International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Hybrid ResNeXt and LSTM Model for Enhanced Deepfake Detection on the FaceForensics++ Dataset

Vaishali Bankar
Research Scholar,
Department of CS and IT,
Dr. Babasaheb Ambedkar
Marathwada University,
Aurangabad, (Maharashtra) India.

Diksha R. Pawar
Research Scholar, Department of CS
and IT, Dr. Babasaheb Ambedkar
Marathwada University,
Aurangabad, (Maharashtra) India.

Pravin L. Yannawar
Professor,
Department of Cs and IT,
Dr. Babasaheb Ambedkar
Marathwada University,
Aurangabad, (Maharashtra) India.

Abstract - Deep-learned technology used to widely spread fast-faking has raised a lot of apprehensions about the possible abuse of this technology to formulate false information and create artificial audio visual imagery. Here, this paper provides a strong architecture of deep-fake video detection, which combines a ResNeXt convolutional neural network of spatial features with a long short-term memory network of temporal inconsistency analysis. The model was trained and tested on the FaceForensics++ dataset and thus, detects minor changes caused during the process of face manipulation. Frames extraction, face detection, crop and resizing of the face were the steps in pre-processing pipeline to normalize the input data. The hybrid model achieved an accuracy of 97.48 % which shows a great deal of accuracy and recall when it comes to genuine and spofer video discrimination. These findings support the effectiveness of CNN and RNN architecture integration to develop better detection performance. Possible future studies will seek to increase generalizability to non-homogenous data as well as add multimodal capabilities of audio streams and biological signals to have an even stronger defence against detection.

Keywords: Deepfake Detection, Neural networks, ResNeXt, LSTM, FaceForensics++.

1. INTRODUCTION

Over recent years the headlong rush in the development of deep learning technologies has seen the creation of very realistic synthetic media, often known as deep fakes. Deep learning has brought about a revolution in the creation of synthetic media, where hyper-realistic AI-generated videos can be created, to manipulate a face, voice, or action. They are mostly based on Generative Adversarial Networks (GANs) and auto encoders that have made tools that can change faces, change speech or simulate a scenario with disturbing realism accessible to anyone. The AI-created videos, including face-swapping or facial re-enactment ones, are such critical issues that they can be misused and used to pass misinformation, develop fake news, and even blackmail people. Although, in the early days, deep fakes were used as entertainment, today, they are a significant threat to society, particularly in the form of political disinformation, financial fraud, and non-consensual explicit content [1]. GANs, that oppose generative and discriminative networks, are better at producing realistic synthetic images, and other auto encoders reduce and re-establish information, and allow facial re-enactment to be seamlessly performed. Applications like Deep Face Lab and Face Swap have been built on top of these networks, making the generation of a deep fake as easy as it can be made by an amateur. Political fake news: doctored speeches of the world leaders, including speeches given by Volodymyr Zelenskyy in the conflict in Ukraine, have been deployed to create division at the expense of unity [3]. Financial fraud: The existence of voice cloning and deep fakes has enabled scams, most recently a 2023 incident where an employee has used a cloned voice to defraud their CEO, convincing staff to use $35 million to hand over [4]. Non-consent pornography: out of the entire deep fakes on the internet, more than 96 % of it is explicit content and it mostly attacks women. As a result, the requirement of well-developed deep fake detection techniques has never been as urgent as ever.

The recognition of deep fakes has been a game of cat and mouse since the use of generative AI is quickly evolving and because of the growing advanced sophistication of artificial media. Stable Diffusion Typical GANs and hybrids also referred to as diffusion models generate high-fidelity forgeries containing fewer artefacts traceable to the model, including irregular lighting or unnatural facial motions that defined previous deep fakes [7]. Also, the advantages of adversarial training allow deep fake generators to learn and avoid the already available detection models, effectively leaving the traditional approaches useless (Nguyen et al., 2023). As an example, the minor discrepancies that can be caused by time, like the eye gaze or micro-expressions being out of sync, are usually removed in newer deep fakes, which makes it uneconomical to analyse frame by frame anymore [8]. Deep fakes can include irregular viewpoints, coverings or substandard inputs. In addition, other ethical and logistical obstacles, such as privacy issues when training data is collected, and the scalability of processing high-, real-time video with high resolution makes the challenge even bigger to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

implement scalable solutions [9]. The given challenges emphasise the importance of flexible, multimodal detection models and joint processes to normalise the measures of evaluation in various settings.

We suggest a deep-learning-based solution, in detecting deep fake videos, with attention to the face forensics plus dataset. We use a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to examine video frames and extract inappropriate artefacts that deep fake generation systems do not remove. Training directly on the Face Forensics++ dataset we will show how well our strategy can succeed even when trained on one, well-curated dataset. The main strength of this work is the creation of a deep fake detector capable of properly differentiating between natural and fake videos based on the frame-level features and time discrepancies. The feature extractor in our model requires a ResNeXt CNN to discover features and a temporal sequence analyser (temporal sequence experiment) based on a Long Short -Term Memory network to detect deep fakes with high precision.

2. LITERATURE REVIEW

Deepfake detection has been a subjective topic of research endeavour in the last few years due to the increased sophistication of synthetic media, and the potential to misuse it. Different methods have been suggested, all of which relate to the process of creating deep fakes in diverse aspects. It is one of the oldest and most powerful approaches that aim to find the artefacts of facial warping that are commonly added to the process of deep fakes generation. The FaceForensics++ dataset by Rossler et al. [2] emerged as a benchmark dataset and has been a standard in Deepfake detection studies. Their work revealed that the tools of Deepfake generations, like Face2Face and Deepfake, leave some insincere artefacts in the nature of irregular facial textures or unnaturally existing deformations. Although these artefacts could not be perceived by human senses, they could be identified with the help of Convolutional Neural Networks (CNNs) that were trained on high-quality datasets. The Figure Forensics package with its images consisting of a large variety of both real and manipulated videos has allowed scholarly community to create and test detection models that are almost perfectly accurate [2], [10].

There is also another interesting method suggested by Li et al., as it was aimed at identifying the inconsistency of eye-blinking patterns within deep fake videos. They employed to process Long-term Recurrent Convolutional Network (LRCN) to remove the changes in time of the blinking behaviour of the eyes, since such behaviour is not natural or is missing in deep fake videos in general [11]. They emphasized the relevance of the time-varying procedure in detecting when content is deceptive using deep fakes because artificial generators of such fakes back then could not duplicate the authentic blinking of a human being [11]. But due to the improvements in deep fake making methods, eye-blinking patterns can no longer be trusted to be enough as a method. It is now possible to produce realistic blinking patterns with modern deep fake methods, including Generative Adversarial Networks (GANs)-based and auto encoders-based methods, and thus decreases the credibility of the methodology [12].

Another potential approach, which has been investigated, is capsule networks when detecting deep fakes. Nguyen et al. have suggested a system that protects against forged pictures and videos by using capsule networks. Capsule networks built to capture hierarchies of relationships among features proved to be promising to detect manipulated media by observing inconsistencies between the facial structure and textures [13]. This, however, was constrained by the necessity to obtain large training data and the difficulty of the network architecture to compute. Also, capsule networks did not generalize effectively to downstream applications of various deep fake generation methods; especially when faced with high-quality fake images generated using state-of-the-art technology [13], [14].

In a more recent writing, scholars have resorted to biological indicators, including heart rate and facial micro-expressions, to identify deep fakes. This paper [15] suggested an algorithm that is an extraction of the biological signals of facial areas in videos by remote photo plethysmography (rPPG). It is a method that tracks the slightest changes to the colour of the skin due to blood flow which is not present in video syntheses. Their method was quite successful in categorising videos as genuine or counterfeit, however, it demanded specific hardware and was computationally expensive, which restrained its application to real-time uses. Nevertheless, the use of biological indicators is an opportune way forward to detect deep fakes, due to its use of biological characteristics that are non-reproducible by the conventional forms of generative models today [10], [15].

Unlike all these approaches, our solution consists in taking the advantages of CNNs and Recurrent Neural Networks (RNNs) to find deep fakes with the help of analysing both spatial and temporal information in video frames. CNNs are very good at crowding spatial information, like facial features, distortions, etc., whereas RNNs are very clear when it comes to finding temporal information that is out of place, like a head twisting unnaturally or a facial expression that does not resemble the one on the original face. With these two architectures mixed together, our model will be able to identify the fake deep images with a lot of accuracy even when the deep plastic early faces a more developed counterfeit. Our model is trained and evaluated using the FaceForensics++ database so that the model can be checked to be robust and applicable to the other deep fake generation methods. Our solution will be to offer

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

a scalable and effective way of detecting deep fakes, such that the current training methods will be overcome, and the performance may be of high quality in the field of real-life applications [13], [17].

Although Deepfake detection has advanced, there are still a number of obstacles. Among the most crucial issues, the quick development of deep fake generation methods should be noted. With new tools and frameworks that include diffusion models and transformer-based models, more realistic forgeries that are less detectable are being generated. Considering the example of diffusion models that create images by progressively adding more details and removing noise, it is possible to create high-quality deep fakes with fewer detectable artefacts. This upward game of developed deep fakes and detectors is forced to develop dynamic and resistant methods of detection to keep up with the technology and improve the game as much as possible [18], [19].

The generalisation of detection models to multiple datasets and realistic situations is another problem. Most of the current models are optimized on such benchmark data as FaceForensics++ or Celeb-DF, which might not comprehensively capture the diversity experienced in practice. As an example, deep fakes produced based on low-quality videos or unusual camera angles can be left without detection because the training data are not diverse. In response to this concern, scholars are investigating multimodal datasets in terms of audio, video, and metadata and transfer-learning methods in order to enhance model generalisation [20], [21].

To summarize, deep fake detection has achieved a lot in recent years, and numerous methods tackle the issue on different levels. Since identifying the artefacts of face warping, the inconsistency of eye blinking, and utilizing biological cues and networks of capsules, the researchers have studied numerous methods of recognizing manipulated media. Nevertheless, the pace at which this technology was developing new tools of deep fake generation and the issue of bias and generalisation of datasets underline the necessity of further advancement in this area. The presence of CNNs and RNNs within our system, that is, the combination of these methods of studying spatial and temporal characteristics, is a forward step towards creating effective and scalable solutions to deep fake detection. With the use of FaceForensics++ dataset and the regardless of the shortcomings of the current techniques, we will make contributions to the current fight to prevent the proliferation of the synthetic media and safeguard the integrity of the digital content [22], [23].

3. METHODOLOGY

3.1 Dataset

In the current study, we exploited the FaceForensics++ corpus, the most used dataset in the area of the deep fake detection studies. The sample size is 2,000 videos, half of whom are authentic samples and the other half are bogus samples [2]. These fabricated instances were created using a repertoire of complex deep-fake generation algorithms, i.e. Face2Face, DeepFake and Neural Textures which capture a variety of approaches to creating believable manipulated media.

Face2Face is a technique that allows for real-time facial re-enactment, Deepfake leverages auto encoders and generative adversarial networks (GANs) to swap faces, and Neural Textures focuses on using texture mapping for face manipulation.

The corpus is carefully divided into training, validation and test groups with 70% of the videos assigned to the training, 15% to the validation and the other 15% to the testing. Such stratification has ensured that one can train, validate and then evaluate the performance of models with regard to generalising to out-of-sample deep-fake content. The heterogeneity of synthesis modalities and visual content makes FaceForensics++ a better testbed to test deep fake detection systems [2], [24].

3.2 Pre-processing

Before training the model, the videos in the FaceForensics++ dataset were pre-processed to ensure that the data could be fed into the deep learning model in a suitable format. The pre-processing steps involved several key stages: video frame extraction, face detection, cropping, resizing, and dataset creation. These steps were essential for extracting the relevant spatial information from each video frame and ensuring consistency across the dataset. Below, we detail each step of the pre-processing pipeline [25].

3.2.1. Video Frame Extraction

The first stage of pre-processing involved dividing all the videos into separate frames, thus converting online video streams into quasi-static images, which can be processed by networks. Since the frame rate of every clip was the same, then we sampled 30 fps. As a result, on a video of length T seconds the total number of frames extracted, N is given by.

$$N=T\times 30$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

For example, a video that is 10 seconds long would yield 300 frames, calculated as:

$$N=10\times30=300$$

This process ensures that the temporal information is preserved by sampling frames at regular intervals throughout the video. This technique is commonly used in video preprocessing for deepfake detection [26], [27].

3.2.2. Face Detection

The next process after extracting the frame was to recognize the area of the face in every frame. Face detection is also essential because it allows the model to focus on the facial regions and in which discriminatory signs of deep-fake detection are most likely to be located. OpenCV library was used and a Haar Cascade classifier that was pre-trained to perform face localisation was utilised. The process searches the image and identifies matching patterns of human faces, marking candidate areas to proceed with analysis further.

Mathematically, the process of detection will look like the process of a sliding window over the picture. A classifier labels each window by giving it a score and when the score becomes greater than some set threshold the window is considered to harbour a face. This can be formalised as having the following relation:

$$f(x,y,w,h)=score(window(x,y,w,h))$$

In which $f(x, y, w, h)$ is the detection function, and (x, y) are the top-left coordinates of the window, and w h its width and height, respectively. The score thus obtained is the probability of the window capturing a face. In case of $f(x, y, w, h)$ being greater than τ, then the region is taken to be a face.

$$If\ f(x,y,w,h)>\tau,\ mark\ as\ face.$$

Other current literature has also advanced the face-detection algorithms to boost the accuracy of algorithms and performance in real applications [27].

3.2.3. Face Cropping

The next thing was to extract the region of interest (ROI) after localisation of faces. We defined a bounding box around the identified face, which is defined by the coordinates of the upper-left and bottom-right points, $(x1, y1)$ and $(x2, y 2)$ respectively. The ROI, the cropped face is therein received as.

$$cropped\_face = frame\ [y1:y2,x1:x2]$$

Where, *cropped_face* is the portion of the frame containing the face. The alignment of the face masks is essential, and it was found that misaligned face masks lower the effectiveness of deep fake detection [27].

3.2.4. Face Resizing

To standardize the input for the deep learning model, we down sampled each cropped facial ROI to a fixed dimension of 112×112 pixels. It is necessary because this standardisation enables repeatable feature extraction over all samples, and that is critical to effective training. The bilinear interpolation was used to perform the resizing operation thus maintaining image fidelity: The resizing operation is done using bilinear interpolation to maintain the quality of the image:

$$resized\_face=resize(cropped\_face,(112,112))$$

Where $(112,112)$ represents the desired output dimensions for each face.

When all the frames were extracted, detected, cropped, and resized, the face images were classified into a new dataset. This was then divided into training, validation and test data which resembled the existing split of the FaceForensics energy: 70% of the samples were used in training, 15% in validation and 15 %in testing. The given approach of partitioning helps to overcome over-fitting and provide sufficient generalisation tests [29]. The ready images were saved in separate files and accompanying annotation files were binary code labels of real (0) or fake (1) provenance of each frame. The binary coding is in harmony with regular binomial classification problems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

3.3 Model Architecture

This paper describes a deep-fake detector, which combines the capabilities of both convolutional neural networks (CNNs) and recurrent neural networks (RNNs). According to this system, a ResNeXt CNN is used to extract spatial features and a Long Short-Term Memory (LSTM) network is used to analyze video sequences at a temporal level. The hybrid model will aim at resolving the spatial and temporal aspect of the deep fake detection hence enhance the accuracy of identification of the manipulated video. The model components will be detailed out below.

3.3.1 ResNeXt CNN

The trained ResNeXt 50 model has been used in the current architecture. ResNeXt-50 architecture consists of fifty convolutional blocks and uses a cardinality of thirty-two with four parallel block paths, and has a fully dimensional configuration of *32 x 4*. This architecture has been proven effective in a body of image-recognition problems due to its computational effectiveness. The ResNeXt50 model produces a 2048 dimensional feature-vector, summarising the high level spatial features, face landmarks, texture details and other salient features which may be distorted in deep- fake videos, which are generated by each input frame [31]. These feature vectors are then fed to LSTM network where they get analysed in a temporal sequence as the ResNeXt model individually analyses each frame [31].

To calculate the feature vector of an image (f) of an image [I] at a given frame one can use the ResNeXt -50 model in the following way:

$$f=ResNeXt(I)$$

Here, the frame size of an image is represented by I and f 2048 refers to the 2048 dimensional output of an architecture comprised of the resNeXt CNN. This vector is the product of the spatial data of the frame and the input to the LSTM to analyse temporal relationships. A pathologically-trained ResNeXt-50 is used to make transfer learning with direct fine-tuning on a deep-fake dataset of moderate size, which contributes to the increased effect of the model on the deep-fake task. Transfer learning is particularly useful in cases of scarce data. Through ResNeXt50, the architecture navigates through the extracting of spatial information relevant at the same time avoiding the chances of overtraining [32]. Recent research revealed that ResNeXt networks are more accurate and use fewer parameters than other conventional CNNs like AlexNet and VGG, especially in combination with data-augmentation method and transfer learning. In addition, a new ResNeXt positions itself as an effective and efficient model in video-analysis, which makes it an apt option in feature extraction in the systems of deep-fake detection [32].

3.3.2 LSTM Network

The second basic element of the model is the LSTM network. Being a subclass of recurrent neural networks (RNNs), it will attempt to reproduce long-term dependencies in sequence data. The videos of deep-faking also show manipulation over time; a face-motion distortion could be the result of the unnaturally fast eye blinking or a badly performed gesture is a marker of manipulation. Thus, LSTM network will be adopted that manipulates the sequence of feature vectors produced by each frame, thus making it possible to recognise such artefacts in time.

In our application, the LSTM network has one layer and 2048 hidden units, this is the same order as the feature vectors produced by the ResNeXt 50 model. The LSTM takes the sequence of 2048-dimensional vectors, which are a frame of a video, and learns the temporal dynamics existing across several frames of a video. This enables the creation of low-level temporal features of abnormalities, including erratic head motion, facial-expression malfunction, or blinking anomalies, which are inherent to deep-fake material. The issue of over-fitting RNNs is mentioned among the major problems when it comes to training such networks because of the large number of parameters. This risk is offset by a dropout rate of 0.4, and LSTM layer stimulates the production of feature robustness that generalises to previously unseen data when testing [33].

The output of the LSTM network is subjected to a fully connected layer containing two output nodes, the classes of the two real and fake. The raw scores are then transformed into class probabilities with the help of a softmax activation function and the choice is made according to the job that is most likely to happen, so that the video is recognized as a real one or a manipulated one. The softmax is the following:

$$P(y = \{k|z\}) = e^{zk} / \sum_{i} e^{zi}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

In which the z refers to the vector of raw outputs of the fully connected layer, and the z k is the score of the class k (either real or fake) [34].

Advantages of Combining ResNeXt and LSTM

The model is successful due to combining the strengths of the two architectures by employing the ResNeXt CNN and LSTM network to combine them. The ResNeXt CNN is very effective in extracting high-level spatial features per frame, it extracts features that can be manipulated in deep-fake videos. In the meantime, the LSTM neural network predicts temporal variations in the video that appear as inconsistencies across frames and represent manipulation. This composite method can result in a deep-learned and effective deep-fake detection system of both spatial and temporal scope of the issue. Moreover, CNNs combined with LSTMs have been demonstrated to be successful in other video-based tasks, e.g., action recognition and video captioning [35]. Empirically, this combination is more successful than method single-models to find temporal anomalies, thus making it an ideal selection in deep-fake detection.

3.4 Training

The model was trained by using an adaptive learning rate optimisation algorithm, the Adam optimizer. Adam is preferred in deep learning regimes due to its efficient speculations and the ability to adjust the learning speed during the learning process. For our model, we set the learning rate to 1e-5 and applied a weight decay of 1e-3. Weight decay is a regularisation technique that penalises large weights and hence promotes generalisation to unseen data in the case of L2 regularisation. A small learning rate makes sure that the model parameters are updated in small steps which helps in achieving stability in the training. The cross entropy loss function was chosen because it can be effective with binary classification problems like deep fake detection. Cross-entropy measures the difference between the predicted class probabilities and the true labels, encouraging the model to output probabilities that match the target distribution.

To avoid overfitting, Overfitting is a common problem in deep learning algorithm training especially when there is limited data, in which case the deep learning model may become too specialised to the training set at the expense of performance when faced with unseen data [33]. We dealt with this by setting the number of epochs to 20 and made sure to use a batch size of four, and by using early stopping to prevent this overfitting from continuing through training if the validation loss was observed to plateau, meaning we would stop training if the loss was no longer decreasing. The optimisation configuration and training protocol took the advantage of data-augmentation techniques to artificially increase the size of the data set by means of random transformations (rotations, flips, scaling). This helped the model learn invariant features, further enhancing its generalization ability. For the performance evaluation, we used the FaceForensics++ test set to test the model. Accuracy, precision, recall and F1-score were calculated to characterise its efficacy in classification. Additionally, we used a confusion matrix to visualize the model's predictions and identify any misclassifications.

4. RESULTS AND DISCUSSION

4.1 Model Performance

The trained deep fake detection model was able to obtain 97.48% overall accuracy for the FaceForensics++ test set, with the results presented in Table 1. This means that the model was successfully able to correctly classify 97.48 percent of the samples in the test, in this case by successfully separating real from manipulated videos. Accuracy is an important precision to test how well the classification is performed, and the high value indicates that the model is very reliable in detecting manipulated content.

Table 1: Model Performance

Metric	Value
Accuracy	97.48%
Precision	97.5%
Recall	97.4%
F1-Score	97.45%

In addition to accuracy, we reported precision, recall and F1 for score. The model's precision is 97.5 percent, which means that 97.5 percent of the videos labelled as "fake" were fake. This low false-positive rate means the model does not miscategorise real videos as fake - a key issue when the model is run in applications where bogus alerts may be a problem. The recall of 97.4 per cent shows that for all actual fake videos, the model found 97.4 per cent of them, and thus reduced false negative. Such a high recall is especially important considering the constant evolution of the techniques of deep fakes, requiring a maximal coverage of detection to ensure

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

that they remain effective. The F1 - score which is a harmonic mean of the precision and recall was 97.45 % indicating well balanced results in both the identification of fake videos and minimising error. These results demonstrate the high effectiveness of our model to detect deep- fakes, even with a single training data set, and attest to the robustness and generalisation ability of our model in terms of various evaluation metrics.

4.2 Confusion Matrix

The confusion matrix shown in Table 2 shows a specific evaluation of the model performance. The model was able to correctly distinguish between real videos and fake videos with 974 and 975 videos, respectively. These results are the figures for true positives (TP) and true negatives (TN) respectively, which demonstrates the strong power of this model to correctly classify both the actual and the fake videos. Nevertheless, there are some misclassifications: i.e., the model labelled 26 real videos as fake (false positives, FP) and 25 fake videos as real (false negatives, FN). These errors were mostly caused by difficulty points like low quality video frames or low lighting conditions, which make it hard to detect subtle artifacts of a deepfake. Despite these inaccuracies, the model did perform overall excellently, with only minor misclassifications in both classes.

Table 2: *Confusion Matrix for the Test Set*

	Predicted Real	**Predicted Fake**
Actual Real	974	26
Actual Fake	25	975

4.3 Comparison with Existing Methods

Our model's performance is on par with current deepfake detection techniques. Rossler et al. [2] achieved an accuracy of 96.5% in the FaceForensics++ dataset with a CNN-based method, while the accuracy achieved with our model is larger, achieving an accuracy of 97.48%, and thus proving the value of using CNNs and RNNs to detect deepfakes.

4.4 Limitations

While our model works very well on the FaceForensics++ data set, it has several limitations. Firstly, the model was trained and evaluated using a single dataset which may limit its ability to generalize to other datasets. Secondly, the model is based on face detection which could fail in videos where faces are partially occluded or are not clearly visible. Finally, the performance of the model can be degraded when it is applied to videos that have low resolution or poor lighting conditions.

5. CONCLUSION

In this paper, we presented a deep learning-based method for detecting the deepfake videos using the FaceForensics++ dataset. Our model combines a ResNeXt CNN for extracting features candidate and an LSTM network for analyzing temporal sequence, and accuracy is 97.48% on the test set! The results show that our approach is very effective in the detection of deepfake videos while only being trained with one news-corporation data-set. Future work will include improving the generalisability of the model by training on different datasets and by adding extra modalities such as audio analysis and biological signals, as well as exploring more advanced architectures such as transformers, to further improve the performance of the model.

6. REFERENCES

[1] [1] Chesney, R., & Citron, D. (2019). Deepfakes and the New Disinformation War. Foreign Affairs.
[2] [2] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11).
[3] [3] Giacomello, A. (2025). *Technological Deceptions: An Analysis of How Deepfakes are Framed in the Press*. Charles University, Prague.
[4] [4] Sherman, J. (2025). A FEAST OF FRAUD: HOW INTERNATIONAL HESITATIONS TO REGULATE DEEPFAKES ARE CREATING A BUFFET FOR FINANCIAL CRIMINALS. *The George Washington International Law Review*, *56*(1/2), 91-117.
[5] [5] Pawar, D. R., & Yannawar, P. (2023, July). Recent advances in audio-visual speech recognition: Deep learning perspective. In *First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022)* (pp. 409-421). Atlantis Press.
[6] [6] Pawar, D., Borde, P., & Yannawar, P. (2024). Generating dynamic lip-syncing using target audio in a multimedia environment. *Natural Language Processing Journal*, *8*, 100084.
[7] [7] Bhattacharyya, C., Wang, H., Zhang, F., Kim, S., & Zhu, X. (2024). Diffusion deepfake. *arXiv preprint arXiv:2404.01579*.
[8] [8] Passos, L. A., Jodas, D., Costa, K. A., Souza Júnior, L. A., Rodrigues, D., Del Ser, J., ... & Papa, J. P. (2024). A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, *41*(8), e13570.
[9] [9] Deng, J., Lin, C., Zhao, Z., Liu, S., Wang, Q., & Shen, C. (2024). A survey of defenses against ai-generated visual media: Detection, disruption, and authentication. *arXiv preprint arXiv:2407.10575*.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

[10] [10] Bankar, V., Pawar, D. R., Yannawar, P. L., & Sambhajinagar, C. Review on Unmasking Deepfake Technology:“Challenges and Solutions for Detection”.
[11] [11] Li, Y., Chang, M. C., & Lyu, S. (2018, December). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)* (pp. 1-7). Ieee.
[12] [12] Yasur, L., Frankovits, G., Grabovski, F. M., & Mirsky, Y. (2023, July). Deepfake captcha: A method for preventing fake calls. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security* (pp. 608-622).
[13] [13] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019, May). Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2307-2311). IEEE.
[14] [14] Roth, T., Gao, Y., Abuadbba, A., Nepal, S., & Liu, W. (2024). Token-modification adversarial attacks for natural language processing: A survey. *AI Communications*, *37*(4), 655-676.
[15] [15] Ciftci, U. A., Demir, I., & Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*.
[16] [16] Wu, J., Zhu, Y., Jiang, X., Liu, Y., & Lin, J. (2024). Local attention and long-distance interaction of rPPG for deepfake detection. *The Visual Computer*, *40*(2), 1083-1094.
[17] [17] Sabir, Ekraam, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. "Recurrent convolutional strategies for face manipulation detection in videos." *Interfaces (GUI)* 3, no. 1 (2019): 80-87.
[18] [18] Tavassi, E. T. Generating Deepfakes with Stable Diffusion, ControlNet and LoRA.
[19] [19] Cardenuto, J. P., Yang, J., Padilha, R., Wan, R., Moreira, D., Li, H., ... & Rocha, A. (2023). The age of synthetic realities: Challenges and opportunities. *APSIPA Transactions on Signal and Information Processing*, *12*(1).
[20] [20] Vyas, K., Pareek, P., Jayaswal, R., & Patil, S. (2024). Analysing the landscape of deep fake detection: A survey. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(11s), 40-55.
[21] [21] Khalid, F., Javed, A., Ilyas, H., & Irtaza, A. (2023). DFGNN: An interpretable and generalized graph neural network for deepfakes detection. *Expert Systems with Applications*, *222*, 119843.
[22] [22] Padmashree, G., & Karunakar, A. K. (2023). Ensemble of Machine Learning Classifiers for Detecting Deepfake Videos using Deep Feature. *International Journal of Computer Science*, *50*(4).
[23] [23] Hong, K., & Du, X. (2021). Self-supervised deepfake detection by discovering artifact discrepancies. In *CEUR Workshop Proceedings*.
[24] [24] Brodarič, M., Štruc, V., & Peer, P. (2024). Cross-dataset deepfake detection: evaluating the generalization capabilities of modern deepfake detectors. In *Proceedings of the 27th Computer Vision Winter Workshop (CVWW 2024). Slovensko društvo za razpoznavanje vzorcev= Slovenian Pattern Recognition Society* (pp. 47-56).
[25] [25] Castell, V. M. (2024). Deep Fake Detection: Evaluation of Several Face Forgery Detectors.
[26] [26] Waseem, S., Bakar, S. A. R. S. A., Ahmed, B. A., Omar, Z., & Eisa, T. A. E. (2023). DeepFake on face and expression swap: A review. *IEEE Access*, *11*, 117865-117906.
[27] [27] Jadhav, A., Patange, A., Patel, J., Patil, H., & Mahajan, M. (2020). Deepfake video detection using neural networks. IJSRD-Int. *J. Sci. Res. Dev*, *8*(1).
[28] [28] Kaur, S., Kumar, P., & Kumaraguru, P. (2020). Deepfakes: temporal sequential analysis to detect face-swapped video clips using convolutional long short-term memory. *Journal of electronic imaging*, *29*(3), 033013.
[29] [29] Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., & Wang, X. (2018). Deep learning for pixel-level image fusion: Recent advances and future prospects. *Information fusion*, *42*, 158-173.
[30] [30] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).
[31] [31] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
[32] [32] Cilivery, N. (2024). Deepfake Detection Using LSTM and RESNEXT50.
[33] [33] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.
[34] [34] Gao, Bolin, and Lacra Pavel. "On the properties of the softmax function with application in game theory and reinforcement learning." *arXiv preprint arXiv:1704.00805* (2017).
[35] [35] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).