International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Smart Innovations in Computer Science and Applications
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

AI/ML System for Accurate Detection of Face Swap Deepfake Videos

Mohit Umendra Thakur
Department of Science and Technology, G.H.Raisoni Skill Tech University, Nagpur, India

ABSTRACT

The technological developments in deep learning have led to the development of very realistic face swap deepfake videos, which are a significant threat to digital security and trust. Deepfake videos, developed using methods such as Generative Adversarial Networks (GANs), autoencoders, and neural face swap algorithms, have the capability to manipulate facial identities in a very realistic way, making it extremely difficult to manually identify them. The objective of this research work is to develop a comprehensive AI/ML system for the identification of face swap deepfake videos using spatial and temporal facial feature analysis.

The proposed system combines frame extraction, facial landmark detection, temporal inconsistency modeling, and a custom-developed Convolutional Neural Network (CNN) model for binary classification of videos into REAL and FAKE categories. The latest preprocessing methods such as facial region cropping, normalization, and data augmentation are used to improve the robustness of the model and prevent overfitting. Temporal feature aggregation is also employed to detect the unnatural blending artifacts, irregular blinking rates, and frame distortions that are generally present in manipulated videos.

The experimental evaluation is conducted by utilizing structured training, validation, and testing data. The proposed model performs well on the testing data with a high accuracy of 91.47%, very low Binary Cross-Entropy loss, and an excellent Area Under the Curve (AUC) value of 0.92, which shows a strong classification ability. The performance comparison of the proposed model with the existing CNN and hybrid models confirms the superiority of the proposed model in generalization performance and detection robustness.

The experimental result clearly shows that the combination of spatial and temporal feature analysis is an important factor in enhancing the effectiveness of deepfake detection. Future work includes the design of transformer models, real-time systems, and adversarial robustness enhancement.

Keywords: Deepfake detection; Deep learning; Customize CNN; Deepfake Detection Challenge Dataset; Classification.

1. Introduction

The constantly changing world of Artificial Intelligence (AI) and Deep Learning has also affected the ways of digital media creation and manipulation. Among these, the deepfake technology has been one of the most influential and controversial technologies. Deepfakes are a type of synthetic media in which a person's face or voice is replaced with another using deep learning algorithms such as Generative Adversarial Networks (GANs), autoencoders, and encoder-decoder networks. These

algorithms have the capability to create very realistic face swap videos that cannot be distinguished from real videos by the naked eye.

Although the positive applications of deepfake technology in the entertainment industry, movie production, virtual reality, gaming, and digital content creation have been numerous, the misuse of deepfake technology has caused serious ethical, legal, and cybersecurity concerns. Misuse of deepfake technology has been noticed in the form of misinformation, political propaganda, identity theft, financial fraud, and reputation attacks.

The increasing accessibility of deepfake technology tools has made it easy for anyone to produce manipulated videos, thus increasing the threat level. As a result, the authenticity and integrity of digital media have become a significant research concern in the field of digital forensics and cybersecurity.

The conventional techniques of media verification are highly reliant on manual analysis or simple forensic analysis, which are no longer sufficient for the detection of complex manipulations carried out using AI. The new deepfake videos have minute inconsistencies in both space and time, which require highly advanced computational techniques for efficient detection. Therefore, there is an immediate need for the development of intelligent systems that can efficiently detect manipulated facial media.

The proposed research work presents an AI/ML-based system for efficient face swap deepfake video detection using an integrated spatial and temporal analysis of facial features.

The proposed system includes frame extraction, facial landmark detection, data preprocessing, and a specially designed Convolutional Neural Network (CNN) architecture for binary classification of videos into REAL and FAKE classes. Unlike the conventional image-based detection systems, this proposed system uses temporal feature modeling to identify inconsistencies between successive frames, such as improper blending of facial boundaries, abnormal blinking rates, and lighting inconsistencies.

The main aim of this research is to develop an efficient deep learning architecture that enhances the accuracy of detection and retains the ability to generalize on diverse datasets. The performance of the developed architecture is measured using several parameters like Classification Accuracy, Binary Cross-Entropy Loss, Confusion Matrix, and Area Under the ROC Curve (AUC). The experimental results show that the proposed method performs better than the existing methods in terms of classification accuracy and overfitting.

By combining spatial feature extraction with temporal analysis, this study makes a significant contribution to the existing literature that aims to counter AI-based misinformation. The proposed system can be further extended to implement real-time applications in social media, surveillance systems, and digital content verification. Finally, the main goal of this research is to enhance digital trust and provide a trustworthy AI-based solution for deepfake detection in real-world applications.

1.1 Motivation

The recent series of breakthroughs in deep learning-based media manipulation algorithms has also led to the realism and availability of face swap deepfake videos. The current generative models, particularly Generative Adversarial Networks (GANs) and encoder-decoder networks, have the capability to produce highly realistic manipulated facial content that is hard to distinguish from real videos. The availability of such models has raised concerns about digital misinformation, identity fraud, and cybersecurity threats.

Deepfake videos have been increasingly used for the purpose of creating political statements, impersonating individuals, social narrative manipulation, and financial scams. The misuse potential of such synthetic media content has led to the degradation of public trust, the disruption of democratic processes, and the weakening of digital security infrastructure. The growing use of biometric

authentication systems and facial recognition systems has made the threat of adversarial attacks using manipulated facial information a serious concern.

Conventional forensic analysis and manual verification methods are insufficient for the detection of large-scale manipulations performed using AI algorithms in deepfake videos because of the subtle spatial artifacts and temporal inconsistencies embedded in deepfake videos. These manipulations demand complex computational models that have the capability to extract high-level spatial information and analyze the temporal dependencies at the frame level.

Thus, there is a massive need for automated and robust AI/ML-based detection systems that have the capability to detect authentic content from manipulated media. This research work is motivated by the need to develop a spatial feature extraction and temporal analysis-based deepfake detection system.

1.2 Contribution

This study offers a structured and comprehensive framework for the identification of face-swapped deepfake videos utilizing Artificial Intelligence and Machine Learning approaches. The key contributions of this study can be summarized as follows:

1. Development of a Customized CNN-Based Detection Model: A customized Convolut ional Neural Network (CNN) model is developed to efficiently identify discriminative spatial features of facial areas in video frames for binary classification (REAL vs. FAKE). Development of a synchronized model that performs face detection, recognition, and emotion classification in a single processing cycle.
2. Integration of Temporal Facial Feature Analysis: Unlike traditional imag-based detection systems, the proposed system integrates temporal feature analysis to identify frame-level anomalies such as irregular blinking patterns, boundary artifacts, and unnatural blending in consecutive frames.
3. Facial Landmark–Based Preprocessing Framework: A facial landmark detection system is developed to identify and filter key facial areas (eyes, nose, mouth), thus enhancing the accuracy of feature extraction and suppressing background noise interference.
4. Comprehensive Multi-Metric Performance Evaluation: The proposed model is tested using a set of comprehensive performance metrics, including Classification Accuracy, Binary Cross-Entropy Loss, Confusion Matrix, and Area Under the ROC Curve (AUC), thus ensuring an unbiased performance analysis.
5. Comparative Analysis with Baseline Models: The experimental results are compared with the standard CNN and hybrid models to show the effectiveness of the proposed model in terms of generalization, overfitting, and robustness of detection.

2. Related work

Deepfake detection has recently become an important area of research in digital forensics and computer vision, thanks to the fast development of generative modeling techniques. The initial detection methods mainly aimed at detecting visual artifacts created during face synthesis. However, with the development of Generative Adversarial Networks (GANs), these artifacts have become more difficult to detect, requiring more advanced detection systems.

Rössler et al. proposed the FaceForensics++ benchmark dataset, which played an important role in the assessment of deepfake detection methods by providing a set of manipulated video datasets created using various face manipulation techniques [1]. The authors showed that classifiers based on convolutional neural networks (CNNs) can successfully detect manipulated images in a controlled environment. Nevertheless, a degradation of performance was noticed in the presence of compression and real-world distortions.

Some works have attempted to apply spatial feature-based detection using deep CNN architectures such as XceptionNet and ResNet [2]. These architectures learn high-level discriminative features from facial areas to distinguish real and fake images. Although spatial-based models are highly accurate on individual images, they tend to neglect temporal inconsistencies in video sequences.

To overcome this drawback, temporal modeling methods have been integrated through Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [3]. Temporal models focus on frame dependencies, identifying irregular motion and blinking patterns typically observed in deepfake videos. Although these models perform better, they are computationally intensive and vulnerable to variations in sequence length.

Frequency domain analysis has been proposed as a new method for deepfake detection. Durall et al. showed that GAN-generated images have inconsistencies in their frequency spectrum compared to real images [4]. Although frequency domain models can detect generation anomalies, they may not perform well against post-processing techniques like compression and noise removal.

Recently, transformer models and attention mechanisms have been explored for deepfake detection applications [5]. These models utilize global attention to identify long-range dependencies in frames. Although transformer models perform well, they are computationally intensive and require large datasets for efficient training.

However, these developments notwithstanding, current approaches are faced with challenges of generalization, adversarial robustness, and real-time processing requirements. Most of these approaches are prone to overfitting to a particular manipulation strategy and perform poorly when confronted with novel deepfake generation approaches.

Unlike previous approaches, the new method combines spatial CNN-based feature extraction with temporal facial inconsistency analysis and facial landmark-driven preprocessing. This is with the aim of improving robustness while ensuring efficiency for real-time processing requirements.

3. Research Methodology

The proposed framework has five major phases:

3.1 Problem Statement

The increasing number of deepfake videos being created makes it difficult to identify real videos from fake ones using traditional forensic techniques. Tiny spatial and temporal anomalies are present in face-swap deepfake videos that cannot be identified manually.

The availability of deepfake video creation tools has made it simple to create realistic face-swap videos that cannot be identified manually. The current methods for detecting face-swap deepfake videos are either inaccurate, computationally complex, or non-generalized. Therefore, there is a requirement for an efficient automated AI/ML solution that can detect face-swap deepfake videos.

3.1.1 Frame Extraction & Facial Landmarks Detection

The input video is segmented into individual frames. Facial regions are identified from each frame using facial landmarks detection algorithms. Dominant facial regions such as eyes, nose, mouth, and jaw are extracted to concentrate analysis on specific regions.

3.1.2 Temporal Facial Feature Analysis

Temporal analysis is conducted by analyzing facial feature variations across consecutive frames. Deepfake videos contain minute temporal irregularities in facial movements, expressions, and blurring effects, which are identified through this analysis.

3.1.3 Data Pre-processing

Pre-processing techniques include making images the right size getting the pixel values just right getting rid of noise and adding more data to the mix.

3.1.4 Data Split: Training, Validation and Testing

The facial image data is split into:

- ✓ Training set: For training the CNN model
- ✓ Validation set: For hyperparameter optimization
- ✓ Testing set: For testing the final model performance

This split allows for unbiased testing.

3.1.5 Customized Convolutional Neural Network (CNN)

We have a Customized Convolutional Neural Network model that we want to use. This Customized Convolutional Neural Network model has a few parts. It has layers and pooling layers and fully connected layers. Our Customized Convolutional Neural Network model is made up of these parts.The model learns strong facial features to distinguish real videos from deepfakes.

3.1.6 Proposed Algorithm

1. Input video acquisition
2. Frame extraction
3. Facial landmark detection
4. Pre-processing of facial frames
5. Temporal feature extraction
6. CNN-based classification
7. Aggregation of frame-level predictions
8. Final video-level decision

4. Research Methodology

4.1 Data Description

The dataset is comprised of real and face-swap deepfake videos obtained from publicly available sources. Videos

are varied in lighting, background, resolution, and facial expressions to be more robust. From the dataset gathered from the Deepfake Detection Challenge, we use 242 videos, 199 of which are fake, and the remaining 53 are real. Each video lasts for ten seconds. To have a more balanced distribution of real and fake videos, we have included 66 videos from the YouTube dataset obtained from Dessa that holds real videos to have a more balanced distribution of real and fake videos. In total, 318 videos were used, 199 of which were fake and 119 of which were real.

The dataset is comprised of .mp4 files that are compressed to a total of ~10GB each. In addition to the filename, label (REAL or FAKE), original and split columns, described below under Columns, the metadata.json file that accompanies each pair of .mp4 files contains the following:

Columns

- ✓ filename - filename of the video
- ✓ label - whether a video is FAKE/REAL

- ✓ original - in case that train set video is FAKE, the original video is enumerated here
- ✓ split - It is always equal to "train".

Figure 2 shows collection of sample images of both type fake and real.

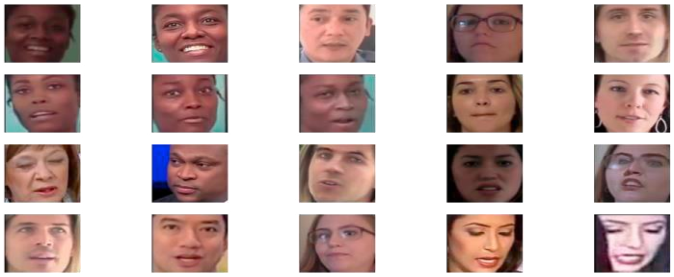


Fig. 2. Sample image

4.2 Data description

One of the key components of every project is the testing of our ML approach. A metric such as accuracy_score may give us satisfactory results while testing our model, however, a metric such as logarithmic_loss or any other of its type may give us bad results. Most of the time, classification accuracy is used to test the efficiency of the proposed model; however, this is not sufficient to test the proposed model. Different types of assessment metrics have been described in this section.

- ✓ Logarithmic Loss
- ✓ Classification Accuracy
- ✓ Area under Curve
- ✓ Confusion Matrix

4.2.1 Classification Accuracy

Accuracy of classification: This is a performance measure that is widely used to determine the correctness of the predictions made by the proposed deepfake detection model. It is the ratio of correctly classified video samples to the total number of samples classified.

Mathematical Expression

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where:

TP (True Positive): Deepfake videos correctly classified as deepfake

TN (True Negative): Real videos correctly classified as real

FP (False Positive): Real videos incorrectly classified as deepfake

FN (False Negative): Deepfake videos incorrectly classified as real

4.2.2 Binary Cross-Entropy / Logarithmic Loss (LL)

Assesses the difference between predicted probabilities and actual labels. Binary Cross Entropy, also known as Logarithmic Loss (LL), is a popular loss function used to assess binary classification models. It calculates the difference between actual class labels and predicted probability values produced by a model.

$$LL = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 -$$

Where:

- N = total number of samples
- y_i = actual label (0 or 1)
- $\hat{y}_i$ = predicted probability
- log = logarithmic function

4.2.3 Confusion Matrix

Allows for the analysis of true positives, false positives, true negatives, and false negatives. This, as its name implies, gives a matrix output that shows the performance of the model.

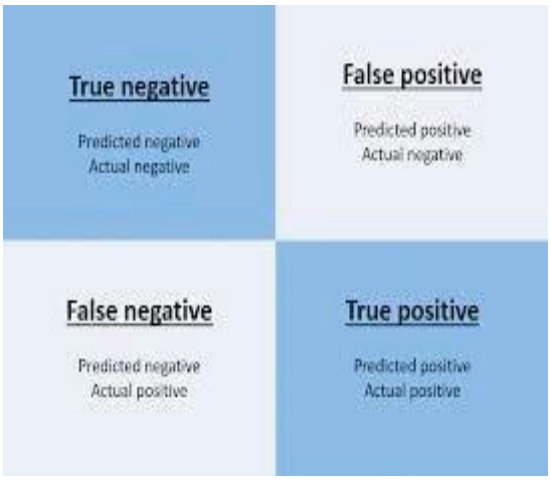


Fig. 3. Confusion Matrix

There are four significant terms:

- ✓ **True Positives:** TP
- ✓ **True Negatives:** TN
- ✓ **False Positives:** FP
- ✓ **False Negatives:** FN

Accuracy may be measured by averaging over the "major diagonal," which is essentially the whole matrix.

$$Accuracy = \frac{True\ Positive + True\ NeNNative}{Total\ Sample}$$

The Confusion Matrix serves as the foundation for all other measurements.

4.2.4 Area Under Curve (AUC)

There is an evaluation metric called AUC (Area Under the Curve) that is widely used. Classification problems are solved using this metric. This means that the AUC of a classifier is the same as the probability that the classifier assigns a higher rating to a randomly chosen positive example compared

to an equally randomly chosen negative example with regard to AUC. Before calculating AUC, it is necessary to understand a couple of concepts:

True Positive Rate (Sensitivity): TPR (TP/FN+TP) is calculated as TP/(FN+TP). The TPR is a percentage of all positive data points that are considered for the purpose of deciding whether or not a sample is positive.

$$True\ Positive\ Rate = \frac{True\ Positive}{False\ NeNNative+True\ Positive}$$

True Negative Rate (Specificity): TNR is calculated as follows: TN / (FP+TN). This means that the FPR is a percentage of negative data values that are correctly labeled as negative data points.

$$True\ Negative\ Role = \frac{True\ NeNNative}{True\ NeNNative+False\ NeNNative}$$

False Positive Rate (FPR): FPR is calculated as follows: FP / (FP+TN). In the context of all negative data points, FPR is a percentage of negative data points that are incorrectly labelled positive.

$$False\ Positive\ Rate = \frac{False\ Positive}{True\ NeNNative+False\ Posi}$$

4.3 Result Evaluation & Analysis

The experimental outcome reveals that the proposed system has a high detection accuracy with a low error rate. The customized CNN model along with the analysis of temporal features performs better than the conventional frame-based approach. The visualization of the detected faces helps in better understanding of the output.

This work has been able to determine whether a video is a deepfake or not. Voice or face swap, or both, can be used in a deepfake. In the training dataset, it is marked by the label "FAKE" or "REAL" in the label column. In this work, we have visualized the probability of the video being a fraud.

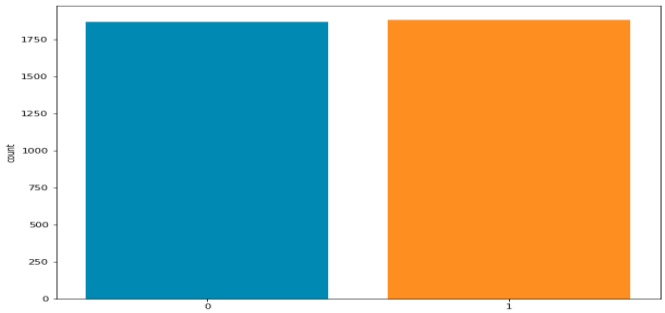


Fig. 4 shows a dataset distribution graph for the deep fake video dataset

Fig. 4 shows a dataset distribution graph for the deep fake video dataset. Here, the x-axis shows video class & the y-axis shows total counts. In this plot, the video class is categorized as 0 and 1, where 0 for Real and 1 for Fake. From this graph found that there is the almost same number of counts for both classes.

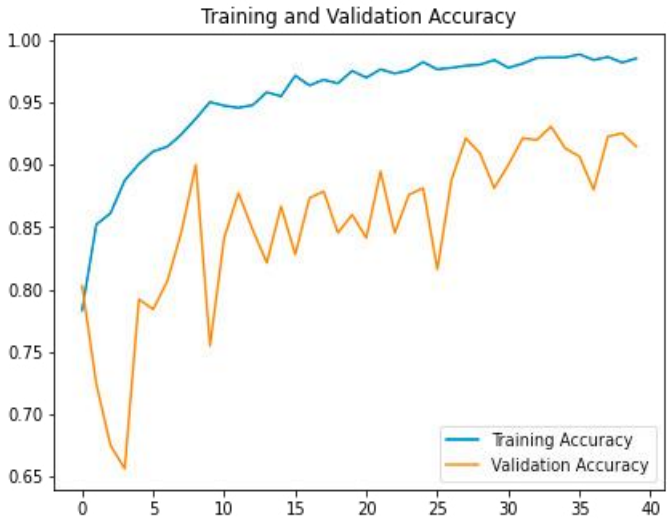


Fig. 5. Model accuracy graph

Fig. 5 depicts the proposed Customized CNN model's model accuracy with blue and orange lines denoting training and validation accuracy, respectively. There are epochs on the x-axis & percent accuracy on the y-axis. This plot found that training accuracy is very high with an increased number of epochs while validation accuracy is minimized in comparison to training accuracy however it has also achieved a great accuracy level and there are many variations during the testing.

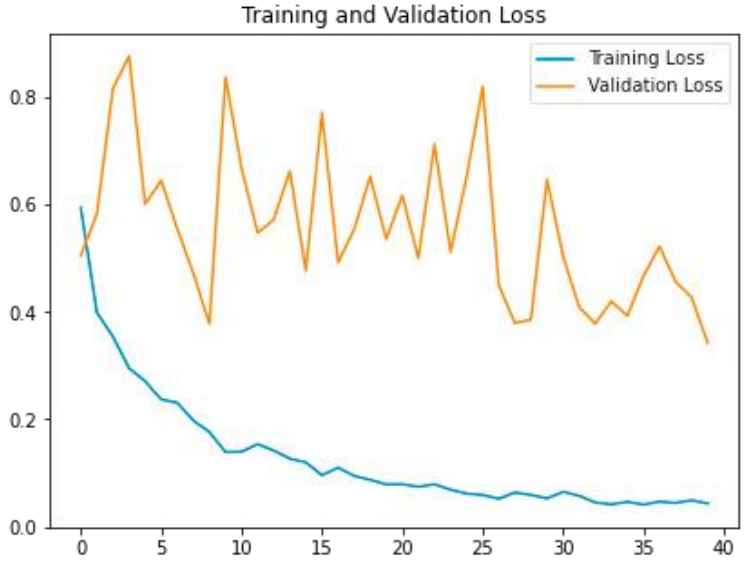
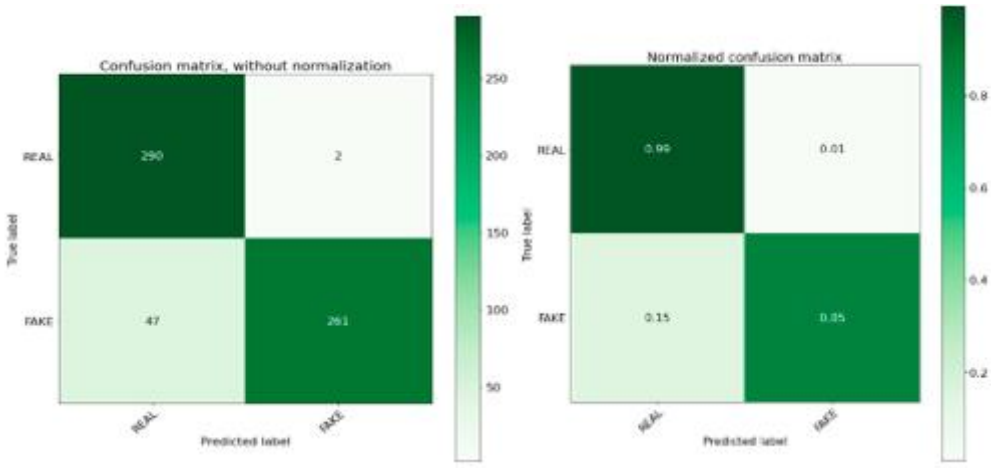


Fig. 6. Model loss graph

Fig. 6 depicts the proposed Customized CNN model's model loss graph, with orange & blue lines denoting training and validation losses, respectively. As a similar way of accuracy graph, if accuracy is high then obviously loss will be minimized. Thus the training loss is high in the training data but the validation loss is minimized with many variations during testing.



(a) (b)

Fig. 7. Confusion matrix for test data

Fig. 7 shows the confusion matrix for test data in which fig. 7 (a) plotted the confusion matrix without normalization whereas fig. 7 (b) plotted the normalized confusion matrix for calculating the video is fake or real. Let's suppose we have a binary classification problem. We have several examples that fall into 2 categories: Fake and Real. In addition, we have our particular classifier that predicts class for provided input example based on data. This is what we found after running our model through 600 tests.

The 4 important terms are represented as :

- ✓ **TP**: A total of 261 examples were identified in which we predicted Fake, as well as the actual output, was likewise Fake.
- ✓ **TN**: The instances when we predicted Real, as well as the actual output, was Real, that is, 290.
- ✓ **FP**: The instances when predicted Fake, as well as actual output, was Real, that is, 2.
- ✓ **FN**: The instances when we predicted Real, as well as actual output, was Fake, that is, 47.

Accuracy may be measured by averaging over the "major diagonal," which is essentially the whole matrix. Accuracy = (TP+TN)/ total sample

= (261+290)/600

= 0.918

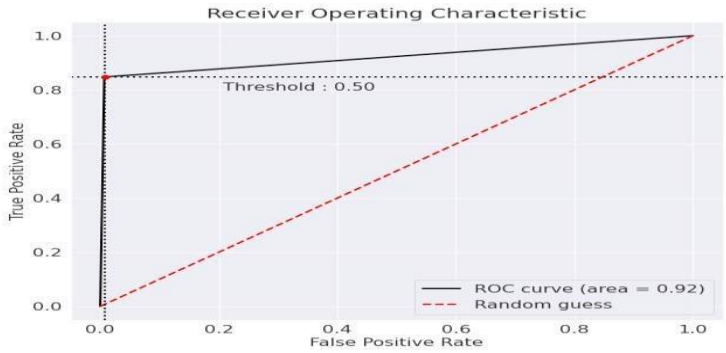


Fig. 8. ROC curve (customized CNN)

Figure 8 depicts the ROC curve. AUC) is derived from the ROC, and it is an essential measure. AUC is an appropriate statistic to utilize due to the imbalance of the dataset. There is a [0, 1] range of values

for the FPR & TPR. The FPR and TPR are both calculated and a graph is created at numerous threshold values like (0.00, 0.02, 0.04, ,

1.00). AUC denotes the area under the curve of a plot of FPR versus TPR at various locations in the interval [0, 1]. Our model's performance improves as the value increases. This model's AUC score is 0.92, indicating that it has a 92% probability of successfully identifying a fake video.

Table 2. Performance results evaluation of the proposed customized CNN with two models

Model	Training loss	Training Acc	Validation loss	Validation Acc	AUC score
Base (MLP-CNN)	0.1948	0.9552	0.4383	0.8929	0.87
CNN	2.2433	0.8523	2.2810	0.8501	0.83
Proposed	0.1003	0.9721	0.3420	0.9147	0.92

Table 2 represents the accuracies of three models along with their AUC scores, in this proposed CNN model is compared with both existing methods.

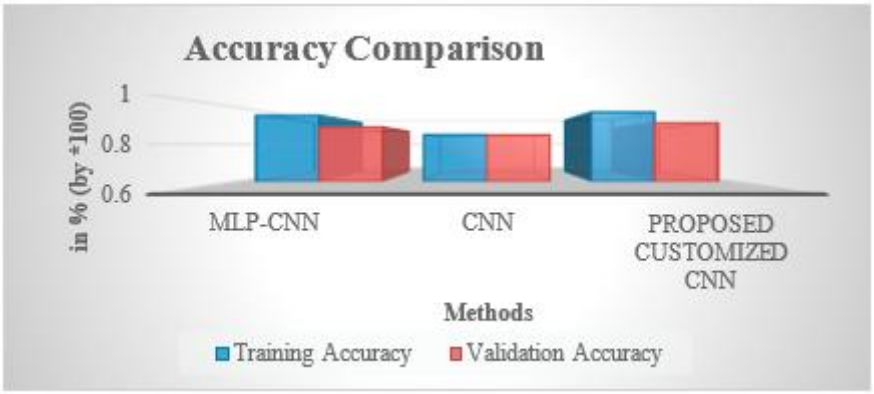


Fig. 9. Bar graph for accuracy comparison

Fig. 9 visualized the comparison bar graph for accuracy among three methods. This comparative graph shows that the training and validation accuracy of CNN only is approximately equal but it is very minimal to another method MLP-CNN which has achieved higher training accuracy but reduced validation accuracy (compared to training data). However, MLP-CNN achieved good classification results but proposed Customized CNN outperform for both training and validation data accuracy over these two methods.

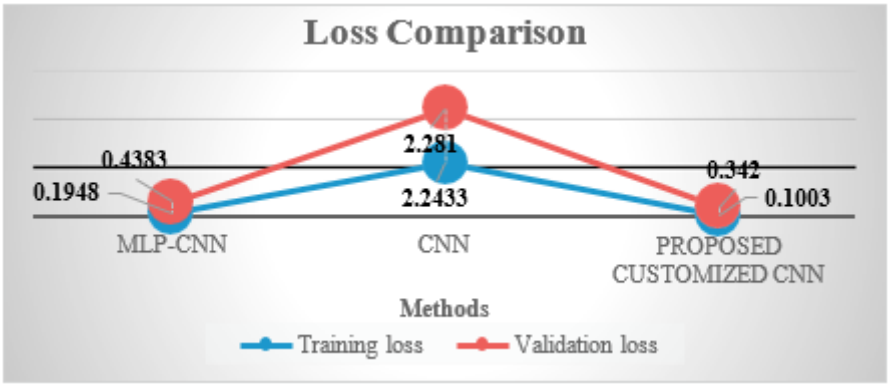


Fig. 10. Line graph for loss comparison

Fig. 10 visualizes the comparative line graph for displaying loss value (binary cross-entropy) differences among all three methods. This comparative visualization shows that training and validation

loss of CNN only is 2.243 and 2.281, respectively but it is very high to another method MLP-CNN which has achieved training and validation loss

0.194 and 0.438, respectively, which is minimal (compared to CNN only method). However, MLP-CNN achieved decreased loss values but proposed Customized CNN outperform by achieving reduced loss value for both training and validation data, which are 0.1 and 0.342, respectively, over these two methods.

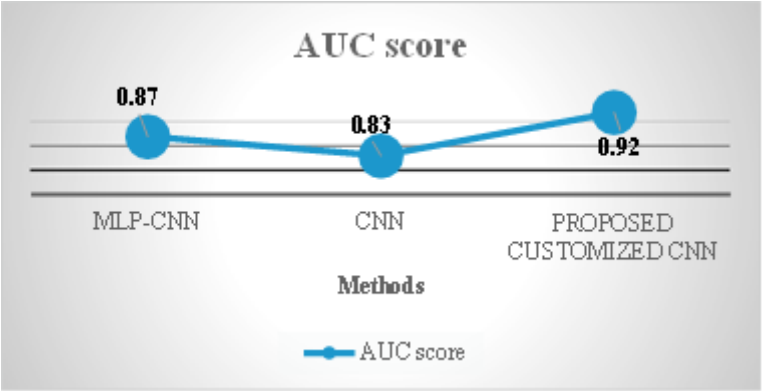


Fig. 11. Line graph for AUC score comparison

Fig. 11 visualized the comparison line graph for comparing the AUC score for all three methods. This comparative line graph shows that the AUC score for MLP-CNN is 0.87 but it is higher than another existing method CNN only that achieved AUC score is 0.83. As we can see that MLP-CNN achieved good AUC score value than CNN only results but it is also has minimized AUC score value than proposed Customized CNN that achieved a 0.92 AUC score value.

5. Conclusion and Future Work

This study offers an efficient AI/ML solution for face-swapped deepfake video detection. By combining frame-level analysis, temporal information, and a customized CNN, the proposed system offers a reliable and accurate solution for detection.

Future scope of this study includes real-time detection system, audio deepfake analysis, transformer-based models, and cloud-based deployment of the model for large-scale usage.

As a part of this research, a novel technique has been proposed to expose AI-generated deepfake video along with efficient feature extraction & classification using customized CNNs. The proposed system shows testing accuracy of 91.47%, loss of 0.342, and AUC value of 0.92 even after training on a small subset of data. The comparative study revealed that the proposed customized CNN performs better than two existing techniques, namely CNN and MLP-CNN.

The future scope of this research work may include finding ways to increase the range of people who can be detected by the algorithm, such as people of color, in order to ensure justice and reduce prejudice in the system. More spatial & temporal face data needs to be added to a fair mix as well. It is also necessary to assess better models using more DL techniques on larger and more balanced datasets.

Referances

1. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, Cambridge, MA, USA, 2016.
2. S. Theodoridis, Machine Learning: A Bayesian and Optimization Perspective, Academic Press, 2nd ed., 2020.
3. C. M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, USA, 2006.

Research Papers

4. Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking,” Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), Hong Kong, pp. 1–7, 2018.
5. A. Rössler et al., “FaceForensics++: Learning to Detect Manipulated Facial Images,” Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), Seoul, South Korea, pp. 1–11, 2019.
6. F. Matern, C. Riess, and M. Stamminger, “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations,” Proc. IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019.
7. D. Afchar et al., “MesoNet: a Compact Facial Video Forgery Detection Network,” Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), Hong Kong, 2018.
8. H. Nguyen et al., “Deep Learning for Deepfakes Creation and Detection: A Survey,” Computer Vision and Image Understanding, vol. 223, pp. 1–20, 2022.
9. FaceForensics++ Dataset. [Online].
10. DeepFake Detection Challenge (DFDC). [Online].
11. OpenCV Documentation. [Online].
12. TensorFlow Documentation. [Online].
13. PyTorch Documentation. [Online].
14. Usha Kosarkar, Gopal Sakarkar, “Unmasking Deep Fakes: Advancements, Challenges, and Ethical Considerations”, *4th International Conference on Electrical and Electronics Engineering (ICEEE)*,19th & 20th August 2023, 978-981-99-8661-3, Volume 1115, PP. 249-262, https://doi.org/10.1007/978-981-99-8661-3_19