataset Description for UCF- Crime Dataset

| | |
|---|---|
| **Number of Classes** | 3 (Shoplifting, abuse, vandalism) |
| **Total Number of video footages** | 150 |
| **Number of video footages in class 1(Shoplifting)** | 50 |
| **Number of video footages in class 2 (Abuse)** | 50 |
| **Number of videos footages class 3 (Vandalism)** | 50 |

(a) (b) (c)

Figure 1. Sample images from UCF crime dataset (a) Vandalism (b) Abuse (c) Shoplifting

### 3.1.2 Body Language Dataset(BoLD)[19] :

The Body Language Dataset (BoLD) focuses on capturing human behavior and emotional expressions through the analysis of body language.This dataset is constructed to be a large scale , fully annotated dataset to study and understand in-the-wild human emotion from body language.

This dataset contains 9827 video clips and 13,239 instances ie., persons of interest in these clips.There are a total of 26 classes of emotions in the body language dataset but only 3 classes of emotions are used. These classes are fear, anger, and suffering.Each video's duration varies from minimum of 3 seconds and maximum of almost 10 seconds.

(a) (b) (c)

Figure 2. Sample images from BoLD dataset (a)Anger (b) Suffering (c) Fear

**Table 2**: Dataset Description for BoLD Dataset

| **Number of Classes** | 3(Anger ,Suffering ,Fear) |
|---|---|
| **Total Number of videos** | 975 |
| **Number of videos in class 1 (Fear)** | 321 |
| **Number of videos in class 2(Anger)** | 424 |
| **Number of videos in class 3(Suffering)** | 230 |

### 3.2 CrimeEmoGuardNet:

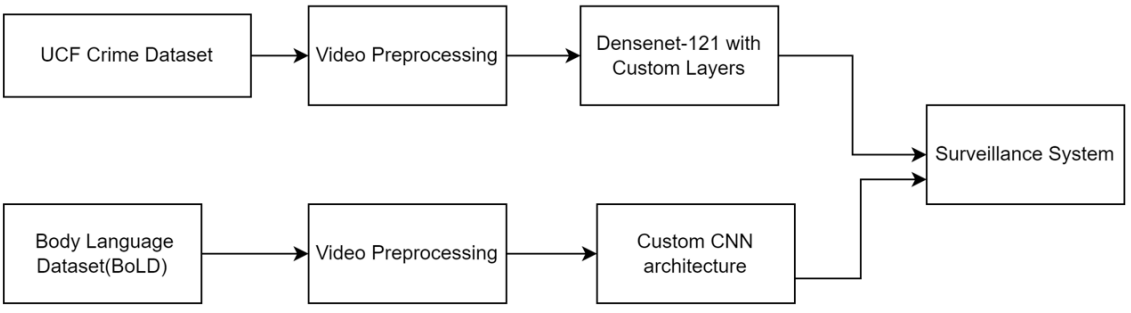


Fig 1 Proposed Framework

One of the main data preprocessing tasks performed on the video dataset before evaluating the Crime detection methods is converting the videos into individual frames. Processing and

analyzing video data pose significant challenges due to the large volume of data and the complexity of video structures hence the video data is converted into individual frames. This process involves extracting each frame from the video stream and saving it as a separate image file.This process is performed using OpenCV (Open Source Computer Vision Library). OpenCV provides a comprehensive set of functions for capturing, processing, and analyzing video data. Using OpenCV, videos can be read frame by frame, and each frame can be saved as an image file in formats such as Joint Photographic Experts Group (JPEG) or Portable Network Graphics (PNG). There are various benefits for converting videos into individual frames, namely the reduced data volume since some of the videos in the dataset are very large in size and by converting videos into frames, the data volume is significantly reduced, as each frame is represented as a separate image file. Another advantage is that the frames can be processed in parallel, enabling efficient utilization of computational resources and reducing processing time. This is particularly advantageous when analyzing large video datasets. Data augmentation is a fundamental technique used in machine learning and deep learning to artificially expand the size of a dataset by generating new, slightly modified versions of existing data samples. This process involves applying a variety of transformations to the original data, such as rotation, scaling, cropping, flipping, and adding noise. The goal of data augmentation is to introduce variability and diversity into the training dataset, thereby improving the robustness and generalization ability of machine learning models. In image classification and computer vision tasks, common augmentation techniques include rotation, translation, scaling, flipping, cropping, shearing, brightness adjustment, contrast adjustment, and adding Gaussian noise. These transformations help create variations in the appearance of objects and scenes, making the model more robust to different viewpoints, lighting conditions, and image distortions.

For the Crime detection task, DenseNet121[19] model has been exploited in consideration of its exceptional performance in image classification tasks.The utilization of DenseNet121, a variant comprising 121 layers, is motivated by its outstanding performance on benchmarks such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Pre-trained on the ImageNet dataset, DenseNet121 serves as a feature extractor, capturing hierarchical representations of input images through its deep architecture. Leveraging transfer learning, the learned features from DenseNet121 are exploited to enhance the discriminative capabilities of the system.

In conjunction with DenseNet121, a custom classification model tailored to the dataset and objectives is constructed. The architecture comprises densely connected layers, interspersed with batch normalization and dropout layers to enhance generalization and mitigate overfitting. The initial layer consists of 64 neurons activated by rectified linear units (ReLU), serving as an intermediate feature representation. The architectural details of DenseNet-121 based crime detection network is presented in Figure.

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_7 (Dense)             (None, 128)               2097280

 batch_normalization_6 (Bat  (None, 128)               512
 chNormalization)

 dropout_6 (Dropout)         (None, 128)               0

 dense_8 (Dense)             (None, 16)                2064

 batch_normalization_7 (Bat  (None, 16)                64
 chNormalization)

 dropout_7 (Dropout)         (None, 16)                0

 dense_9 (Dense)             (None, 2)                 34

 batch_normalization_8 (Bat  (None, 2)                 8
 chNormalization)

 dropout_8 (Dropout)         (None, 2)                 0

 dense_10 (Dense)            (None, 3)                 9

=================================================================
Total params: 2099971 (8.01 MB)
Trainable params: 2099679 (8.01 MB)
Non-trainable params: 292 (1.14 KB)
_________________________________________________________________
```

Fig 8 : Architectural details of DenseNet-121 based Crime Detection Model

The CNN model for video emotion recognition adopts a sequential architecture, consisting of densely connected layers with rectified linear unit (ReLU) activation functions. The input layer has a shape defined by the feature vector extracted from the last layer of a pre-trained neural network, specifically 2048-dimensional. This feature vector encapsulates high-level visual representations learned from the input video frames. Following the input layer, batch normalization layers are introduced to stabilize the learning process by normalizing the activations of each layer. Dropout layers with a dropout rate of 0.5 are incorporated to prevent overfitting by randomly deactivating neurons during training. Subsequently, the model comprises several dense layers, each comprising 128, 64, and 4 units, respectively, with ReLU activation functions. These layers facilitate the aggregation of spatial and temporal features learned from the convolutional layers. The final dense layer employs a softmax activation function to output class probabilities, enabling the model to predict the

likelihood of each emotion class given the input video sequence.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_3 (Dense)             (None, 128)               524416

 batch_normalization_3 (Bat  (None, 128)               512
 chNormalization)

 dropout_3 (Dropout)         (None, 128)               0

 dense_4 (Dense)             (None, 64)                8256

 batch_normalization_4 (Bat  (None, 64)                256
 chNormalization)

 dropout_4 (Dropout)         (None, 64)                0

 dense_5 (Dense)             (None, 4)                 260

 batch_normalization_5 (Bat  (None, 4)                 16
 chNormalization)

 dropout_5 (Dropout)         (None, 4)                 0

 dense_6 (Dense)             (None, 3)                 15

=================================================================
Total params: 533731 (2.04 MB)
Trainable params: 533339 (2.03 MB)
Non-trainable params: 392 (1.53 KB)
_________________________________________________________________
```

Fig 9 : Architectural details of Emotion Recognition Models

Upon training completion, the model's performance is evaluated on the test data, computing the test loss and accuracy. This comprehensive approach ensures the robustness and effectiveness of the CNN model for emotion recognition in videos.

### 3.4 Evaluation Metrics

The evaluation of classification models is crucial in assessing their performance and effectiveness in real-world applications. Among the fundamental tools for evaluation, the confusion matrix stands as a cornerstone, offering a comprehensive summary of a model's performance across different classes.

### 3.4.1 Confusion Matrix Overview:

The confusion matrix is a square matrix that compares the actual class labels of a dataset with the predicted class labels generated by the classification model. It organizes predictions into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing insights into the model's ability to correctly classify instances.

#### 3.4.1.1 Key Components:

True Positives (TP): Instances correctly predicted to belong to a specific class.

True Negatives (TN): Instances correctly predicted to not belong to a specific class.

False Positives (FP): Instances incorrectly predicted as belonging to a class they do not belong to (Type I error).

False Negatives (FN): Instances incorrectly predicted as not belonging to a class they do belong to (Type II error).

Performance Metrics:

From the confusion matrix, various performance metrics can be derived to evaluate the model's effectiveness, including:

Accuracy: The proportion of correctly classified instances over the total, calculated as (TP + TN) / Total.

Precision: The proportion of true positive predictions out of all instances predicted as positive, calculated as TP / (TP + FP).

Recall (Sensitivity): The proportion of true positive predictions out of all actual positive instances, calculated as TP / (TP + FN).

Specificity: The proportion of true negative predictions out of all actual negative instances, calculated as TN / (TN + FP).

F1 Score: The harmonic mean of precision and recall, offering a balanced measure, calculated as 2 * (Precision * Recall) / (Precision + Recall).

The confusion matrix provides valuable insights into the strengths and weaknesses of a

classification model, enabling practitioners to refine its performance and address specific challenges. It serves as a fundamental tool for model evaluation and validation, guiding decisions in machine learning applications.

## 4. Results and Discussion

### 4.1 Experimental Setup:

The dataset used for Crime detection consists of three classes of anomalies, namely shoplifting, vandalism and abuse with each class containing around 30 videos each.The videos are divided into train and test set with 75% is used for training and 25% is used for testing.The videos are converted to individual frames to reduce space and computational duration for analysis. The converted dataset is then normalized for feeding into the model for training.

The dataset used for Body Language classification consists of video files and corresponding joints files which consists of 2D pose estimation results.The npy files are used to identify the joints present in the video such as nose , neck , right shoulder , left shoulder etc.There are also two csv files train and test where train file consists of information related to 80% of video data while the test file consists of information from the rest.The two files categorize the videos into a total of 26 emotions out of which 3 emotions are taken into consideration , namely anger,fear and Doubt/Confusion.

#### 4.1.1 Involved Software and Hardware:

The primary objective for evaluation in Crime detection for videos is to attain an optimally trained model for each Crime detection method. This entails optimizing both the hyperparameters of the model as well as the trainable parameters.

#### 4.1.1.1 Software:

The entire algorithms have been implemented in python.The deep learning approaches have been mostly implemented using Tensorflow and Keras libraries.The data processing steps were implemented using OpenCV , numpy and pandas.

#### 4.1.1.2 Hardware:

The training and testing of the models were performed on the following hardware:

**Table 3** : Hardware Specifications

| **Artifact** | **Value** |
|---|---|
| CPU Model name | 12th Gen Intel(R) Core(™) i7-12650H 2.30GHz |
| RAM | ~15.7 |

All computation has been performed on a single process and a single thread.

### 4.1.2 Hyperparameter Tuning:

Some of the hyperparameters are general and dependent on the datasets used and others are listed for a specific approach.These are the following:

#### 4.1.2.1 UCF-Crime Dataset:

**Table 4** : Hyperparameters for UCF Crime Dataset

| **Hyperparameter** | **Value** |
|---|---|
| Batch size | 5 |
| Epochs | 22 |
| Learning rate | 0.0001 |
| Optimizer | Adam |
| Regularization | L1,Dropout |

Training and Testing data is split in such a way that 75% is used for training and 25% is used

for testing.

#### 4.1.2.2 Body Language Dataset(BoLD):

**Table 5** : Hyperparameters for Body Language Dataset(BoLD)

| **Hyperparameter** | **Value** |
|---|---|
| Batch size | 4 |
| Epochs | 50 |
| Learning rate | 0.0001 |
| Optimizer | Adam |
| Regularization | L1,Dropout |

Training and Testing data is split in such a way that 75% is used for training and 25% is used for testing.

## 4.2 Results:

In this chapter , the results of the approaches used are presented. The following sections list the confusion matrices and the other evaluation criterias for all the approaches.

### 4.2.1 Results for Crime Detection:

**Table 6** : Classification report for UCF Crime Dataset

| Dataset | Architecture | Precision | Recall | FI-score |
|---|---|---|---|---|
| UCF-crime | CNN Densenet121 with 3 dense layer with 128,16,2 neurons in each respectively.All layers have dropout rate 0.5 | 0.75 | 0.75 | 0.74 |
| UCF-crime | CNN Densenet121 with 3 dense layer with 64,32,1 neurons in each respectively.All layers have dropout rate 0.5 | 0.68 | 0.66 | 0.66 |

### 4.2.2 Results for Body Language Detection:

**Table 7** : Classification report for Body Language Dataset(BoLD)

| Dataset | Architecture | Precision | Recall | FI-score |
|---|---|---|---|---|
| BoLD | CNN Densenet121 with 3 dense layer with 128,64,32 neurons in each respectively.All layers have dropout rate 0.5 | 0.34 | 0.35 | 0.35 |
| BoLD | CNN , 2 Dense Layer 64,32 neurons in each respectively.All layers have dropout rate 0.5 | 0.26 | 0.38 | 0.35 |

4.2.3 Confusion Matrix:

CNN Densenet with 3 dense layer with 128,16,2 neurons in each respectively

| | | Predicted | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| Actual | 0 | 1573 | 339 | 622 |
| | 1 | 561 | 1963 | 1 |
| | 2 | 1040 | 62 | 1493 |

Fig 9 Confusion Matrix for CNN Densenet Model 1

CNN Densenet with 3 dense layer with 32,8,1 neurons in each respectively

| | | Predicted | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| Actual | 0 | 893 | 849 | 792 |
| | 1 | 378 | 2145 | 2 |
| | 2 | 1065 | 110 | 142 |

Figure 10 Confusion Matrix for CNN Densenet Model 2

CNN Model for Body Language Dataset

| | | Predicted | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| Actual | 0 | 182 | 173 | 125 |
| | 1 | 199 | 167 | 154 |
| | 2 | 74 | 43 | 63 |

Figure 11 Confusion Matrix for CNN model for body language dataset

Validation for Emotion Recognition in UCF- Crime Dataset

| Actual \ Predicted | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 67 | 413 | 0 |
| 1 | 19 | 501 | 0 |
| 2 | 2 | 178 | 0 |

Figure 12 Confusion Matrix for Validation of Emotion recognition in Body Language Dataset

### 4.2.4 Training vs Validation Graph:

This graph illustrates the evolution of the model's performance in the training process, comparing the accuracies in training and validation.This graph illustrates the accuracy of the model on both the training and validation datasets across successive epochs. The x-axis represents the number of epochs and the y axis indicates the accuracy of the model, ranging from 0 to 1.0. As the model undergoes training iterations , the training accuracy shows the performance on the training data and the validation accuracy shows its performance on unseen validation data.

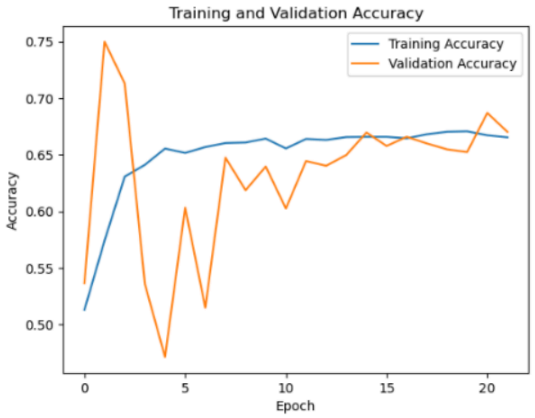


Figure 13 Accuracy vs Epoch for training and validation UCF-crime Dataset

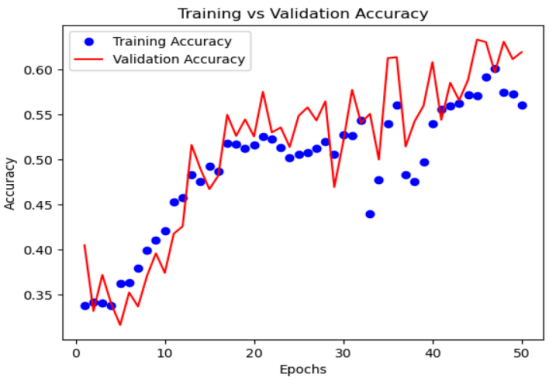


Figure 14 Accuracy vs Epoch for training and validation for BoLD Dataset

## 4.2 Discussions

The comparison of the classification reports obtained from the two different models evaluated with the UCF-crime dataset reveals distinct performance characteristics, highlighting significant differences in their precision, recall, and F1-scores across various classes, as well as their overall accuracy.

### 4.2.1 Crime Detection Results:

Model 1, utilizing DenseNet121 architecture with 3 dense layers having 64, 32, and 1 neurons respectively, demonstrates higher precision values across all classes compared to Model 2, which employs DenseNet121 with 3 dense layers having 32, 8, and 1 neurons in each, respectively. This suggests that Model 1 tends to make fewer false positive predictions, indicating a more accurate identification of relevant instances for each class. As shown in the previous section, Model 1 exhibits precision scores of 50%, 83%, and 71% for Classes 0, 1, and 2, respectively, whereas Model 2 shows lower precision values of 38%, 69%, and 64% for the same classes.

Moreover, Model 1 achieving higher recall scores implies that it effectively captures more relevant instances of each class. This is demonstrated by its recall values of 62%, 78%, and 58% for Classes 0, 1, and 2, respectively, which are higher than the corresponding recall scores of 35%, 85%, and 55% obtained by Model 2.

Additionally, Model 1 shows higher F1-scores across all classes as well as in the overall evaluation. Model 1's higher F1-scores indicate a better balance between precision and recall compared to Model 2. For instance, Model 1 has F1-scores of 55%, 80%, and 63% for Classes 0, 1, and 2, respectively, which outperforms Model 2's F1-scores of 37%, 76%, and 59% for the same classes.

In terms of overall accuracy, Model 1 outperforms Model 2, with an accuracy of 66% compared to Model 2's accuracy of 58%. This indicates that Model 1 can provide more accurate predictions across all classes, reflecting its overall superior performance in the evaluated task.

4.2.2 Emotion Recognition Results:

The results of emotion recognition depict a comprehensive analysis of the model's performance in accurately identifying and classifying emotional states conveyed through video data. The confusion matrix provides a detailed breakdown of the model's predictions across different emotional classes, offering insights into its effectiveness in distinguishing between various emotional expressions.

From the confusion matrix, it is evident that the model achieved varying levels of accuracy in recognizing different emotional states. For instance, the model correctly identified 182 instances of Class 0 which represents Fear , while misclassifying 173 instances as Class 1 which is Anger and 125 instances as Class 2 which is suffering. Similarly, it correctly classified 167 instances of Class 1 but misclassified 199 instances as Class 0 and 154 instances as Class 2. Moreover, for Class 2, the model struggled to make accurate predictions, with 74 instances being incorrectly classified as Class 0 and 43 instances being incorrectly classified as Class 1. The model was able to classify 63 instances correctly as class 2 which is suffering.

Overall, the confusion matrix serves as a valuable tool for evaluating the model's performance, providing a granular understanding of its strengths and weaknesses in recognizing different emotional expressions. By analyzing the distribution of correct and incorrect predictions across various classes, researchers can identify areas for improvement and refine the model to enhance its accuracy and robustness in emotion recognition tasks.

When testing for both Crime detection and emotion recognition with videos from the UCF-crime dataset, it has been noticed that while the model for Crime detection is able to classify the videos correctly , the emotion recognition model is classifying the majority of the videos under the class Anger.

In conclusion, while the confusion matrix offers valuable insights into the model's performance, it is essential to complement these findings with additional evaluation metrics such as precision, recall, and F1-score to obtain a comprehensive understanding of its overall effectiveness in emotion recognition. Through continuous refinement and optimization, researchers can develop more reliable and accurate models capable of discerning subtle nuances in human emotional expressions, paving the way for applications in diverse domains such as mental health assessment, human-computer interaction, and personalized user experiences.

## 5. Conclusion:

Based on the results and findings from the Crime detection and emotion recognition models, it is evident that Model 1 consistently outperforms Model 2 across various evaluation metrics. In both tasks, Model 1 demonstrates higher precision, recall, and F1-scores, as well as overall accuracy, indicating its superior performance and effectiveness in identifying anomalies and recognizing emotional states from video data.

Specifically, in Crime detection, Model 1 exhibits higher precision values, indicating fewer false positive predictions, and higher recall values, implying a better capture of relevant instances across different classes. This is further supported by the higher F1-scores achieved by Model 1, reflecting a better balance between precision and recall. Similarly, in emotion recognition, Model 1 showcases superior performance in accurately classifying different emotional expressions, as evidenced by the confusion matrix analysis.

These findings suggest that Model 1, with its tailored architecture and optimized parameters, is better equipped to handle the complexities inherent in both Crime detection and emotion recognition tasks. Its ability to effectively capture and interpret spatial and temporal features from video data enables more accurate predictions and classifications compared to Model 2.

In the case of emotion recognition, the model used is found to have an F1-score of 0.35 with a particular preference for Class 1(Anger) which it classifies the most. The reason for this is the imbalance dataset.The number of videos for suffering is much lesser compared to the number of videos corresponding to Anger and Fear.

This can be further seen when testing the emotion recognition model with UCF-crime dataset , where the majority of the videos used show Anger as the emotion which is predominant.

In conclusion, the results underscore the importance of model architecture and parameter optimization in achieving optimal performance in tasks such as Crime detection and emotion recognition. Model 1 emerges as the preferred choice for both tasks, offering promising avenues for future research and application development in areas such as surveillance, mental health assessment, and human-computer interaction. Further refinements and enhancements to the model may lead to even better performance, highlighting the potential for continued advancements in the field of deep learning-based video analytics.

**References**

[1]As retail thefts continues to surge, crime-fighting fog hits shoplifters. (2023,September,4).TheTimesofIndia.https://timesofindia.indiatimes.com/business/international-business/as-retail-thefts-continues-to-surge-crime-fighting-fog-hits-shoplifters/articleshow/103369009.cms

[2]Mishra, S. (2024, April 20). Graffiti on road surface huge risk for commuters: Experts. TheTimesofIndia.https://timesofindia.indiatimes.com/city/bhubaneswar/graffiti-on-road-surface-huge-risk-for-commuters-experts/articleshow/109467883.cms

[3] Nigam, N., & Dutta, T. (2022). Emotion and gesture guided action recognition in videos using supervised deep networks. IEEE Transactions on Computational Social Systems.

[4] Zeng, H., Shu, X., Wang, Y., Wang, Y., Zhang, L., Pong, T. C., & Qu, H. (2020). Emotioncues: Emotion-oriented visual summarization of classroom videos. IEEE transactions on visualization and computer graphics, 27(7), 3168-3181.

[5] Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music videos. Multimedia Tools and Applications, 80, 2887-2905.Wang, L., Tan, H., Zhou, F., Zuo, W., & Sun, P. (2022).

[6] Unsupervised Crime video detection via a double-flow ConvLSTM variational autoencoder. IEEE Access, 10, 44278-44289.

[7] Pelvan, S. Ö., Can, B., & Ozkan, H. (2023). A hierarchical approach for improved Crime detection in video surveillance. IEEE Access.

[8] Bajgoti, A., Gupta, R., Balaji, P., Dwivedi, R., Siwach, M., & Gupta, D. (2023). SwinCrime: Real-Time Video Crime Detection using Video Swin Transformer and SORT. IEEE Access.

[8] Singh, D., & Mohan, C. K. (2018). Deep spatio-temporal representation for detection of road accidents using stacked autoencoder. IEEE Transactions on Intelligent Transportation Systems, 20(3), 879-887.

[9]Yu, S., Wang, C., Mao, Q., Li, Y., & Wu, J. (2021). Cross-epoch learning for weakly supervised Crime detection in surveillance videos. IEEE Signal Processing Letters, 28, 2137-2141.

[10]Santhosh, K. K., Dogra, D. P., Roy, P. P., & Mitra, A. (2021). Vehicular trajectory classification and traffic Crime detection in videos using a hybrid CNN-VAE Architecture. IEEE Transactions on Intelligent Transportation Systems, 23(8), 11891-11902.

[11]Wu, L., Yang, Z., He, J., Jian, M., Xu, Y., Xu, D., & Chen, C. W. (2019). Ontology-based global and collective motion patterns for event classification in basketball videos. IEEE Transactions on Circuits and Systems for Video Technology, 30(7), 2178-2190.

[12]Yousaf, K., & Nawaz, T. (2022). A deep learning-based approach for inappropriate content detection and classification of youtube videos. IEEE Access, 10, 16283-16298.

[13]Savchenko, A. V., Savchenko, L. V., & Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Transactions on Affective Computing, 13(4), 2132-2143.

[14]Samadiani, N., Huang, G., Hu, Y., & Li, X. (2021). Happy emotion recognition from unconstrained videos using 3D hybrid deep features. IEEE access, 9, 35524-35538.

[15]Viegas, C. (2020, November). Two stage emotion recognition using frame-level and video-level features. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 912-915). IEEE.

[16]Cevikalp, H., Yavuz, H. S., & Triggs, B. (2019). Face recognition based on videos by using convex hulls. IEEE Transactions on Circuits and Systems for Video Technology, 30(12), 4481-4495.

[17]Waqas Sultani, Chen Chen, Mubarak Shah, "Real-world Crime Detection in Surveillance Videos"IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

[18]Yu Luo, Jianbo Ye, Reginald B. Adams, Jr., Jia Li, Michelle G. Newman and James Z. Wang, ARBEE: Towards Automated Recognition of Bodily Expression of Emotion In the Wild," International Journal of Computer Vision, vol. 128, no. 1, pp. 1-25, 2020.

[19]Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).