International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Real-Time Violence Detection and Alert System

Neeraj Upadhyay
Department of Information Technology
Inderprastha Engineering College
Ghaziabad,India
neerajofficialnu@gmail.com

Mohit Karakoti
Department of Information Technology
Inderprastha Engineering College
Ghaziabad,India
mohitkarakoti7777@gmail.com

Asst. Prof. Tanya Sharma
Department of Information Technology
Inderprastha Engineering College
Ghaziabad,India
tanya.sharma@ipec.org.in

Abstract —Personal safety and quality of life are seriously threatened by the violent acts. In order to curb the hostile conduct, numerous other strategies have been implemented, such as installing and maintaining surveillance systems. It will be extremely relevant if the monitoring systems are able to recognize violent conduct on their own and immediately send warning or alarm signals to the authorities. New researchers are drawn to the active study fields of computer vision and image processing for the detection of aggressive and aberrant conduct. In order to detect violence in video streams, this research suggests a real-time violence detection and alarm system that uses MobileNetV2 along with LSTM. The system offers a scalable, precise, and flexible public safety solution by utilizing a Telegram Bot for real-time alerting message.

Keywords — MobileNetV2, Realtime detection, LSTM, Transfer learning, Neural network, callbacks, RLVS.

I. Introduction

The growing issues surrounding public safety have prompted the development of new technological interventions that can reduce risk and offer protection to communities. This research offers a new real-time violence detection and alert system, drawing on the power of mobilenetv2,lstm and telegram bot, to offer a robust and responsive real-time platform for violence reportage and detection.new and new frameworks for first detection, reportage, and intervention in such incidents.

In an atmosphere of mounting fear on public security, safety, and prevention of violence .technological development has been a major benefactor in the persistent commitment to safeguarding both groups and individuals .the pressing concerns brought up by various types of violence, whether in public or in private settings have driven the study of innovative and adaptive frameworks for the first identification, documentation, and response in such incidents. It is within this backdrop that the current research employs the procedure of using a better and comprehensive real-time violence alert system,which effectively integrates mobilenetv2, lstm, and a telegram bot to create a strong and effective system for violence detection andincident management.

At the core of our real-time violence alert system is the strong model based on mobilenetv2 and lstm, a deep learning model that is known for its incredible speedaccuracy, and efficiency. Under real-time video stream analysis, we undertake the daunting task of detecting and classifying violent behavior, thereby allowing the early detection of likely incidents.the ability to detect images not only provides a high degree of accuracy while at the same time enabling the continuous unification of the system with ongoing surveillance networks and thereby boosting security measures and enabling the rapid response to threats [1]. This is indicative of the transformatory role that deep learning technology can play in addressing current social issues.

Telegrams, with their huge user base, privacy-centric approach, and real-time messaging features, have established themselves as a best communication channel to disseminate warnings to concerned authorities. This mechanism essentially acts as the medium by which the process of converting detection into action is made functional so that the response to violence is immediate and firm. The telegram bot has therefore emerged as a core part of the system to establish the necessary communication link in ensuring public safety[2]. In our relentless quest for a more advanced anda complete real-time violence notification framework,always focus on scalability, accuracy, and flexibility.

Our system is developed to support a wide variety of environments, e.g., public areas, private property and businesses, institutions. Thus, its adaptability makes it a useful tool for public security and safety augmentation in multicultural settings, thereby supporting a wide variety of security demands. Our proposed system was tested using several datasets of violence and actual surveillance systems. The result indicated that the system was capable of detecting violent crimes correctly in real time. The system is scalable and installable in a wide number of cameras.

The rest of the paper is structured as follows:section ii comprises comparative studies of similar research studies. The chosen methodology is outlined in detail in section iii, and the proposed frameworkand the data set, and section iv givesexperimental procedures and outlines the approach's assessment. It finishes the paper by offering remarks on potential future research.

II. Related Work

Current methods to detect violence are categorized into three groups: visual-based methods, audio-based methods, and combined systems. Visual approach gather visual information and then that information is presented with proper characteristics. Features can be local or global. Global features consist of average speed, region occupancy, relative positioning fluctuations, and object-background interactions, while local attributes consist of location, velocity, shape, and color. Auditory data is used by the audio-based technique to classify violent behavior .A hierarchical method based on Gaussian mixture models and hidden Markov models is employed to differentiate between explosions, gunshots, and car braking in audio .

The hybrid method puts a lot of focus on the integration of both visual and audio elements. Some methods identify violent moments in recordings using the identification of blood and fire, the intensity of movement, and the audio related to them. CASSANDRA technology identifies aggression in CCTV recordings using motion features associated with articulation and scream-like acoustical signals [3].

A. Modelling with 2D CNN

In [4] M.S kang made a pipeline for real-time on-device violence recognition that utilizes a 2D CNN. There proposed pipeline have 3 major components: frames grouping, a spartial attention module known as Motion Saliency Map (MSM), and a Temporal attention module known as Temporal Sequence and Excitation (T-SE) block.
They used a technique that averages the channel of the input frames, Three sequential channel averaged frames were pooled to serve as input for 2D CNN, MSM identifies significant region feature maps derived from motion boundries by utilizing the difference between successive frames.The TSE block can naturally highlight the time interval associated withan event interest.The proposed pipeline provide significant improvement in computational complexity over exsisting 3D-CNN models.The model was tested on 6 different dataset and the result revealed better performance in terms in terms of accuracy and speed from existing Pytouch served as a foundation for network implementation .The suggested method [4] works fine in real time scenario..

B. Spartio tempored features with 3D CNN

In [5] we have seen a model that utilizes low level functionality method for identify violent frames. The proposed model leverage spatio-temporal interest & K-Means clustering to group similar interest point. After calculating the cluster average for every significant cluster, author apply these finding to classify an event with an unfamiliar sequence.
They discovered that the spatio temporal interest point remain unaffected by variation in scale rotation & lighting in the proposed approach.The author had used the Laplacian of Gaussian blur to identify interest points. After detecting interest points, each is described using a spatio-temporal descriptor that encodes the local structural information surrounding the point. K-Means clustering is then applied to group similar interest points into distinct clusters.
The cluster means are subsequently used to classify unseen sequence events.The proposed method surpass existing cutting-edge algorithm in terms of accuracy on 2 publicly available dataset. This process is computationally very intensive and depends high parameters, making it unsuitable for real-time applications and day-to-day use [3].

C. CNN &LSTM

The paper proposed a model that detect violence using deep learning techniques. CNN is used as the spartial feature extraction, while LSTM serve as a method for learning temporal data. The proposed model is mode on hockey fight dataset achieving an accuracy of 98% at 131 frames per second. This model surpassed every other model at that time in terms of accuracy & precision.The proposed model was more efficient at detecting violence between 2 individual but violence generally involve a large group of people. Detecting violence in large group of people is relatively tough because it has multiple features that are difficult to capture. This whole model was implemented in python. It show high accuracy with benchmark dataset but it struggle to perform generalization in real current time[6].

D. Deep Autoencoder and CNN

[7] proposed a CNN-based system that was designed to process data streams in real time, which is captured using an optical sensor in dynamic scenarios. 1st deep features are extracted from the frames by utilizing transfer learning or a pre-trained CNN, then these features are processed via a deep autoencoder, which helps to capture temporal data from the activities in CCTV footage. To classify human behaviour, a quadratic SVM is trained as a non-linear learning approach. While testing, an iterative fine-tuning method updates the model parameter using the latest data from the evolving environment. Experimental results show promising results, and the proposed model outperforms the current SOTA method in terms of both processing time and accuracy of that time.
In our paper, we build a hybrid neural network model that show high generalization in detecting violence and send real time alert to the user via telegram bot

.

III. Architectural Design

The recommended technique is used for recognizing violent and non-violent incidents in real-time security camera footage is thoroughly examined in this section. In our system, we use MobileNetV2 and LSTM, along with hyperparameters that accurately identify the presence of violence in videos. When violence is detected, messages are

sent to a telegram group via a telegram bot for real real-time

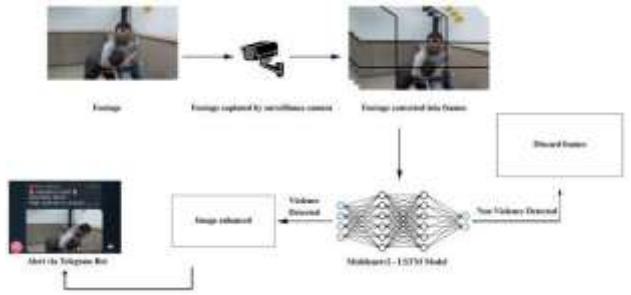

Fig.1 Proposed Model Architecture

IV. Methodology

A. Dataset Collection

To test our methodology, we work with these two datasets, Hockey Fight Dataset [8], which is standard in SOTA of violence detection, and Real Life Violence Situations Dataset, which is a relatively new dataset but has shown some promising results[9].

- Hockey Fight Dataset (HF)

The dataset consists of 1000 video clips of hockey games of the National
Hockey League (NHL). Each clip is restricted to 50 frames and has a resolution of 720×576 pixels. 500 of the clips are labeled as ‘fight’, and the other 500 are labeled as ‘non-fight'. Fig represents some sample clips of HF Dataset.

- Real Life Violence Situations Dataset(RLVS)

This dataset contains 2000 video clips of real-life violence situations.1000 clips are labelled as 'NonViolence' and 1000 are labelled as 'Violence'.

B. Data Preprocessing

Video clips are converted into frames and now data augmentation is applied. Techniques such as cropping, randomrotate, horizontalflip, motionblurr,gaussianblur, etc, are applied via Albumentations Augmentation Pipeline. Also if images are in BGR they are converted to RGB. If frame ID is 7, then frames are skipped to avoid duplication. Also the both datasets are splitted into training set and testing set in 80% -20% ratio where 80% clips are used to train model and rest 20% clips are used to test model .This split have been performed via scikit-learn library.
The proposed model is written in python using Keras library [10] with TensorFlow [11] backend and some helper libraries like OpenCV [12] and matplotlib [13].Adam optimizer is also used in this model [14].I Regarding the hardware used; the system is run using Kaggle [15] which have GPU 2XP100 NIVIDIA GPU , and a CPU ryzen5 5600U along with 16 GB RAMs and 500 GB as ssd .

C. Transfer Learning

The suggested model is a binary image classification model based on MobileNetV2 and LSTM to learn and understand spatial-temporal features in images. We have used MobileNetV2 which is a successor of MobileNetV1[16]. It is trained on the ImageNet dataset. The model takes an input of a shape picture (IMG_SIZE,IMG_SIZE,ColorChannels) and extracts dense spatial information through a pre-trained MobileNetV2 backbone (with frozen weights). These features are reduced through a GlobalAveragePooling2D layer prior to being reshaped into a 3D tensor that can be utilized through a Long Short-Term Memory (LSTM) layer. The LSTM consists of 64 units that reflects temporal correlations among spatial information[17]. In order to retain regularization, sequential output is provided through a thick layer that contains ReLU activation function and a dropout layer[18]. Lastly, a sigmoid-activated dense neuron provides output as a single probability score that reflects the probability that the input is from class 1 (for instance, violence). The model is constructed along the Adam optimizer and binary cross-entropy loss, thus it is suitable for binary classification problems.

D. Hyperparameter initialization

The use of hyperparameter configurations and training procedures was optimized in view of enhancing the training procedure while achieving maximum generalization. Input images were resized to dimensions of 120×120 pixels, while ColorChannels were set to 3 (RGB), which was a balanced choice, maximizing the utilization of computational resources and sufficient feature representation. The model was trained for 50 epochs, with a batch size of 4, allowing frequency updates of weights for enhanced generalization over a small dataset. The initial learning rate (start_lr) was set to 0.00001, whereas the maximum learning rate (max_lr) used was 0.00005 and minimum learning rate (min_lr) was kept at 0.00001. A learning rate schedule was implemented using a custom schedule that employed a warm-up policy, with rampup_epochs=5, sustain_epochs=0, and exponential decay factor (exp_decay)=0.8, for a gradual decline of learning rate across epochs. To counteract overfitting, kernel_regularizer=regularizers.l2(0.005) was applied to the dense layers in addition to dropout layers incorporated within the model [19]. Training employed various callbacks: EarlyStopping with a patience parameter of 3, which stops training when there is no improvement in validation performance; ReduceLROnPlateau with an adaptive decreasing learning rate scheme if validation loss is not decreasing. ModelCheckpoint was used to track the weights of the best model in terms of validation accuracy for an optimal model.

E. Performance Evaluation

For performance evaluation, the system uses confusion

matrix, accuracy, precision, recall, and F1-score, AUC score and MCC score. For calculating the accuracy, precision, recall, F1-score,AUC Score and MCC score the following formulas have been used[20][21][22].

$$\text{Accuracy} = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad \text{.....(1)}$$

$$\text{Precision} = \frac{Tp}{Tp+FP} \quad \text{..... (2)}$$

$$\text{Recall} = \frac{p}{TP+Fn} \quad \text{.....(3)}$$

$$\text{F1 Score} = 2\times \frac{(Precision\times Recall)}{(Precision+Recall)} \quad \text{.....(4)}$$

$$\text{AUC} = \frac{total\ \ positives\times total\ \ negatives}{no\ of\ times\ positive\ is\ ranked\ heigher\ than\ negatives} \quad \text{....(5)}$$

$$\text{MCC Score} = \frac{Tp\times Tn - Fp\times Fn}{\sqrt{(Tp+Fp)(Tp+Fn)(Tn+Fp)(Tn+Fn)}} \quad \text{.....(6)}$$

where TP, TN, FP, FN represent true positives, true negatives,false positives and false negatives of a confusion matrix respectively

. (a) (b)

Fig.2 Samples from the RLVS dataset, (a) non violence samples,(b)violence samples.

V. RESULT

In this section we have discussed about the results of our model's .Due to OOM(out of memory) error, 1000 and 1600 clips were used from hocket fight dataset and real life violence situation dataset respectively.

When we trained our base model on both dataset and we achieved an accuracy of 89% on training accuracy and accuracy of 88% on validation accuracy . Due to low accuracy we do fine tunning in our model by changing our L2 regularization from .0005 to 0.0003, patience was changed form 5to 3,for early stopping, patience was changed to 3, min learning rate was sent to 1e-6 for reduce on plateau, also validation loss is monitored. Sustain epoch was set to 0 ,batchsize changed from 4 to 16, exponential decay of 0.8 was also added. After doing this fine-tuning, our accuracy on train and accuracy on test come to be 95.11% and 93.62% respectively.

AS for the 'hockey fight' dataset presented in figure, accuracy on test and train were 95.26% and 94.12% respectively. Also losses on train and test 14.64% and 16.33% respectively.Confusion Matrix of HF dataset is shown below.

precision,recall,f1-score all were 94% . ALong with that MCC score was 88.24% and AUC was 98.39 %. ALl these score suggest strong generalisation and no overfitting .

As for RLVS dataset presented in figure accuracy on test and train were 95.11 and 93.62 respectively. losses of train and test were 14.58% and 18.66% respectively.confusion matrix is shown below.Below is the confusion Matrix of the model. Best epoch was 33. Also, our model makes 9380 correct predictions and 639 wrong predictions.Our model was showing 94% of precision in detecting both violence as well as non-violence class.F1 score of Non violence was recorded to be 92%and for violence class it was 94% . 4% difference in recall was seen where violence class show 95% of recall and non violence show 92%. Our model make 9380 correct prediction and 639 wrong predictions. model reflect very good accuracy but a difference of 4% was seen in between test loss and train loss.may be that can be because of our slight more samples in violence class than non violence .

Although our model show very strong generalization and no symptom of overfitting .

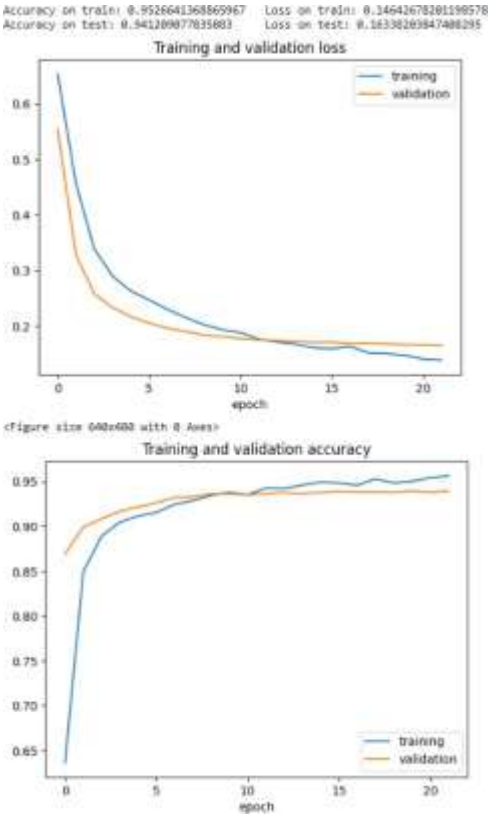

Fig.3 Accuracy and losses for Hockey fight Dataset

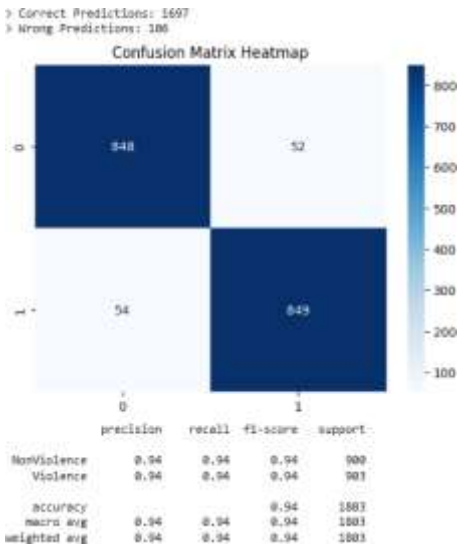

Fig.4 Confusion Matrix for Hockey fight Dataset

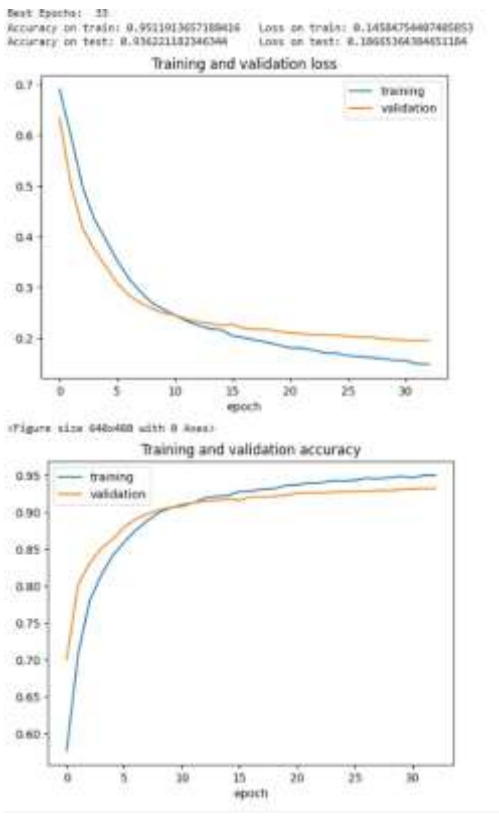

Fig.5Accuracy and losses for RLVS Dataset

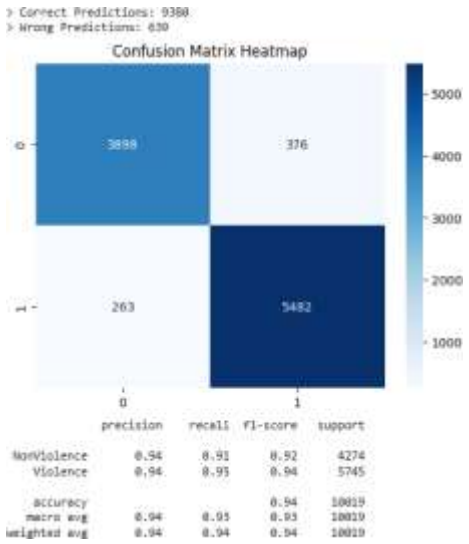

Fig.6Confusion Matrix for RLVS Dataset

VI. Conclusion

Detection of violent clashes and disruptive activity in video footage is a significant research area. In our research, we have created a system that can identify violent incidents within video clips. Initially, all the frames are extracted from each segment of the video, followed by applying data augmentation methods to the frames. Then, the augmented frames are subsequently passed through the mobilentV2-LSTM model, utilizing transfer learning and hyperparameters that improve the model's generalization. Apart from this, the model is coupled with a Telegram bot that sends notifications whenever violent frames are detected. With the progress in technology, the model demonstrates how AI-based solutions can be utilized towards improving public safety. Although the system has promising outcomes, there is still sufficient scope for improvement and enhancement. Future research can explore improvement in violent pattern detection, real-time decision-making, and coupling with other upcoming technologies such as transformers. Through exploring these aspects, we are able to take the boundaries of innovation further, ultimately resulting in safer communities and a safer future.

VII. References

[1]Khan SU, Haq IU, Rho S, Baik SW, Lee MY. Cover theViolence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies. Applied Sciences. 2019;9(22):4963. https://doi.org/10.3390/app9224963J

[2]R. Parlika and A. Pratama, "The Online Test Application Uses Telegram Bots Version 1.0," Journal of Physics: Conference Series, vol. 1569, p. 022042, 2020. doi: 10.1088/1742-6596/1569/2/022042.

[3]W. Zajdel, J. D. Krijnders, T. Andringa and D. M. Gavrila, "CASSANDRA: audio-video sensor fusion for aggression detection," 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, London, UK, 2007, pp. 200–205, doi: 10.1109/AVSS.2007.4425310.

[4]M. -S. Kang, R. -H. Park and H. -M. Park, "Efficient Spatio-Temporal Modeling Methods for Real-Time ViolenceRecognition," in IEEE Access, vol. 9, pp. 76270-76285,2021, doi: 10.1109/ACCESS.2021.3083273.

[5]Ullah FUM, Ullah A, Muhammad K, Haq IU, Baik SW.Violence Detection Using Spatiotemporal Features with 3DConvolutional Neural Network. Sensors (Basel). 2019 May30;19(11):2472. doi: 10.3390/s19112472.

[6]A. -M. R. Abdali and R. F. Al-Tuma, "Robust Real-TimeViolence Detection in Video Using CNN And LSTM," 20192nd Scientific Conference of Computer Sciences (SCCS),2019, pp. 104-108, doi: 10.1109/SCCS.2019.8852616.

[7]Ullah A, Muhammad K, Haq IU, Baik SW. Actionrecognition using optimized deep autoencoder and CNN forsurveillance data streams of non-stationary environments.
Future Generation Computer Systems. 2019 Jul 1;96:386-97.

[8]Nievas, Enrique Bermejo and Suarez, Oscar Deniz and Garcia, Gloria Bueno and Sukthankar, Rahul, "Hockey Fight Detection Dataset", 2016, hosted on bittorent.

[9]M. Soliman, M. Kamal, M. Nashed, Y. Mostafa, B. Chawky, D. Khattab, " Violence Recognition from Videos using Deep Learning Techniques", Proc. 9th International Conference on Intelligent Computing and Information Systems (ICICIS'19), Cairo, pp. 79-84, 2019

[10]https://keras.io/

[11]https://www.tensorflow.org/

[12]https://opencv.org/

[13]https://matplotlib.org/

[14]D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, Jan. 2017

[15]ttps://www.kaggle.com/

[16]L. Torrey and J. Shavlik, "Transfer learning," in Handbookof research on machine learning applications and trends:algorithms, methods, and techniques, IGI Global, 2010, pp.242–264.

[17]R. C. Staudemeyer and E. R. Morris, "A Tutorial into Long Short-Term Memory Recurrent Neural Networks," arXiv preprint arXiv:1909.09586, Sep. 2019.

[18] J. He, L. Li, J. Xu, and C. Zheng, "ReLU Deep Neural Networks and Linear Finite Elements," arXiv preprint arXiv:1807.03973, Jul. 2018. [Online]. Available: https://arxiv.org/abs/1807.03973

[19] C. Cortes, M. Mohri, and A. Rostamizadeh, "Learning Non-Linear Combinations of Kernels," Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009), Montreal, Canada, pp. 109–116, 2009. [Online]. Available: https://ppl-ai-file-upload.s3.amazonaws.com/web/directfiles/attachments/46387881/b7278f80-9185-4cb1-ac23-7e5b7ac05200/l2.pdf

[20] C. X. Ling, J. Huang, and H. Zhang, "AUC: A Better Measure than Accuracy in Comparing Learning Algorithms," in *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence (AI 2003)*, Lecture Notes in Artificial Intelligence, vol. 2671, Y. Xiang and B. Chaib-draa, Eds. Berlin, Germany: Springer, 2003, pp. 329–341.

[21] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 6, pp. 1–13, Jan. 2020. [Online]. Available: https://doi.org/10.1186/s12864-019-6413-7

[22] Z. Karimi, "Confusion Matrix," unpublished manuscript, Kharazmi University, Oct. 2021. [Online]. Available: https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/attachments/46387881/c30b0477-f8cb-41ca-86c5-53cd84601eaf/ConfusionMatrix-for-final-paper.pdf