A Comparative Analysis of State-of-the-Art Algorithms for Robust Deep Fake Detection

Zehra Mogulkoc, Beyzanur Yuce
Department of Computer Engineering
Abdullah Gul University, Kayseri, Turkey
{zehra.mogulkoc, beyzanur.yuce}@agu.edu.tr

***Abstract*— It's becoming more and more important to discern real information from modified fabrications in an era where deepfakes, or sophisticated artificial media made using advanced AI techniques, are becoming increasingly popular. Using an extensive analysis of three state-of-the-art algorithms—Gate Recurrent Unit (GRU), Long Short-Term Memory networks (LSTM), and Convolutional Neural Networks (CNN)—this research investigates the critical requirement for efficient deepfake identification. Beyond their superficial resemblance to real people, deepfakes represent a variety of risks, ranging from the spread of misinformation and violations of personal privacy to the degradation of societal trust. This paper carefully compares and contrasts the performance of CNN, LSTM, and GRU in terms of accuracy, robustness against new deepfake approaches, and computing efficiency. This analysis is based on carefully selected datasets that are representative of real-world situations and seeks to reveal subtle differences between the unique advantages and disadvantages that each algorithm has. The implications of this extensive study transcend far beyond academia, potentially directing advances in media forensics, policy frameworks, and the ongoing fight against the destructive spread of misinformation and digital manipulation.**

***Index Terms*— DeepFake, LSTM, GRU, CNN**

The complete implementation code, detailed experiments, and results can be found in this Colab notebook: `https://colab.research.google.com/drive/1kHOhbP-pvyCFtB-I_6S4AKbItew1AEYL?usp=sharing`.

I. INTRODUCTION

In the era of rapidly evolving technology, deepfake productions are becoming more and more common raising concerns about the potential harm they may cause—from spreading disinformation to compromising privacy. Artificially generated media, or "deep fakes," present a significant problem since they may mislead visual content, such as images and videos, by utilizing advanced machine learning techniques. It is essential to detect these deepfakes to prevent their misuse and preserve the integrity of digital assets. [1]

Addressing the challenge of deep fake detection requires a nuanced understanding of the underlying algorithms and their capabilities. Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU) are among the key architectures in deep learning that have proved effectiveness in various sequential and spatial tasks. Motivated by the critical need for reliable deep fake detection mechanisms, this paper presents a comprehensive comparative analysis of the performance of LSTM, CNN, and GRU algorithms in the context of deep fake recognition.

Our motivation arises from the growing societal impact of deep fakes, as well as the need to build detection approaches that can keep up with generative models' increasing sophistication. Through an examination of the advantages and disadvantages of the LSTM, CNN, and GRU architectures, we hope to provide insightful information that will help designers create deeper fake detection systems that are more reliable and accurate. We want to identify the subtleties in these algorithms' performance through methodical testing and assessment, which will serve as a basis for developing deep fake detection as a field and creating a more secure cyberspace.

A. Related Work

Growing concerns have led to an increase in research projects aiming identifying DeepFakes. Many research treat DeepFake detection as a binary classification problem and use deep neural networks to detect retouched or altered videos. For this purpose, Convolutional Neural Networks are widely used and have proven impact on solving computer-aided detection problems. As an example, MesoNet [1] uses a specially-crafted CNN architecture, focusing on the mesoscopic properties (small-scale patterns) of DeepFake pictures. Another illustration is Xception [3], which trains an XceptioNet [2] model using the traditional cross-entropy loss. [4] demonstrated that CNNs can be used to detect image modifications with the help of high-pass filters. Moreover, the incorporation of dynamic dense blocks into a CNN addresses optimization issues in deep neural networks, and the addition of an attention mechanism improves generalization performance overall.[5]

Besides CNN-based approaches, LSTM and GRU are among the most popular algorithms used for detection of DeepFakes. While the CNN module focuses on capturing particular characteristics and patterns inside individual frames, LSTM analyzes the temporal correlations between frames to identify patterns that indicate deepfake manipulation.[6] Additionally, a study introducing a Convolutional LSTM-based Residual Network for the purpose of identifying deepfake movies presented how effectively LSTM can be combined with deep learning architectures to improve detection accuracy.[7] Similar to LSTM, GRU is also useful for examining temporal correlations and identifying long-range relationships in video sequences when applied to deepfake contents which are critical for identifying the subtle modifications found in deepfake content.[8]

Furthermore, there are some other studies using Hybrid methods for identification of manipulations in images and videos. The majority of these methods also include the implementation of LSTM or CNN algorithms. For example, the study [9] showcased the efficacy of employing a combination of Long Short-Term Memory (LSTM) networks and a CNN model for detecting deepfakes. Additionally, GRU and convolutional neural networks (CNN) have been combined to create a comprehensive deepfake detection system. This application showcases GRU's versatility and effectiveness in handling the complexities of deepfake detection.[10] The study's findings, which involved integrating GRU into CNN, showed promise in improving the precision and resilience of deepfake detection techniques.[11]

II. MATERIALS AND METHODS

A. Dataset

The dataset utilized in this study originates from Meta for the Deep Fake Detection Challenge (DFDC) and is available on both Meta's official website and Kaggle [12]. Alongside the dataset, we created a CSV file named df_metadata.csv from the provided metadata, which contains essential labels and annotations crucial for model training.

Strategically partitioned, the dataset consists of distinct training and test sets. About 80% of the data is dedicated to the training set, ensuring a diverse array of examples for robust model learning, while the remaining 20% is reserved for the test set, serving as an independent benchmark for model evaluation.

For a brief statistical overview, the dataset includes a total of 3293 videos, comprising 1727 authentic videos and 1566 deepfake videos.

B. Preprocessing

A novel preprocessing approach was implemented, featuring a two-step process. The Frame Count Analysis ensured videos from the "Real videos" directory contained a minimum of 150 frames, establishing a robust temporal context. The calculated average frame count per video, at 147.62 frames, reflects the dataset's temporal diversity. Subsequently, Face Video Generation used advanced face detection and recognition models (MTCNN and Inception Resnet V1) to align, extract, and recognize faces. These meticulous preprocessing steps resulted in a standardized and face-focused dataset, providing a strong foundation for developing and evaluating deepfake detection models.

C. Proposed Methods

In this study, our main objective is to undertake a thorough examination and comparison of prevalent methodologies within the realm of deepfake detection. Confronting the escalating challenges presented by the widespread proliferation of highly realistic fake videos, commonly known as deepfakes, our emphasis is on the meticulous evaluation and juxtaposition of established algorithms. Drawing insights from an extensive literature review, we specifically compared three of the most widely utilized models: Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM). By the findings of Yesugade et al. in 2022, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models exhibited an accuracy rate of approximately 88% [13]. Building upon the research conducted by B. Dolhansky et al., their study reveals an algorithm incorporating ResNet and GRU components, boasting an impressive success reaching a rate of 99% [12]. Afchar et al. proposed MesoNet, a CNN-based model with fewer layers to focus on mesoscopic features in images. They introduced MesoInception-4, incorporating Meso-4 and Inception modules that use convolutional neural network layers. The models were evaluated on internet-collected deepfake videos, achieving an average detection accuracy of about 98% for Deepfake videos [1]. Yang et al. compared head poses using central regions and facial features to detect fake and original images or videos. Their SVM model achieved approximately 84% accuracy[15].

Our developed CNN-based deepfake detection model, Xception2, incorporates a variant of the Xception architecture. The model consists of a sequence of convolutional layers, each followed by batch normalization and ReLU activation, capturing hierarchical features in the input data. The architecture culminates in a global average pooling layer to obtain a comprehensive spatial representation. The final output is generated through dropout regularization and a linear layer. This developed CNN model achieved an accuracy rate of approximately 60% in the DFDC dataset.

Our developed LSTM-based deepfake detection model utilizes a pre-trained Residual Network CNN, specifically the ResNeXt50 model, for effective feature extraction. The architecture includes an LSTM layer, leaky ReLU activation and dropout regularization enhance the model's robustness, followed by a linear layer for classification. During the forward pass, the input data undergoes reshaping, processing through the ResNeXt50 backbone, and adaptive average pooling to capture spatial features. The LSTM layer captures temporal information, and the final output is obtained through dropout and a linear layer for classification. This developed model achieved an accuracy rate of approximately 83% in the DFDC dataset which is the highest accuracy among the algorithms we compared.

Our last developed model integrates a Gated Recurrent Unit (GRU) with a ResNet-50 backbone for efficient spatial and temporal modeling in deepfake detection. The GRU layer captures temporal dependencies bidirectionally, while batch normalization enhances model stability. Dropout regularization, applied before the linear classification layer, ensures robust performance. This model adeptly discerns patterns in both spatial and temporal dimensions, achieving an accurate result in differentiating real and fake videos within the DFDC dataset with an accuracy of 85%. [1-2] [3-5]

Model Type	Architecture and Components	Accuracy
CNN+LSTM (Yesugade et al., 2022)	Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM)	88%
CNN + GRU (B. Dolhansky et al.)	Residual Network CNN (ResNeXt50) + Gated Recurrent Unit (GRU)	99%
CNN (Afchar et al.)	MesoNet with MesoInception-4 (CNN-based)	98%
SVM (Yang et al.)	Support Vector Machine (SVM) on head pose features	84%
Xception2 (Developed CNN)	Xception architecture with convolutional layers	60%
LSTM (Developed Model)	ResNeXt50 + LSTM	83%
GRU (Developed Model)	ResNet-50 + GRU	85%

TABLE I

SUMMARY OF DIFFERENT DEEP LEARNING MODELS FOR DEEP FAKE DETECTION.

D. Experiments

In our experiments, we conducted our research on the Google Colab platform, leveraging the computational power of an NVIDIA T4 GPU. The software environment was configured with Python version 3.10.12. Specifically, we employed PyTorch (torch) for implementing deep learning models, opencv-python for efficient video and image processing, seaborn for data visualization, and numpy for fundamental numerical operations. Running each model takes around an hour, for a total of three hours.

E. Findings

We have gained important insights by conducting a thorough comparison of our generated models, all of which use common structures in deepfake detection. With an accuracy rate of 85%, the Gated Recurrent Unit (GRU) model stood out as the most successful. The Long Short-Term Memory (LSTM) model came in second and performed admirably, with an impressive success rate of 83%. Conversely, the Convolutional Neural Network (CNN) model produced a significantly lower accuracy rating of 60% when used on its own. Each model's outcomes and additional information are explained below.

F. Gated Recurrent Unit (GRU)

The GRU model demonstrated a clear improvement in performance during the training process over 20 epochs. Both the training and validation sets exhibited a consistent decrease in loss values, indicating effective learning. Concurrently, accuracy values exhibited a notable increase, reaching a final accuracy of 85.80%.

The confusion matrix provides additional insights into the model's performance:

Metric	Value
True Positive (TP)	247
False Positive (FP)	59
False Negative (FN)	34
True Negative (TN)	315

TABLE II

CONFUSION MATRIX METRICS FOR GRU.

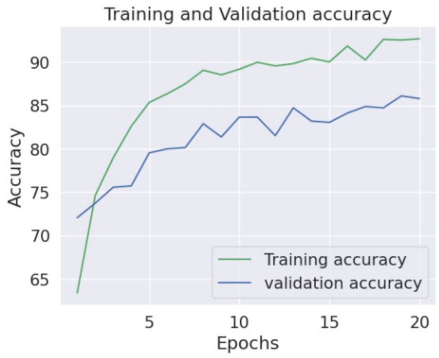

Fig. 1. Accuracy per epoch for GRU

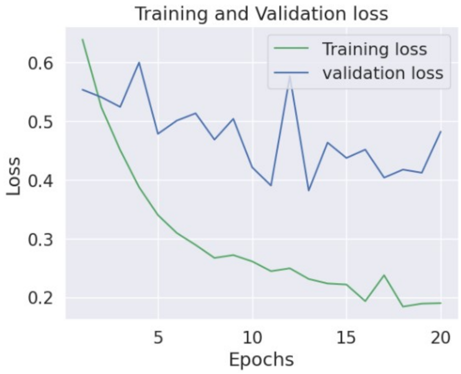

Fig. 2. Loss per epoch for GRU

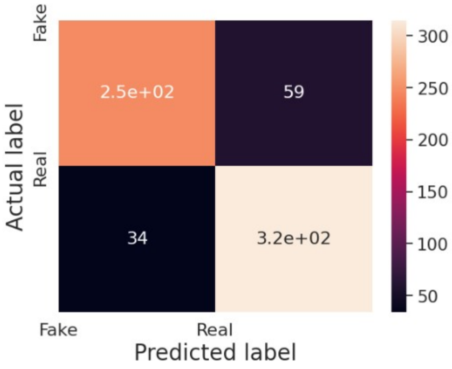

Fig. 3. Confusion matrix for GRU

G. Long Short Term Memory (LSTM)

The training loss of the created LSTM model decreased as expected, while the testing loss rose. It is evident that the accuracy chart is a successful model because it displayed an improving trend in both groups. The following was the outcome of the confusion matrix:

Metric	Value
True Positive	281
False Positive	23
False Negative	88
True Negative	263

TABLE III
CONFUSION MATRIX METRICS FOR LSTM.

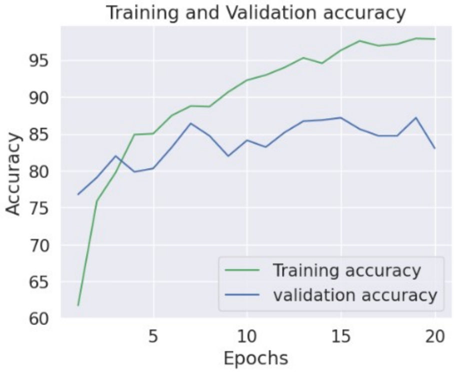

Fig. 4. Accuracy per epoch for LSTM

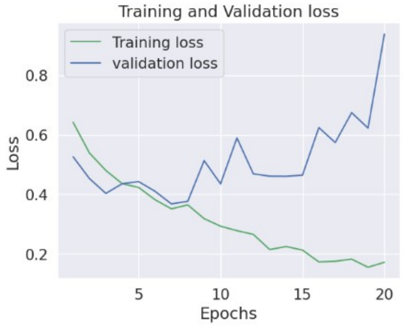

Fig. 5. Loss per epoch for LSTM

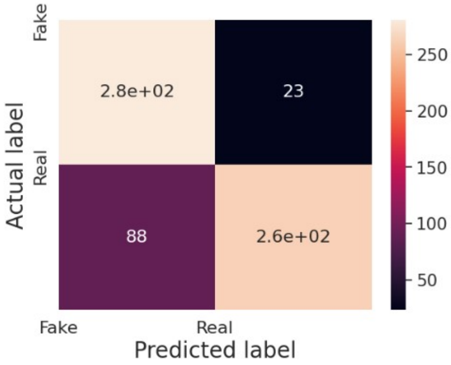

Fig. 6. Confusion matrix for LSTM

H. Convolutional Neural Network (CNN)

In the developed CNN model, the training set exhibited a decreasing trend in the loss graph. Despite fluctuations in the test set, a downward trend in the loss was observed. Moreover, fluctuations in accuracy also increased, indicating the model's success. The confusion matrix results are as follows:

Metric	Value
True Positive	240
False Positive	64
False Negative	193
True Negative	158

TABLE IV
CONFUSION MATRIX METRICS FOR CNN.

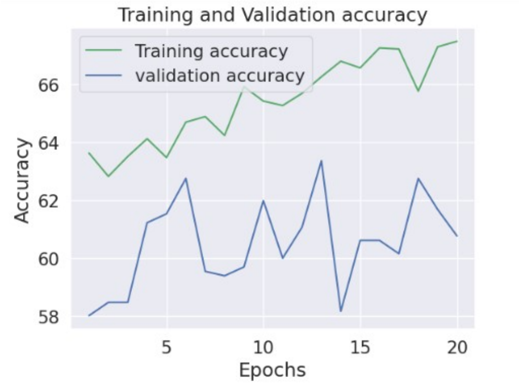

Fig. 7. Accuracy per epoch for CNN

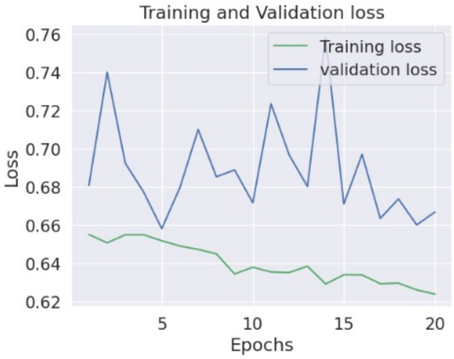

Fig. 8. Loss per epoch for CNN

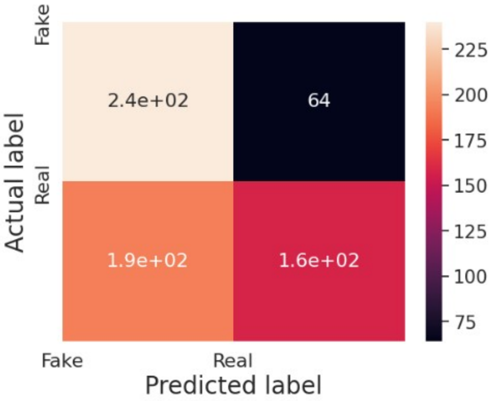

Fig. 9. Confusion matrix for CNN

III. Conclusions

In order to effectively detect deep fakes, we conducted a comprehensive analysis of three prominent deep learning algorithms in this study: the Gate Recurrent Unit (GRU), Long Short-Term Memory networks (LSTM), and Convolutional Neural Networks (CNN). The use of algorithms becomes increasingly important as deepfakes become more common. Our results highlighted specific performance attributes.

With an accuracy rate of 85%, the GRU model stood out and demonstrated steady improvement after training. With 247 true positives, 59 false positives, 34 false negatives, and 315 true negatives, the confusion matrix confirmed effectiveness. With 83% accuracy, the LSTM model produced a confusion matrix with 281 true positives, 23 false positives, 88 false negatives, and 263 true negatives. It also showed expected trends in training and testing losses. When the CNN model was utilized independently, it produced a confusion matrix with 240 true positives, 64 false positives, 193 false negatives, and 158 true negatives. This accuracy resulted in fluctuations but ultimately showed progress in training.

Our findings have consequences that reach beyond the boundaries of academia, including the fields of media forensics, policy frameworks, and the continuous fight against the spread of false information and online manipulation. It is vital to continuously improve and develop detection techniques as we navigate the rapidly changing deepfake technological ecosystem in order to defeat generative models that are becoming more and more complex. In order to enhance transparency in deep fake detection systems, future research in this area may investigate hybrid approaches that take advantage of the complimentary qualities of various algorithms as well as clarity and interpretability issues. Furthermore, tackling scalability and real-time detection problems would help to improve these algorithms' practical usability in a variety of contexts.

In conclusion, this research offers a thorough comparison of the differences in performance of the most advanced deep fake detection algorithms available, setting the stage for further developments in the continuous search for reliable and safe digital content.

IV. Acknowledgement

We would like to express our deepest gratitude to our instructor Dr. Rifat Kurban, for all of his help and assistance throughout this project. Their knowledge and support were invaluable to our success, which made the study rewarding.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, "MesoNet: a compact facial video forgery detection network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (2018), pp. 1-7, DOI: 10.1109/WIFS.2018.8630761.

[2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, "FaceForensics++: learning to detect manipulated facial images," in Proceedings of the IEEE/CVF International Conference on Computer Vision (2019), pp. 1-11.

[3] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 1251-1258.

[4] D. H. Kim and H. Y. Lee, "Image manipulation detection using convolutional neural network," in International Journal of Applied Engineering Research, vol. 12, no. 21, pp. 11640–11646, 2017.

[5] X. Mao, L. Sun, Z. Hongmeng, S. Zhang, "A deepfake compressed video detection method based on dense dynamic cnn", International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2022), 2023. https://doi.org/10.1117/12.2674838.

[6] M. Shaikh, L. Nirankari, V. Pardeshi, R. Sharma, P. Kale, "Deepfake detection using deep learning (cnn+lstm)", Interantional Journal of Scientific Research in Engineering and Management, vol. 07, no. 11, p. 1-11, 2023. https://doi.org/10.55041/ijsrem26808.

[7] S. Tariq, S. Lee, S. Woo, "A convolutional lstm based residual network for deepfake video detection", 2020. https://doi.org/10.48550/arxiv.2009.07480.

[8] R. Ram, M. Kumar, T. Al-shami, M. Masud, H. Aljuaid, M. Abouhawwash, "Deep fake detection using computer vision-based deep neural network with pairwise learning", Intelligent Automation Amp; Soft Computing, vol. 35, no. 2, p. 2449-2462, 2023. https://doi.org/10.32604/iasc.2023.030486.

[9] D. Guera and E. J. Delp, “Deepfake video detection using recurrent ¨ neural networks,” in 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2018, pp. 1–6.

[10] S. Tariq, S. Lee, S. Woo, "One detector to rule them all: towards a general deepfake attack detection framework", 2021. https://doi.org/10.48550/arxiv.2105.00187.

[11] A. Pishori, B. Rollins, N. Houten, C. Nisha, O. Uraimov, "Detecting deepfake videos: an analysis of three techniques", 2020. https://doi.org/10.48550/arxiv.2007.08517.

[12] B. Dolhansky et al., "The DeepFake Detection Challenge Dataset," 2020. [Online]. Available: arXiv:2006.07397.

[13] Yesugade, T., Kokate, S., Patil, S., Varma, R., Pawar, S. (2022). Deepfake detection using lstm-based neural network. Object Detection by Stereo Vision Images, 111–120. https://doi.org/10.1002/9781119842286.

[14] T. Fernando, C. Fookes, S. Denman and S. Sridharan, "Detection of Fake and Fraudulent Faces via Neural Memory Networks," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1973-1988, 2021, doi: 10.1109/TIFS.2020.3047768.

[15] Y. Li, X. Yang, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), pp. 8261–8265, 2019.