Unveiling Anomalies In Surveillance Videos through Various Transfer Learning Models

Sachin Solanki
Computer Science & Engineering Department,
Devang Patel Institute of Advance Technology and Research
(DEPSTAR), Faculty of Technology and Engineering (FTE),
Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
codesachin501@gmail.com

Yash Shah
Computer Science & Engineering Department,
Devang Patel Institute of Advance Technology and Research
(DEPSTAR), Faculty of Technology and Engineering (FTE),
Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
shahyash452@gmail.com

Deep Rohit
Computer Science & Engineering Department,
Devang Patel Institute of Advance Technology and Research
(DEPSTAR), Faculty of Technology and Engineering (FTE),
Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
deeprohit2163@gmail.com

Dipak Ramoliya
Computer Science & Engineering Department,
Devang Patel Institute of Advance Technology and Research
(DEPSTAR), Faculty of Technology and Engineering (FTE),
Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
dipakramoliya.ce@charusat.ac.in

***Abstract*— Video anomaly detection touches upon the process of automatically identifying abnormal events or behaviors in video streams. It plays a crucial role in various domains, including security, surveillance, and safety monitoring. Traditional methods of video analysis often rely on manual inspection, which often requires a significant amount of time and is susceptible to human error. The concept of video anomaly detection aims to overcome these limitations by leveraging advanced technologies such as computer vision and machine learning. This research study performs a relative analysis on the models of deep learning such as VGG16, Resnet50, VGG19, DenseNet121 to deal with video anomaly detection. This study uses subsets of various classes of UCF crime datasets. The UCF Crime Dataset contains 1903 video clips captured from real-world surveillance cameras. Results show Densenet-121 with an ROC-AUC score of 0.85 performs better when vis-à-vis the other models used in this study.**

Keywords*— *Artificial Neural Networks, Anomaly Detection Deep learning, Transfer learning, Computer Vision

I. Introduction

Video anomaly detection [1] refers to the automated recognition of unusual events, behaviors, or activities captured in video footage that deviate from expected or normal patterns. It is a crucial task in surveillance and security systems as it aims to detect and alert the presence of such anomalies.

For performing video anomaly detection, advanced technologies like deep learning, computer vision, and machine learning are utilized. These techniques analyze the temporal and spatial information in video frames to identify deviations from normal behavior. By learning and recognizing patterns that indicate of anomalies, the system can automatically flag potential security threats or safety concerns. [2]

The practical applications of video anomaly detection are diverse. In security and surveillance, it helps identify criminal activities such as theft, vandalism, or robbery, enabling timely interventions and enhancing public safety. Recent advancements in machine learning and computer vision, specifically deep learning models like convolutional neural networks [3] have shown encouraging results in capturing spatial and temporal patterns for effective anomaly detection.

Video anomaly detection is a challenging problem because of the vast amount of data that requires processing and the variety of factors that can give rise to anomalies. [20] However, the advancement of deep learning methods has led to significant improvements in the accuracy and performance of video anomaly detection systems.

Overall, video anomaly detection significantly contributes to enhancing security, safety, and operational efficiency across various domains. By enabling automated monitoring, analysis, and alerting in video surveillance systems, it empowers security personnel and decision-makers to proactively respond to events that are abnormal, mitigate risks, and maintain a secure environment.

II. Approaches for Transfer Learning

Transfer learning is a subset of machine learning that involves using cognizance gained from one task to enhance the performance of a different task. Within the domain of deep learning, there are two primary techniques for transfer learning: [9] feature extraction and fine-tuning. These approaches leverage pre-trained models, typically trained on [6] large-scale datasets like ImageNet, to accelerate the training process and improve performance on new, specific tasks. [7]

A. Feature Extraction

In the feature extraction technique, a pre-trained model is utilized for the purpose of feature extraction. The initial layers of the model, which have learned general visual features from the original dataset, are kept intact. However, the top layer, responsible for classification, is removed or replaced with a new classifier that suits the specific task at hand. Data preprocessing is a crucial step in transfer learning which includes tasks such as resizing images and normalization to efficiently provide preprocessed image. Pooling layer plays a vital role in CNNs and are generally used in transfer learning to extract and condense crucial information from the input data before sending it to next layers i.e. fully connected layers or other convolutional layers.

B. Fine-tuning

In the fine-tuning approach, not only the top layer but also some of the lower layers of the pre-trained model are modified during the training process. The pre-trained model's weights are initialized with the values learned from the original dataset, and these weights are updated and adjusted during training on the target dataset. [19]

By employing transfer learning, deep learning models can benefit from the knowledge and representations learned from large-scale datasets, saving significant time and computational resources while achieving better performance on new tasks. In this comprehensive study, fine-tuning of the deep learning [11] models— VGG16, DenseNet121, VGG19, and ResNet50 has been done.

- *DenseNet121:* DenseNet-121 is a deep neural network comprising 121 layers that have over eight [14] million parameters. It follows a unique architecture with Dense Blocks, where feature maps within each block maintain same dimensions while the number of filters varies. These Dense Blocks have transformation layers that use batch normalization to down-sample the feature maps. The main objective of batch normalization is to accelerate and stabilize the training by performing input normalization of each layer. This strategy encourages feature reuse, and efficient parameter sharing, and helps the model better capture graphically complex patterns.

- *VGG-16:* A deep convolutional neural network which has 16 layers and comprises of thirteen convolutional [12] layers, max pooling for dimension reduction, and three fully connected layers. SoftMax classifier is the last layer that is used for classification tasks. VGG-16 is widely used for image recognition. The main objective of softmax is to convert the model's original predictions into probability distribution

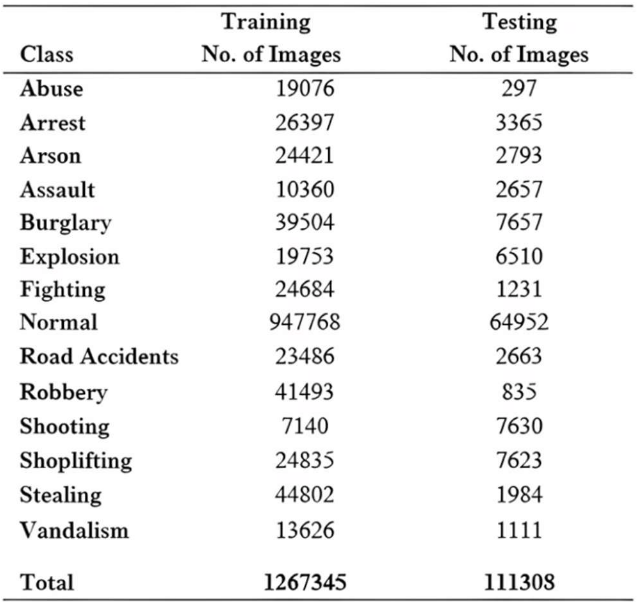
- *VGG-19:* VGG-19 is a deep convolutional [13] neural network with nineteen layers. It contains 16 convolutional layers, max-pooling for dimension reduction, and 3 fully connected layers. The last layer is a SoftMax classifier for classification tasks.

- *ResNet50:* ResNet-50 is an altered version of the Residual Network (ResNet) architecture. It contains 48 convolutional, one MaxPool layer, and one [15] average pool layer. Each block within the network consists of 3 convolutional layers, and the identification block also includes 3 convolutional layers. With over 23 million trainable parameters, ResNet-50 is designed to learn intricate patterns in data effectively.

III. MAIN DATASET

We have used subsets of the UCF crime dataset for evaluating the models. The University of Central Florida (UCF) Crime Dataset is a collection of videos depicting various criminal activities, designed for research and development in the domain of surveillance of videos and activity recognition. The UCF Crime dataset represents a groundbreaking collection of 128 hours of videos, making it the first dataset of its kind. It comprises 1900 raw real-world security camera footage, which features 13 diverse and realistic anomalies, such as Abuse, Assault, Arrest, Road Accident, Arson, Burglary [21], Fighting, Explosion, Robbery, Shoplifting, Shooting, Vandalism, and Stealing. The selection of anomalies is based on their notable influence on public safety. The anomaly detection task involves two approaches. Firstly, a general anomaly detection method groups all anomalies together and treats normal activities separately. Secondly, a specific anomaly recognition approach is used to individually recognize each of the 13 anomalous activities. From each complete-length video, the tenth frame is being extracted and is integrated with every video in that specific class. The images in their raw format is of size 64*64 and in the .PNG format

Table 1. No. of images used for the purpose of testing and training data



Class	Training No. of Images	Testing No. of Images
Abuse	19076	297
Arrest	26397	3365
Arson	24421	2793
Assault	10360	2657
Burglary	39504	7657
Explosion	19753	6510
Fighting	24684	1231
Normal	947768	64952
Road Accidents	23486	2663
Robbery	41493	835
Shooting	7140	7630
Shoplifting	24835	7623
Stealing	44802	1984
Vandalism	13626	1111
Total	**1267345**	**111308**

The "Abuse" category showcases instances of cruelty or violence directed towards vulnerable individuals, such as children, the elderly, animals, and women. "Burglary" [24] includes videos capturing thieves entering buildings or houses to commit theft without the use of force. "Robbery" depicts thieves unlawfully taking money through force or threats, excluding shootings. In "Stealing," individuals are shown taking money or property without asking for permission. "Shooting" encompasses videos where people use guns to shoot others. "Shoplifting" displays people swiping goods from the shops while acting to be customers.

The "Assault" category captures sudden and violent physical attacks where the victim will not retaliate, while

"Fighting" shows a dispute involving more than one individual attacking each other. "Arson" features videos depicting deliberate acts of setting property on fire. "Explosion" involves destructive [23] events where something blows apart, excluding cases where a person intentionally sets off the explosion or fire.

In "Arrest," videos capture police capturing individuals, while "Road Accident" showcases traffic [22] accidents indulging vehicles and pedestrians. "Vandalism" displays intentional destruction [25] or damage to public or private property. Lastly, the "Normal Event" consists of videos depicting everyday scenes, both indoors and outdoors.

IV. Model Traning

For the objective of carrying out research, the deep learning models such as VGG16, ResNet50, DenseNet121, and VGG19, were being trained using a dedicated NVIDIA RTX A4000 GPU with 16 GB of memory. The images included in the dataset were rescaled to 255 × 255 pixels. For the purpose of carrying out development [8] and implementation of the CNN algorithm, TensorFlow [5] library version 2.6.0 was used.

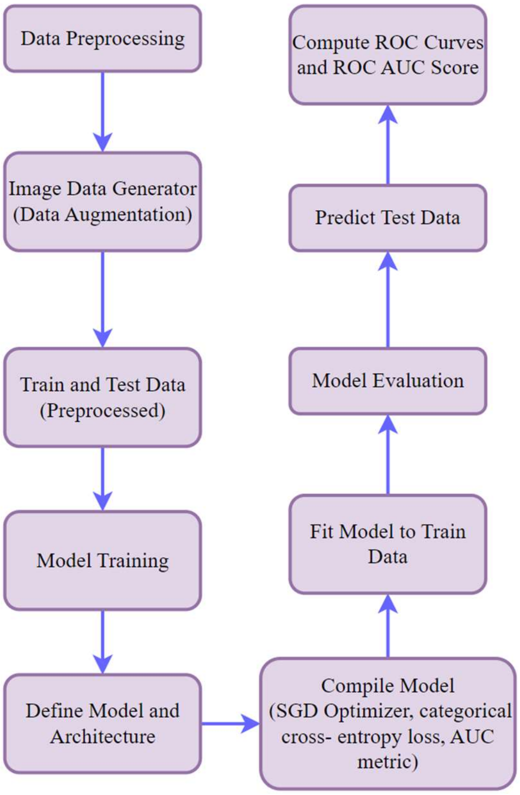

Data Preprocessing
Image Data Generator (Data Augmentation)
Train and Test Data (Preprocessed)
Model Training
Define Model and Architecture
Compile Model (SGD Optimizer, categorical cross- entropy loss, AUC metric)
Fit Model to Train Data
Model Evaluation
Predict Test Data
Compute ROC Curves and ROC AUC Score

Fig. 1. Conceptual representation of data

V. Evaluation Metrics

When models were trained, they were evaluated on the test data of around 1 gigabyte. The models were trained on around 1.25 lakh images which belonged to 14 different classes. For analyzing the performance of various models used, we have applied the concept of the ROC-AUC curve. The ROC-AUC curve often referred to as the Receiver Operating Characteristic [16] curve and Area Under the Curve, is a visual representation used to assess the performance of models for binary classification. In the area of machine learning, to evaluate how well a model can differentiate between two classes (positive and negative classes), this concept is used.

TPR (True Positive Rate) [17] and FPR (False Positive Rate) are two important performance measures used to evaluate the effectiveness of a binary classification model. They are derived from the confusion matrix and are fundamental components of the ROC curve.

A. True Positive Rate (TPR)

It represents the proportion of positive cases that the model accurately classifies as positive among all the positive cases present in the dataset. The equation for TPR is calculated as:

$$TPR = \frac{True\ Positives}{(True\ Positive\ +\ False\ Negative)}$$

- *True Positive (TP)*: When the result of the actual class is positive and model also predict it(class) as a positive.
- *False Negative (FN)*: When the result of the original class will be positive but the result predicted by the model will be negative. (Incorrect prediction).

B. False Positive Rate (FPR)

It represents the percentage of negative instances that the model incorrectly classifies as positive of all the negative instances in the dataset. The equation for FPR is calculated as:

$$FPR = \frac{False\ Positives}{(True\ Negatives + False\ Positives)}$$

- *False Positive (FP):* [18] When the result of the original class will be negative and the result predicted by the model will be positive.
- *True Negative (TN)*: When the result of the actual class will be negative and result predicted by the model will be negative.

A high TPR and a low FPR are desired characteristics of a good classification model, as it indicates that the model correctly identifies positive instances while minimizing false alarms (misclassifying negatives as positives).

The plot generated will have following:

- *Multiple ROC curves:* The plot will have many ROC curves, one for each class in the multi class problem. Each ROC curve depicts the performance of the model for a specific class. The curves x-axis will show the False Positive Rate, [4] and the y-axis will show the True Positive Rate.
- *AUC scores:* Each ROC curve will be labelled with the name of the corresponding class and its corresponding AUC score. The AUC score represents the area [10] under the ROC curve. Higher AUC

scores indicate better performance in classifying the instances of that class.

- *Dashed line:* The plot will also have a black dashed line, representing the ROC curve for random guessing. This line is a baseline that represents the performance of a classifier that makes random predictions. ROC curve above this line indicates that the model is performing better than random guessing.

Following are the results obtained on graphs:

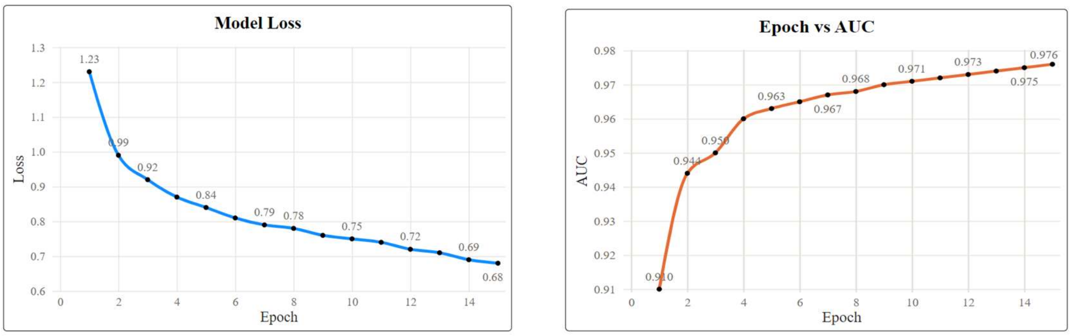


Fig. 2. (a) DenseNet121- Loss graph (b) DesnseNet121- AUC graph

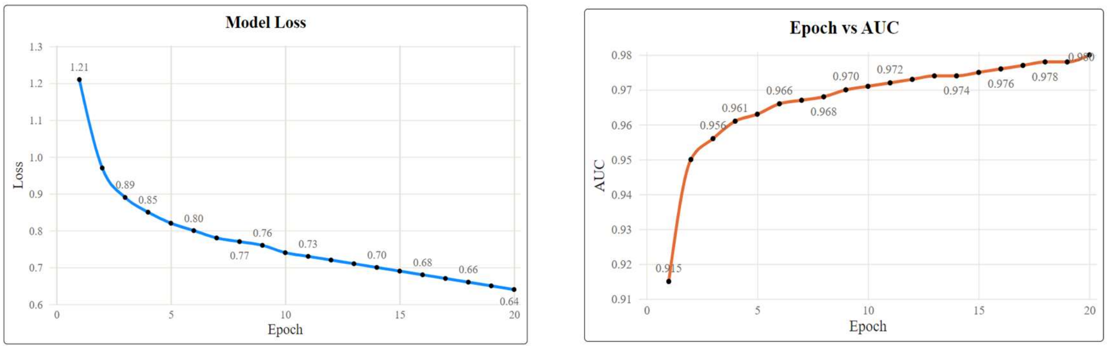


Fig. 3. (a) ResNet50- Loss graph (b) ResNet50- AUC graph

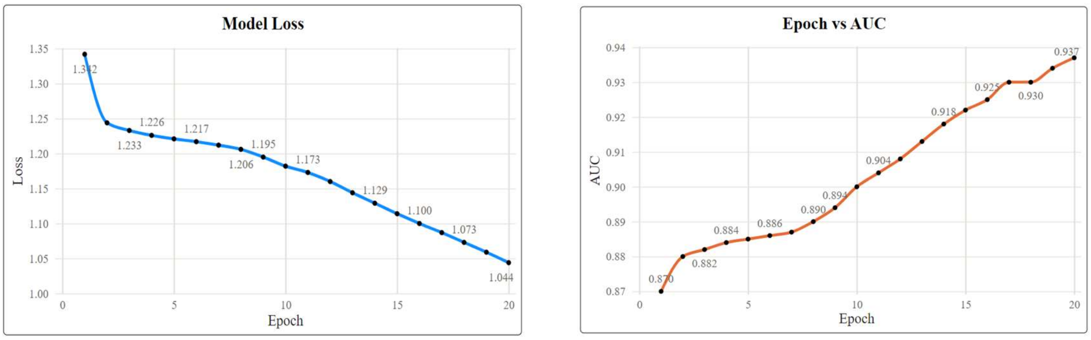


Fig. 4. (a) VGG-16- Loss graph (b) VGG-16- AUC graph

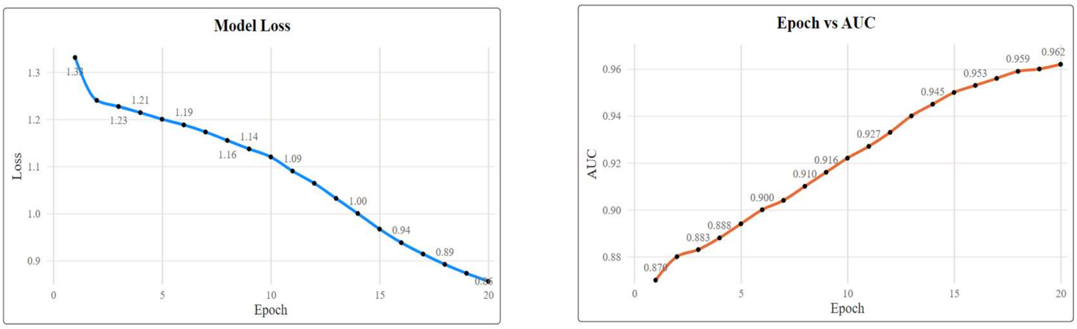


Fig. 5. (a) VGG-19- Loss graph (b) VGG-19- AUC graph

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

Figures 2, 3, 4, and 5 present the relationship between epoch and AUC, as well as epoch and loss, for each model. Upon analyzing these figures, certain patterns emerge, highlighting the models' performance characteristics.

The results specifically indicate that densenet-121 exhibits a notable decline in loss as early as epoch 15 when compared to the other models. Conversely, VGG-16 demonstrates a comparatively higher loss after epoch 20, implying slower convergence and potentially less effective learning during this phase.

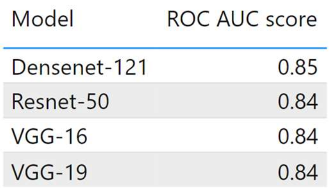
Furthermore, the figures reveal that Densenet-121 and Resnet-50 achieve the highest AUC scores after Epoch 15 and Epoch 20, respectively. This suggests that these models excel in quickly learning distinctive features among different classes. Their ability to attain high AUC scores relatively early in the training process indicates efficient differentiation abilities.

Table 2. AUC score of multiclass based on models

Model	Dense Net121	ResNet50	VGG-16	VGG-19
Assault	0.83	0.79	0.84	0.69
Burglary	0.78	0.72	0.78	0.72
Explosion	0.78	0.78	0.74	0.71
Fighting	0.44	0.42	0.25	0.39
Road Accidents	0.7	0.76	0.48	0.57
Robbery	0.75	0.66	0.53	0.44
Shooting	0.62	0.61	0.65	0.68
Shoplifting	0.78	0.8	0.61	0.67
Stealing	0.6	0.59	0.65	0.64
Vandalism	0.52	0.62	0.74	0.64

Table 2. illustrates that the Densenet-121 model consistently achieves higher AUC scores compared to the other three models across a majority of classes, such as Abuse, Burglary, Arrest, Explosion, Arson, Assault, Fighting, Normal, and Robbery. Notably, even for the 'Normal' class, which contains the highest number of images by a significant margin, Densenet-121 demonstrates a superior AUC score compared to the other models. These findings suggest that Densenet-121 exhibits strong performance across various classes, and its superiority becomes evident as early as epoch 15. This signifies that Densenet-121 is proficient of effectively learning and discriminating distinctive features, leading to better classification results, especially for the "Normal" class which has large number of images.

From figure 6, it could be illustrated that among the four models evaluated, Densenet-121 achieved the highest overall ROC-AUC score of 0.85, which indicates its superiority in performance compared to Resnet-50, VGG-16, and VGG-19. Despite their strengths, they achieved slightly lower ROC-AUC scores compared to Densenet-121, making Densenet-121 the preferred choice in this evaluation.



Model	ROC AUC score
Densenet-121	0.85
Resnet-50	0.84
VGG-16	0.84
VGG-19	0.84

Fig. 6. The ROC-AUC score of trained models

The figure 7,8,9 and 10 depicts the AUC score of the fourteen classes of the UCF crime dataset which assesses the overall performance of the curve.

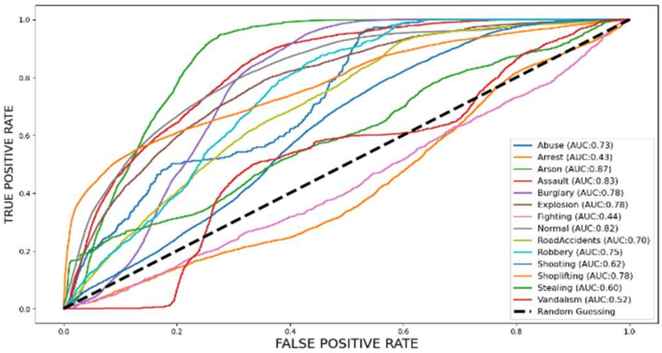


Fig. 7. Class wise ROC-AUC score of DenseNet121 model

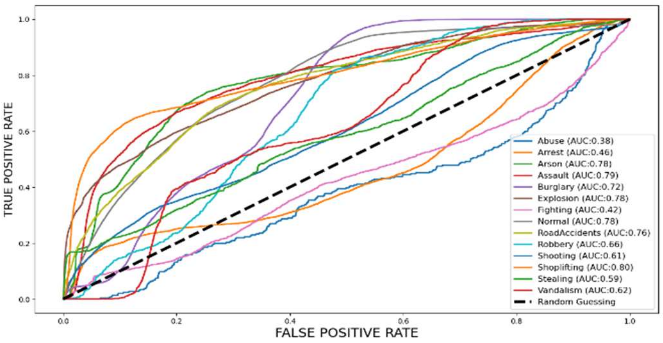


Fig. 8. Class wise ROC-AUC score of ResNet50 model

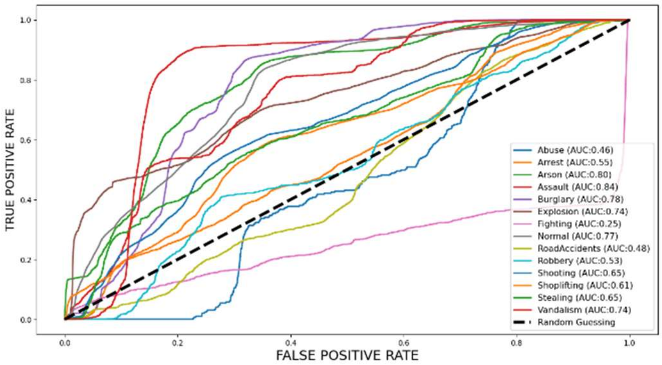


Fig. 9. Class wise ROC-AUC score of VGG16 model

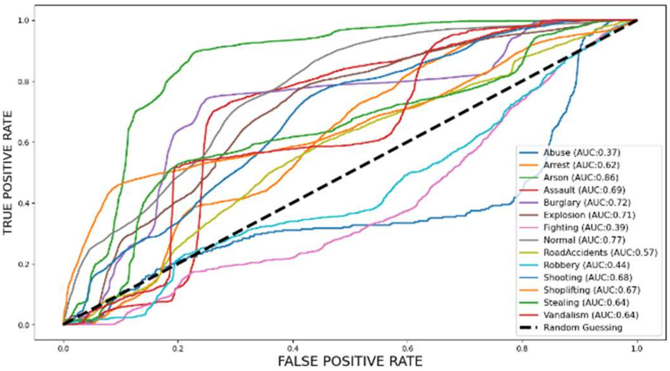


Fig. 10. Class wise ROC-AUC score of VGG19 model

VII. CONCLUSION

In this research study, a comparative and comprehensive analysis was conducted to assess the performance of transfer learning models, namely Densenet-121, VGG-19, Resnet-50, and VGG-16, for detecting anomalies in videos. The study utilized the UCF-crime dataset which comprised approximately 128 hours of surveillance footage. The dataset was divided into 14 distinct classes, and models were trained on these classes. To prepare the data for training, image processing techniques were applied. After, the pre-trained models were used to classify the processed images by passing them through the final layer for classification. The objective of the study was to multi-class classification, including categories like Abuse, Arrest, Road Accidents, Fighting, and Normal behavior. After a thorough examination, Densenet-121 was the most effective model across various classes, including the normal behavior class, which contained the largest number of images. The model achieved an impressive ROC-AUC score of 0.85, indicating its efficiency when handling large datasets. These findings propose that Densenet-121 suits better real-time applications, which makes it a potential choice for classifying anomalies at a large scale in the future.

REFERENCES

[1] Ramachandra, Bharathkumar, Michael J. Jones, and Ranga Raju Vatsavai. "A survey of single-scene video anomaly detection." IEEE transactions on pattern analysis and machine intelligence 44.5 (2020): 2293-2312.

[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[3] Patrikar, Devashree R., and Mayur Rajaram Parate. "Anomaly detection using edge computing in video surveillance system." International Journal of Multimedia Information Retrieval 11.2 (2022): 85-110.

[4] Hand, David J., and Robert J. Till. "A simple generalisation of the area under the ROC curve for multiple class classification problems." Machine learning 45 (2001): 171-186.

[5] Ren, Jing, et al. "Deep video anomaly detection: Opportunities and challenges." 2021 international conference on data mining workshops (ICDMW). IEEE, 2021.

[6] Bansod, Suprit, and Abhijeet Nandedkar. "Transfer learning for video anomaly detection." Journal of Intelligent & Fuzzy Systems 36.3 (2019): 1967-1975.

[7] Shazia, Anis, et al. "A comparative study of multiple neural network for detection of COVID-19 on chest X-ray." EURASIP journal on advances in signal processing 2021 (2021): 1-16.

[8] Abbas, Zainab K., and Ayad A. Al-Ani. "A comprehensive review for video anomaly detection on videos." 2022 International Conference on Computer Science and Software Engineering (CSASE). IEEE, 2022.

[9] Bozinovski, Stevo. "Reminder of the first paper on transfer learning in neural networks, 1976." Informatica 44.3 (2020).

[10] Rosset, Saharon. "Model selection via the AUC." Proceedings of the twenty-first international conference on Machine learning. 2004.

[11] Mascarenhas, Sheldon, and Mukul Agarwal. "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification." 2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON). Vol. 1. IEEE, 2021.

[12] Dhuri, Vighnesh, et al. "Real-time parking lot occupancy detection system with vgg16 deep neural network using decentralized processing for public, private parking facilities." 2021 international conference on advances in electrical, computing, communication and sustainable technologies (ICAECT). IEEE, 2021.

[13] Butt, Umair Muneer, et al. "Detecting video surveillance using VGG19 convolutional neural networks." International Journal of Advanced Computer Science and Applications 11.2 (2020).

[14] Sri Jamiya, S. "An efficient algorithm for real-time vehicle detection using deep neural networks." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.11 (2021): 2662-2676.

[15] Grabowski, Dariusz, and Andrzej Czyżewski. "System for monitoring road slippery based on CCTV cameras and convolutional neural networks." Journal of Intelligent Information Systems 55.3 (2020): 521-534.

[16] Pillai, Manu S., et al. "Real-time image enhancement for an automatic automobile accident detection through CCTV using deep learning." Soft Computing (2021): 1-12.

[17] Ramoliya, Dipak, and Amit Ganatra. "Insights of Deep Learning-Based Video Anomaly Detection Approaches." Intelligent Communication Technologies and Virtual Mobile Networks. Singapore: Springer Nature Singapore, 2023. 663-676.

[18] Narkhede, Sarang. "Understanding auc-roc curve." Towards Data Science 26.1 (2018): 220-227.

[19] Too, Edna Chebet, et al. "A comparative study of fine-tuning deep learning models for plant disease identification." Computers and Electronics in Agriculture 161 (2019): 272-279.

[20] Lv, Hui, et al. "Localizing anomalies from weakly-labeled videos." IEEE transactions on image processing 30 (2021): 4505-4515.

[21] Kamijo, Shunsuke, et al. "Traffic monitoring and accident detection at intersections." IEEE transactions on Intelligent transportation systems 1.2 (2000): 108-118.

[22] Zhao, Bin, Li Fei-Fei, and Eric P. Xing. "Online detection of unusual events in videos via dynamic sparse coding." CVPR 2011. IEEE, 2011.

[23] Bermejo Nievas, Enrique, et al. "Violence detection in video using computer vision techniques." Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14. Springer Berlin Heidelberg, 2011.

[24] Zhu, Yi, and Shawn Newsam. "Motion-aware feature for improved video anomaly detection." arXiv preprint arXiv:1907.10211 (2019).

[25] Ghazal, Mohammed, Carlos Vázquez, and Aishy Amer. "Real-time automatic detection of vandalism behavior in video sequences." 2007 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2007.