Received: 11 December 2023 | *Revised:* 31 March 2024 | *Accepted:* 9 June 2024
DOI: 10.37917/ijeee.22.1.20 *Vol. 22 | Issue 1 | June 2026*

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Open Access
IJEEE University of Basrah College of Engineering

Fusing Spatial and Temporal Features Extracted Using Convolutional Neural Networks and Gated Recurrent Units for Improved Deepfake Detection

Mohamed Abdulrahman Abdulhamed*[1], Asaad Noori Hashim[2]
[1]Department of Computer Science, Computer Science and Information Technology college, University of Basra, Basra, Iraq
[2]Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq

Correspondance
*Mohamed Abdulrahman Abdulhamed
Department of Computer Science, Computer Science and Information Technology college,
University of Basra, Basra, Iraq
Email: mohammed@uobasrah.edu.iq

Abstract
Deep falsification of multimedia content, especially videos and photos, threatens social cohesion (e.g., rumour propagation, extortion, and truth distortion) and must not be ignored. In some cases, this issue requires effective detection solutions. Most studies suggest that convolutional neural networks (CNNs) may not be able to extract complex features like those used in deepfake production. Thus, hybrid approaches that can capture complex features and act as powerful descriptors for binary classification are needed to separate bogus from true content. In this paper, a hybrid algorithm is developed to combine gated recurrent units (GRU) and CNN. The proposed model aims to improve the extraction of complex features by simultaneously capturing instantaneous and spatial features. This approach permits the extraction of implicit features that are vital to the final classification process, especially when dealing with a sequential series within video content. Finally, a dense neural network is used to classify these features. Practically, two data sets were used to train the proposed model: the FaceForensics++ (FF++) and DeepFake Detection Challenge (DFDC) datasets. The evaluation results of the proposed model on the FF++ dataset for the Area Under the Curve (AUC) and F1-score metrics reached 0.88% and 0.85%, respectively. While DFDC is 0.95% and 0.86% for the same metrics, respectively.

Keywords
Deepfake Detection, Media Forensic Detection, Deep Learning, Gated Recurrent Unit.

I. Introduction

Deepfake videos produced using advanced artificial intelligence technologies have recently experienced a significant surge in popularity, attracting much attention from various sources. This advancement facilitated the use of a method known as deep-fake, which allows for the alteration of facial features with a remarkable level of realism. Currently, a substantial amount of fake videos is being disseminated across the Internet, with a considerable portion of these movies specifically targeting individuals of prominence, such as celebrities and politicians. These recordings are frequently used to tarnish the reputation of important figures and manipulate public sentiment, thus posing a huge threat to social cohesion. The deepfake technique uses neural networks, namely, autoencoders or generative adversarial networks, to substitute faces in the target video with faces from the source video. The use of these technologies facilitates the creation of facial alteration videos, contingent on the availability of substantial datasets [1, 2]

Deepfake detection approaches can typically be classified into several overarching methodologies, each using distinct features or concepts to identify altered information. Among the most widely used techniques are [3] image and video quality analysis, temporal and spatial feature analysis and deep

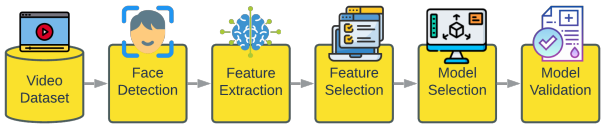

Fig. 1. Deep fake detection building steps.

learning approaches. Recently, many researchers have turned to using deep learning algorithms in the process of detecting deep falsification [4]. Video fake detection approaches use deep convolutional neural networks (CNNs) to identify visual spacing within frames and deep recurrent neural networks (RNNs) to calculate temporal spacing between frames. Deepfake detection presents interpretability, robustness, and generalizability as the three most challenging problems to solve.

Deep learning techniques can use several structures to develop different models, such as CNNs and RNNs [5]. CNNs have demonstrated impressive capabilities and scalability in applications for efficient image and video processing. They are primarily used for image processing and for automatic segmentation of images and their associated data. CNNs have the unique ability to extract image features that can then be applied to various classification models. The primary structure of these neural networks consists of input and output layers, as well as hidden layers for processing pixel data. Thus, the extracted features from these layers can be used to develop more accurate and efficient models to detect deepfake in images [6, 7] . In contrast, the RNN architecture uses the previous layer's output as input for the subsequent layer. Consequently, this neural network architecture can process (instantaneous) time series data. Thus, it contains types, such as gated recurrent units (GRUs), that can control the transmission of only essential data to subsequent layers. GRUs are a unique form of RNN due to their multiple gates, such as the update gate and reset gate. In practice, these gates allow the GRU to represent short- and long-term dependencies in sequence data with greater precision than conventional RNNs [8]. Intuitively, hybrid methods that combine CNN and GRU must be developed to create an efficient model to extract complex features within deep-fake segments [9].

Several databases are available for training and testing models for deepfake detection, including DFD [10], FaceForensics++ [11] and DFDC [12]. Based on sources [13, 14], the stages of building a deep fake detection model can be generally summarised into five stages; as shown in Fig. 1, these steps are followed to complete the design of a deep fake detection model.

The primary focus of this research study is to present an interesting method for extracting intricate features present in deceptive visual content using deepfake algorithms. The study emphasises the introduction of a hybrid algorithmic model that combines a convolutional neural network (CNN) architecture to capture spatial features and a series of continuous frames that rely on gated recurrent units (GRU) to extract temporal patterns. In summary, this study focused on building a hybrid algorithm, i.e., ConvGRU model, for fake video detection. Overall, the study presents the following contributions: Improved the dataset's quality by applying Contrast Limited Adaptive Histogram Equalisation (CLAHE) and Implemented a hybrid methodology that integrates CNN with GRU to enhance the detection of fabricated content embedded within videos.

The remainder of this paper is structured as follows: Section II. presents some studies related to our research. Section III. offers a summary that is theoretically relevant to the subjects of our study. Then, the proposed algorithm to create the classification model is presented in Section IV. . Section V. evaluates the proposed model and discusses the results. Finally, the conclusion and future directions are presented in Section 6.

II. Related Works

Current research in the field of fake media detection mostly uses CNN architectures in conjunction with various techniques, including transformer models, recurrent networks, and multimedia features. These approaches are utilised to identify and discern deep false images and videos effectively. This section discusses a selection of related studies.

In [15], the authors presented a new image representation technique known as 'face X-ray', which was designed to detect instances of fraud in facial photographs. The greyscale image provides insight into the potential decomposability of the input image, indicating if it may be effectively separated into a blend of two distinct images originating from disparate sources. The use of face X-ray technology has proven to be a reliable method for identifying counterfeit photographs produced by various face alteration algorithms. This technique operates under the assumption that a blending process is involved in the creation of such forgeries. Furthermore, the training of face X-ray models does not necessitate the use of fabricated images. The effectiveness of the subject was demonstrated by a series of comprehensive experiments. The research may be limited by the time available for the study, which may affect the depth and breadth of the research. This makes the method incompatible with modern deepfake generation methods.

The primary objective of [16] was to investigate the identification of facial alteration in video sequences using contemporary methodologies, such as Face Swap and deepfake. The study used an ensemble approach by combining many trained

CNN models to identify modifications in films. The suggested approach uses attention layers and Siamese training, yielding encouraging results on two publicly accessible datasets that include more than 119,000 videos. The model's ability to detect deep fakes made using different methods may be hindered by the study's use of a specific deepfake generation strategy.

In [17], a novel approach was introduced to detect deepfake content by leveraging audio and video modalities, together with the analysis of perceived emotions, to differentiate between authentic and manipulated media. The approach demonstrated a performance of 84.4% in terms of AUC on the DeepFake-TIMIT dataset and 96.6% on the DF-TIMIT dataset, drawing inspiration from the Siamese network architecture and incorporating the triplet loss. The efficacy of the approach relies on the presence of extensive datasets for training. This could provide a constraint in situations where such datasets are not easily accessible or are challenging to acquire.

In, [18] presented a novel approach for detecting deepfake videos utilising synthetic facial regions and 3D posture estimation techniques. The proposed system was designed to identify manipulated videos that have been artificially created. The algorithm produced incongruent facial landmarks and traits, resulting in subtle disparities between counterfeit and genuine faces. The implemented system used the Dlib library for face recognition and uses OpenFace2 to construct a conventional three-dimensional facial model. The system uses unstructured annotated data feature vector data and uses a support vector machine classifier, which exhibits an area below the receiver operating characteristic (AUROC) value of 0.89. Model performance may be affected by the quality and diversity of training data. The model's real-world performance may be affected by the training data not covering all deepfake approaches.

In [19], a model that combines YOLO, CNN, and XGBoost was proposed to find deepfakes. In this work, the YOLO face detector was used to pull faces from video frames. The InceptionResNetV2 CNN was then used to extract facial features from the pictures that have already been extracted. The CelebDF-FaceForencics++ (c23) dataset was used for the study. It is a combination of two well-known datasets, Celeb-DF and FaceForensics++ (c23). In academic studies, evaluation factors, such as accuracy, specificity, precision, recall, sensitivity, and F1 score are often used. The results showed that the CelebDF-FaceForensics++ merged CelebDF-FaceForensics++ dataset (c23) gives an area under the curve (AUC) value of 90.62%. The XGBoost model's shortcomings in managing high-dimensional video data are not addressed in the research.

The primary objective of LipForensics [20] is to identify manipulated facial recordings by detecting considerable semantic irregularities in mouth movements. The study uses the FaceForensics++ dataset, together with DeepFakes, Face Swap, Face2Face and NeuralTextures, for the purposes of training and testing. The resulting accuracies obtained are 82.4%, 73.5%, 97.1% and 97.6%, respectively. However, the study does not disclose the approach's limitations. It does not answer how the system would handle low-quality photographs or videos, lighting, angles, or facial expressions. Further research could improve system performance in different settings.

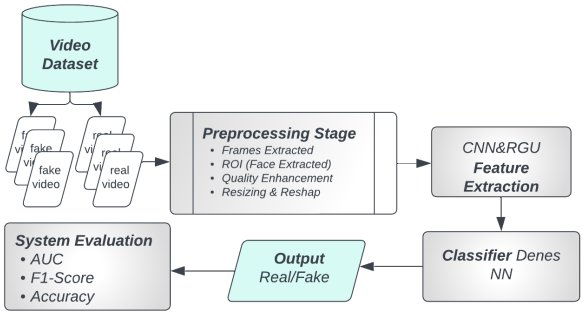

Fig. 2. Proposed Deepfake Detection Algorithm Pipeline.

III. Proposed Method

The algorithmic pipeline suggested in the deepfake detection model is depicted in Fig. 2. The framework comprises a sequence of steps accompanied by specialised functions, as

Shown in the diagram. The initial stage involves performing input preprocessing operations, whereby video frames are utilised as input and the facial location is restored, alongside various optimisation techniques, such as rescaling and other relevant procedures. In the subsequent phase, the extracted

Sequences are inputted into the hybrid model, known as the Conv&GRU neural network, to capture spatial and temporal aspects effectively based on the facial feature information. The extracted characteristics are subsequently inputted into dense neural networks to execute the ultimate classification procedure using the sigmoid function. In the subsequent sections, a comprehensive breakdown of each phase within the proposed deepfake detection process is presented.

A. Preprocessing Stage

The preprocessing stage is crucial to the development of computer vision models, particularly in the realm of system design that heavily relies on advanced artificial intelligence techniques, such as machine learning and deep learning. The process begins with the initialisation and optimisation of the dataset, as shown in Fig. 3. The key proposed treatments can

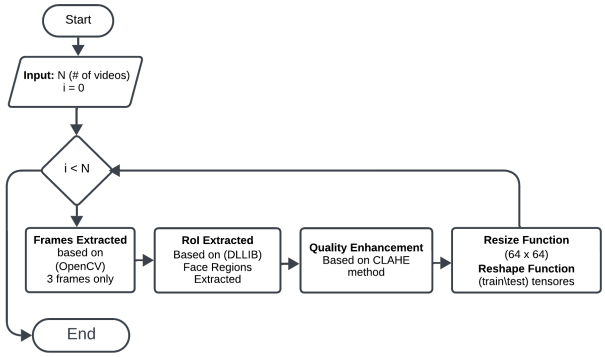

Fig. 3. Pre-processing stage of the proposed model.

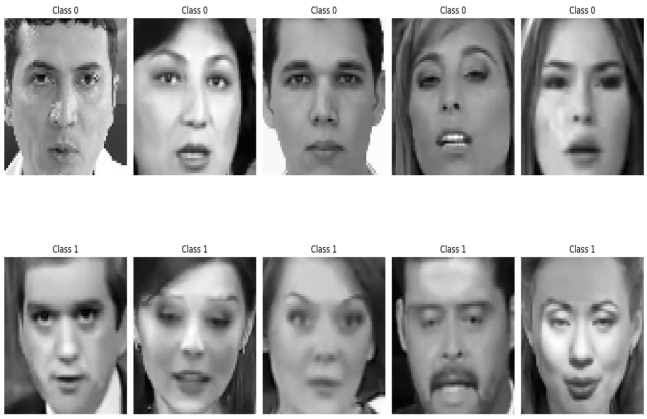

Fig. 4. FF++ training dataset samples: class 0 for real faces and class 1 for fake faces.

be succinctly summarised as follows. At first, the procedure of extracting frames from every video within the dataset is executed, whereby only three frames, which are subsequently utilised in following procedures, are extracted.The process depends on the utilisation of OpenCV libraries [20]. Additionally, extraction of the region of interest relies on a specific library to accomplish this operation, with the face region being the targeted area in the context of deep-fake issues [21].

Afterwards, the strong image processing approach, i.e. Contrast-limited adaptive histogram equalisation (CLAHE), enhances detail visibility in frames with uneven or inadequate lighting. CLAHE dynamically adjusts the visual contrast in small areas. CLAHE enhances the detail of the image without noise or artificial effects by limiting the enhancement of contrast to particular locations and constraining the amplification. This technology is used in medical imaging, satellite images, and computer vision to improve subtle feature visibility for correct analysis and interpretation. CLAHE enhances images and advances image processing in several fields [22]. Finally, we work to make the extracted samples the same size as (64×64) and shape so that they are prepared for the feature extraction step using the deep learning algorithm. This step keeps things consistent and gets the samples ready. To make them a tensor structure, they must be changed. Fig. 4 shows a selection of training samples, showcasing five samples from each class within the FF++ training dataset.

B. Feature Extraction

Deep learning techniques, specifically CNNs [23], have achieved remarkable achievements in various computer vision tasks, such as image classification, object detection, and face recognition, since their inception. A CNN architecture typically has convolutional and pooling layers, followed by a stack of fully connected (FC) layers. CNN convolutional layers extract local features at various levels. During the initial phases, these layers are responsible for capturing the essential elements, such as edges, contours, and gradients. However, as the process progresses, these layers proceed to extract more advanced information that plays a critical role in the classification of images. In parallel, the use of max/average pooling operations is implemented to decrease the resolution of features. In the end, the FC layers execute nonlinear operations on the extracted image features.

By contrast, GRUs signify a notable progression within the field of RNNs. The method was first proposed by Cho et al. In 2014 to mitigate the issue of the disappearance gradient in conventional RNNs. RNNs can selectively update or forget information from the preceding time step, hence facilitating improved modelling of long-term dependencies in sequential data. These capabilities render them proficient in tasks, such as language modelling, speech recognition, and machine translation.

However, in some complex tasks of computer vision, especially the deep-fake problem, it is necessary to find instantaneous features in addition to the spatial features that are extracted on the basis of CNN. Therefore, some techniques must be used to extract features capable of distinguishing distinct faces from reality. This structure proposal utilises a hybrid deep learning network with CNN and GRU to capitalise on the aforementioned factors. The diagram depicted in Fig. 5 illustrates the comprehensive architecture that has been proposed for the feature extractor in our research. The CNN module is tasked with extracting spatial features from the input data, whereas the GRU module is designed to capture temporal correlations within the data. Through the integration of these two components, our hybrid network demonstrates the capability to capture spatial and temporal information efficiently, resulting in enhanced performance in our research investigation. Moreover, the suggested architecture facilitates rapid and precise feature extraction, rendering it appropriate

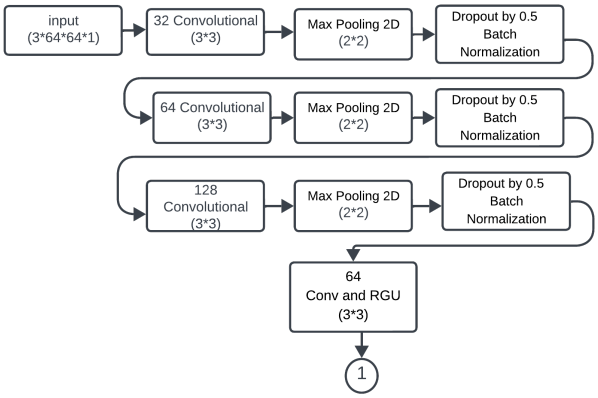

Fig. 5. Proposed hybrid deep learning network (Conv and GRU) for feature extraction.

TABLE I. ARCHITECTURE OF CNN AND GRU PROPOSED FOR FEATURE EXTRACTION

Layer name	Output	Number of Parameters
Input Layer	(None, 3, 62, 62, 32)	320
Conv_0 (2D)	(None, 3, 31, 31, 32)	128
Conv_1 (2D)	(None, 3, 29, 29, 64)	18,496
Conv_2 (2D)	(None, 3, 14, 14, 64)	256
Conv_3 (2D)	(None, 3, 12, 12, 128)	73,856
Conv_4 (2D)	(None, 3, 6, 6, 128)	512
GRU_layer	(None, 4, 4, 64)	442, 624
Flatten	(None, 1024)	-
Dense_1	(None, 128)	131, 200
Dropout	64	-
Dense_2	(None, 1)	129

for many applications in the field of deep learning. Table 1 presents a comprehensive overview of the proposed network designed for feature extraction. It includes information regarding the outputs of each convolutional network layer, as well as the corresponding number of parameters.

C. Classifier Stage

The important task of the FC layer in our proposed model is to serve as a classifier for binary classification. Situated at the terminal point of the network, this layer, commonly known as the dense layer, plays a crucial role in the process of mapping the retrieved high-level information from the preceding layers to binary outcomes. Consisting of neurons that are fully associated with the output of the preceding layer, it forms a densely coupled network. The weights linked to these connections are modified during the training process, allowing the model to represent complex relationships within

Fig. 6. Classifier of the proposed model.) for feature extraction.

the feature space effectively. The use of the sigmoid activation function in the FC layer enables the conversion of the extracted characteristics into a probability score ranging from 0 to 1, which facilitates the binary decision-making process of the model.

The proposed classifier in our model can be summarised as follows: The 'flatten' function is applied to the last layer of the hybrid neural network (Conv and GRU) to create two dense networks, where the output shape for this layer is (53824). The first dense layer, consisting of 128 neurons, is improved by batch normalisation using the ReLU activation function. To avoid overfitting, the 'dropout' mechanism is used with arguments of 0.5. Then comes the turn of the final classification network, which uses the activation function of the signed type to give the final output of the classification. Fig. 6 shows the scheme of the two dense neural networks for the final classification process.

IV. EXPERIMENTS AND RESULTS

In this part, the datasets utilised to evaluate the model are outlined. Additionally, a thorough examination of the implementation details is engaged in. Ultimately, a comparison was made between the results and the current approaches described in the available academic literature.

A. Dataset Used

This section addresses the datasets used in our study for training the suggested model, as well as their examination and evaluation. Two data sets were used to tackle the deepfake issue, as outlined below.

First, the DeepFake Detection Challenges (DFDC) dataset [12] is an important resource in the field of computer vision and deep learning. It was developed to explicitly address the increasing difficulty of identifying deep-fake videos. This data set is distinguished by the wide variety of videos it contains. These videos feature a diverse range of performers, including professionals and amateurs, and cover a wide range of scenarios and contexts. Details are presented in Table 2. Secondly, the FaceForensics++ (FF++) dataset [11] constitutes a substantial addition to the realm of computer vision and deep learning, specifically within the context of identifying deep faces and detecting image and video manipulation in facial content. The provided data set encompasses a wide range of

TABLE II. DATASETS USED IN THE PROPOSED MODEL.

Property	DFDC	FF++	Description
No. Of real videos	5244	1000	Number of real videos in the dataset
No. Of fake videos	5244	5000	Number of fake videos in the dataset
Video Resolution	(Commonly 720p or 1080p)	(Commonly 720p or 1080p)	Resolution of the videos in the dataset
Annotations	Binary labels (real or fake) for each video	Binary labels (real or fake) for each video	Type of annotations provided for each video
Subjects	66	977	Number of subjects or actors in the dataset
Video Categories	Various scenarios, actors, and contexts	Different actors, scenes, and manipulation types	Categories of videos in the dataset

modified videos that have been made using various strategies for deepfake creation; additional information is provided in Table II.

As a result, the dataset serves as a complete benchmark for assessing the effectiveness of deep-fake detection algorithms in terms of their ability to withstand different types of manipulations.

B. Metrics of Evaluation

Three crucial metrics, namely, the F1 score, AUC and accuracy, are of utmost importance in the assessment of classification machine learning models. The F1 score is a valuable metric for datasets with imbalanced class distributions because it effectively combines a precision (the proportion of correctly predicted positive instances) and recall (the proportion of correctly predicted positive instances out of all actual positive instances) into a unified measure. In contrast, the AUC is a crucial metric for assessing a classification model's discriminatory power, ranging from 0 to 1, and is particularly useful for comparing models across different threshold settings and imbalanced datasets. Lastly, precision, the ratio of accurately predicted instances of the total number of instances, is a key statistic of the performance of the classification model. Measures a model's data classification accuracy simply and intuitively. Accuracy is extensively used and understood; however, it might be biased towards the dominant class and fail to account for false positives and negatives in imbalanced datasets where one class considerably dominates the other. Therefore, the proposed model was evaluated on the basis of these metrics to obtain accurate results.

C. Results and Comparison

As stated in paragraph (4.1), the FF++ and DFDC datasets were used in the evaluation of our proposed model. Based on the hyperparameter and augmentation configuration in Table III, the most remarkable results derived for both datasets are presented in this section. The findings of the model examination can be succinctly summarised as follows.

The beginning is with the DFDC dataset, which used 500 videos for the training set. For the validation and examination sets, we used 120 samples. An accuracy of approximately 85% was obtained. Fig. 7a shows the accuracy results for the training set and the test set. Fig. 7b shows the loss values plot for training and validation, where a value of 0.5443 was obtained from the loss function of the test.

On the contrary, a training set of 500 frames was utilised for the FF++ dataset, with the model being trained on around 1500 frames. The testing set consisted of 200 videos. The accuracy achieved was 82.14%, as shown in Fig. 7c, which illustrates the training and validation accuracy graph. Furthermore, the error function for the test reached a value of 0.504, as demonstrated in Fig. 7d. Table IV provides a condensed description of the most important results that were obtained

TABLE III. HYPERPARAMETER AND AUGMENTATION SETTINGS

Parameter	Value	Description
Min learning rate	1e-6	Minimum learning rate for the optimizer
Batch sizes	32	Number of samples in each batch
Epochs	100	Number of training epochs
FRAME PER VIDEO	3	Number of frames extracted from each video
Rotation range	15	Range of rotation for data augmentation (degrees)
(Width and Height) Shift Range	0.1	Range of width and height shift for data augmentation (proportion)
(Shear and Zoom) range	0.1	Range of shear and zoom for data augmentation (proportion)

after studying both data sets. Training the proposed model using the DFDC dataset outperforms that with the FF++ dataset; this superiority can be observed in a 2D bar graph in Fig. 8. According to Table V and TableVI, our proposed model is superior to some relevant studies based on both datasets. The comparison was conducted using two fundamental metrics commonly employed to assess machine learning models: accuracy and area under the curve.

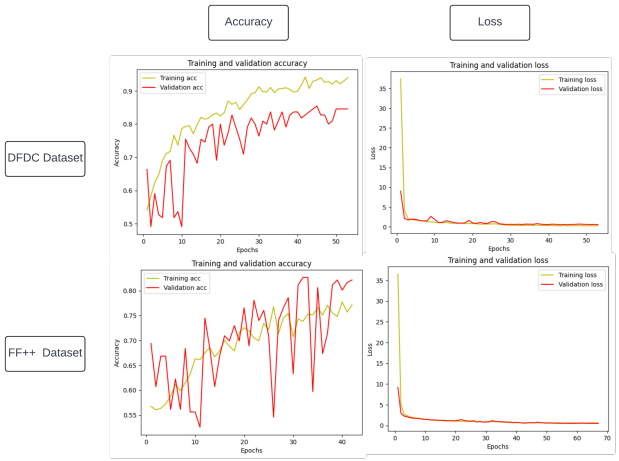

Fig. 7. Proposed model accuracy and loss plots for DFDC and FF++ datasets

TABLE IV. RESULTS OF THE PROPOSED MODEL.

Dataset	F1 score	AUC	Accuracy
DFDC	0.857	0.938	85 %
FF++	0.834	0.88	82 %

TABLE V. COMPARISON OF OUR PROPOSED METHODOLOGY WITH OTHER RELATED STUDIES BASED ON FACEFORENSICS++ DATASET.

Method	Dataset	Acc (%)	AUC
[24]	FaceForensics++	81	0.658
[25]	FaceForensics++	84	0.763
[26]	FaceForensics++	80	0.722
[27]	FaceForensics++	84	0.8
Our Proposal	FaceForensics++	82	0.880

The AUC scale comparison results above show that our proposed model is superior, which enhances its generalization efficiency and enhances its ability. Consequently, analyzing deepfake dataset samples through the use of CNN and the GRU is what makes our method effective. The CNN component excels at identifying spatial patterns in the data, and it

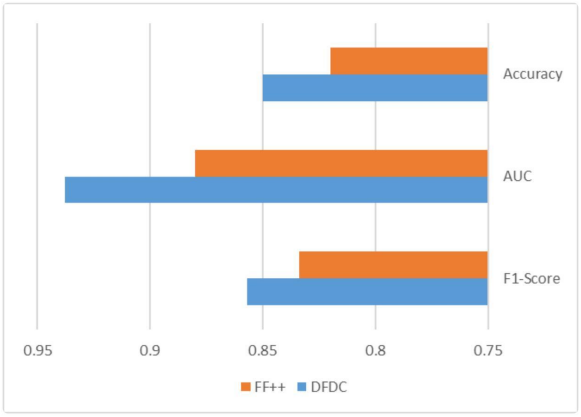

Fig. 8. Comparison of training and testing of the proposed model based on the DFDC and FF++ datasets.

TABLE VI. COMPARISON OF OUR PROPOSED METHODOLOGY WITH OTHER RELATED STUDIES BASED ON DEEPFAKE DETECTION CHALLENGE DATASET.

Methods	Accuracy (%)	AUC
[28]	76	0.924
[29]	93	0.533
[17]	-	0.844
[30]	-	0.832
Our Proposal	85	0.938

is able to capture the visual features of the videos effectively. Conversely, the GRU component excels at recognizing temporal dependencies and comprehending the sequence and timing of the frames in the videos. The efficacy of this technology is further demonstrated by its ability to accurately differentiate counterfeit videos from authentic ones, showcasing its skill in collecting intricate characteristics within deepfake dataset samples.

Moreover, it is crucial to recognise the fundamental challenges and constraints inherent in this research, which can be concisely stated as follows: One of the primary obstacles to tackling the deepfake phenomenon is the scarcity of an extensive data set that effectively captures its intricate characteristics. This difficulty stems from the inherent weakness of the dataset, which contains genuine information, to potential manipulation and distortion. As a result, the absence of diversity poses a hindrance to the model's capacity to capture major patterns that are crucial for the process of generalisation. As an illustration, a subset of 1000 authentic films sourced from the FF++ dataset is employed to generate an estimated count of 4000 counterfeit videos through the utilisation of various deepfake methodologies.

V. Conclusion

This paper presents a new method for detecting fake material in video footage through the utilisation of convolutional neural networks (CNN) and gated recurrent units (GRU). This paper aims to propose a very efficient approach for constructing deep convolutional neural networks (CNNs) and integrating them with gated recurrent units (GRUs) of the type known as GRU-RNN. The objective is to create a complex feature extractor capable of simultaneously estimating spatial and instantaneous characteristics. Subsequently, the characteristics are transmitted to the fully connected network in order to carry out the ultimate classification. The results obtained showed the possibility of contributing to the development of deep fake detection models concerning generalization efficiency, as the results of the F1 score metric reached 0.857 and 0.834 for the DFDC and FF++ datasets, respectively. One big limitation we faced was the lack of a comprehensive data set that could accurately depict the different forms of deepfake technology. However, we would be keen to use a larger dataset to train the proposed model, with the aim of improving its efficiency in detecting different types of deepfakes. Future work also includes the integration of the suggested architecture as a lightweight approach within social media applications.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "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, 2020.

[3] M. Masood, M. Nawaz, K. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.

[4] A. Tiwari, R. Dave, and M. Vanamala, "Leveraging deep learning approaches for deepfake detection: A review," *arXiv preprint arXiv:2304.01908*, 2023.

[5] D. Weimer, B. Scholz-Reiter, and M. Shpitalni, "Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection," *CIRP Annals*, vol. 65, no. 1, pp. 417–420, 2016.

[6] M. Taye, "Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions," *Computation*, vol. 11, no. 3, p. 52, 2023.

[7] S. Cong and Y. Zhou, "A review of convolutional neural network architectures and their optimizations," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 1905–1969, 2023.

[8] F. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis of deep learning models: Cnn, rnn, lstm, gru," *arXiv preprint arXiv:2305.17473*, 2023.

[9] Y. Dong, S. Patil, B. Van Arem, and H. Farah, "A hybrid spatial–temporal deep learning architecture for lane detection," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 1, pp. 67–86, 2023.

[10] N. Dufour *et al.*, "Deepfakes detection dataset by google & jigsaw." Deepfakes detection dataset by google & jigsaw, 2019.

[11] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

[12] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[13] M. Rana, M. Nobi, B. Murali, and A. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[14] L. Stroebel, M. Llewellyn, T. Hartley, T. Ip, and M. Ahmed, "A systematic literature review on the effectiveness of deepfake detection techniques," *Journal of Cyber Security Technology*, vol. 7, no. 2, pp. 83–113, 2023.

[15] L. Li *et al.*, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001–5010, 2020.

[16] S. Suratkar and F. Kazi, "Deep fake video detection using transfer learning approach," *Arab Journal of Science and Engineering*, vol. 48, no. 8, pp. 9727–9737, 2023.

[17] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 2823–2832, 2020.

[18] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265, IEEE, 2019.

[19] A. Ismail, M. Elpeltagy, M. Zaki, and K. Eldahshan, "A new deep learning-based methodology for video deepfake detection using xgboost," *Sensors*, vol. 21, no. 16, p. 5413, 2021.

[20] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5039–5049, 2021.

[21] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[22] G. Yadav, S. Maheshwari, and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," in *2014 international conference on advances in computing, communications and informatics (ICACCI)*, pp. 2392–2397, IEEE, 2014.

[23] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*, pp. 1–6, IEEE, 2017.

[24] S. Suratkar and F. Kazi, "Deep fake video detection using transfer learning approach," *Arab Journal of Science and Engineering*, vol. 48, no. 8, pp. 9727–9737, 2023.

[25] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)*, pp. 1–8, IEEE, 2019.

[26] D. Zhang, C. Li, F. Lin, D. Zeng, and S. Ge, "Detecting deepfake videos with temporal dropout 3dcnn.," in *IJCAI*, pp. 1288–1294, 2021.

[27] F. Alanazi, G. Ushaw, and G. Morgan, "Improving detection of deepfakes through facial region analysis in images," *Electronics (Basel)*, vol. 13, no. 1, p. 126, 2023.

[28] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "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, 2020.

[29] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," *arXiv preprint arXiv:1910.12467*, 2019.

[30] J. Hu *et al.*, "Recap: Detecting deepfake video with unpredictable tampered traces via recovering faces and mapping recovered faces," *arXiv preprint arXiv:2308.09921*, 2023.