RESEARCH ARTICLE OPEN ACCESS

Analyzing Fake Audio using Hybrid CR-NN Approach

Ishita Upadhyay[1], Daksh Kalia[2], Sawinder Kaur[3], Nancy[4]
upadhyayishita21@gmail.com[1] dakshkalia016@gmail.com[2] sawinderkaurvohra@gmail.com[3]
nancyverma16@gmail.com[4]
[1,2,3]AmitySchool of Engineering and Technology, [4]Amity School of Mathematics.

Abstract
Deep learning has recently made significant strides towards solving a wide range of challenging issues, from computer vision and human-level control to large data analytics. Deepfake technologyis one of them; it poses a serious risk to national security, democracy, and privacy. Deepfake filmsare extremely lifelike digitally altered audio of humans saying and doing things that never happened. The negative uses of this technology on social media platforms, such as defacing individuals, outweigh the benefits of deep learning applications. Early in the identification of deepfakes, conventional tools such as signal processing, image processing, and lip-syncing were utilized, but when combined with more modern deep learning technologies, the accuracy of the results is very low. In this research, a method has been proposed that automatically identify deepfakes in media audio filesis proposed here. A hybrid CNN and RNN model that can identify phone audio has been proposed in this study. The accuracy, precision and f1-score achieved by the proposed model are 86.42%, 80.5% and 89.18%.

Keyword: Deep Learning, Deepfake, CNN, RNN, Detection.

Date of Submission: 16-04-2024 Date of acceptance: 28-04-2024

I. Introduction

Fake audio detection, also known as audio forensics or audio authentication, is the process of determining the authenticity or integrity of an audio recording. This field has gained significant attention in recent years due to the proliferation of audio manipulation techniques, such as deepfake technology, that can be used to create fabricated or misleading audio recordings [1]. Detecting fake audio is crucial for ensuring the trustworthiness of audio evidence in legal cases, journalism. Even while the modern internet has transformed our way of life, it still lacks the security to give each user safe access and defense against harmful attacks. This presents difficulties for the biometric authentication technique that is currently in use. There are several categories in which employing biometrics carries danger such as deepfake sounds, spoofing sensors, data and network manipulation, and inaccurate sensors. Professionals in digital forensics must be up to dateon technological advances in order to have a competitive advantage over attackers. There are new guidelines for digital forensics because of a recently resurrected debate concerning the reliability of several conventional forensic methods. Some of the shortcomings of the new voice biometrics technology are the voice replication tools.Finding these artificial voices would make the evidence admissible in a court of law, provided the scientifically sound methods follow established procedures and demonstrate their accuracy, research potential, and academic community acceptance [2]. Successful voice deepfake identification is inherently lacking in testing methods, because deepfake voices are an emerging technology that is always changing, there is limited research on the subject and only a few practical solutions available Kaur S, Kumar P, Kumaraguru P (2020) *et al* [11]. This could lead to cybercrimeslike scamming and the exploitation of personal data. The task of determining if a given media fileis authentic or not falls under the umbrella of the digital forensic discipline, in particular multimedia forensics Kaur S, Kumar P, Kumaraguru P (2020) *et al* [12]. The analysis procedure constitutes a significant aspect of digital forensics. A rigorous study of the component of such an audio file is necessary for forensic investigators to accurately assess the validity of a given false audio multimedia file, especially deepfakes that use sophisticated machine learning techniques to create a fake audio component. To assist digital forensic investigators in identifying voice cloning or deepfake audio for the purpose of gathering evidence, this article assesses current methods for deep learning-based deepfake audio detection Kaur S, Kumar P, Kumaraguru P (2020) *et al* [13]. This research attempts to understand how to help an investigator with deepfake identification of audio using hybrid C-RNN deep learning approach and different pre-processing approaches.

Enhancing the Detection and Mitigation of Fake Audio in the Digital Age With the development of deep learning and advanced AI, it is now possible to produce synthetic audio that is almost identical to real recordings [3]. The growing advent over false audio, which includes produced or modified audio content including audio forgeries, deepfakes, and misleading audio information, is the main focus of this study problem. Maintaining confidence and authenticity in avariety of industries, from media and entertainment to security and forensics, requires an understanding of and response to the spread of fake audio.

II. Literature Review

Different models have been proposed for the identification of fake audio using deep learning approaches. In this section related research work is discussed in which all the tools and techniques which they have used in there respective research work is discussed.

Akash Chintha *et al.(2020)* introduced the XcepTemporal convolutional recurrent neural network framework for deepfake detection. We use an XceptionNet CNN as a salient and efficient facial feature representation. In this system, the researcher extracted features using MFCC and CNN. They extracted 26features from MFCC and 1024 features from last layer of CNN model [4]. Ahmed J. Obaid et al.(2022) , conducted a review of previous studies and what researchers dealt with on the subject of deepfakes. Explainthe concepts of deepfakes. Counterfeiting methods and techniques and patterns through the techniques andalgorithms used in counterfeiting [3]. Nan Yan et al.(2021) , presented a stereo faking corpus which is created using the Haas effect technique. Two identification algorithms for fake stereo audio are proposed. One is based on Mel-frequency cepstral coefficient features and support vector machines. The other is based on a specially designed five-layer convolutional neural network [5]. Hira Dhamyal *et al.(2021)* , suggest the microfeatures as standalone features for speaker-dependent forensics, voice biometrics, and for rapid pre-screening of suspicious audios, and as additional features in bigger feature sets for computationally intensive classifiers. This uses an image processing approach combined with deep learning which detects the inconsistency that exists in fake media [6]. Dora M. Ballesteros *et al.*(2021) , proposed a solution based on a Convolutional Neural Network (CNN), using image augmentation and dropout[2].

In previous studies it can be seen that various techniques, including speech synthesis, deepfake, and audio manipulation, can be employed to produce fake sounds. The designs are not expressly designed to withstand each kind of attack, CNNs and RNNs might not be equally effective in tackling the mall. Adversarial assaults, in which minute skillfully constructed changes to the input cause misclassifications, can affect both CNNs and RNNs. This vulnerability could be used to produce misleading audio that avoids detection in the context of fake audio detection.

According to previous research, CNNs typically require a fixed size input. If the size varies, resizing or cropping may be required, resulting in information loss. Also, when the receptive field is too small, CNNs struggle to capture global context in images. RNNs process input sequentially, limiting their parallelization capabilities and resulting in slower training than other architectures. CNN and RNN both require large amounts of labelled data for effective training and mayunderperform when data is scarce.

III. Problem Statement

To analyze fake audio fetched from different online resources. The proposed model will be a hybrid CR-NN model created by combining the limitations of each model to achieve a higher accuracy in detecting fake audio files. Combining CNNs and RNNs improves fake audio detection by leveraging CNNs for spatial features and RNNs for temporal dependencies, resulting in a comprehensive and adaptable solution for effectively capturing manipulation patterns.

The mathematical notation for Convolutional Neural Network is as following:

$$(f*g)(I,j) = \Sigma_m \Sigma_n f(m,n) \cdot g(i-m,j-n) \qquad Eq(1)$$

In the above equation *Eq(1)* where f is the input image, g is the convolutional kernel (filter), (i,j) are the spatial coordinates, and * denotes the convolution operation.

The mathematical notation for Recurrent Neural Network is as following:

$$h_t = \tanh(W_{hh} \cdot h_{t-1} + W_{xh} \cdot x_t + b_h) \qquad Eq(2)$$

In the above equation *Eq(2)* where h_t is the hidden state at time t. W_{hh} is the weight matrix for the hidden state. h_{t-1} is the hidden state at the previous time step. W_{xh} is the weight matrix for the input. X_t is the input at time t. b_h is the bias term for the hidden state. tanh is the hyperbolic tangent activation function.

IV. Proposed Methodology

A hybrid model proposed to overcome the limitations of the CNN and RNN models. The proposed model works in four steps. The data is collected, then the collected data preprocessed to remove any unwanted values. The useful features from the dataset are extracted. After preprocessing the dataset is trained and tested according to the hybrid model to get the required results. The results are then evaluated on the basis of different evaluation metrics such as accuracy, precision and f1 score.

4.1 Phase 1: Data Collection

The proposed hybrid is used to perform analysis using Real-time dataset ASVspoof(2019) collated from Kaggle. The ASVspoof 2019 consists of 14,000 audio clips, each which 1 second ofaudio data. The audio clips are divided into two categories namely bona fide (real) clips and spoofed (fake) clips.

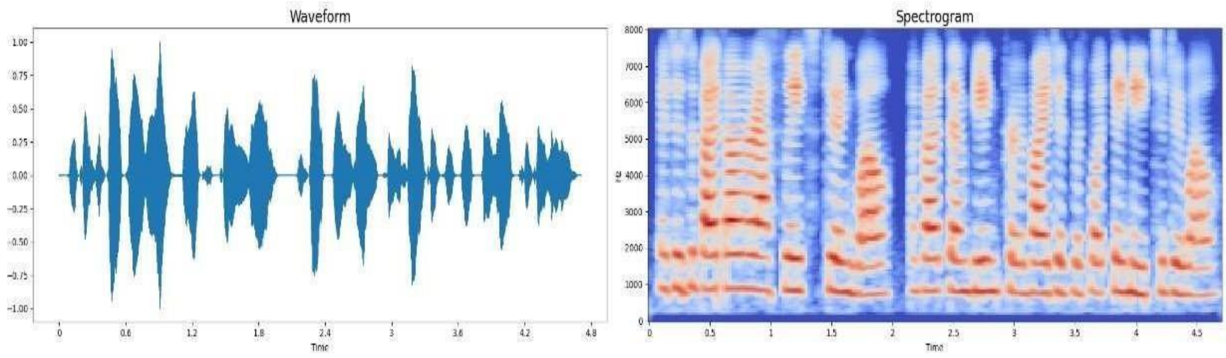
AudioFile	Size	Duration	Category
LA_0039	16 KHz	20 seconds	Spoof
LA_0069	16 KHz	20 seconds	Bonafide
LA_0014	16 KHz	20 seconds	Spoof
LA_T_236175	16 KHz	20 seconds	Bonafide

Table 1: Details of the ASVspoof(2019) dataset

4.2 Phase 2: Data preprocessing

This phase involves various techniques to clean, transform, and standardize the data to enhance its quality and consistency. Exploring the dataset to understand its structure, data types, distribution of features, and presence of missing values and outliers [7] Outliers are data points that deviate significantly from the majority of data in a dataset. This initial analysis helps identifypotential issues that need to beaddressed. Handle missing values by either removing incomplete samples, imputing missing values using appropriate techniques, or encoding them as a separate category [8]. Address outliers by either removing them or applying outlier detection algorithms like Z- score, isolation forest etc. Normalize numerical features to a common scale, such as min- max scaling or z-score normalization.

The following Fig.(1) was analyzed for fake audio detection during the preprocessing of the dataset as thefigure shows the lag between the time and the data is not grouped together. The X-axis depicts the time and the Y axis shows the amplitude. In the real audio it can be seen that the amplitude has greater number of high peaks as compared to the fake audio.

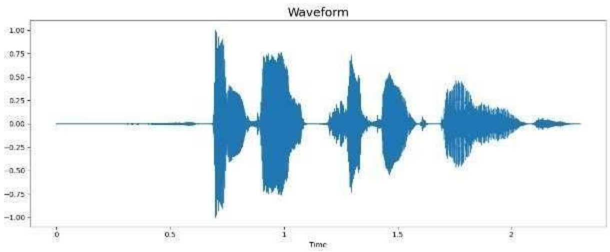


(a) Waveform (b) Spectrogram

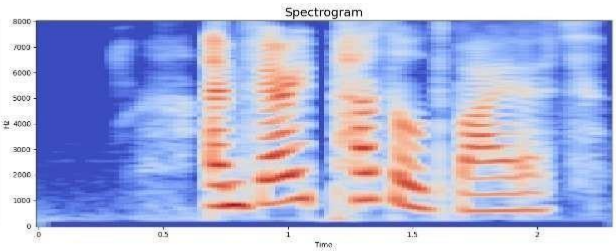
Fig.1 Fake Audio Features

This ensures that features with larger scales don't dominate the model's learning process. Extract relevant features from the raw audio data. Common feature extraction methods include Mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLP), and frequency-domain features. In this model MFCCS and audio waveforms are used . MFCCs capture the spectral characteristics of audio signal which represent the distribution of energy across frequency bands. Whereas audio waveforms provide a time domain view of the audio signal i.e., how the amplitude changes over time. Selecting the most relevant and informative features using techniques like feature correlation analysis or machine learning-based feature selection methods [9]. This helps reduce dimensionality and improve modelperformance. Divide the pre-processed data into training, validation, and testing sets. The training set is used to fit the machine learning model, the validationset is used for hyperparameter tuning and model selection, and the testing set is used for unbiased evaluation of the model's performance [10]. Consider applying data augmentation techniques to artificially increase the size and diversity of the training data. This can be particularly beneficial for datasets with limited samples.

The following Fig.(2) was analyzed for fake audio detection during the preprocessing of the data set as the figure shows not having lag in the time and the data clustered. The X- axis depicts the time and the Y axis shows the amplitude.



(a) Waveform



(b) Spectrogram

Fig.2 Real Audio Features

The results have been analyzed on the basis of experiment performed. The following Fig 3and Fig. 4 are Mel-frequency cepstral coefficients (MFCCs) and Audio waveform respectively. The X- axis depicts the time and the Y axis shows the amplitude.

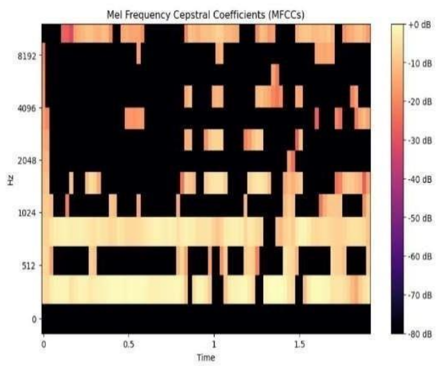


Fig.3 Mel-frequency cepstral coefficients

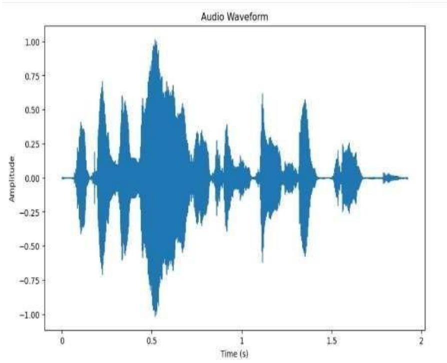


Fig.4 Audio Waveform

4.3 Phase 3: Proposed Hybrid CR-NN model

Recurrent Neural Networks (RNNs) are a type of neural network architecture designed for sequence data, where the output not only depends on the current input but also on the previous inputs in the sequence. Thebasic equation for a simple RNN can be expressed as follows: *Eq.(3)* and *Eq.(4)*

$$h_t = f(W_{hh} * h_{t-1} + W_{xh} * x_t + b_h) \quad Eq.(3)$$

$$y_t = g(W_{hy} * h_t + b_y) \quad Eq.(4)$$

Where Whh is the weight matrix that governs the transition from one hidden state to the next, W_{xh} is the weight matrix that shows input to hidden state transition, x_t is the input, b_h is the bias term for hidden state and f is the activation function. W_{hy} is the weight matrix transforming hidden state to output, h_t represents the hidden state, b_y represents the bias term for output and g is the activation function.

The equation $h_t = f(W_{hh} * h_{t-1} + W_{xh} * x_t + b_h)$ represents the update of the hidden state based on the current input x_t and the previous hidden state h_{t-1}
The equation $y_t = g(W_{hy} * h_t + b_y)$ represents the computation of the output y_t based on the current hidden state h_t.
CNN is a type of neural network architecture commonly used for image recognition and computer vision tasks. The basic building blocks of a CNN include convolutional layers, pooling layers, and fully connected layers. Here's a simplified equation to represent the forward by an activation function can be represented as: *Eq.(5)*

$$Z=f(W*X+b) \qquad Eq.(5)$$

For a pooling layer, the equation is simpler, typically involving a max or average pooling operation shown by *Eq.(6)*

$$Y=\text{Pooling}(Z) \qquad Eq.(6)$$

Where X is the input vector, W is the weight matrix, b is the bias vector, f is the activation function and Z is the output.
This process is repeated through multiple convolutional and pooling layers, and eventually, the outputis flattened and passed through one or more fully connected layers. The equations for fully connected layers are similar to those in a standard neural network: *Eq.(7)*

$$A=f(W*X+b) \qquad Eq.(7)$$

W is the weight matrix, X is the input vector, b is the bias term, and f is the activation function.

4.4 Phase 4: Evaluation

The results are classified into two categories fake audio and real audio. The evaluation metrics used toevaluate the model are accuracy, precision , recall, f1-score.
The degree to which a value is accurate in respect to the information is measured by its accuracy.The formula of accuracy is shown by *Eq(8)*

$$\text{Accuracy}=\frac{\text{Number of Correct Predition}}{\text{Total Number of Predition}} * 100 \qquad Eq.\ (8)$$

The degree of detail that a value conveys is known as precision. The formula of precision is shown by *Eq(9)*

$$\text{Precision}=\frac{\text{True Positives}}{\text{False Positives+True Positives}} \qquad Eq.(9)$$

Where Ture Positive is the number of true positives (instances correctly predicted as positive),False Positive is the number of false positives (instances incorrectly predicted as positive). High precision indicates that the model has a low rate of false positive predictions, meaning that when it predicts a positive outcome, it is predicted to be correct.

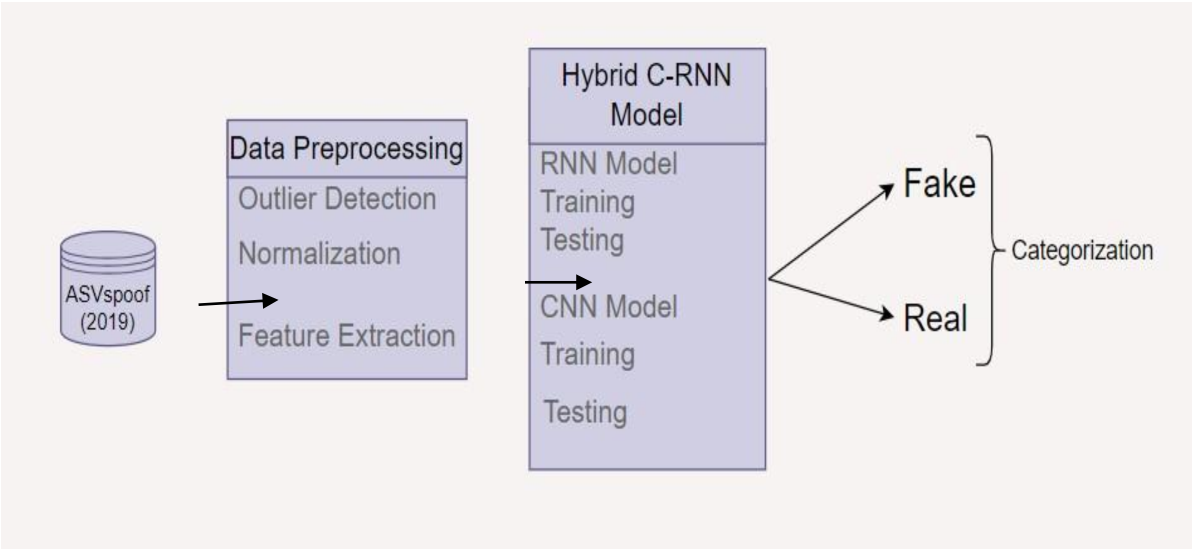


Fig . 5 Proposed Model

V. Experimental Results

In this section, results have been analyzed on the basis of the experiments performed. Training Accuracy: Indicates how well the model learned from the training data by measuring performanceon the training set, validation accuracy aids in hyperparameter tuning and helps prevent overfittingand testing accuracy assess performance on separate testing set not seen by the model during training . The accuracy of R- NN classifier model is 88% as shown in Fig.5 precision is 80% as shown in Fig.6 and F1-score is 89.041%as shown in Fig.7 respectively. In Fig. 5 the X_1 axis represents epochs the Y_1 axis shows loss the X_2 axis shows epochs and Y_2 axis represents accuracy.

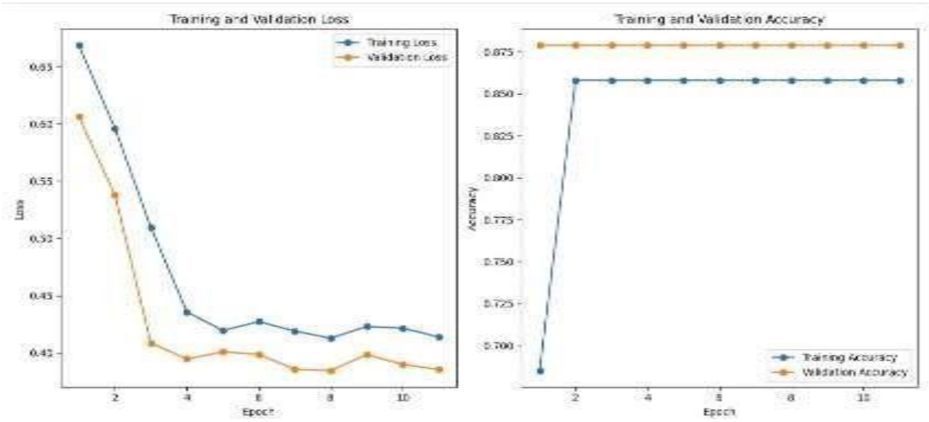


Fig.6 Accuracy of RNN

In Fig.6 the X_1 axis represents epochs, the Y_1 axis shows loss the X_2 axis shows epochs and Y_2 axis represents precision.

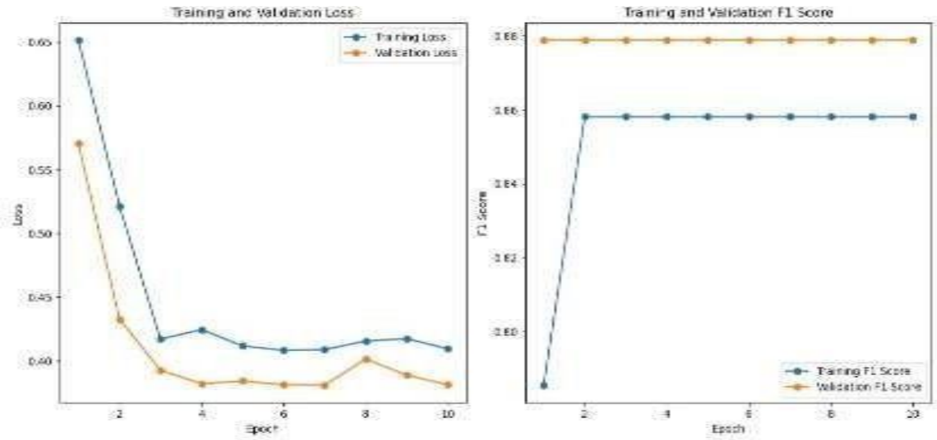


Fig.7 Precision of RNN

In Fig.7 the X1 axis represents epochs, the Y1 axis shows loss the X2 axis shows epochs and Y2 axis represents F1 score.

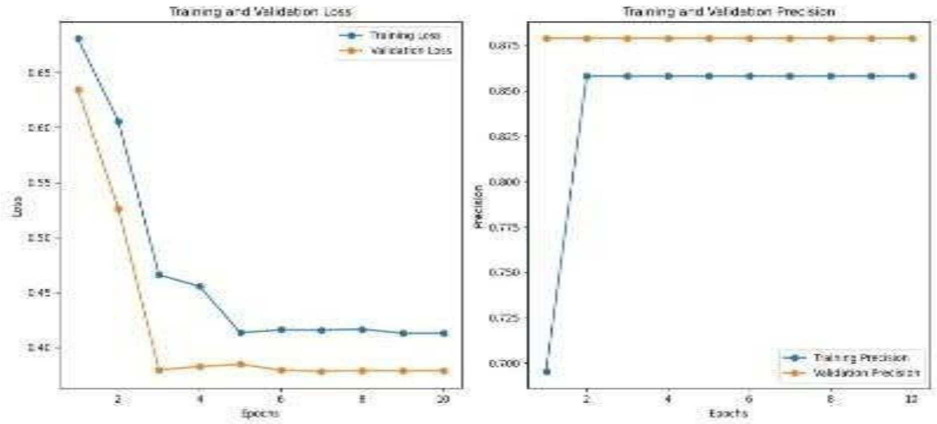


Fig.8 F1-score of RNN

The accuracy of CNN is 84.29 % as shown in Fig.8, precision is 82 % as shown in Fig.9and F1- score is 90.66 % as shown in Fig.10 respectively. In Fig.8 the X1 axis represents epochs, the Y1 axis shows loss the X2 axis shows epochs and Y2 axis represents accuracy.

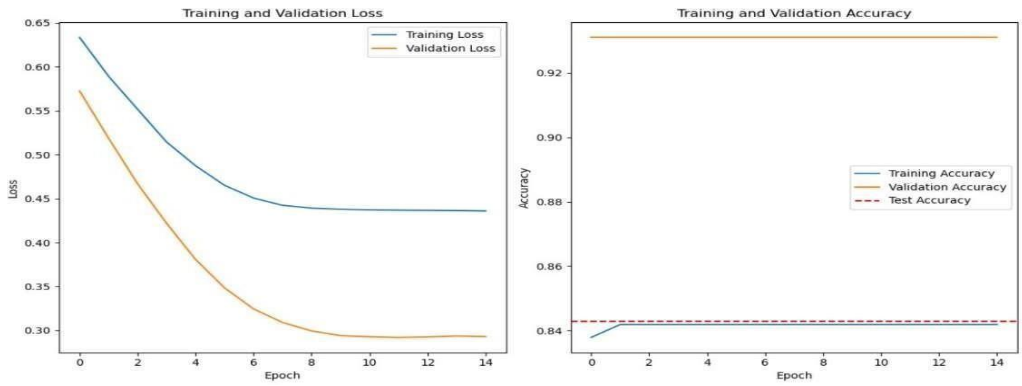


Fig.9 Accuracy of CNN

In Fig.9 the X1 axis represents epochs, the Y1 axis shows loss the X2 axis shows epochs and Y2 axisrepresents precision.

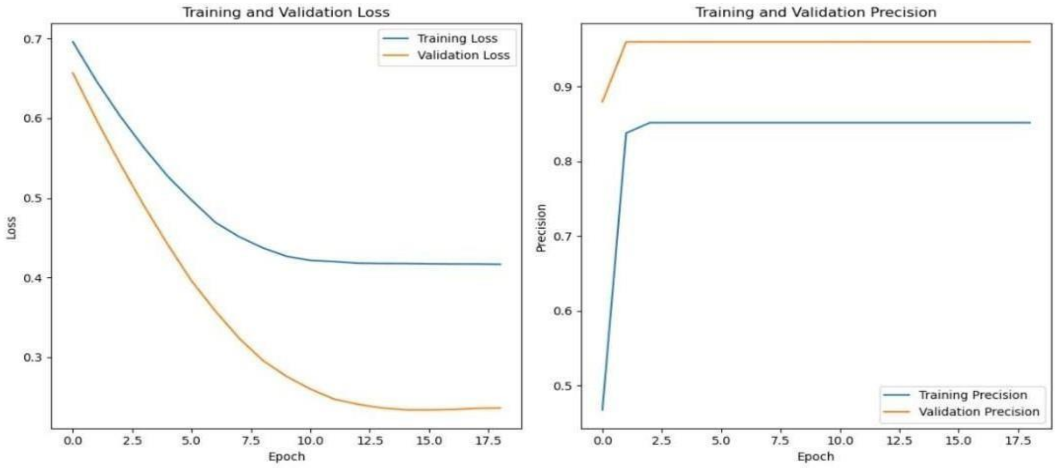


Fig.10 Precision of CNN

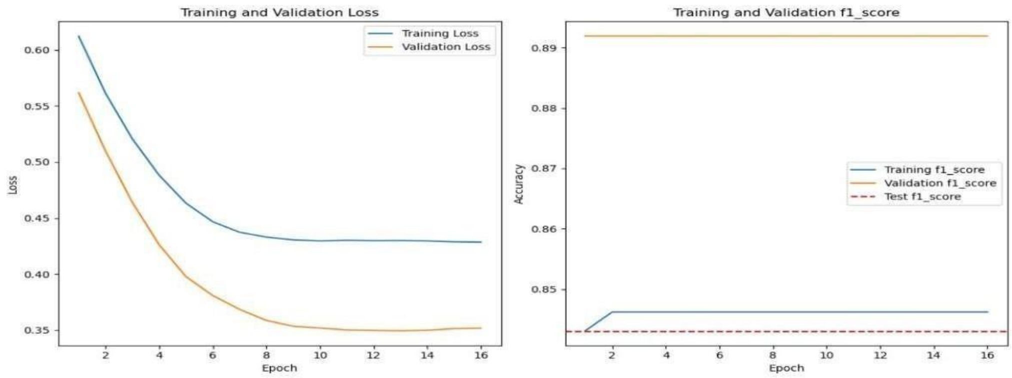


Fig.11 F1-score of CNN

In Fig.10 the X_1 axis represents epochs, the Y_1 axis shows loss the X_2 axis shows epochs and Y_2 axis represents F1 score.

The accuracy of the proposed hybrid CR-NN model is 86.42% as shown in Fig.11, precision is 80.5% as shown in Fig.12 and F1-score is 89.18% as shown in Fig.13. In Fig. 11 the X- axis shows the epochs and Y-axis shows the accuracy.

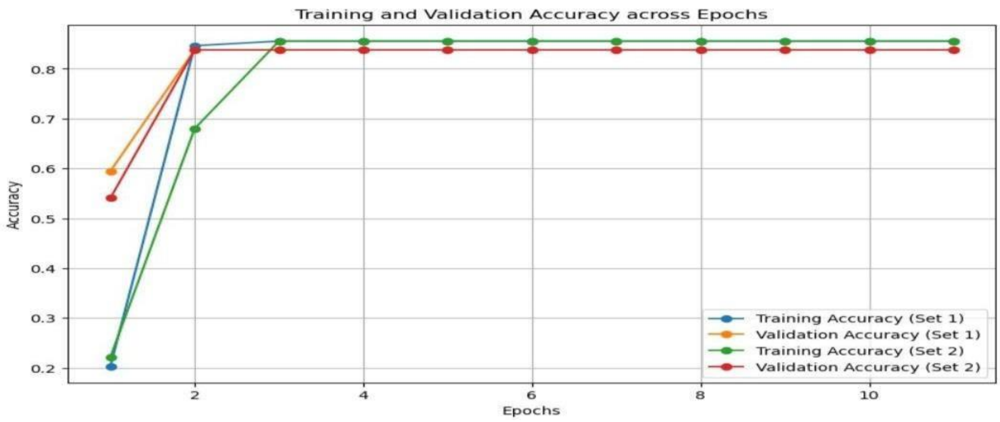


Fig.12 Training and validation score of hybrid CR-NN Model.

In Fig. 12 the X- axis shows the epochs and Y- axis shows the precision.

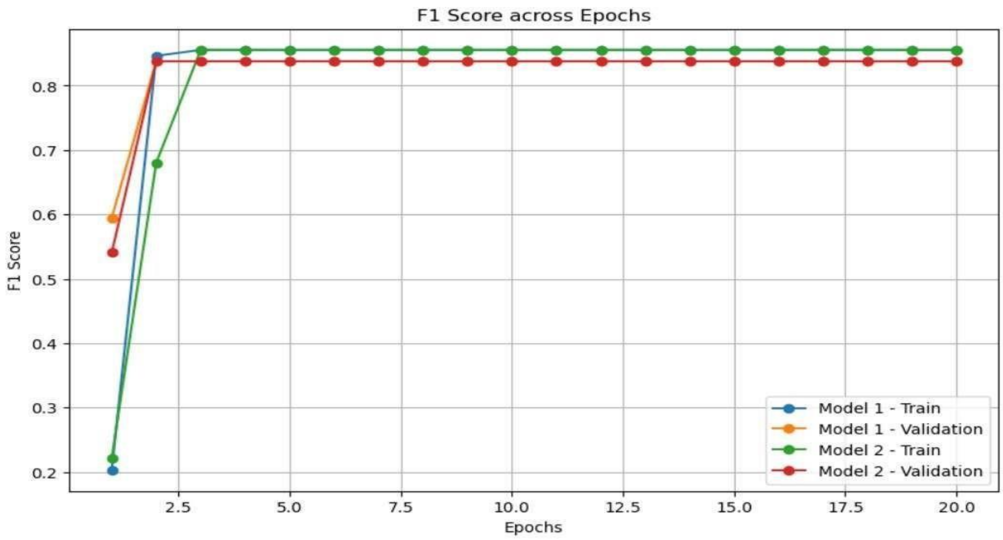


Fig.13 Precision of hybrid CR-NN model In Fig. 13 the X- axis shows the epochs and Y- axis shows the F1 score.

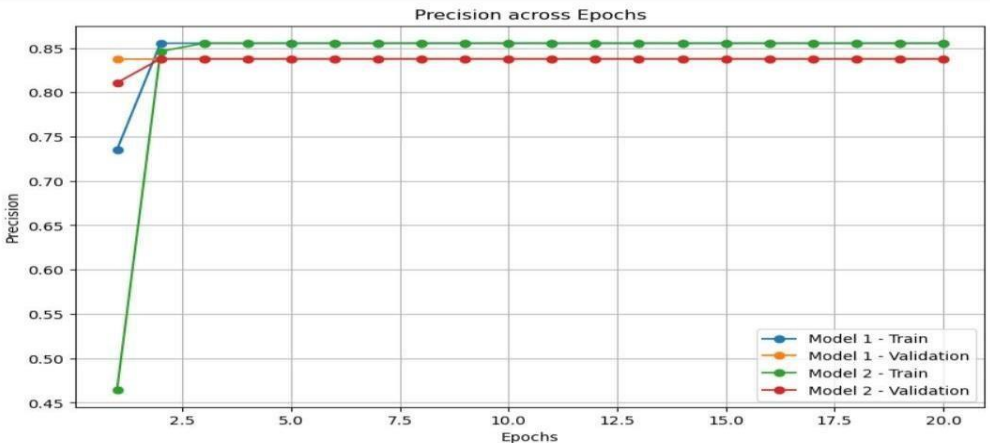


Fig.14 F1-score of hybrid CR-NN model

VI. Conclusion and future scope

The hybrid CR-NN approach helped to overcome the limitations of CNN and RNN which was used for deepfake detection. The accuracy achieved by the proposed model is 86.42% , precision is80.15% and recall rate is 89.18%.

The hybrid CNN/RNN approach can be useful in tasks with both spatial and temporal dependencies. Further advancements and applications for this hybrid model are likely to emerge across various domains as technology and research progress.

References

[1]. Frank J, Schönherr L. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. 2021;

[2]. Ballesteros DM, Rodriguez-Ortega Y, Renza D, Arce G. Deep4SNet: deep learning for fake speechclassification. Expert Syst Appl 2021; 184.

[3]. Abdulreda AS, Obaid AJ. A landscape view of deepfake techniques and detection methods. International Journal of Nonlinear Analysis and Applications 2022; 13: 745–755.

[4]. Chintha A, Thai B, Sohrawardi SJ *et al.* Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. IEEE Journal on Selected Topics in Signal Processing 2020; 14:1024–1037.

[5]. Liu T, Yan D, Wang R, Yan N, Chen G. Identification of fake stereo audio using svm andcnn.Information (Switzerland) 2021; 12.

[6]. Dhamyal H, Ali A, Qazi IA, Raza AA. Fake audio detection in resource-constrained

settings using microfeatures. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, International Speech Communication Association 2021, 3646–3650.

[7]. Nasar BF, Sajini T, Lason ER. Deepfake Detection in Media Files - Audios, Images and Videos. 2020IEEE Recent Advances in Intelligent Computational Systems, RAICS 2020, Institute of Electrical andElectronics Engineers Inc. 2020, 74–79.

[8]. Wijethunga RLMAPC, Matheesha DMK, Noman A Al, De Silva KHVTA, Tissera M, Rupasinghe L. Deepfake audio detection: A deep learning based solution for group conversations. ICAC 2020 - 2nd International Conference on Advancements in Computing, Proceedings, Institute of Electricaland Electronics Engineers Inc. 2020, 192–197.

[9]. Sohrawardi SJ, Seng S, Chintha A *et al.* Poster: Towards robust open-world detection of deepfakes. Proceedings of the ACM Conference on Computer and Communications Security, Association for Computing Machinery 2019, 2613–2615.

[10]. Ali M, Sabir A, Hassan M. Fake audio detection using Hierarchical Representations Learning and Spectrogram Features. 2021 International Conference on Robotics and Automation in Industry, ICRAI 2021, Institute of Electrical and Electronics Engineers Inc. 2021.

[11]. S. Kaur, P. Kumar, and P. Kumaraguru, “Automating fake news detection system using multi-level voting model,” *Soft comput*, vol. 24, no. 12, pp. 9049–9069, Jun. 2020, doi: 10.1007/s00500-019-04436-y.

[12]. S. Kaur, P. Kumar, and P. Kumaraguru, “Detecting clickbaits using two-phase hybrid CNN-LSTM biterm model,” *Expert Syst Appl*, vol. 151, Aug. 2020, doi: 10.1016/j.eswa.2020.113350.

[13]. S. Kaur, P. Kumar, and P. Kumaraguru, “Automating fake news detection system using multi-level voting model,” *Soft comput*, vol. 24, no. 12, pp. 9049–9069, Jun. 2020, doi: 10.1007/s00500-019-04436-y.