Development Of Artificial Intelligence-Based Model for Forensic Analysis of Cross-Platform Deepfakes

*Joseph, C. C[1], Ojeniyi, J. A[2], Noel, M. D[3], Ahmad, S[4], Fasola, O. O[5], Uduimoh, A. A[6]

Department of Cyber Security Science, Federal University of Technology, Minna, Nigeria

DOI: https://doi.org/10.51244/IJRSI.2025.1210000285

Received: 31 October 2025; Accepted: 07 November 2025; Published: 19 November 2025

ABSTRACT

Synthetic media, a product of advancements in artificial intelligence (AI) and machine learning, represents a transformative innovation that reshapes how content is created, manipulated and consumed. One of the most advanced technology in synthetic media is deepfakes. The misuse of deepfakes poses significant threats to privacy, security and societal trust. Cybercriminals exploit this technology for phishing scams, identity theft and spreading misinformation. This study developed an AI-based hybrid model for forensic analysis of cross platform deep fakes to address these gaps. The study developed a hybrid model that integrates 2D CNN, 1D CNN and RNN framework capable of isolating platform-specific spatial and temporal features for forensic analysis of cross-platform deepfakes. The platforms include social media such as Facebook, Instagram, Youtube, Tiktok and Twiter. The research used Celeb DF dataset for training, validation and testing of the hybrid model. Result from the evaluation metrics showed that the model achieved an accuracy of 99 % on the training data and 93.5 % on the test data. Result equally showed that precision, recall and MCC value were 96.8 %, 90 % and 87.2 % respectively on the test data. The model outperformed single CNN and RNN models and some hybrid models reported. This study demonstrated a reliable hybrid model with high detection accuracy for the forensic analysis of deepfakes. Hence, the model has the potential to address the problem of misinformation caused by synthetic media.

Keywords: Artificial Intelligence, forensic analysis, 2D CNN, RNN, Deepfakes, Hybrid model

INTRODUCTION

In recent years, there has been a significant increase in the creation and dissemination of deepfake videos across various platforms [1,2]. This creation and dissemination of deepfake videos constituted both detection and security threats considering that deepfakes are manipulated videos from sophisticated deep learning techniques including Autoencoders and Generative Adversarial Networks (GANs) to produce contents that appear authentic but fabricated [3,4]. These deepfakes pose significant threats to digital authenticity, individual privacy, and public trust. Although, video manipulation is not a new phenomenon, however, the use of advanced technologies such as deep learning has revolutionised the process [5]. The misuse of deep learning in misinformation campaigns, cybercrimes and blackmail has raised global concerns. For instance, deepfakes are used to impersonate high-profile individuals. Typical example is a 2018 viral deepfake video featuring former President Barack Obama insulting President Donald Trump and making other controversial statements [3]. Additionally, cybercriminals have exploited deepfakes to impersonate and defraud individuals and corporate organisations, a typical case is a 2024 video conference where an employee was tricked into transferring $25 million during team meeting [6].

Digital forensics involves the collection, verification and analysis of digital evidence in a manner that adhere to both legal and scientific standards [7]. Forensic tools and techniques must undergo peer review and be scientifically validated [8]. Large Language Models (LLMs), however, present challenges due to their lack of transparency regarding training data and internal processes, complicating the verification of their outputs' scientific reliability. Additionally, data confidentiality is critical in digital forensics and reliance on cloud-based

LLMs raise concerns about exposing sensitive case information to third parties [9]. The role of digital evidence in law enforcement investigations has gained significant attention. This heightened focus stems from the pervasive nature of digital data in incidents requiring police scrutiny, as well as its capacity to provide detailed insights into the actions of involved parties. In the context of deepfake forensics, this trend has shown growing need for sophisticated digital forensic techniques to authenticate and analyse potentially manipulated media. As deepfakes become more prevalent and sophisticated, law enforcement agencies increasingly rely on advanced digital forensic methods to discern genuine content from artificial creations [10]. This is to ensure the integrity of evidence in criminal investigations

Despite advancements in detection methods, existing forensic tools require complex models to address the challenges posed by differences in how videos or images are saved and compressed across devices and applications [11]. Most detection methods rely heavily on few metrics, but this study explored quite a number of metrics including the recent underutilised metric like the Matthews Correlation Coefficient (MCC) which are essential for evaluating forensic reliability in real world applications [12]. Furthermore, existing models are static and lack adaptability to evolving deepfake techniques and platform updates, which lead to rapid obsolescence
[13].

This study developed an AI-based hybrid model for forensic analysis of cross platform deep fakes to address these gaps. The model aims at enhancing detection accuracy by utilising hybrid 2D CNN, 1D CNN and RNN framework capable of isolating platform-specific artifacts. It also incorporated comprehensive evaluation metrics such as Accuracy, Area under the Curve (AUC), Precision, Recall, F1 Score, Specificity and the Matthews Correlation Coefficient (MCC). These metrics were chosen to ensure a balanced evaluation covering accuracy, correctness and practical usability. This addressed the existing limitations and provided a scalable solution for combating the growing threat of deepfakes across digital ecosystems.

Related Work

2.1 Customized deepfakes detection models

Customized deepfakes detection models are AI systems tailored to identify manipulated images or videos with higher accuracy than generic models. They achieve this through: architecture customization, data-driven tailoring, and training techniques. One of such studies is the development of M-Task-SS (customised CNN) deepfake detection model that uses multi-task learning, self-supervision and incorporated additional layers such as dense layers, Max Pooling layers, and dropout layers to enhance detection accuracy to improve generalisation across different datasets [14]. Kosarkar *et al*. [15] explores the detection of deepfake images and videos using a customised Convolutional Neural Network (CNN) algorithm and compares its performance against two other CNN models. The customised CNN model, which includes additional layers such as a dense layer, Max Pooling, and a dropout layer. Almestekawy *et al*. [16] introduces a novel deepfake detection method focused on enhancing model stability and ensuring reproducible results, which are often overlooked in existing research. The technique combines multiple spatiotemporal textures with deep learning-based features, utilising an enhanced 3D Convolutional Neural Network (CNN) within a Siamese architecture. This enhanced 3D CNN incorporates a spatiotemporal attention layer designed to filter out irrelevant input sections, allowing the model to concentrate on critical regions. The feature extraction module uses Gray Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP) to extract texture features, which are then combined with deep learning features using a Single-Layer Perceptron (SLP) in the feature fusion module. Jafar *et al*. [17] develop the DFT-MF model, a deep learning approach using mouth features to detect Deepfake videos by analysing lip/mouth movement. The methodology involves utilising CNNs to process video frames, focusing on mouth movements as key indicators of manipulation. The model analyses these features to classify videos as either fake or real. This approach capitalises on the inconsistencies and artifacts introduced during the deepfake creation process, particularly in the lip and mouth areas.

2.2 Hybrid CNN-RNN

This model integrates a Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN), specifically combining Efficient Net as the CNN and Long Short-Term Memory (LSTM) as the RNN. The RNN layer, LSTM, is placed on top of the CNN to process sequential data by taking feature vectors extracted by the CNN as input. Efficient Net, pre-trained on the ImageNet dataset, is used to extract spatial features from input video sequences through its convolutional layers. These features are then normalised to enhance processing speed, activated to introduce non-linearity, and downsized using max pooling to retain essential information while reducing data dimensions [18]. Suratkar and Kazi [19] introduced a novel framework for detecting deep fake videos using transfer learning with autoencoders and a hybrid Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) model. The study analyses the generalisability of convolutional autoencoders and CNN-RNN models on benchmark datasets such as DFDC, Face Forensics, Face-Forensics++, and DFD. Petmezas *et al.*, [20] presented a hybrid model with deep learning framework that combines CNN and Long short-term memory (LSTM) and transformer architecture to detect video based deepfakes. The CNN extracts the spatial facial features from the individual frames, while the LSTM captures temporal relationships between frames to recognize motion inconsistencies. The transformer module enhances attention to subtle feature and variation, improving detection accuracy and robustness.

Sharma *et al*. [21] introduces a GAN-CNN ensemble model designed to improve deepfake detection in social media images while minimising the catastrophic forgetting issue common in CNN-based detectors. The methodology involves using a Generative Adversarial Network (GAN) to generate and store samples from previous tasks, which are then replayed during the training of new tasks. This approach aims to make the CNN more robust and adaptive in identifying deepfakes in dynamic environments.

MATERIALS AND METHODS

3.1 Data collection

The data for this research was downloaded from Celeb-DF (v2) dataset. The integration of the Celeb-DF datasets provides a robust foundation for developing advanced deepfakes forensic analysis, prioritising high-quality synthetic videos that minimise detectable artifacts [22]. Celeb-DF contains high-quality deepfake videos of celebrities, generated using an improved synthesis process that reduces visual artifacts like colour mismatches, splicing boundaries and inconsistent face orientations. It includes 590 authentic videos of celebrity, additional 300 youtube real video and 5639 fake video [23]. Celeb-DF addressed limitations of earlier datasets (e.g., DFD, DFDC-P, FF++) by producing synthetic videos that closely resemble real-world deepfakes. Individual frames were extracted from the video for further processing**.** Video frames were generated by breaking down a video into individual images [2].

3.2 Pre-processing steps

To enhance the effectiveness of the forensic analysis of deepfakes, it is essential to first improve the image quality through pre-processing. A critical step involves face detection, which allows filtering out frames or parts of frames without faces [24]. To isolate the face Region of Interest (RoI), computer vision techniques were used to automatically detect faces in images. The Dlib classifier was employed to identify 68 facial landmarks. This detector ensembled regression trees to predict the placement of facial features based on pixel intensities within the images. This effectively identified faces by mapping 68 key points by eliminating unnecessary frames.

The image were resized. This involves adjusting the dimensions of an image to either enlarge or reduce its size without altering its content. In this approach, the image resolution was standardised to 224 × 224 pixels to ensure consistency across the dataset. The dataset was divided into three subsets: training, validation, and testing. The validation phase is used to determine the most effective architecture. During model training, the validation set helps select the best-performing architecture, while the combined training and test sets are used to evaluate the model's performance after training.

3.3 Model Development, Training and Testing

A hybrid model that consists of 2D CNN, 1D CNN and RNN for the forensic analysis of cross-platform deepfakes was developed. Pre-processed video data was fed into this model, which was then trained to learn patterns from the input. Following training, the model was tested to evaluate its performance.

The model employs a deep learning approach with a hybrid 2D CNN, 1D CNN and RNN to determine whether a video is real or fake. This classification depends on whether the number of manipulated frames in the video exceeds a predefined threshold. The model analyses some key variables to make this decision such as patterns, texture, artifacts, and temporal dynamics of consecutives frames (e.g. unnatural blinking, facial expression temporal incoherence). Integration of these metrics helps the model identifies inconsistencies typical of synthetic media, ensuring robust detection of deepfakes. The eye shape detector uses coordinates (points 37–46) from facial landmarks to calculate the Euclidean distances between the endpoints of both the left and right eyes, denoted as d1 and d2, respectively. Similarly, the lip shape detector uses lip coordinates (points 49–68) to compute the length of the inner lips (d3) and the outer lips (d4) by calculating the Euclidean distances among relevant lip coordinates. The nose shape detector utilises facial landmark data (points 28–36) to determine the nose's shape. It calculates the base width of the nose (d5) and the width at its highest point (d6) using Euclidean distances. These measurements - eye widths (d1, d2), inner and outer lip lengths (d3, d4), and nose widths (d5, d6) was used as shape features to train a classifier that can differentiate between real and manipulated videos based on variations in facial features across frames.

3.4 Model Performance Evaluation Metrics

This study enhanced the AI-based model performance by leveraging relevant metrics. By integrating these approaches, the researcher developed robust frameworks for improving AI models across diverse applications, ensuring higher-quality outputs and adaptability in real-world scenarios. Below are the performance metrics:

3.5 How the model works

The study developed a hybrid model which consists of CNN (2D and 1D), RNN for the forensic analysis of cross platform deep fakes, with each contributing to the model’s efficacy and efficiency. CNN models detect deepfakes by extracting hierarchical features to spot subtle artifacts and inconsistences in deepfakes [25]. CNN focuses on different facial region, while RNNs analysed the temporal sequence of video frames to detect subtle inconsistencies introduced in manipulated videos. 2D CNNs capture spatial features in individual frames for deepfakes detection, especially by analysing individual frames or images. A 2D CNN processes spatial features by applying convolutional filters over the two-dimensional pixel data of each frame. This extracts patterns that distinguish real from fake images. It is often integrated into larger video detection frameworks for better performance. The 1D CNN identifies local and short range pattern such as blinking anomalies, lip-sync errors and flickering.

This hybrid model leverages CNN strength in spatial analysis and RNN strength in temporal sequence modeling. RNNs improve the accuracy of deepfakes detection by learning temporal correlations, but for its limitation in the detection of spatial features, the 2D CNNs was integrated to analyse single-frame spatial features, which cannot capture temporal inconsistencies across video frames. This are often useful cues in deepfake videos. The 1D CNN identify inconsistences in short sequence making it easier and faster for RNN by reducing the sequence length through temporal pooling.

In the hybrid model, preprocessed video frames were first imputed into 2D CNN where individual frames were examined to extract spatial features. Each frame is then compress into feature vector. The vector is then fed to a 1D CNN that extracts temporal features over a short window by applying convolutional filters. This convolutional filters detect short-range and local inconsistences. The temporal pattern from 1D CNN is then analysed by RNN for long-range dependencies across the entire video. The result is passed through a classifier, where the output is classified as “real or fake”. This combination is effective because deepfakes often introduce temporal artifacts that static frame-based models might miss.

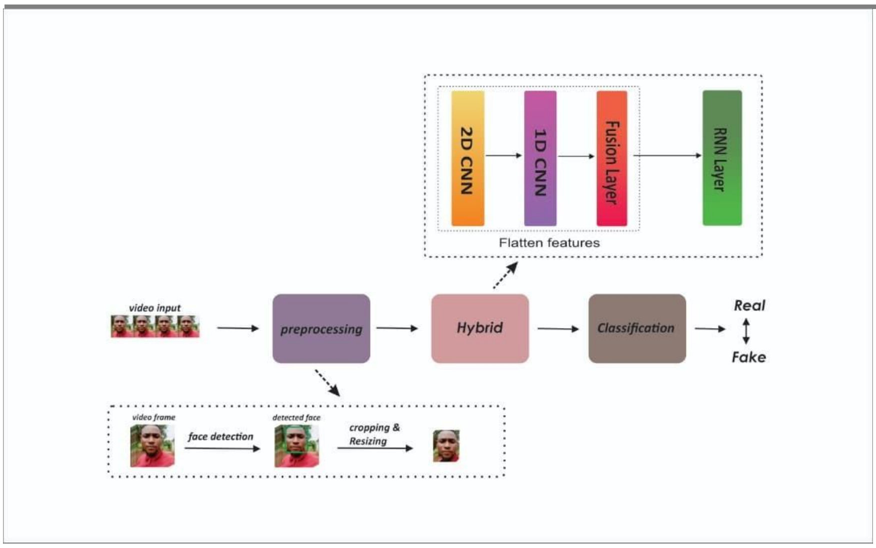

Figure 1: Architecture of the model 4

RESULTS AND DISCUSSION

The model was developed using Jupyter Lab and machine learning libraries such as Numpy, Pandas, Tensorflow, Matplotlib, Hashlib, Sklearn and Seaborn for the preprocessing and training of the hybrid model. The choice of JupyterLab is predicated on its ability to provide an integrated platform combining code execution, efficient data handling, data visualisation, cleaning and documentation. The process facilitates flexible model development and experimentation. The collective libraries support the entire training process in Jupyter Lab from data manipulation it support the model construction, data visualisation and evaluation.

The model was thoroughly trained to obtain the balance between performance and computational efficiency. Training and validation of the model were conducted systematically using data subset obtained from Celeb DF with the training taking placed on the selected training set and valuation of the set for validation; this is to help prevent overfitting and accurate assessment of the model. The model was trained for 100 epoch to monitor peak of the model performance.

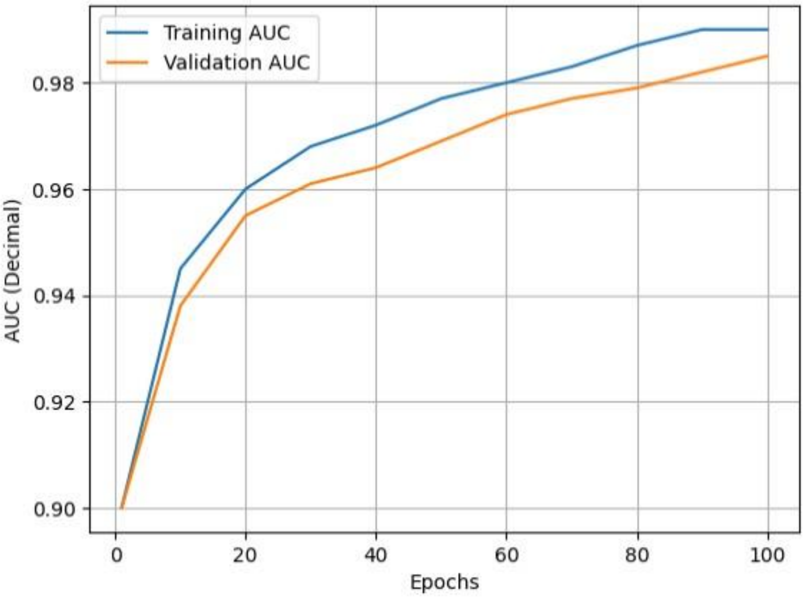

Figure 2: Training and Validation AUC

The model AUC climbs steadily from 0.90 at epoch 1 and reaches its peak at 0.99 for epoch 100. This implies that the model fits better with the training leading to the standardisation of the model at 100 epoch. Figure 2 describes the model AUC with the blue line representing the training AUC sets and the yellow line indicating the validation AUC. The increasing AUC value indicate an excellent discriminative power, which is the ability of the model to distinguish between real and fake videos.

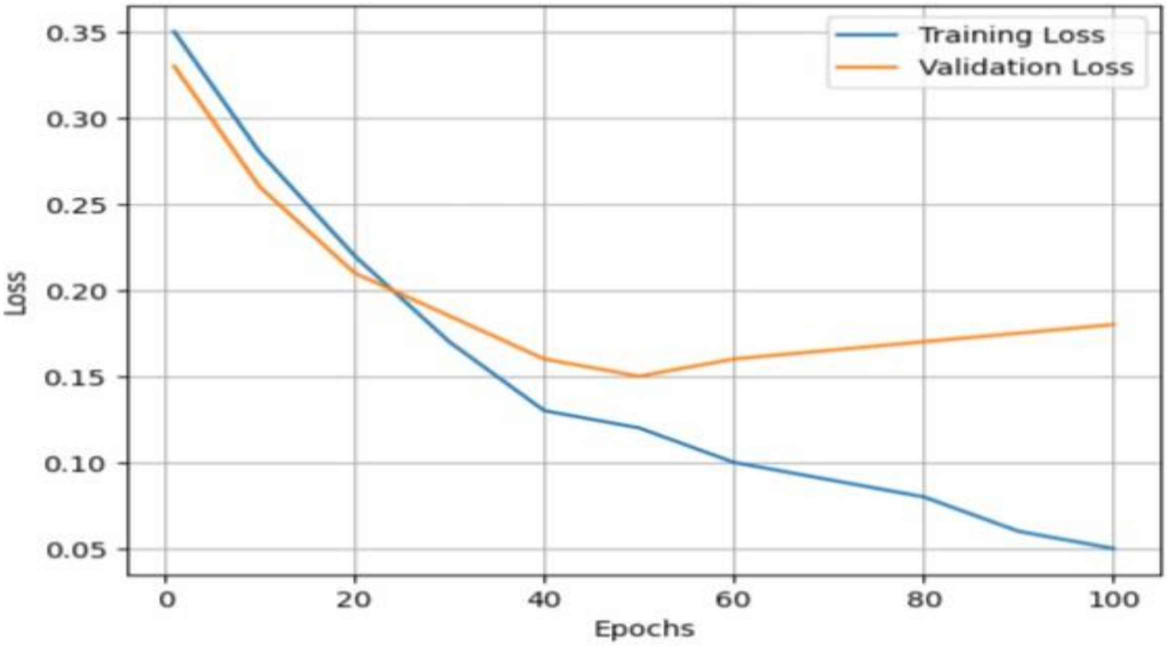

Figure 3: Training and Validation Loss

Figure 3 shows the loss in the training and the validation sets. The result revealed that the value loss of training set at 100 epoch was 0.05, while that of the validation set was 0.18. This indicates that there is no overfitting in model. The value loss indicates how well the training and validation datasets were predicted by the model. Lower value loss is better. The model training loss decrease generally and the training accuracy increases, meaning the model is learning from the data.

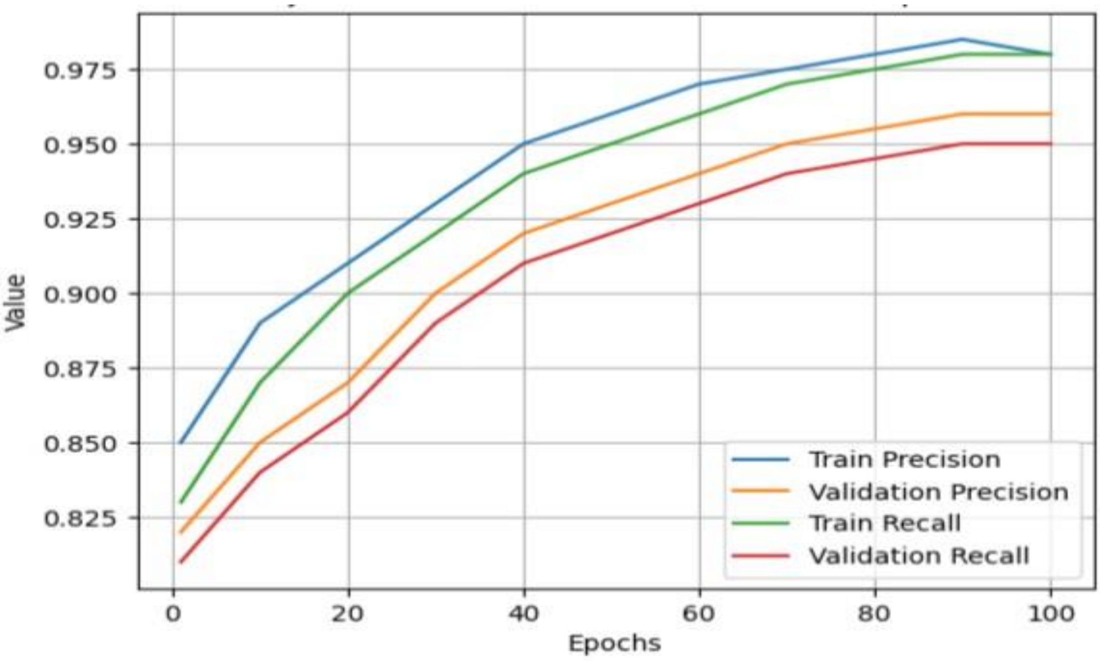

Figure 4: Training and Validation Precision/Recall

Precision measures the accuracy of positive prediction while Recall measures how well the model captures actual positives. Figure 4 shows steady improvement over 100 epochs for both training and validation sets. Training precision and recall reached about 0.98 while validation precision and recall stabilised around 0.95. The close range between training and validation metrics indicates good generalisation and balanced performance, with the model effectively managing false positive and false negatives. Overall, the model demonstrates strong and reliable learning with well-balanced precision and recall reflecting a successful training process and architecture.

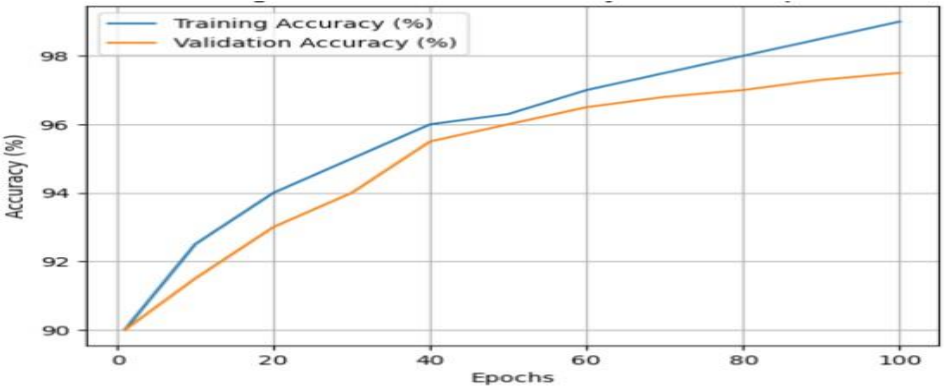

Figure 5: Training and Validation Accuracy

Accuracy is a common metric used to evaluate the performance of a classification model. It measures the proportion of correctly predicted instances out of the total instances evaluated. Figure 5 shows improvement in both training and validation accuracy across 100 epochs. Initially, the model starts with about 90% accuracy on both training and validation data, indicating a reasonable baseline performance from the early stages. As the epochs progress, the training accuracy rises smoothly, eventually reaching approximately 99%. This continuous increase reflects the model's growing ability to learn and fit the training data by effectively capturing the relevant features and patterns. The validation accuracy also improves steadily, increasing from around 90% to nearly 99% by the end of 100 epochs. The close tracking of validation accuracy alongside training accuracy suggests the model generalises well to unseen data and does not merely memorise the training set. The relatively small gap between training and validation accuracy indicates limited overfitting which is a desirable traits because it shows the model predictions are robust and reliable when applied to new examples.

4.1 Model testing

The model was tested on a balanced dataset of 200 videos, consisting of 100 real and 100 fake videos to evaluate the performance of the hybrid deepfake video detection model. The objective was to assess the model's effectiveness in accurately classifying videos as either fake or real and this objective was achieved as the model was able to identify video of fraudulent origin.

Figure 6 shows the confusion matrix of the text data. The result falls into both fake and real.

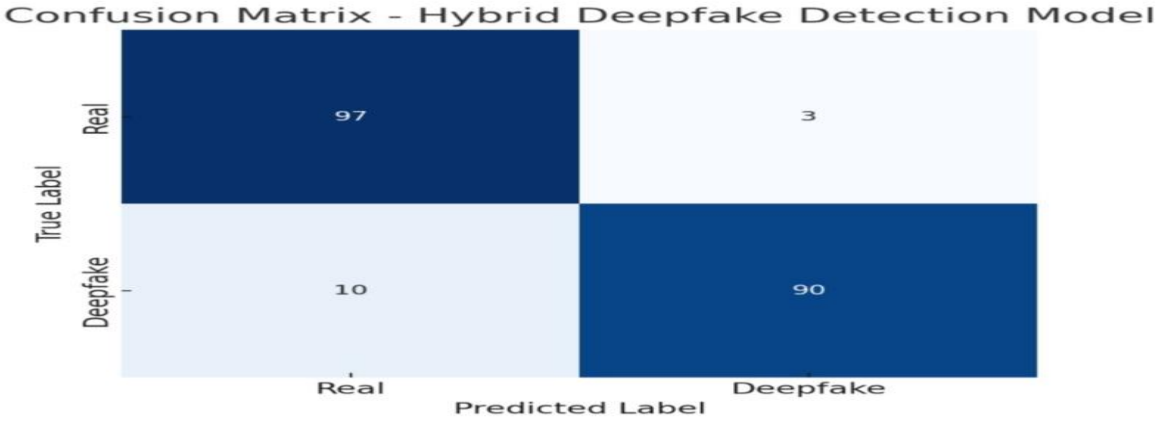

Fig 6: Confusion matrix for test data

The results below show what was found after testing 200 videos of fake and real videos:

True Positive (TP) is the number of cases that the model was able to predict the positive class in respect to the ground truth “Fake” therefore TP are videos that are actually predicted by the model as fake.

TP = 90

True Negative (TN): is a situation where the model predicts a negative outcome correctly. True negative in this context is when the model correctly predict real video as real when it is actually real.

TN = 97

False Positive (FP) is when the model predicts the positive outcome incorrectly. In this context, it is when the model predicts a real video as fake.

FP = 3

False Negative (FN) is when the model predicts a negative outcome incorrectly. In this case, it occurs when the model predicts fake video as real.

FN = 10

The performance of the developed model was evaluated based on different metrics. The model classified 187 out of the 200 test data correctly, thereby achieving an overall accuracy of 93.5%

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \frac{90+97}{90+97+3+10} \quad \frac{187}{200} = \qquad = \quad = 0.935$$

The model achieved a precision of 96.8%, which means that almost all the videos it identified as deepfakes were truly manipulated. However, there were a few false positives. Similarly, the model reached a recall of 90%, which shows that most of the deepfake videos were correctly detected, with only 10% incorrectly classified as real.

$$Recall = \frac{TP}{TP+FN} \quad \frac{90}{90+10} \quad \frac{90}{100} = \qquad = \qquad = 0.90$$
(2)

$$Precision = \frac{TP}{TP+FP} \quad \frac{90}{90+3} \quad \frac{90}{93} == = 0.968 \qquad (3)$$

$$False\ Positve\ Rate = \frac{FP}{FP+TN} \quad \frac{3}{3+97} \quad \frac{3}{100} = \qquad = \ = 0.03 \qquad (4)$$

The F1 Score which is 93.3% shows the balance between precision and recall, this is a proof that the classifier is reliable and can effectively handle false positives and false negatives.

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad \frac{2 \times (0.968 \times 0.9)}{0.968 + 0.90} == 0.933 \qquad (5)$$

With a Specificity of 97%, the model has shown its ability to accurately differentiate between real and fake videos. The balanced accuracy of 93.5% makes it clear that the model’s performance is consistently good across both classes. Hence, ensuring there is no partiality between real and fake videos.

TN

$Specificity = \_\_\_\_ \quad \frac{97}{97+3} \quad \frac{97}{100} \quad = \qquad = \qquad = 0.97$
(6)

$TN+FP$

$Recall(Sensitivity)+Specificity$

$Balanced\ Accuracy = \_\_\_\_\_\_\_\_\_\_\_\_ \qquad (7) \quad 2$

$Balanced\ Accuracy = {}^{0}\frac{90+0.97}{2} \cdot = 0.935 \qquad (8)$

The Matthews Correlation Coefficient (MCC) helps to get a balanced evaluation. The MCC of 0.872 confirms that the model is reliable. Since the MCC is close to 1, it simply means the model performed greatly in predicting real and fake videos

$MCC\overline{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} = \overline{\sqrt{(90+3)(90+10)(97+3)(97+10)}} = 0.872 \quad (TP\times TN)-(FP\times FN) \qquad ({}^{90\times 97})-({}^{3\times 10})$

(10)

Table 4.1 is a summary of the metrics used to evaluate the performance of the model test data and their respective values?

Figure 7 shows the ROC curve. AUC is derived from the ROC, and it is an essential metrics. AUC is an important statistical metrics utilise due to the imbalance of the dataset. AUC denotes the area under the curve of a plot of False Positive Rate versus True Positive Rate at various locations in the interval [0, 1]. As the value increases the model's performance improves. This model's AUC score is 0.935, indicating that it has a 93.5% probability of successfully identifying a fake video.

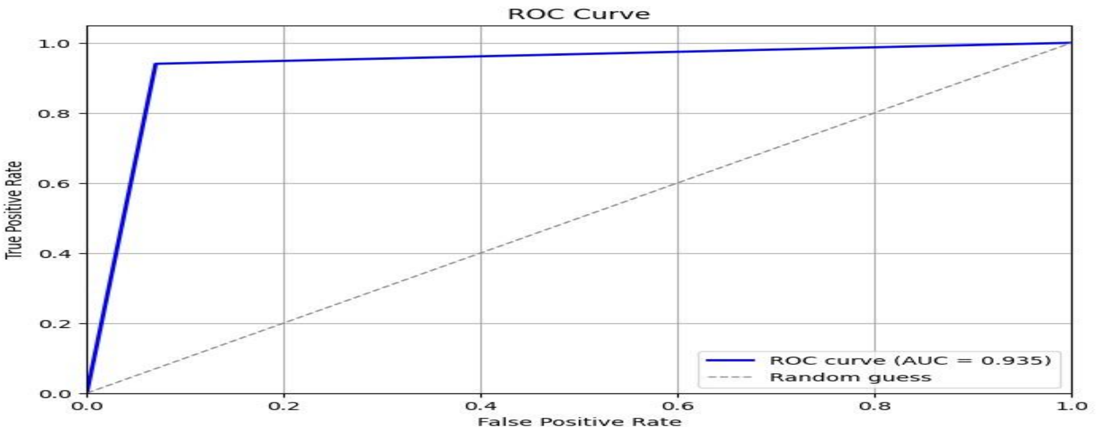

Figure 7: Receiver Operating Characteristic Table 1 Summary of the Model's Performance Metrics

Table 1 Summary of the Model's Performance Metrics

S/No	Metric	Value
1	Accuracy	93.5%
2	Precision	96.8%
3	Recall ivity) (Sensit	90.0%
4	Specificity	97.0%

5	F1 Score	93.3%
6	Balanced Accuracy	93.5%
7	MCC	0.872
8	AUC	0.935

DISCUSSION OF THE FINDINGS

The hybrid model was developed with the intention of capturing spatial and temporal features in deepfakes. Deepfakes often leave a trace of artifacts such as inconsistences in colour, texture, and irregular edges. There are also inconsistences in facial movement and unnatural blinking pattern [26,4]. Considering the training and validation performance as indicated from result of metrics, it can be deduced that the model learned the spatial and temporal features effectively from the dataset (Celeb DF) and there is no evidence of overfitting. The near perfect training AUC and total accuracy reflect the strong classification power of the model to distinguish between real video and deepfakes. The training precision of 98% indicates that the model is reliable in predicting true positive. According to Cuellar [27], precision is a very important forensic metric where false positives undermine confidence score and lead to waste in the investigation resources. The model training and validation precision demonstrate a high ability to minimise false alarm. The high training and validation recall rate shows the ability of the model to detect manipulation in deepfakes while maintaining a high precision. The architectural design, the training and validation metrics align with superiority of hybrid models over CNN or RNN single model. The result of the various metrics shows the superiority of the CNN-2D-RNN hybrid model over individual model and similar hybrid model [20].

The hybrid model achieved an accuracy of 93.5%, precision of 96.8%, recall of 90%, F1 score of 93.3%, MCC of 0.872 and AUC of 0.935 on the test dataset obtained from Celeb DF. These results show that the model performed better than similar CNN, RNN and a CNN-RNN hybrid model reported [20]. The AUC and accuracy value on the test data (Celeb DF) reflect the strong classification power of the model to distinguish between real video and deepfakes. The precision of 96.8% on the test data indicate that the model is reliable in predicting true positive, precision is a very important forensic metric where false positives undermine confidence score and lead to waste in the investigation resources. The model testing precision demonstrates a high ability to minimise false alarm as the false positive rate is only 3.2%. The recall rate shows the ability of the model to detect manipulation in deepfakes while maintaining a high precision. Matthew Correlation Coefficient (MCC) has not been given much attention in the past, it is a strong way to measure binary classification because it incorporates all confusion matrix [12]. It has been argued to be better than F1 and accuracy in terms of imbalance. The F1 score of 93.3% shows a balance between precision and recall. In some cases, high recalls are achieved at the expense of precision and vice versa. The reported MCC of 0.872 shows that the developed model is better at balancing true and false predictions. The use of MCC metric makes the model more reliable than many other methods that have not reported. The model specificity is 97% meaning the correctly classifying 97% of real video. This reduces the risk of labelling real video as fake. Unlike some models, this model maintains a balance between specificity and recall.

The current results show that the developed hybrid CNN / 2D-CNN / RNN model not only outperforms previous CNN-based and hybrid models in accuracy and precision but also offers a stronger basis for real-world deployment in forensic and security-sensitive environments.

5.1 Limitation

This research encountered a few limitations. The research was carried out on a limited dataset using only data from Celeb DF for training validation and testing. The high cost of hardware and lack of funding were critical challenges in carrying out this research.

CONCLUSION

In this research work, we developed, validated and tested an AI-based hybrid model consisting of 2D CNN, 1D CNN and RNN architecture for a reliable and improved cross-platform forensic analysis of deepfakes with high accuracy. The application of CNN component of the hybrid extracts the spatial features frame by frame while the RNN component captures the temporal features, the model was able to capture effectively the temporal inconsistency and subtle artifacts introduced during the synthesis of the deepfakes. The hybrid model performed highly in metrics like accuracy 93.5 %, Precision 96.8% Recall 90 % and MCC 87.2 %. The application of many metrics particularly the MCC demonstrates the balanced and unbiased nature of the hybrid model and this presents the quality and reliability of the model for forensic admissibility. This model outperformed individual CNN and RNN models in adaptability and generality in addressing emerging deepfakes technology. Worthy of note in this research is finding on the importance of involving hybrid architecture in digital forensics for robust and trustworthy results. It can be deduced that this study established a solid foundation for employing hybrid model for digital forensics in solving the security posed by deepfakes.

Future study will focus on testing this hybrid model on other databases and comparing it with other transformer based models.

Funding: No funding was received for this research, as the research was self-funded.

Conflict of Interest: There is no conflict of interest.

Dual Publication: No dual publication submission.

Ethical Approval: Not applicable

REFERENCES

1. Agarwal, H., & Singh, A. (2021). Deepfake detection using svm. In 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) 1245-1249.
2. Reis, P. M. G. I., & Ribeiro, R. O. (2024). A forensic evaluation method for DeepFake detection using DCNN-based facial similarity scores. Forensic Science International, 358, 111747.
3. Chen, H. S., Rouhsedaghat, M., Ghani, H., Hu, S., You, S., & Kuo, C. C. J. (2021). Defakehop: A lightweight high-performance deepfake detector. In 2021 IEEE International conference on Multimedia and Expo (ICME) 1-6.
4. Sunil, R., Mer, P., Diwan, A., Mahadeva, R., & Sharma, A. (2025). Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation. Heliyon e42273.
5. Choi, Y., Uh, Y., Yoo, J., & Ha, J. W. (2020). Stargan v2: Diverse image synthesis for multiple domains.
   In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 81888197).
6. Chen, H., & Magramo, K. (2024). Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’(2024). URL https://www. cnn. com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intlhnk/index. html.
7. Arshad, H., Jantan, A. B., & Abiodun, O. I. (2018). Digital forensics: Review of issues in scientific validation of digital evidence. Journal of Information Processing Systems, 14(2).
8. Sharma, B., Ghawaly, J., McCleary, K., Webb, A. M., & Baggili, I. (2025). ForensicLLM: A local large language model for digital forensics. Forensic Science International: Digital Investigation, 52, 301872.
9. Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analysing leakage of personally identifiable information in language models. In 2023 IEEE Symposium on Security and Privacy (SP) 346-363.
10. Horsman, G., & Dodd, A. (2024). Competence in digital forensics. Forensic Science International: Digital Investigation, 51, 301840.

11. de Rancourt-Raymond, A., & Smaili, N. (2023). The unethical use of deepfakes. Journal of Financial Crime, 30(4), 1066-1077.
12. Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining, 16(1), 4.
13. Viola, M., & Voto, C. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. Synthese, 201(1), 30.
14. Batagelj, B., Kronovšek, A., Štruc, V., & Peer, P. (2025). Robust cross-dataset deepfake detection with multitask self-supervised learning. ICT Express.
15. Kosarkar, U., Sarkarkar, G., & Gedam, S. (2023). Revealing and classification of deepfakes video's images using a customize convolution neural network model. Procedia Computer Science, 218, 2636-2652.
16. Almestekawy, A., Zayed, H. H., & Taha, A. (2024). Deepfake detection: Enhancing performance with spatiotemporal texture and deep learning feature fusion. Egyptian Informatics Journal, 27, 100535.
17. Jafar, M. T., Ababneh, M., Al-Zoube, M., & Elhassan, A. (2020). Forensics and analysis of deepfake videos. In 2020 11th international conference on information and communication systems (ICICS) 053058.
18. Koritala, S. P., Chimata, M., Polavarapu, S. N., Vangapandu, B. S., Gogineni, T. K., & Manikandan, V. M. (2024, June). A Deepfake detection technique using Recurrent Neural Network and EfficientNet. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
19. Suratkar, S., & Kazi, F. (2023). Deep fake video detection using transfer learning approach. Arabian Journal for Science and Engineering, 48(8), 9727-9737.
20. Petmezas, G., Vanian, V., Konstantoudakis, K., Almaloglou, E. E., & Zarpalas, D. (2025). Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification. Multimedia Tools and Applications, 1-20.
21. Sharma, P., Kumar, M., & Sharma, H. K. (2024). GAN-CNN ensemble: a robust deepfake detection model of social media images using minimised catastrophic forgetting and generative replay technique. Procedia Computer Science, 235, 948-960.
22. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3207-3216).
23. Vamsi, V. V. V. N. S., Shet, S. S., Reddy, S. S. M., Rose, S. S., Shetty, S. R., Sathvika, S., ... & Shankar, S. P. (2022). Deepfake detection in digital media forensics. Global Transitions Proceedings, 3(1), 74-79.
24. Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5203-5212).
25. Dong, F., Zou, X., Wang, J., & Liu, X. (2023). Contrastive learning-based general Deepfake detection with multi-scale RGB frequency clues. Journal of King Saud UniversityComputer and Information Sciences, 35(4), 90-99.
26. Sohail, S., Sajjad, S. M., Zafar, A., Iqbal, Z., Muhammad, Z., & Kazim, M. (2025). Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning. Information, 16(4), 270.
27. Cuellar, M. (2024). The Neglected Error: False Negatives and the Case for Validating Eliminations. arXiv preprint arXiv:2412.05398.