Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal

Vol. 11 No. 2, 2025

# Development of a Dual CNN-LSTM Model for The Detection of Deepfake Video Cyber Phishing Attacks

**Akinwumi D.A., Orojo, T. E., Akingbesote. B.O. & Aliyu, E. O.**
[1]Dept. of Cyber Security; [2&4]Dept. Computer Science; [3]Dept. Data Science
Adekunle Ajasin University
Akungba-Akoko, Nigeria
**E-mails**: david.akinwumi@aaua.edu.ng, temitopeorojojr@gmail.com, bakingbesote@gmail.com, olubunmi.aliyu@aaua.edu.ng

## ABSTRACT

Deepfake technology has introduced new complexities into cyber-phishing attacks by enabling cybercriminals to create authentic-looking and artificial intelligence altered videos to impersonate trusted individuals. These synthetic videos can deceive users more effectively than traditional phishing techniques, which rely on text or email phishing attacks. As such, existing cybersecurity mechanisms, primarily designed to detect textual or URL-based threats, fall far short when it comes to video-based deception detection. This research developed a dual CNN-LSTM model for the detection of deepfake video cyber phishing attacks. To achieve this, a dataset of 5,000 labeled videos were collected from the Cyber Institute of Atlanta, USA. The ResNeXt CNN were applied to extract the features, capturing spatial patterns within the extracted frames. The dataset was splitted into 70% training and 30% test sets. The model was trained using a merged benchmark dataset of both the Deepfake Detection Challenge and the FaceForensics++ and a dropout rate of 0.4 was applied to mitigate overfitting. The experimental results showed that the model had 99.81% detection accuracy, indicating a high overall correctness in its predictions. The precision score of 99.73% reflects the model's strong ability to correctly identify deepfake phishing content without misclassifying genuine videos. The recall rate of 99.87% confirms that the model effectively detects the majority of deepfake instances, minimizing false negatives. The combined performance is further substantiated by an F1-Score of 99.79%, which balances the trade-off between precision and recall, showcasing the robustness of the model. The results demonstrated that the proposed model had high reliability when differentiating between real and altered video content. The model is therefore recommended for the detection of deepfake videos to prevent phishing attacks in a computer network environment.

**Keywords:** Deepfake Technology, Security, Phishing Attacks, Artificial Intelligence, Convolutional Neural Networks, Long Short-Term Memory.

**Aims Research Journal Reference Format:**
Akinwumi D.A., Orojo, T. E., Akingbesote. B.O. & Aliyu, E. O. (2025): Development of a Dual CNN-LSTM Model for The Detection of Deepfake Video Cyber Phishing Attacks. Advances in Multidisciplinary Research Journal. Vol. 11 No. 2, Pp 57-67
www.isteams.net/aimsjournal. dx.doi.org/10.22624/AIMS/V11N2P5

## 1. INTRODUCTION

Over the years, cyber-phishing attacks have changed from simple email forgery to more complex attacks, like video impersonation. As technology gets better, so do the tools that cybercriminals can access, which makes phishing attacks more successful. Now, deepfake cyber-phishing is a global security challenge (Murugesan, 2025). A deepfake is a fake video where someone's appearance or voice is changed to mimic another person for bad purposes. Deepfakes are a recent advance in artificial intelligence (AI) related to forgery and they have caused worry in digital communication and security (Goodfellow et al., 2014).

This type of forgery is often done using Generative Adversarial Networks (GANs), which can make very realistic faces and voices. Because it is easier to create deepfakes, cybercriminals are using them more often for phishing attacks that take advantage of people's trust in visual communication (Nguyen et al., 2020). Many cybercrime groups use deepfakes to create fake faces of people, such as CEOs, public figures or family members to make victims divulge important information or send money (Chandrasekaran et al., 2021).

Currently, many deepfake attacks use video content sent via emails, online calls, or messaging apps. These attacks often try to make the victim think they are interacting with someone they know. Unlike text-based phishing, which depends on grammar and suspicious URLs, video deepfakes are visually convincing and harder to confirm. The problem with video deepfakes is not just technical but also psychological. Victims may find it hard to doubt something that looks like live video (Korshunov & Marcel, 2020). Deepfake phishing poses a major threat to business operations, cybersecurity, and trust in institutions. We are already seeing the consequences of deepfake phishing, as phishing attempts, data breaches, and damage to reputation rise in both public and private sectors. As deepfake attacks become more common and sophisticated, there is a need for a smart model that can analyze video content and detect subtle signs of tampering (Tolosana et al., 2020).

To address this problem, some researchers have contributed valuable solutions. For instance, Handrasekaran et al. (2021) worked with temporal modeling. Mittal et al. (2022) used frame-based analysis, and Verdoliva (2020) suggested an attention mechanism to fight deepfake attacks. Other researchers have tried different approaches like convolutional networks (Zhou et al., 2018), face landmark detection (Li et al., 2018), and identifying artifacts in images (Matern et al., 2019). As Nguyen et al. (2020) pointed out, many current deepfake detection systems only spot inconsistencies in single images and don't look at changes over time. Some methods do offer good detection by checking many visual details in each frame. While these authors have created or modified some models that have helped improve deepfake detection research, spotting phishing-related changes to video using both spatial and time-based clues is still a challenge (Ganiyusufoglu et al., 2020). This paper solves this by creating a dual CNN-LSTM model to detect deepfake video cyber phishing attacks.

The structure of this paper is as follows Chapter two discusses the existing research on deepfake detection and phishing attacks. In Chapter three, the designed hybrid deep learning model is discussed and chapter four discusses on the experimental results. Chapter five concludes the paper with a summary and suggestions for future work.

## 2. LITERATURE REVIEW

Deepfakes have emerged as a pressing concern for modern day-to-day life in the digital world because of their growing use in the field of fraud, impersonation, and disinformation (Nguyen et al., 2020). A deepfake refers to a type of synthetic media that often employs deep learning methods such as Generative Adversarial Networks (GANs), which can produce very convincing fake audio, video, or image content (Goodfellow et al., 2014). The growing abuse of deepfakes as a means of choice for phishing-based cyber-attacks is of great concern. This has given attackers to make malicious use of manipulated videos to impersonate trusted individuals for fraudulent purposes (Chandrasekaran et al., 2021). Studies revealed that the degree of trust given to video and voice makes victims vulnerable to being misled, especially when deepfakes become part of an email or social engineering campaigns.

To address this challenge, many researchers have proposed some methods in the area of Deep learning. Deep learning is one of the most promising technologies for media forensics, especially for the analysis of deepfakes. Some Deep learning techniques include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Convolutional Neural Networks tends to get applied to extracting spatial features from images or frames of video, whereas Recurrent Neural Networks (RNNs) come in handy for modeling sequential patterns over time (Nguyen et al., 2021). Both CNN and RNN structures have become popular in deepfakes classification because they can learn both motion-based and appearance-based features. Though some of these methods use only static frame analysis, newer method like, 3D Convolutional Neural Networks (3D-CNNs) use spatiotemporal learning for enhanced performance. However, such methods tend not to be specific enough for phishing contexts or generalize poorly under practical conditions (Tran et al., 2015). A well-structured hybrid model can examine inconsistencies of facial movement, lip sync, and eye blinking, all of which have the potential for being weaknesses of deepfake videos for the purpose of impersonation.

Some of the researchers have attempted deepfake detection using several deep learning methods, for example, Bayar and Stamm (2016) created a universal image forgery detector with CNNs that employed a special constrained convolutional layer for the purpose of enhancing manipulation artifacts. The model proved to have robust detection properties for a variety of image forgeries but had limited generalization with the use of particular datasets. Also, its viability for video-based attacks wasn't addressed, which made its usage in the case of real-time phishing detection less effective. In Zhou et al. (2018), the authors designed a two-stream Faster R-CNN structure for detecting tampered image areas by fusing the RGB and noise stream inputs. The model produced state-of-the-art performance on image datasets and was found to resist compression artifacts. Nevertheless, it dealt only with static images and didn't expand its scope to detecting manipulations in video or dynamic phishing attacks.

The work of Afchar et al. (2018) presented MesoNet, a lightweight CNN model that specializes in the detection of deepfake and Face2Face manipulations. MesoNet worked well on uncompressed video but lacked temporal modeling and thus performed poorly when facing forgeries that preserve frame-level appearance but diverge in motion patterns. Its performance dropped remarkably when faced with compression or tested on unseen deepfake styles. The work of Zhou et al. (2019) improved face forgery detection by employing CNNs that were trained on face image databases. They enhanced tampering accuracy by letting the network learn features particular to the facial areas. However, their image-based classification constrained the relevance of their model for video-based manipulations, for which spatial and temporal information play a prominent role. Qais et al. (2020) dealt with detecting deepfake audio using CNNs that were trained using the spectral audio features of the spectral centroid and zero-crossing rate. The model worked effectively for voice forgeries and emphasized the significance of audio features for identifying synthetic speech. However, it didn't deal with deepfake video forgeries or their phishing aspects.

Oscar de Lima et al. (2021) created spatiotemporal models such as I3D and R3D for the detection of deepfakes from video. These models integrated the analysis of movement and performed better than one-frame detectors on the Celeb-DF v2 benchmark. They were trained on one data set and performed poorly when tested on new or unseen manipulations, a shortcoming that compromises real-world reliability. Almars (2021) surveyed the deepfake detection methods and identified typical issues, including limited diversity of datasets and poor scalability. The author observed that majority of the methods do not generalize across the categories of manipulations and media sources, especially phishing. The generalization issues arise from the lack of widely standardized, cross-domain datasets.

Hamza et al. (2022) employed deep learning and machine learning-based models for detecting forgery in audio. With the use of MFCCs and the VGG-16, their model had a high accuracy but had varied performance across audio subsets. While their work promoted audio forgery detection, they did not provide a method of detecting embedded deepfakes in a video. Medical deepfakes with an EfficientNet-V2-powered model called Med NetAlbahli et al. (2023). While performance on the CT-GAN datasets was high, their efforts were limited to a specific domain and did not address the general risks of deepfake impersonations across communication channels, such as email or video conferencing.

Fathima et al. (2024) presented a CNN-based method based on Mel Spectrograms to detect deepfake audio. The method worked for synthetic voice datasets but not for multimodal or video situations, thereby not being useful for face and voice-based phishing attacks. Arshed et al. (2024) investigated the use of Vision Transformers (ViTs) and CNNs for detecting forged medical images to detect insurance fraud. While the model worked well with diffusion-rendered images, its adaptability was limited, and the amount of training data needed was substantial, as with most deep learning-based deepfake detection methods.

While all these authors have applied deep learning approaches for detecting manipulated media, the specific challenge of identifying deepfake content used in phishing-based video attacks remains insufficiently addressed. For instance, in Oscar de Lima et al. (2021) and Fathima et al. (2024), the authors demonstrated the potential of spatial or audio-based detection models. However, many fall short in adapting to novel, real-world manipulation strategies that were not captured during training. Furthermore, the absence of a robust model that simultaneously captures both spatial and temporal inconsistencies in deepfake phishing videos presents a major research gap. This paper addresses the gap by developing a dual CNN-LSTM model for the detection of deepfake video cyber phishing attacks.

## 3. DESIGN OF THE PROPOSED DEEPFAKE CYBER PHISHING ATTACKS MODEL

The proposed Dual CNN-LSTM model's architecture consists of several linked modules: the dataset, data pre-processing, feature selection, model training and testing, and the classification module. Figure 1 shows a conceptual diagram of the model. The dataset includes both real and deepfake videos, which are used to train and test the models. The goal is to distinguish real videos from manipulated ones. The pre-processing module includes frame capture, face detection, cropping, and video reconstruction. For details on these steps, see (Rössler, 2019) and (Dolhansky, 2020). The next module, feature extraction, is the crux of our work and is discussed in section 3.1.

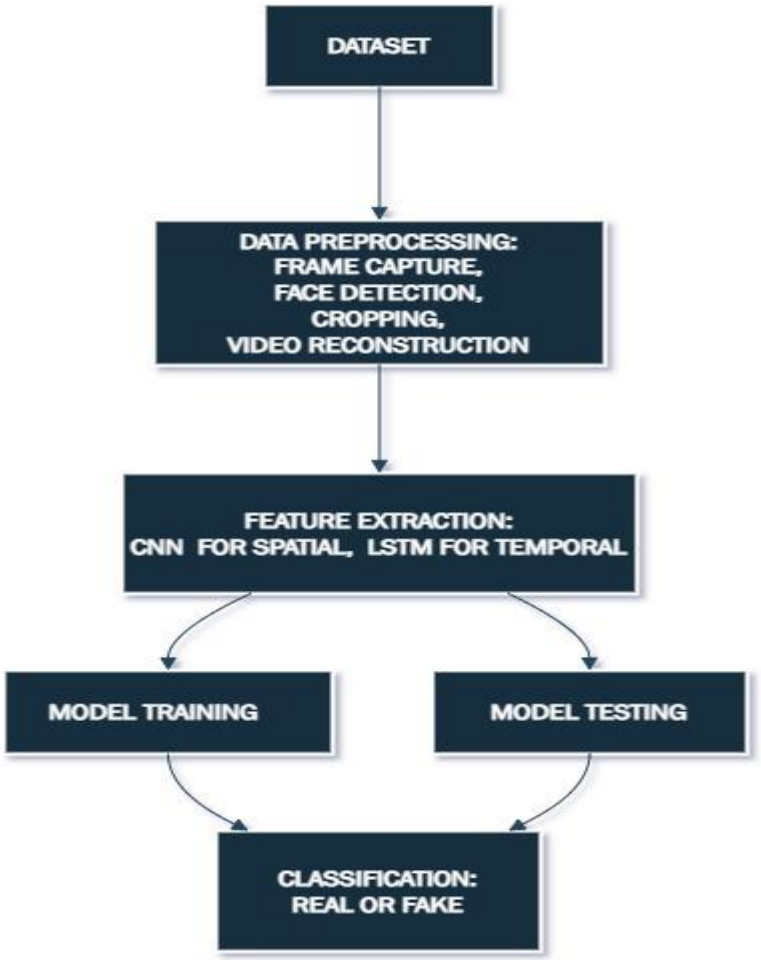



Figure 1: Architecture of the Proposed Deepfake Cyber-Phishing Attacks Detection Model

### 3.1. Dual CNN-LSTM model for Feature Extration

As shown in Figure 2, we extracted features with dual Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. The CNNs captured spatial aspects, like strange mixing or facial landmark distortions. The LSTM networks then took these feature maps and learned the order of movements and frame changes. This setup let the system detect strange facial movements and transition issues, which are common in deepfakes.

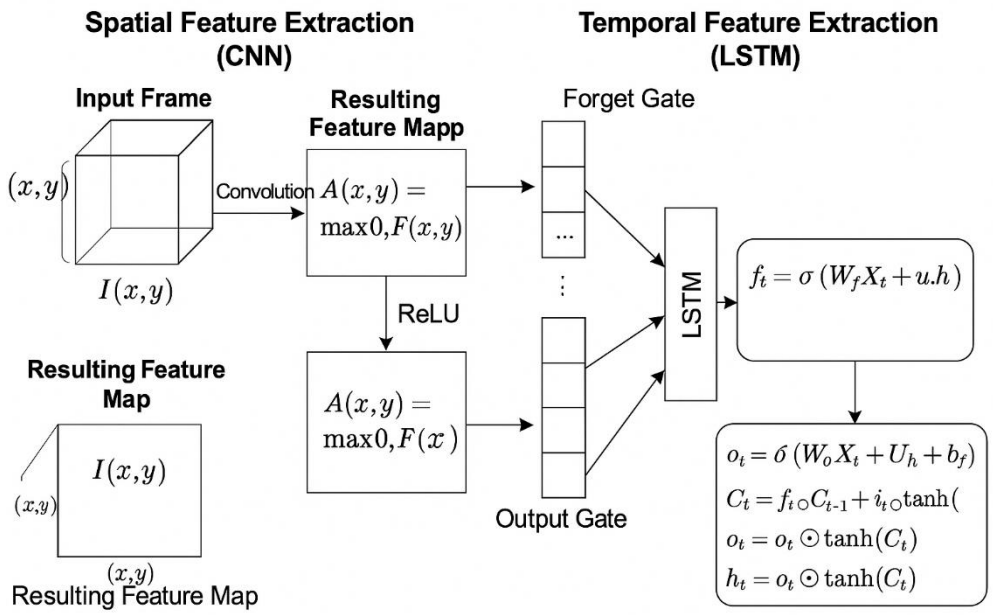



Figure 2: Architecture of the Dual CNN-LSTM model

The FaceForensics++ and Deepfake Detection Challenge (DFDC) datasets were used for this model, as shown in Figure 3. These datasets were chosen because they have a lot of different real and fake videos. They also contain different face swaps and facial reenactments that mimic typical deepfake conditions, as Rössler (2019) noted. The varied lighting, ethnicities, expressions, and backgrounds in these datasets support a detection model that works in different manipulation situations. The datasets include labeled real and fake video samples. The videos were preprocessed by extracting frames, detecting and cropping faces, and reconstructing the videos. This made sure the input data for the learning model was consistent and high-quality. Face regions are chosen from the video frames because most deepfake techniques change facial features. This lets spatiotemporal inconsistencies caused by forgeries be learned by the model.

This paper looks at phishing using fake videos that impersonate people to steal data or trick victims. The fake videos in our dataset are digitally doctored to look like other individual, which makes it easy for attackers to impersonate people. The experiment to spot these deepfakes was built to handle tasks that use a lot of resources, like picking video frames, finding faces, prepping data, and training complex machine learning models. Because processing 5000 videos and training these models takes a lot of power, the cloud platform with GPUs were adopted.

In addition, Google Colab Pro was selected as the experimental platform because of its strong computing power, adaptable Python language support, and compatibility with key deep learning frameworks. The system was set up with a Tesla T4 GPU and high-speed VRAM to speed up calculations and make model training faster. Python 3.11 was employed to build the system, enabling active execution and real-time visualization throughout the entire process for developing the models. During training, 70% of the dataset was used to learn the model's parameters, and the remaining 30% was used to test the model's accuracy. Convolutional Neural Networks (CNNs) were used to extract features for the spatial domain from the video frames, and the resulting sequences were fed into a Long Short-Term Memory (LSTM) network to learn the temporal patterns. The combined structure used spatial and temporal information to differentiate between real and fake videos.

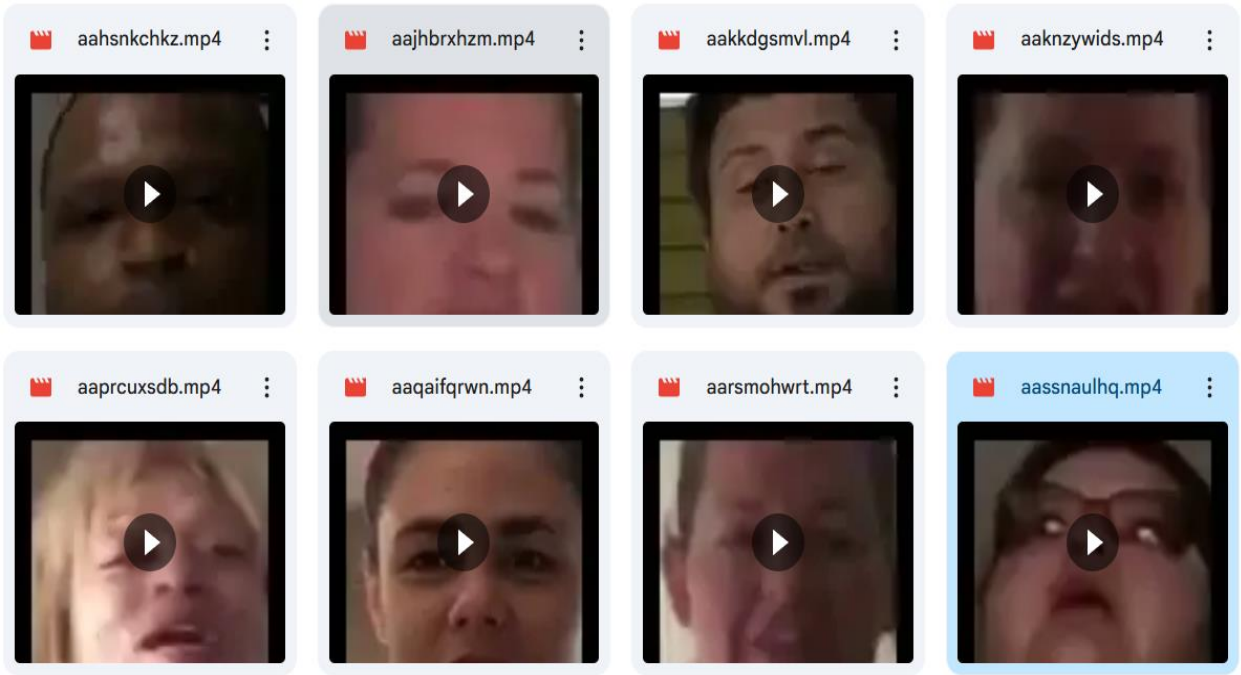



Figure 3: Sample of Deepfake Dataset

The trained model was tested against a video test set for validation. To assess how well the deepfake detection models worked, several evaluation measures were used. These metrics provide different perspectives on the model's accuracy, robustness, and ability to correctly classify both real and manipulated videos. For example, accuracy showed how many samples (both real and fake) were correctly identified out of the total number of samples, showing the model's overall reliability. Precision measured the number of correctly identified deepfakes, showing how often it made right positive predictions. Higher precision means fewer false alarms. Recall is the ratio of true positive predictions to the total actual positives, and it evaluates the model's skill in spotting real deepfakes. The F1-score balances precision and recall for a more trustworthy assessment. The trained model was made to be reliable when used in real-time, given that deepfake attacks are growing and becoming more complex.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositve + TrueNegative + FalsePositive + FalseNegative}$$

$$Precision = \frac{TruePositvie}{(TruePositive + FalsePositive)}$$

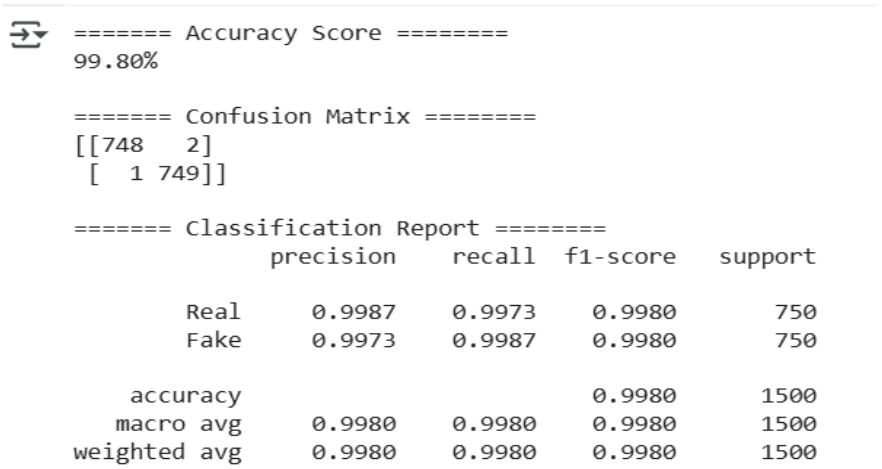
$$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)}$$

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

These metrics together show a full picture of how well the model works and how good it is at detecting deepfake video cyber phishing attacks.

## 4. RESULTS AND DISCUSSION

As earlier mentioned, the model was trained using 70% of the complete dataset and then tested on the remaining 30% (1,500 videos) to see how well it worked. The dual CNN and LSTM model reached 99.81% accuracy, which means it was generally correct in its predictions. Its precision score was 99.73%, showing it could identify deepfake phishing content well without incorrectly labeling real videos. At 99.87%, the recall rate confirms that the model was able to spot most deepfakes, reducing the number of false negatives. Finally, the F1-Score was 99.79%, meaning the model balanced precision and recall well and was good at detecting deepfake cyber phishing videos, as shown in Figure 4.



```
======= Accuracy Score ========
99.80%

======= Confusion Matrix ========
[[748   2]
 [  1 749]]

======= Classification Report ========
              precision    recall  f1-score   support

        Real     0.9987    0.9973    0.9980       750
        Fake     0.9973    0.9987    0.9980       750

    accuracy                         0.9980      1500
   macro avg     0.9980    0.9980    0.9980      1500
weighted avg     0.9980    0.9980    0.9980      1500
```

Figure 4: Output of the Test Dataset

To evaluate the dual CNN-LSTM model, we compared its performance to that of baseline models, as shown in Figure 5. The baseline models, CNN, LSTM, and XceptionNet Detector, are commonly used in deepfake detection and video analysis. These baselines show different ways to approach the problem, letting us assess our model's strengths. We trained and tested the models using the preprocessed DFDC and FaceForensics++ datasets. Table 1 and Figure 5 show the key performance metrics for each model, including accuracy, precision, recall, and F1-score.

Table 1: Performance Comparison of Deepfake Detection Models

| Model | Accuracy (%) | Precision (%) | \| Recall (%) | F1-score (%) |
|---|---|---|---|---|
| CNN-only Baseline | 82.5 | 80.1 | 85.3 | 82.6 |
| LSTM-only Baseline | 78.9 | 76.5 | 81.2 | 78.8 |
| XceptionNet-based Detector | 88.2 | 87.5 | 89.0 | 88.2 |
| Dual CNN-LSTM Model (Proposed) | 99.81 | 99.73 | 99.87 | 99.79 |

In the Table 1 and Figure 5 earlier mentioned, the dual CNN-LSTM model does better than the baseline models in every metric we looked at. It has the highest accuracy, at 99.81%, which means it's really good at telling real videos from deepfakes. Its high precision of 99.73% means it doesn't often flag real videos as fake ones, which is important because false alarms can be a problem in situations like cyber phishing. The model also has good recall of 99.87%, meaning it finds most of the actual deepfakes, lowering the chance that threats will go unnoticed. The F1-score, 99.79%, backs up the idea that the model does a good job balancing precision and recall, which matters when working with uneven datasets.

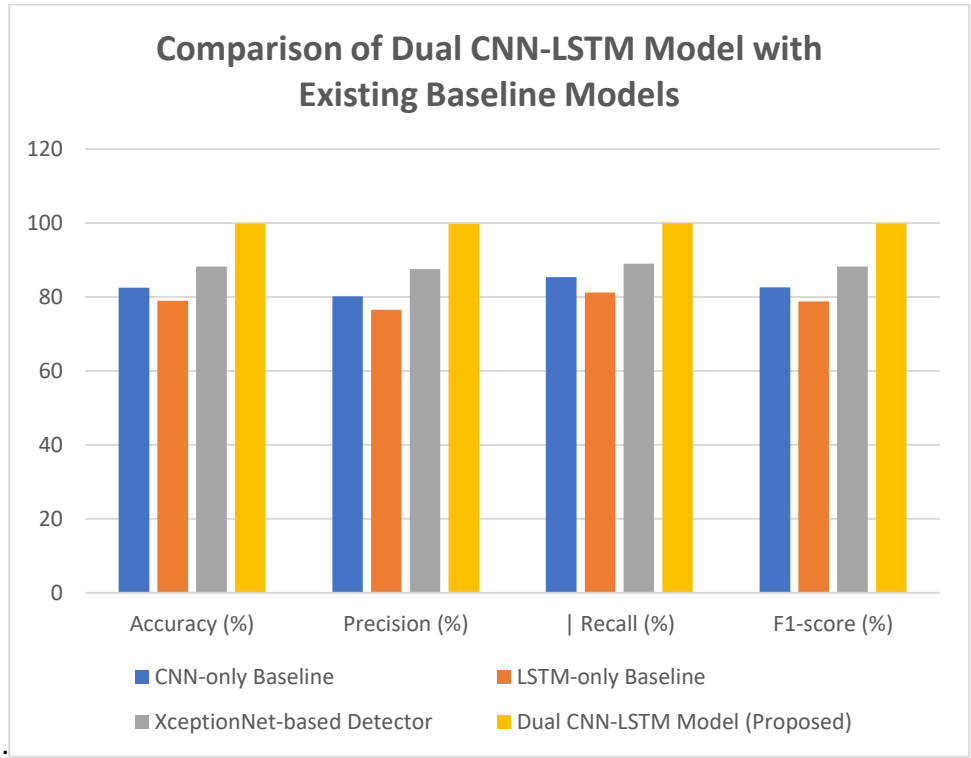



Figure 5: Comparison of Dual CNN-LSTM Model with Existing Baseline Models

## 5. CONCLUSION

This paper introduces a CNN-LSTM model to detect deepfake video cyber phishing attempts. The CNN extracts spatial features from frames, and the LSTM spots time-based problems across frame sequences, creating a well-rounded approach to finding deepfakes. The model, which was trained using the Deepfake Detection Challenge (DFDC) and FaceForensics++ data, uses a structured preprocessing pipeline for reliable input. The design makes use of both visual and time-based hints to identify deepfakes. The training included optimized loss functions and regularization techniques to enhance performance.

Testing showed that the CNN-LSTM model perform better than other baseline methods when looking at key metrics like accuracy, precision, recall, and F1-score. The model perform very well, scoring a test accuracy of 99.81%, a precision of 99.73%, a recall of 99.87%, and an F1-score of 99.79%. These numbers show that the model is good at identifying real and fake video content. Looking at false positives and negatives showed where the model falls short, mostly when dealing with high-tech deepfakes and real videos that have strange features. This points to main areas to work on in the future.

Future research will focus on improving the current visual model by adding audio analysis. This will allow for a more complete, multimodal way to detect deepfake phishing attacks. We plan to keep retraining the model and coming up with new methods to make sure it can adapt to new threats. This CNN-LSTM model is a good move toward strengthening cyber defenses against deepfake video phishing, which is becoming more common and realistic. It helps make the digital environment more secure.

## REFERENCES

1. Abdulqader M. Almars (2021). Deepfakes Detection Techniques Using Deep Learning: A Survey. Journal of Computer and Communications, 2021, 9, 20-35. https://www.scirp.org/journal/jcc
2. Ameer Hamza, Abdul R. Javed, Farkhud Iqbal, Natalia Kryvinska, Ahmads.Almadhor, Zunerajalil2, and Rouba Borghol. (2022). Deepfake Audio Detection via MFCC features using Machine Learning
3. Chandrasekaran, B., Upadhyay, R. K., & Singhal, A. (2021). Multi-modal deepfake detection: Combining visual and audio cues for improved forgery classification. Conference on Computer Vision and Pattern Recognition Workshops, 1460–1470.
4. Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: A compact facial video forgery detection network. https://doi.org/10.1109/WIFS.2018.8630761
5. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. arXiv preprint arXiv:2006.07397. [https://arxiv.org/abs/2006.07397](https://arxiv.org/abs/2006.07397)
6. Fathima G, Kiruthika S, Malar M, and Nivethini T. (2024). Deepfake Audio Detection Model Based on Mel Spectrogram Using Convolutional Neural Network. International Open Access, Refereed Journal-ISSN: 2320-2882
7. Ganiyusufoglu, I., Ngô, L. M., Savov, N., Karaoglu, S., & Gevers, T. (2020). Spatio-temporal Features for Generalized Detection of Deepfake Videos. arXiv preprint arXiv:2010.11844. Retrieved from [https://arxiv.org/abs/2010.11844](https://arxiv.org/abs/2010.11844)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. Neural Information Processing Systems (NIPS).
9. Korshunov, P., & Marcel, S. (2020). Deepfakes: A new threat to face recognition? Assessment and detection. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(2), 123–133.
10. Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1–7). IEEE. https://doi.org/10.1109/WIFS.2018.8630761
11. Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) (pp. 83–92). IEEE. https://doi.org/10.1109/WACVW.2019.00020
12. Mittal, T., Oh, J., Lee, F., & Chandrasekaran, B. (2022). Audio-visual deepfake detection method using affective cues. Winter Conference on Applications of Computer Vision, 2570–2579.
13. Murugesan, B. (2025, March 10). AI-Driven Phishing and Deep Fakes: The Future of Digital Fraud. Forbes Technology Council. Retrieved from [https://www.forbes.com/councils/forbestechcouncil/2025/03/10/ai-driven-phishing-and-deep-fakes-the-future-of-digital-fraud/](https://www.forbes.com/councils/forbestechcouncil/2025/03/10/ai-driven-phishing-and-deep-fakes-the-future-of-digital-fraud/)
14. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2021). End-to-end attention-based deepfake detection. 2021 IEEE International Workshop on Information Forensics and Security (WIFS), 1–6.
15. Rössler, 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 (ICCV), 1–11. [https://arxiv.org/abs/1803.09179](https://arxiv.org/abs/1803.09179)
16. Saleh Albahli1 and Marriam Nawaz. (2023). MedNet: Medical deepfakes detection using an improved deep learning approach. https://doi.org/10.1007/s11042-023-17562-5

17. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–148.
18. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 4489–4497.
19. Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.
20. Zhou, P., Han, X., & Morariu, V. I. (2018). Two-stream neural networks for tampered face detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1831–1839). IEEE. https://doi.org/10.1109/CVPR.2017.199