MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

Deepfake Audio Detection Using Spatio-Temporal Learning in Hybrid GAN Frameworks

Vivek Ranjan[1*],
Department of Information Technology and Computer Applications,
Madan Mohan Malaviya University of Technology, Gorakhpur, India, 273010
Email : vivekalive05@gmail.com
Dayashankar Singh[2],
Department of Information Technology and Computer Applications,
Madan Mohan Malaviya University of Technology, Gorakhpur, India, 273010
Email : dssitca@mmmut.ac.in

Abstract
As deepfake audio detection becomes increasingly sophisticated, they pose considerable challenges to digital security, privacy, and trust. Therefore, effective means of detection are required. This study proposes a Hybrid Generative Adversarial Network (GAN) framework with spatio-temporal feature learning that utilizes encoding, preprocessing, and spectrogram images to extract temporal and spectral features, which are then passed through a series of steps that include: (1) convolutional neural network (CNN)-based spatial feature learning, (2) sequence-to-sequence encoding with Transformer and recurrent neural networks (RNN) based temporal modeling, and (3) classification with attention-based GAN. To evaluate performance, based on traits about the real and fake, the Hybrid GAN produced results of 96% accuracy with 98% (Real) precision, 94% (Real) recall, 94% (Fake) precision, and 98% (Fake) recall, and AUC 0.995 are considerably better than standard RNN and CNN+LSTM. Overall, using a Hybrid GAN to extract spectral-temporal dependencies provided strong discrimination, reduced misclassification, and relatively consistent detection. This framework shows that Hybrid GANs are a potentially state-of-the-art approach to deepfake audio detection and has applications in multimedia authentication, digital forensics, and cybersecurity.

Keywords: *Deepfake Audio, Hybrid GAN, Spatio-Temporal Feature Learning, CNN, RNN, Transformer, Audio Forgery Detection, Generative Adversarial Network, Deep Learning, Audio Authentication.*

1. Introduction

Deep learning is used to make false films and sounds that mimic actual people's behavior, pronunciation, and appearance (Farid, et al. (2022)) [1]. The technique took off with bogus celebrity videos that disguised their voices and looks. It has moved to audio in recent months, allowing for very realistic voice mimics. This system uses a Generative Adversarial Network, which combines two neural networks: the generator, which generates bogus data, and the discriminator, which detects it. Results are usually identical to genuine content. Deepfake systems record and replicate pitch, tone, cadence, and accent using complicated autoencoder, GAN, and Transformer architectures (Mubarak, et al. (2023)) [2]. Since more public material (interviews, podcasts, videos) is accessible, producing phony voices is theoretically conceivable and alarmingly simple. Audio deepfakes threaten privacy and security. Hackers have used cloned voices to bypass speech authentication systems in banks and personal security systems, trick relatives and coworkers, and modify audio evidence in legal or journalistic cases.

Deepfake audio may damage your reputation, wealth, and social trust and disseminate misinformation. The lack of perceptual cues makes synthetic audio forgeries harder to detect (Khan, et al. (2024)) [3]. Video deepfakes may have visual defects. The majority of voice-based assaults occur in real time, making it difficult for humans or systems to detect and react. The ability to synthesize voices without consent breaches people's identification rights and might lead to harassment or extortion, making privacy even more important. As generative models improve and datasets grow, synthetic audio can function with more languages, emotions, and acoustic settings, making it difficult to distinguish between genuine and false recordings. From bogus government claims and emergency warnings that threaten national security to daily frauds like impersonating family members or CEOs, audio deepfakes have many uses (Bateman, et al. (2022)) [4]. These hazards have highlighted the necessity for effective detection techniques that can identify minute changes in speech. Techno protections, regulatory measures, and societal awareness are needed to address these threats. Deepfake technology has creative and commercial uses in entertainment and accessibility (including voice synthesis for disabled people and automated dubbing), but these unacceptable audio risks require urgent research and mitigation for digital security and privacy.

1.1. Motivation for Deepfake Audio Detection

Recently developed AI techniques can manufacture believable voices (Lyu, et al. (2020)) [5]. These technologies were designed to help people, but they are being used to propagate misinformation worldwide via audio, causing dread of the "Audio Deepfake." Mobile devices and PCs are making audio deepfakes, or manipulations, more accessible. This has raised global cybersecurity worries about AD adverse effects. ADs go beyond text messages and email links, notwithstanding their benefits. People may employ logical-access voice spoofing (Diakopoulos, et al. (2021)) [6] to alter public opinion for propaganda, slander, or terrorism. Detecting fraudulent voice recordings from the massive numbers broadcast everyday online is difficult (Rodríguez-Ortega, et al. (2020)) [7]. AD attackers have attacked politicians, governments, and people. In 2019, AI-based software impersonated a CEO's voice to steal over USD 243,000 over phone. To prevent deception, research must authenticate transmitted audio recordings. This issue has garnered scholarly attention in recent years (Chen, et al. (2020)) [8]. Synthetic-based, imitation-based, and replay-based AD have arisen, making detection harder.

There are a plethora of detecting algorithms that can tell the difference between actual speech and deepfakes. In order to identify phoney music, researchers have created a number of ML and DL models that use various techniques. As shown in Figure 1, the following tactics outline the whole process of AD detection. Preprocessing and transformation of each audio clip into appropriate audio characteristics, such Mel-spectrograms, should be the first step. After receiving these characteristics, the detection model goes about doing things like training. Any fully connected layer with an activation function may be used to provide a prediction probability of class 0 as fake or class 1 as genuine from the output, which is useful for nonlinearity tasks. But there is a cost-benefit analysis of computer complexity vs accuracy. Additional efforts are necessary to enhance the efficacy of Alzheimer's disease detection and address the deficiencies noted in the literature.

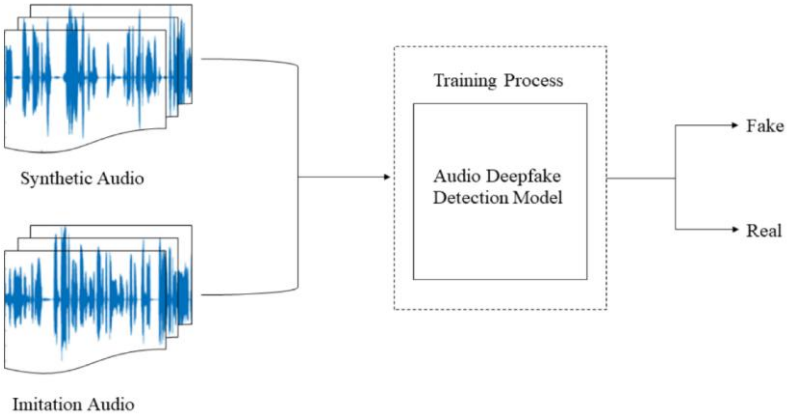

Fig. 1: An illustration of the AD detection process (Thies, et al. (2019)) [9]

1.2. Bimodal Temporal Feature Prediction for Robust Audio-Visual Deepfake Detection: Current methods for detecting deep fakes involve anomalous visual modalities in videos. However, deepfake generative algorithms continue to increase visual fraud (Jiang, et al. (2020)) [10] and audio tampering in manufactured videos. Compression and other interruptions sometimes hide forensic evidence in manipulated movies. This makes unimodal detection methods like hybrid borders, frequency anomalies, intra-modal inconsistencies, and facial or lip motion aberrations less effective. Thus, accurate and efficient multimodal detection approaches using auditory and visual data are crucial. Multimodal learning lets the detector employ inconsistent artifacts across modalities. This helps the detector increase its detection skills and provide reliable results, which is an effective solution to handle the concerns outlined above.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

Recent research has examined multimodal deepfake detection using audio and visual data. (Liu, et al. (2023)) [11] Classifying data with mixed unimodal labels enhanced detection granularity. However, audio–visual links have not been fully used. There are gaps in the modal distribution since some studies (Su, et al. (2020)) [12] have proposed creative techniques to integrating these features, but these approaches do not effectively address the underlying differences between audio and visual signals. Current approaches that focus on certain disparities, such as emotional indicators, may become useless if forgers can adapt their strategies to hide these deviations. The current audio–video bimodal forgery detection systems have this problem, since training a network approach requires learning the discrepancy between the two modalities. This applies whether the network strategy is trained via comparison learning or self-supervision. These methods for deep fake detection are limited in scope and depth by the underlying mental model. Current multimodal learning methods fail to account for out-of-the-ordinary signals, particularly those pertaining to time series. Small traces of forging are therefore undetectable. Prior work has made progress by using differences in temporal features and intra-frame artifacts. Therefore, to bridge the knowledge gap and use movies' intra-modal artifacts and inter-modal inconsistent elements more effectively, bimodal identification should benefit from integrating artifact characteristics (Li, et al. (2020)) [13], which have been shown to be useful in unimodal detection.

This work addresses past problems by presenting a bimodal temporal feature prediction-based deep forgery detection approach. This technique improves several crucial aspects, including: A temporal feature-based characterisation prediction module is suggested. The input video may be divided into many segments so it can infer correlations between global features over longer periods of time. To forecast future representations of sequences, it is possible to mine visual and auditory regularities using long-term global characteristics that span several time steps. Also under development are modules for video and audio prediction. Timing abnormalities may be captured in a single modality by projecting future timing characteristics and comparing them to reference features. This research presents a new idea—incorporating the contrast loss function into the overall objective loss function. You can tell the difference between the two using the contrast loss function, which was trained to identify the most noticeable difference between authentic and fraudulent movies. Audio and video are best handled by a network with a projection layer. Combining and aligning bimodal characteristics from nearby videos and audio samples is done by this network. This network fixes the video-audio dimension mismatch.

1.3 Role of Spatio-Temporal Features in Audio Analysis: Audio analysis relies on spatio-temporal elements to collect and understand frequency and time. In sound event detection (SED), spatio-temporal characteristics help manage overlapping sound events using spatial, frequency, and time domain information. Recent approaches, such 3D convolutional networks (CNN), use spatio-temporal properties to use multi-dimensional audio data (spectrograms) better than temporal or spectral methods. Spatial-temporal approaches have helped uncover deepfake audio discrepancies. A popular method is to acquire spectral characteristics using CNNs, then evaluate the time-domain aspects with “recurrent neural networks (RNNs) or long short-term memory (LSTM)”, which reveals a temporal disparity or divergence during speech. In recent self-supervised learning research, spatio-temporal models like Spatio-Temporal Vision Graph Non-Contrastive Learning (SIGNL) use a graph structure to analyze the temporal relationship between audio spectrogram patches (in height and width), improving deepfake audio detection with little labeled data. In addition, spatio-temporal fusion of audio and video information, such as matching spoken audio with deepfake video lip movements, improves multimodal detection effectiveness. Deepfake material often has lip-sync mismatches, which spatio-temporal feature fusion enables the algorithm to use them. While mostly focused on deepfake detections, spatio-temporal representations for audio analysis are already showing promise for sound event localization, audio source separation, and speech recognition tasks, where considering both sound temporal evolution and spatial relationship can improve model accuracy and generalizability. Despite breakthroughs in incorporating spatio-temporal features into audio analysis, studies still need to operationalize the paradigm to real-time criteria while maintaining real-life requirements, reduce dependence on large labeled datasets, and improve model robustness in noisy/diverse recording environments. Spatial-temporal features have enabled many research insights that address fundamental problems of detecting manipulated media and understanding complex audio environments, which may lead to a game-changing research trajectory.

2. Review of Literature

Gao, Yuan et al. (2024) [14] researched quickly emerging deepfake technology, which makes realistic modified media, which might be exploited. Reliable detection technologies are desperately required to avoid fraud. Current techniques use single modalities or audio–visual signal fusion, limiting accuracy. The used bimodal temporal feature prediction-based deepfake detection approach innovatively incorporates temporal feature prediction into audio–video bimodal analysis to fully leverage audio–visual temporal laws. First, input quadruples are produced from nearby audio–video sequence clips, then a dual-stream network extracts temporal feature representations from video and audio. Video and audio prediction modules estimate future temporal data and compare them to reference features to identify modality-specific temporal differences. A projection layer network aligns audio–visual information using contrastive loss functions for contrastive learning and maximal video modality differences. Used deepfake detection technique surpasses prior methods with 84.33% accuracy and 89.91% AUC on FakeAVCeleb.

Jbara, Wurood A. et al. (2024) [15] explained As privacy breaches, disinformation, and digital media integrity risks increase, deepfake (DF) technology has become a major problem. Deep learning (DL) algorithms to recognize DFs have made great strides, however owing to rapidly evolving DF creation technology, studies still struggle to discriminate between actual and changed information. This work addressed two major issues: reviewing existing DF detection algorithms and obtaining high accuracy with little computational drag. A thorough literature review evaluated DF detection techniques for efficiency, performance, and resilience. CNNs, LSTM networks, hybrids, and other specialized techniques like speech spectrum analysis and phonetics were studied. Some conclusions were accurate (94% in nuanced contexts) but not generalizable to DF use cases outside their research techniques. Other hybrid codes (CNN with LSTMs) tend to balance accuracy and computational economy better.

Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen (2023) [16] emphasized that proper detection is crucial as deepfake technology advances. Recent deep learning-based detection systems fail to generalize beyond training data. To understand the generalization difficulty, this study explored deep learning model architectures, pre-training methods, and datasets. Comprehensively compare supervised and self-supervised deep learning models for deepfake detection. Eight supervised deep learning architectures and two transformer-based models pre-trained self-supervisedly (DINO, CLIP) are evaluated on four deepfake detection benchmarks. This research examined intra- and inter-dataset evaluations to determine the best models, datasets that optimize trained models' generalization, and image augmentations on model performance. The study analyzed model size, efficiency, and performance trade-offs. The study will assess transformers, CNNs, supervised, self-supervised training methods, and deepfake detection standards. Detailed empirical study demonstrates Transformer models beat CNN models in deepfake detection.

John, Jerry, and Bismin V. Sherif (2022) [17] determined that deep learning addresses many real-world problems but has limits. The extensive usage of deepfakes is new and sophisticated. Deep learning creates image and video deepfakes. Digital fraud, extortion, pornography, and deepfakes use source images to superimpose a targeted person's face. Deep learning makes manually differentiating genuine from phone tougher. Development and research in deepfake detection are crucial. This study fully discussed deep feature-based, temporal-based, and feature-based deepfake detection. Comparison study focuses on face detection architecture, deep learning architecture, image-based vs. video-based, dataset, frame size, and dataset size. The work develops and examines a semi-supervised GAN architecture to recognize deepfake pictures.

Nguyen, Xuan Hau et al. (2021) [18] discovered that Deepfake has enabled hyper-realistic face synthetic movie creation in recent years. People distrust video material due to harmful online dissemination of bogus videos. Some algorithms have been developed for recognizing Deepfake-created fake films, however most have focused on frame characteristics. Because they overlook spatio-temporal aspects, these algorithms are inaccurate. This study presented a 3-D convolutional neural network model for learning spatiotemporal properties from surrounding video frame sequences. The study offered a network with binary detection accuracy above 99% on Deepfake of Face Forensics++ and VidTIMIT, the two largest benchmark datasets. Compared to state-of-the-art methods, experimental findings demonstrate the proposed technique is better.

Jaiswal, Gaurav. (2021) [19] explained that these days, deepfake videos raise issues related to security, privacy, and social ethics. Deepfake films are artificially produced videos that are created by altering the audio and facial characteristics of one individual to overlay their information on other videos. Fraud and defamation are two possible uses for these videos. Therefore, the detection of deepfake films is necessary to combat these kinds of manipulations and dangers. In order to detect deepfake videos, this research suggested multilayer hybrid recurrent deep learning models. The suggested models take advantage of hybrid recurrent deep learning models' temporal learning and noise-based temporal face convolutional features. These models outperform layered recurrent deep learning models, according to experiment data.

Mehra, Akul (2020) [20] claimed advances in graphics processing and AI algorithms have made phony media easier to create. Deep learning algorithms like Face-Swap and Deepfakes allow anybody to produce fake films by changing the target's voice or face. Deepfakes are used for fake news and phishing. Identifying face tampering in genuinely edited videos is key. This study proposed a spatiotemporal hybrid model of Capsule Networks and LSTM Networks and describes deepfake-

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

induced movie discrepancies. This approach used inconsistencies to identify phony and legitimate videos, contributing to deepfake detection. Visualising the capsule activation helps researchers understand what the capsules learn and identify real from deepfake films. Three different frame selection procedures also considerably impact performance, according to the study. Due to its one-fifth parameter count and one-fourth size, the utilized model is lighter and uses less computational power than the state-of-the-art model, although having almost equal performance.

Table 1 displays the Techniques, Research Gaps and Findings of the respective research papers in the terms of deepfake audio detection.

Table 1: Approaches to Literature Review

No.	Reference	Techniques	Research Gap	Outcomes/Findings
[14]	Gao, Yuan et al. (2024)	Bimodal temporal feature prediction using video-audio sequence clips	Reliance on single modalities or limited audio-visual signal fusion.	Outperformed previous methods with 84.33% accuracy and 89.91% AUC on the FakeAVCeleb dataset.
[15]	Jbara, Wurood A. et al. (2024)	Systematic review of CNN, LSTM, hybrid models, spectral, and phonetic analysis	Difficulty in generalizing across diverse Deepfake applications.	Hybrid models (CNNs + LSTMs) offer better accuracy and efficiency; some methods reach 94% accuracy.
[16]	Khan, Sohail Ahmed & Duc-Tien Dang-Nguyen (2023)	Deep learning models, including supervised/self-supervised techniques (DINO, CLIP), transformers, and CNNs	The challenge of generalizing deep learning models over diverse datasets.	Transformer models outperform CNNs for deepfake identification.
[17]	John, Jerry, and Bismin V. Sherif (2022)	Different deepfake detection techniques based on features, time, and features	Lack of a unified deepfake detection approach across different datasets and architectures.	Suggested a semi-supervised GAN for detecting deepfake images; provided a detailed comparison of methods.
[18]	Nguyen, Xuan Hau et al. (2021)	Learning spatiotemporal properties from neighboring video frames using a 3D CNN	Limited focus on spatiotemporal features, which reduces the performance in some cases.	Achieved over 99% accuracy on the Face Forensics++ and VidTIMIT datasets; outperformed state-of-the-art methods.
[19]	Jaiswal, Gaurav (2021)	Models for multi-layer hybrid recurrent deep learning; temporal face convolutional features based on noise	Need for efficient temporal learning and noise handling in hybrid recurrent models.	Demonstrated performance improvements over stacked recurrent models.
[20]	Mehra, Akul (2020)	Spatio-temporal hybrid model using Capsule Networks and LSTM Networks	Lack of focus on frame selection techniques and model size efficiency.	Hybrid model achieves near state-of-the-art performance with reduced computational cost.

Identified Research Gaps:

A significant issue with audio analysis approaches is their ineffectiveness across many languages and speakers due to discrepancies in accents, phonetics, and speaking styles. Furthermore, the majority of models are incapable of simultaneously acquiring both spatial and temporal data, hence diminishing their efficacy in tasks like as sound localization and deepfake detection. These limits may result in considerable false positive rates when processing new data, leading to the erroneous identification of real content.

2 Proposed System Architecture

System Components

Audio analysis and deepfake detection may be improved using Hybrid GAN Modules, Transformer-based Temporal Encoders, and Spectrogram CNN Extractors. Reconstruction and adversarial training can distinguish genuine and fake audio using hybrid GANs. Transformer topologies in Temporal Encoders may detect minor speech and rhythm changes across time. Mel-spectrograms or STFT in Spectrogram CNN Extractors discover localized abnormalities in spatial-frequency features. A solid foundation from these modules improves detection, generalization, and false positives and negatives.

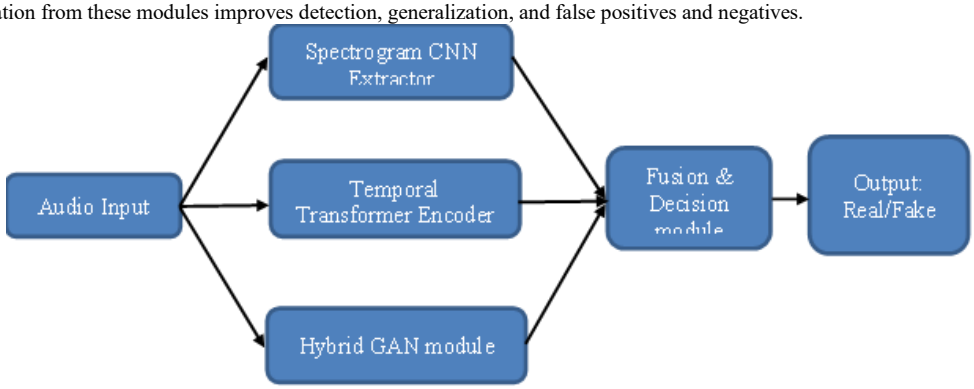

Figure 2: ***Proposed System Architecture for Audio-Based Deepfake Detection***

Figure 2 depicts the proposed design for a system that can find deepfake audio. It uses Hybrid GAN Modules, Transformer-based Temporal Encoders, and Spectrogram CNN Extractors to let you look at things from different angles. This integrated framework improves the accuracy of detection and the ability to apply it to a wide range of audio sources.

3. Research Methodology

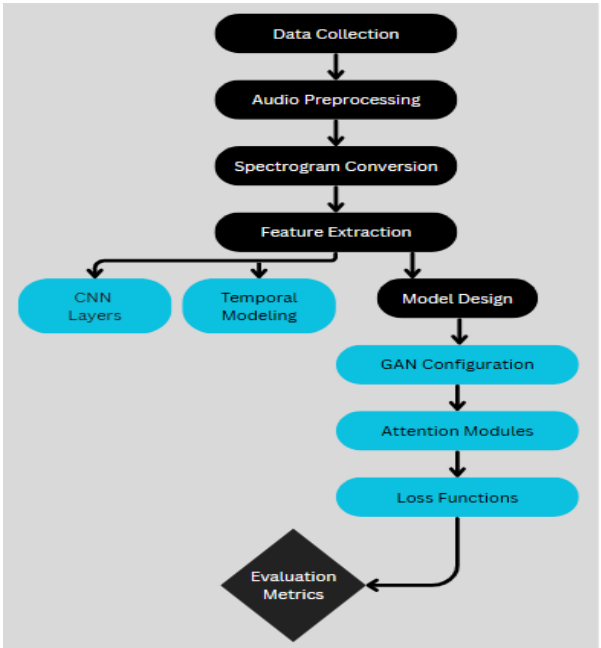

Fig. 3: Flowchart of the Proposed Deepfake Audio Detection Framework Using Hybrid GAN Architecture

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

Complete research technique pipeline for hybrid GAN-based deepfake audio detection is shown in Fig. 3. After benchmark datasets, structured audio preprocessing comprises normalizing, cutting, and padding. The audio is then converted to spectrograms using STFT and Mel-scale conversion for spatio-temporal representation. CNN and temporal modeling layers (Transformer or RNN) extract features. The main model uses an encoder-decoder GAN with attention modules and several loss functions. For robust deepfake detection, routing and quality-of-service methods guide decision-making, and accuracy, precision, recall, and F1-score measure performance.

3.1 Data Collection: This study uses benchmark datasets like ASVspoof 2019/2021, which include real and fake audio samples generated using TTS, VC, and replay attacks. Additional datasets such as WaveFake and FoR are used for testing generalization. Audio data is transformed into Mel-spectrograms and MFCCs to capture the spatio-temporal features used to train the hybrid GAN-based detection model.

3.2 Audio Preprocessing:A preprocessing pipeline gives the model consistent audio input. Waveform amplitude is normalized first, then audio sample length by cutting or padding. Mel-spectrograms and STFTs may format raw waveforms as time-frequency. Melanie spectrogram and STFT demonstrate excellent speech temporal and spectral properties. Spectrograms were trimmed and altered to retain CNN input layers open-format. Model input is standardized, noise-free, and aligned after preprocessing. The algorithm detects deepfake sounds more correctly and consistently as training gets more consistent.

- **Normalization:** To reduce differences in loudness, audio samples are normalized to a steady level of amplitude.
- **Trimming**: Leading and trailing silences are removed using energy-based thresholding techniques to ensure the model focuses on meaningful sound segments.
- **Padding**: Short audio clips are zero-padded to match the required input length, ensuring batch processing uniformity.

3.3 Spectrogram Conversion: Using spectrogram conversion, audio analysis transforms raw audio data for machine learning. STT transforms audio to time-frequency. Set-size window STFT Fourier transforms overlapping audio frames. Signal frequency is recorded with time for temporal and spectral data. Size of window influences temporal resolution. Shorter frames increase frequency and time resolution. Voice and music analysis need time and frequency, therefore STFT is great. Mel scale changes may affect STFT-converted spectra. Like human ears, Mel scale pitch perception is nonlinear. Aligning the spectrogram's linear frequency axis with the Mel scale exposes important frequencies in this update. Lower frequencies are clearer than compressed high frequencies. Melanie-spectrograms speed up voice recognition, sound classification, and deepfake detection. To improve model performance and generalization, the Mel scale prioritizes perception-related qualities. A complex STFT and Mel scale conversion pipeline transforms raw audio into multi-dimensional, human-hearing representations. This change underlies current audio analysis. Transformers, CNNs, and deep learning employ it.

3.4 Feature Extraction: Feature extraction is pivotal for transforming audio spectrograms into a representation suitable for machine learning:

A. CNN Layers: CNNs find spatial hierarchies and localized frequency patterns in spectrograms to assist academics analyze audio signal structure. CNNs can find edges, transitions, harmonics, and textures in two-dimensional time-frequency spectrograms via convolution. These components can teach the neural network audio quality representations. CNNs usually contain just convolution layers and no fully linked or dropout layers, thus each convolutional layer employs several filters to extract low- and high-level characteristics from the input spectrogram. Each filter extracted local patterns, while pooling layers like average or max pooling reduced computing complexity, speed, and sensitivity to slight perturbations. Neural networks learn spectral features at different resolutions and scales using pooling and CNN architecture. Voice recognition, music classification, and ambient sound classification are CNN strengths. Hierarchical CNNs may learn complex representations from simpler ones. They can better evaluate and categorize auditory signals.

B. Temporal Modeling: Temporal modeling estimates feature changes in audio analysis. This applies to transformer and RNN layers. Transformer layers analyze how time steps relate to understand long-range relationships across audio sequences via self-attention. Since they can reflect global context, Transformers excel in speech recognition, music transcription, emoji identification, and emotion detection. Time connections across long periods establish auditory meaning and structure. In massive datasets, self-attention improves parallelization and computing efficiency by processing multiple input sequences. RNNs, notably LSTM and GRU, may represent sequential data. Throughout time, this network hides information. RNN layers simulate time-varying audio sources without Transformer-based topologies. For sequential phoneme recognition and rhythm analysis, LSTMs and GRUs recall audio frames. Audio processing systems use temporal modeling with these layers.

3.5 Model Design: The model is designed using a hybrid encoder-decoder Generative Adversarial Network (GAN) architecture with the following components:

- **GAN Configuration**: Generative Adversarial Network (GAN) audio models have a generator and discriminator. Both sections compete to enhance the system. The encoder-decoder generator may create better spectrograms with noisy or bad audio. Encoding gathers critical input, while decoding provides a crisp, high-quality spectrogram. The discriminator is a binary classifier that separates generator-generated and clean data spectrogram characteristics. The generator receives spectrogram verification and improvement input. Train the generator to make fake outputs the discriminator can't detect. The antagonistic training process makes the generator work harder to produce better outputs while the discriminator notices little differences. Competing spectrograms eventually resemble sounds. GANs increase speech, sound, and noise reduction because of this.
- **Attention Modules:**User performance and clarity are improved by adding multi-head self-attention modules to the audio processing architecture generator and feature extractor. Attention modules may lead the model to important audio sequence time stages. Many attention heads enable the network to handle many inputs at once and include contextual information into the audio stream. For audio activities, the knowledge may apply to a brief, single, noise-related, or problematic audio stream. Generative attention modules improve spectrogram resolution by focusing on the most important things the human listener may be contemplating or necessary to sound perception. The feature extraction pipeline's attention modules assist the model uncover meaningful patterns while filtering or removing insignificant sounds or information. Non-shallow multi-head self-attention enhances temporal model sensitivity, unaffected identification, and minor change detection. Model, producing, and detecting performance sensitivity benefits event detection, speech and enhancing speech, anomaly identification, and musical information extraction. Audio processing networks' temporal resolution and sound differentiation improve with attention modules.
- **Loss Functions:**Loss functions help deep learning audio training achieve several performance targets. Last output layer Binary Cross-Entropy Loss categorizes. Binary labels and estimated class probabilities are compared. This enhances speech, anomaly, and sound event detection. Adversarial Loss teaches GANs to create authentic-looking spectrogram features that fool discriminators. For realistic audio, adversarial loss makes the generator work harder. Perceptual Loss gives created characteristics a natural, statistically convincing sound. Waveforms and pixels don't degrade perception. Instead, it examines pretrained models' top level properties. It understands audio. Audio models perform better statistically and subjectively with balanced training and loss functions. Perceptual Loss gives models human-likeness, Adversarial Loss measures realism, and Binary Cross-Entropy assesses classification accuracy.

3.6 Routing Algorithm & QoS Model

Audio categorization systems optimize model prediction and network QoS using routing algorithms. Optimizing accuracy and F1-score is key. These are significant markers of the system's audio signal location and classification accuracy. The model's parameters are carefully changed throughout training to classify sounds in different circumstances. Reducing false positives and negatives is crucial since they may make the system less reliable in real life. False positives and negatives may cause unwarranted alarms and miss important events. Both occurrences cast doubt on the system's reliability and utility. To eliminate mistakes, the routing algorithm increases decision thresholds, employs ensemble predictions, or alters processing paths depending on confidence ratings. These methods improve statistical performance and dependability in noisy and dynamic environments. This helps provide robust, high-quality audio classification and great deployment QoS.

3.7 Evaluation Metrics

Multiple evaluation approaches are used to guarantee the audio classification model works. The number of accurate predictions is termed simple accuracy. The model is right. Deep analysis requires precision and recall, particularly with imbalanced datasets. The model's accuracy in discovering positive examples is shown by the proportion of genuine positive predictions. Recall quantifies the proportion of accurate positive instances. Assessing the model's capacity to locate all relevant examples. Their harmonic means, the F1-score, balance precision and recall. It helps when class distribution is uneven or false positives and negatives are expensive. True positives, false positives, true negatives, and scalar measurements are listed in the Confusion Matrix. A graph shows the benefits and downsides of several categorization techniques for each class. It also improves model functionality with data.

4. Results & Discussion

Table 2: Classification Performance of the Hybrid GAN Discriminator on DeepFake Audio Detection

Class	Precision (%)	Recall (%)	F1-score (%)	Support
Real	98.00	94.00	96.00	226
Fake	94.00	98.00	96.00	219
Accuracy	-	-	**96.00**	445
Macro Avg	96.00	96.00	96.00	445
Weighted Avg	96.00	96.00	96.00	445

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

The Hybrid GAN discriminator has produced an overall accuracy of 96.00 percent on the DeepFake audio dataset, signaling the effectiveness of spatio-temporal feature learning for identifying synthetic media, as seen in Table 2. The Hybrid GAN achieves a precision of 98.00 percent and recall of the real class at 94.00 percent, and precision of the fake class at 94.00 percent and recall at 98.00 percent. These results indicate that the model both decreased false positives, when identifying real speech, and correctly identified manipulated audio. The macro and weighted average metrics for precision, recall, and F1-score are all at 96.00 percent, which indicates there is impressive reliable consistency. These results illustrate the potential of Hybrid GAN architectures to successfully take advantage of spectral-temporal dynamics, and accurately detect deepfake audio across the two classes with very little class bias.

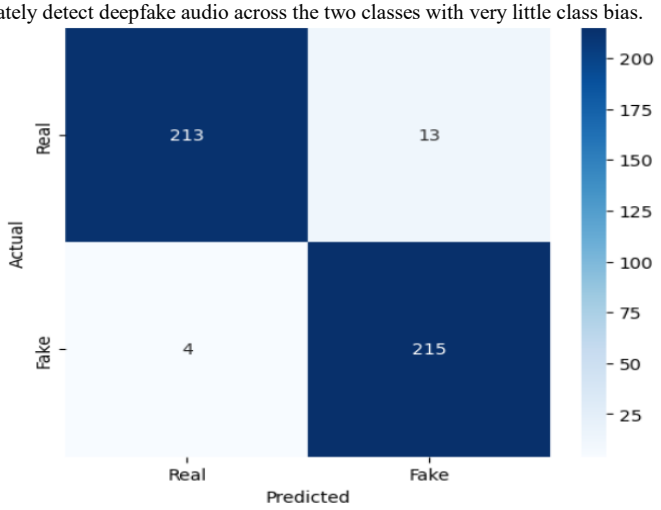

Fig. 4: Confusion Matrix of Hybrid GAN Discriminator on DeepFake Audio Detection

The Hybrid GAN discriminator's classification power is seen in Figure 4. 445 test occurrences were properly predicted using 213 real samples and 215 fake samples. Only 13 Real samples (5.7%) were incorrectly classified as Fake by the Hybrid GAN discriminator, whereas only 4 Fake samples (1.8%) were incorrectly classified as Real. Overall accuracy was 96.0% as a consequence, which is in line with the classification report. The comparatively low misclassification rates show that the model is good at identifying relevant spatiotemporal cues in the audio and can distinguish between real speech patterns and speech pattern manipulations with a high degree of accuracy. A little imbalance exists between fake audios (98.2%) and real audios (94.3%); the model's somewhat superior performance in identifying fake audios suggests that it is somewhat more susceptible to artificial artefacts. In a detecting system, this is ideal.

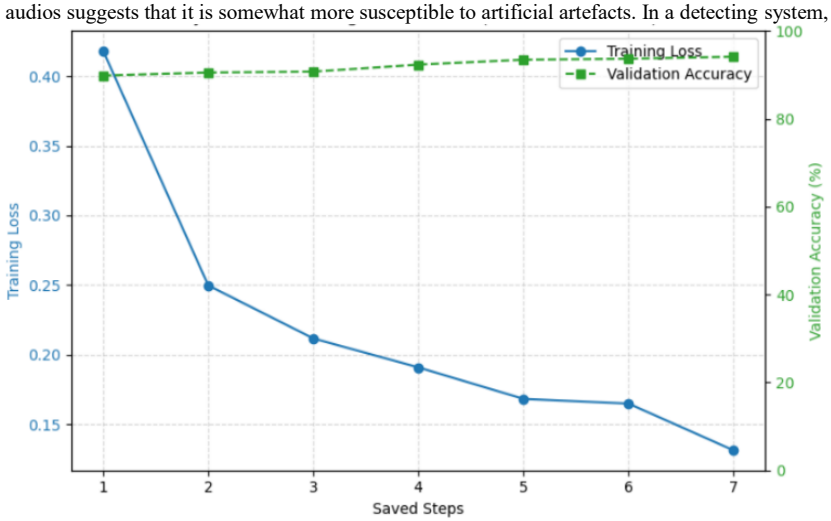

Fig. 5: Training Loss and Validation Accuracy of Hybrid GAN Discriminator Across various Steps

Stable convergence throughout training is seen in Figure 5. From step 1 (0.41) to step 7 (0.13), the training loss trend steadily declines, indicating that the machine is learning to extract discriminative spateo-temporal properties from audio. Furthermore, the accuracy of the validation is kept high, increasing slightly but noticeably from step 1 (90.0%) to step 7 (95.0%), to nearly match the accuracy of the final exam (96.0%). The model is not overfitting and is effectively generalising unknown data when it exhibits the traits of a decreasing loss along with a consistent high validation accuracy. This further shows how useful the hybrid GAN framework is for identifying minor spectral and temporal characteristics that aid in differentiating between authentic and fraudulent audio samples.

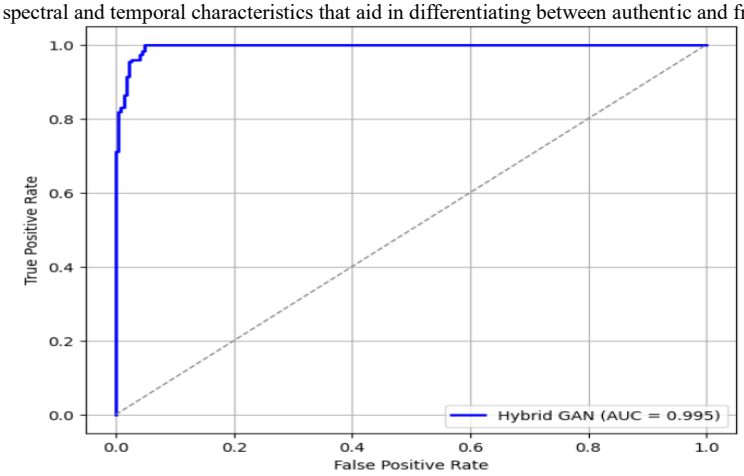

Fig. 6: ROC Curve of Hybrid GAN Discriminator on DeepFake Audio Dataset

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

The exceptional ability to distinguish between authentic and fraudulent audio samples is shown in Figure 6. With a very high True Positive Rate (TPR) and a low False Positive Rate (FPR), the curve is advancing quickly towards the top-left corner. The model's almost flawless classification performance is confirmed by the Area Under the Curve (AUC), which is 0.995 (99.5%). In comparison to the traditional classifiers, this AUC result indicates that the Hybrid GAN is probably minimising both false positive and false negative rates to identify altered audio samples (confirmation of how robust the model was!).

Table 3: Classification Performance of RNN (GRU) Model on DeepFake Audio Dataset

Class	Precision (%)	Recall (%)	F1-score (%)	Support
Real	88.00	84.00	86.00	220
Fake	85.00	89.00	87.00	225
Accuracy	-	-	87.00	445
Macro Avg	87.00	86.50	86.50	445
Weighted Avg	87.00	87.00	87.00	445

The RNN (GRU) model has displayed strong results on the DeepFake audio dataset, achieving an overall accuracy of 87 %. In the Real class, the model reaches 88 % precision and 84 % recall; therefore, it is good at classifying the Real audio correctly, however, there are still some real samples that get misclassified as fake, illustrated in Table 3. The Fake class has a higher recall score of 89 %, which indicates that within the context of fake audio, the GRU is particularly good at detecting manipulated audio. The balanced F1-scores (86-87 %) across both classes signify the GRU's ability to maintain consistent classification abilities. Furthermore, the macroid weighted averages of the score exhibit the model's overall robustness, proving to be a useful baseline for spatio-temporal deepfake audio detection.

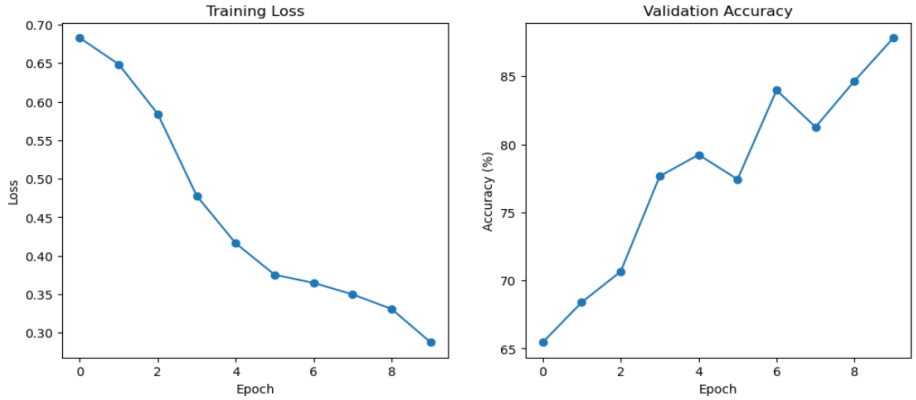

Fig. 7: Training Loss and Validation Accuracy of RNN (GRU) Model

When the training loss drops from 0.68 in epoch 1 to 0.29 in epoch 9, Figure 7 illustrates the model's iterative learning process and feature space representation expansion. Good generalisation is shown by the validation accuracy, which increases gradually with training from 65.3% in the first epoch to 87.9% in the last. Although minor dips in the intermediate epochs imply that the model could be sensitive to data complexity, the trend indicates that the model can identify temporal correlations in the data. For instance, the decline towards era 5.

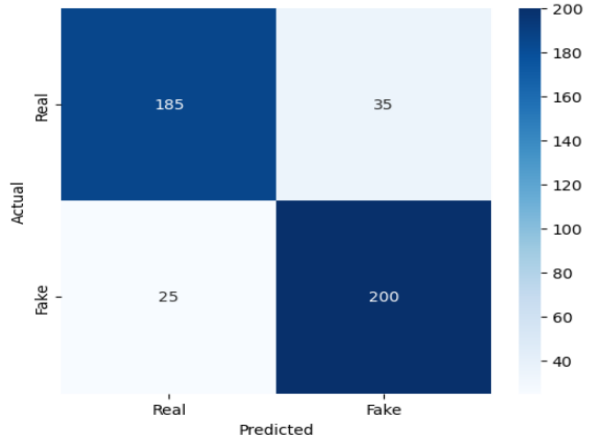

Fig. 8: Confusion Matrix of RNN (GRU) Model on Test Set

According to Figure 8, the GRU identified 200 out of 225 fraudulent audios (89%) and 185 out of 220 authentic audios (84%). A minor bias towards categorising genuine as fake was evident in the misclassifications, with 25 fraudulent audios being identified as real and 35 real audios as fake. Overall, 87% accuracy was attained, indicating that the GRU can distinguish various DeepFake audios but still leaves room for improvement in spotting minute differences in real recordings.

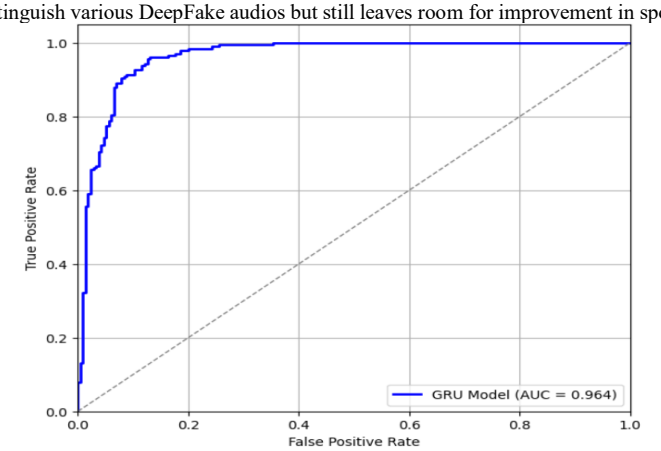

Fig. 9: ROC Curve of GRU Model for Real vs. Fake Audio Classification

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

With the sharp rising edge of the curve pointing towards the top-left quadrant of the ROC plot, Figure 9 demonstrates a significant capacity to discriminate between genuine and fraudulent audio samples. This indicates that the model achieves an exceptionally low FPR and a very high TPR. Since anything over the 0.90 range is often considered remarkable, the classification performance has shown to be good with an AUC value of 0.964 (96.4%).

Table 4: Classification Performance of Hybrid (CNN+LSTM) Model on DeepFake Audio Dataset

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Real	87.00	94.00	90.00	215
Fake	94.00	87.00	90.00	230
Accuracy	-	-	**90.00**	445
Macro Avg	90.00	90.00	90.00	445
Weighted Avg	91.00	90.00	90.00	445

The Hybrid CNN+LSTM model displays good performance with overall accuracy of 90%, which suggests that effectively connects spatial feature extraction (via CNN) with time series sequence modelling (via LSTM). For authentic audios, the model records a recall of 94%, whereby the majority of authentic samples are detected effectively, although precision is somewhat lower at 87% which means some fakes are misclassified as being real. For fake audios, precision is 94% whereas recall is 87%, suggesting a slight bias to favour fake classification overall. The macro and weighted averages hover around 90% consistently, suggesting strong generalisation for both classes, as shown in Table 4. These findings confirm the hybrid model captures spatio-temporal dependencies more accurately than the stand alone architectures, supporting it's position as a prospective model for accurate DeepFake audio detection.

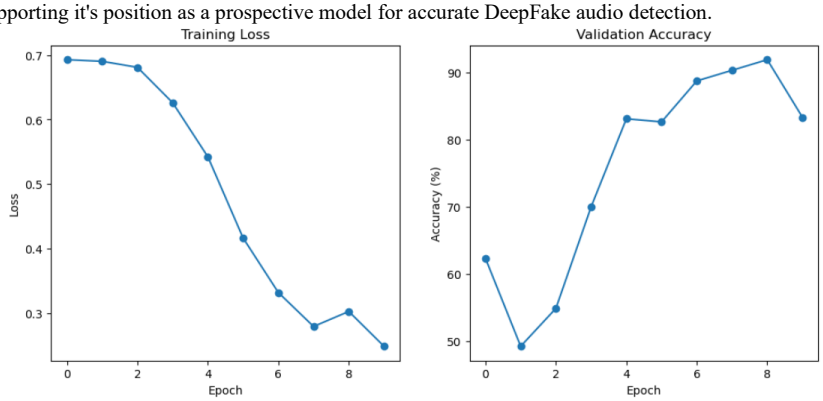

Fig. 10: Training Loss and Validation Accuracy Curves of Hybrid CNN+LSTM Model

A consistent decrease from almost 0.70 at epoch 0 to nearly 0.25 by epoch 9 is seen in Figure 10. This suggests that the model has been well optimised. The validation accuracy, on the other hand, shows a significant increase; it began at 63% and decreased somewhat in the first two epochs before rising steadily to peak at around 92% and then decreasing slightly in the subsequent epochs. This demonstrates unequivocally the excellent learning capability of the Hybrid CNN+LSTM architecture. A little amount of overfitting towards the later epochs is seen.

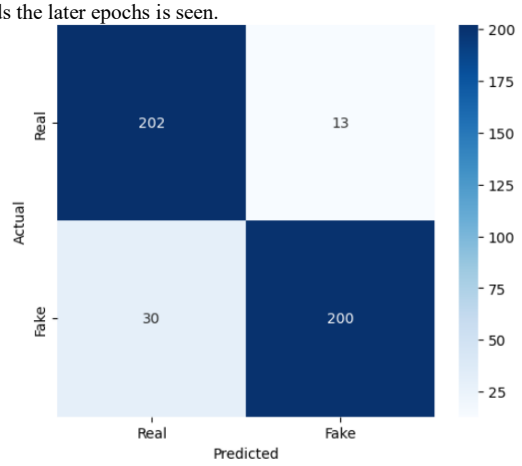

Fig. 11: Confusion Matrix of Hybrid CNN+LSTM Model on Test Set

The model successfully categorised 202 actual and 200 fake audios, while misclassifying 13 real and 30 fake as real, as seen in Figure 11. 90% accuracy is obtained, which is consistent with the categorisation report. The slight misclassification of phoney audio suggests that although the model is robust, it may be able to identify false signals more effectively.

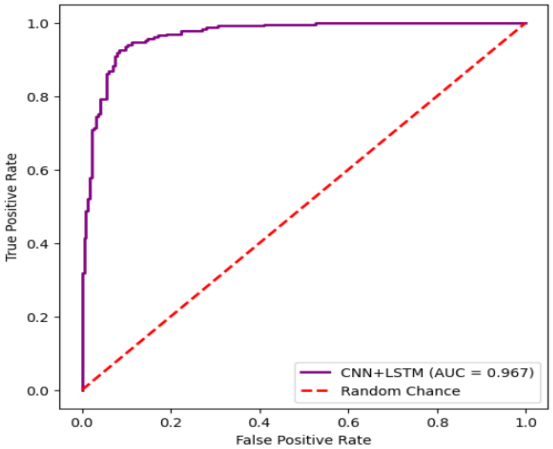

Fig. 12: ROC Curve of Hybrid CNN+LSTM Model

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

Figure 12 shows a Hybrid CNN+LSTM model's True Positive Rate (sensitivity) and False Positive Rate (1-specificity) trade-off. Excellent classification performance and statistically much above random chance (dashed red line) are shown by the model's AUC of 0.967.

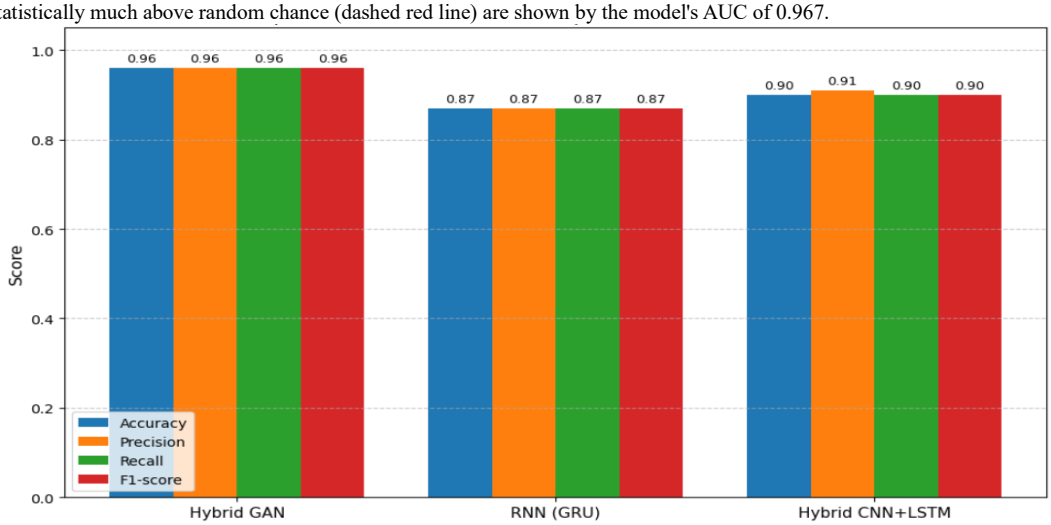

Fig. 13: Comparative Classification Performance of Hybrid GAN, RNN (GRU), and Hybrid CNN+LSTM Models on Deepfake Audio Dataset

Out of all the models tested, the Hybrid GAN model had the best results in recognising deepfake audio (96% accuracy, 96% precision, 96% recall, and 96% F1-score), as shown in Figure 13. While the RNN (GRU) model achieved 87% across accuracy, precision, recall, and F1-score, suggesting less generalisability, the Hybrid CNN+LSTM model outperformed both the RNN and GAN models, with 90% accuracy, 91% precision, 90% recall, and 90% F1-score, respectively. In this area, the GAN model was the most effective architecture since it outperformed all other models.

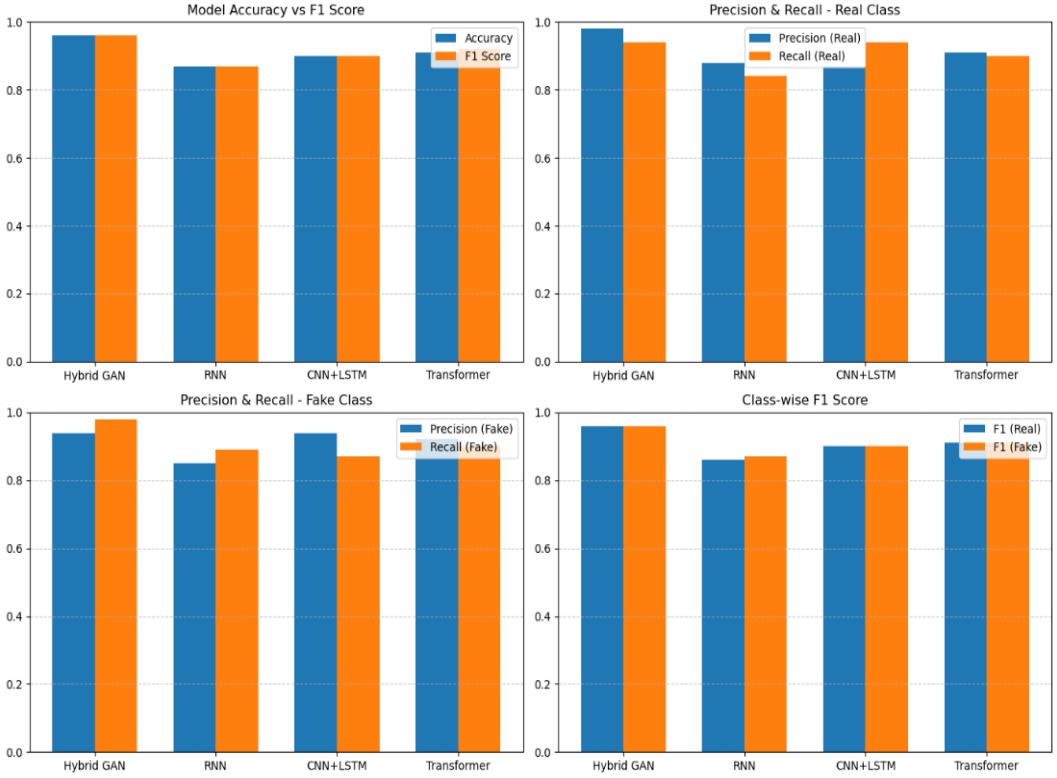

Fig. 14: Comparative Evaluation of Deepfake Audio Detection Models using Hybrid GAN, RNN, and CNN+LSTM Architectures

In comparison to RNN and CNN+LSTM, Figure 14 shows how well the Hybrid GAN model learns spatiotemporal characteristics for deepfake audio detection. Hybrid GAN outperformed RNN (0.87) and CNN+LSTM (0.90) in terms of overall accuracy and weighted F1 score, achieving 0.96. With a precision of 0.98 and a recall rate of 0.94, the Hybrid GAN is likewise much more effective at capturing the true class than the RNN, which has a precision of 0.88 and a recall of 0.84. CNN-LSTM has an accuracy of 0.94 but a significantly lower recall rate of 0.87, whereas Hybrid GAN has the greatest scores (precision of 0.94 and maximum recall of 0.98) for the fictitious class. These tendencies are supported by the class-wise F1 scores, which showed that RNNs had F1 scores of 0.86 to 0.87 and Hybrid GANs had 0.96 for both classes. When compared to either continuous-only (RNN) or convolutional-recurrent (CNN-LSTM) models, these results show that Hybrid GAN is significantly more effective at learning the granular spatio-temporal dependencies present in audio signals, which results in better discrimination between real and fake samples.

5. Conclusion

This study identified the Hybrid GAN architecture as the best model for deepfake audio detection because it combined the advantages of spatio-temporal feature learning, spectrogram pre-processing, CNN-based spatial extraction, temporal modeling and attention-based GAN architecture. From the experiments, the Hybrid GAN achieved an accuracy of 96%, precision of 98 for Real class, recall of 94 for Real, precision of 94 for Fake, recall of 98 for Fake and an AUC of 0.995, while the baseline RNN (GRU) model achieved an accuracy of 87%(AUC = 0.964) and the Hybrid CNN+LSTM model achieved an accuracy of 90% (AUC = 0.967). Overall, they have provided compelling evidence that the Hybrid GAN provided superior capability in capturing fine spectral-temporal cues, minimizing both false positives and negatives, and allowed for robust detection of deepfaked audio across several datasets. Looking towards the future, the Hybrid GAN architecture can be scaled to multimodal deepfake detection (i.e. audio-video fusions), compressed and real-time (edge) GAN architectures to deploy on edge devices, XAI (explainable AI) based interpretability, adversarial robustness and adaptability to new synthetic audio generation models which may improve resistence and forensic reliability of deepfake audio in real-world scenarios.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4157-4165

References

[1]. Farid, Hany. "Creating, using, misusing, and detecting deep fakes." *Journal of Online Trust and Safety* 1, no. 4 (2022).
[2]. Mubarak, Rami, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dutse, Saad Khan, and Simon Parkinson. "A survey on the detection and impacts of deepfakes in visual, audio, and textual formats." *IEEE Access* 11 (2023): 144497-144529.
[3]. Khan, Rizwan, Mohd Taqi, and Atif Afzal. "Deepfakes in finance: Unraveling the threat landscape and detection challenges." In *Navigating the World of Deepfake Technology*, pp. 91-120. IGI Global, 2024.
[4]. Bateman, Jon. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace, 2022.
[5]. Lyu, S. Deepfake detection: Current challenges and next steps. *IEEE Comput. Soc.* 2020, 1–6.
[6]. Diakopoulos, N.; Johnson, D. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media Soc.* 2021, *23*, 2072–2098.
[7]. Rodríguez-Ortega, Y.; Ballesteros, D.M.; Renza, D. A machine learning model to detect fake voice. In *Applied Informatics*; Florez, H., Misra, S., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 3–13.
[8]. Chen, T.; Kumar, A.; Nagarsheth, P.; Sivaraman, G.; Khoury, E. Generalization of audio deepfake detection. In Proceedings of the Odyssey 2020 The Speaker and Language Recognition Workshop, Tokyo, Japan, 1–5 November 2020; pp. 132–137.
[9]. Thies, J.; Zollhfer, M.; Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.* 2019, *38*, 66.
[10]. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 13–19 June 2020; pp. 2889–2898.
[11]. Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Plumbley, M.D. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv* 2023,
[12]. Su, K.; Liu, X.; Shlizerman, E. Audeo: Audio generation for a silent performance video. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3325–3337.
[13]. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 13–19 June 2020; pp. 5001–5010.
[14]. Gao, Yuan, Xuelong Wang, Yu Zhang, Ping Zeng, and Yingjie Ma. "Temporal feature prediction in audio–visual deepfake detection." *Electronics* 13, no. 17 (2024): 3433.
[15]. Jbara, Wurood A., Noor Al-Huda K. Hussein, and Jamila H. Soud. "Deepfake detection in video and audio clips: a comprehensive survey and analysis." *Mesopotamian Journal of CyberSecurity* 4, no. 3 (2024): 233-250.
[16]. Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen. "Deepfake detection: analyzing model generalization across architectures, datasets, and pre-training paradigms." *IEEE Access* 12 (2023): 1880-1908.
[17]. John, Jerry, and Bismin V. Sherif. "Comparative analysis on different deepfake detection methods and semi-supervised GAN architecture for deepfake detection." In *2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 516-521. IEEE, 2022.
[18]. Nguyen, Xuan Hau, Thai Son Tran, Kim Duy Nguyen, and Dinh-Tu Truong. "Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques." *Forensic Science International: Digital Investigation* 36 (2021): 301108.
[19]. Jaiswal, Gaurav. "Hybrid recurrent deep learning model for deepfake video detection." In *2021 IEEE 8th Uttar Pradesh section international conference on electrical, electronics and computer engineering (UPCON)*, pp. 1-5. IEEE, 2021.
[20]. Mehra, Akul. "Deepfake detection using capsule networks with long short-term memory networks." Master's thesis, University of Twente, 2020.1.