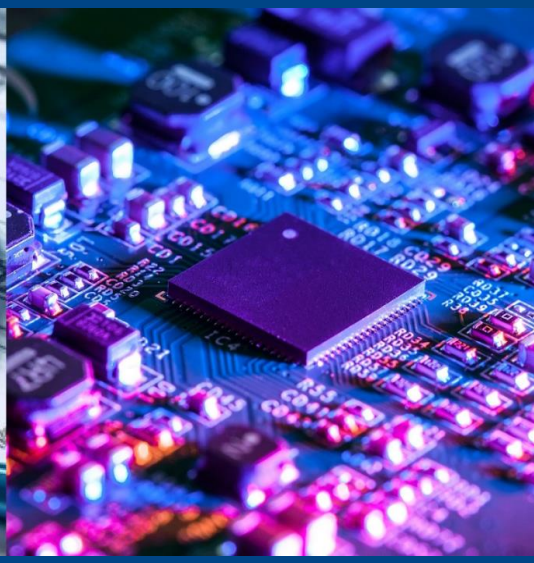
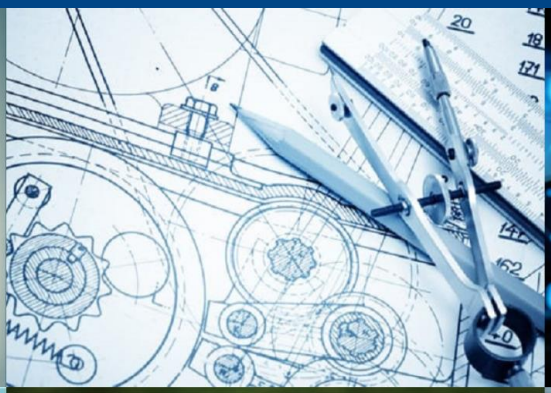
ISSN: 2582-7219

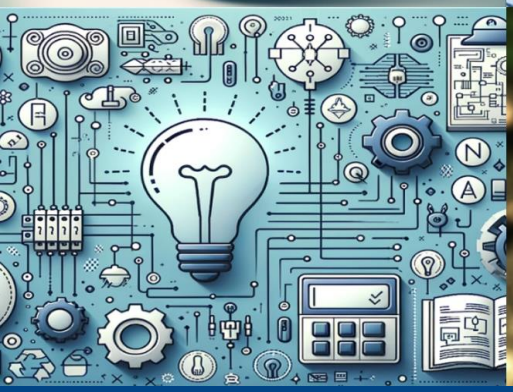




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 3, March 2026**

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI:10.15680/IJMRSET.2026.0903088

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Deepfake Audio Detection

Dr. A.Jeyalakshmi[1], Mr.I.K.Ajay Kiran[2]

Department of Information Technology, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India[1]

PG & Research, Department of Information Technology, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: The rapid advancement of artificial intelligence has enabled the generation of highly realistic synthetic media, including deepfake audio that can convincingly imitate human speech. While such technologies offer benefits in areas such as entertainment, accessibility, and automation, they also pose serious security threats including voice impersonation, financial fraud, misinformation, and privacy violations. Detecting deepfake audio has therefore become a critical challenge in modern digital communication systems. This project presents a deepfake audio detection system that combines traditional machine learning and deep learning approaches to improve reliability and robustness. The system utilizes Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction and implements three classification models: Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Transformer-based neural networks. The models are trained and evaluated using real and AI-generated speech samples. A Flask-based web application is developed to provide a user-friendly interface that allows users to upload audio files and select a detection model dynamically. Experimental results demonstrate that deep learning models, particularly LSTM and Transformer architectures, outperform traditional approaches by effectively capturing temporal and spectral characteristics of synthetic speech. The proposed system enhances trust and security in audio-based communication systems.

KEYWORDS: Deepfake Audio Detection, Synthetic Speech, Voice Impersonation, Mel-Frequency Cepstral Coefficients (MFCC), Support Vector Machine (SVM), Long Short-Term Memory (LSTM), Transformer Neural Networks, Machine Learning, Deep Learning, Audio Forensics, Speech Security, Flask Web Application

I. INTRODUCTION

The legal system is a fundamental pillar of any society, responsible for maintaining order, ensuring justice, and safeguarding the rights and responsibilities of individuals and organizations. Laws regulate almost every aspect of human activity, including employment, education, business transactions, property ownership, family matters, cyber activities, and criminal behaviour. Despite its importance, legal information is often difficult for the general public to access and understand due to the use of complex terminology, lengthy procedures, and formal documentation. This complexity creates a significant barrier for individuals seeking quick and reliable legal guidance for everyday issues. In many cases, people require only preliminary legal information rather than full-scale legal representation. However, traditional methods of obtaining legal assistance, such as consulting lawyers or visiting legal offices, can be expensive, time-consuming, and geographically restrictive. Individuals living in remote or rural areas often face additional challenges due to limited access to legal professionals. Furthermore, legal institutions and help desks are frequently overwhelmed by large volumes of repetitive queries, which reduces their efficiency and increases response delays. These challenges emphasize the growing need for an automated, intelligent system capable of delivering basic legal information in a timely and accessible manner. The rapid evolution of Artificial Intelligence (AI), particularly in the areas of Natural Language Processing (NLP) and Machine Learning (ML), has significantly improved the ability of computer systems to understand and process human language. AI-driven conversational agents, commonly known as chatbots, have demonstrated remarkable success in domains such as customer support, healthcare advisory systems, banking, and education. These systems are designed to simulate human-like conversations, interpret user intent, and provide accurate responses in real time. The application of chatbot technology in the legal domain offers a promising solution to bridge the gap between complex legal knowledge and user-friendly access.An AI-based legal query answering chatbot aims to provide users with instant legal information by analyzing natural language queries and mapping them to relevant legal concepts. Such a system can assist users in understanding their legal rights, obligations, and procedural steps without requiring immediate human intervention. By leveraging NLP techniques such as tokenization, intent classification, and semantic analysis, the chatbot can interpret diverse user queries and deliver

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

structured, meaningful responses. Additionally, integrating machine learning models allows the system to improve its performance over time by learning from user interactions. Designing an effective legal chatbot presents unique challenges, including the need to maintain accuracy, reliability, and consistency in legal responses. Unlike general-purpose chatbots, legal systems must ensure that the information provided is precise and aligned with established legal sources. The chatbot must also handle ambiguous or incomplete queries, manage contextual understanding, and present information in a clear and non-misleading manner. Addressing these challenges requires a carefully designed architecture that combines AI techniques with a well-organized legal knowledge base. This project focuses on the design and development of an AI-based legal query answering chatbot that serves as an intelligent virtual assistant for legal information retrieval. The system is intended to act as a first-level legal support tool, reducing the workload on legal professionals while empowering users with accessible legal knowledge. By improving response time, enhancing accessibility, and simplifying legal information delivery, the proposed chatbot contributes to the modernization of legal assistance systems and promotes greater legal awareness among the public

II. LITERATURE REVIEW

The detection of spoofed and synthetic speech has been an active area of research in audio forensics for more than a decade. Early work in this domain primarily focused on detecting replay attacks and basic speech synthesis using traditional signal processing techniques. Wu et al. [1] provided a comprehensive survey of spoofing attacks and countermeasures for speaker verification systems, highlighting the limitations of classical approaches when faced with advanced synthetic speech.Handcrafted audio features such as MFCCs, linear prediction coefficients, spectral centroid, and pitch-related features have been widely used in early detection systems. Kinnunen and Li [2] demonstrated that MFCC-based features combined with statistical classifiers could effectively represent speaker characteristics and detect certain spoofing attacks. However, these methods were sensitive to noise and lacked robustness against modern neural speech synthesis techniques.

Support Vector Machines (SVMs) became a popular classifier in early deepfake and spoofing detection systems due to their strong generalization capability on small datasets. Studies showed that SVMs trained on MFCC features could achieve reasonable detection accuracy for replay and synthetic speech attacks [1]. Despite their effectiveness, SVM-based systems are limited by their inability to model temporal dependencies in speech signals, which are essential for detecting subtle inconsistencies in deepfake audio. To address these limitations, researchers began exploring deep learning-based approaches. Convolutional Neural Networks (CNNs) were initially applied by treating spectrograms as images and learning spatial patterns associated with fake speech. While CNNs improved detection performance, they did not explicitly capture temporal dynamics. This led to the adoption of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, which are designed to process sequential data. Lavrentyeva et al. [3] demonstrated that LSTM-based models significantly outperform traditional classifiers in detecting replay and synthetic speech by capturing long-term temporal dependencies. Their results highlighted the importance of modeling speech dynamics rather than relying solely on static features. LSTM-based approaches have since become a standard baseline in deepfake audio detection research. More recently, Transformer-based models have gained attention in speech processing tasks due to their ability to model global dependencies using self-attention mechanisms. Vaswani et al. [5] introduced the Transformer architecture, which has since been adapted for speech recognition and spoofing detection. Transformer-based systems have shown improved generalization and robustness compared to RNN-based models, particularly in handling long audio sequences.

The introduction of self-supervised pretrained models such as Wav2Vec 2.0 further advanced the field. Baevski et al. [6] showed that pretrained speech representations learned directly from raw audio significantly improve performance on downstream tasks, including fake audio detection. However, these models often require substantial computational resources, limiting their deployment in lightweight or CPU-based systems.

Large-scale evaluation campaigns such as ASVspoof 2019 emphasized the growing sophistication of deepfake generation techniques and the need for robust detection systems [7].

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. SYSTEM ANALYSIS

3.1 Existing System

Existing deepfake audio detection systems are primarily based on traditional speech processing and classical machine learning techniques. These systems typically rely on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, and spectral statistics to represent speech signals. The extracted features are then classified using algorithms such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM), or Random Forests. While these methods were effective for early spoofing and replay attack detection, they face significant challenges when applied to modern deepfake audio generated using advanced neural speech synthesis techniques.

One of the major limitations of existing systems is their inability to model the temporal dynamics of speech signals. Human speech is inherently sequential, and deepfake generation models often introduce subtle inconsistencies across time. Traditional classifiers, which operate on aggregated or static feature representations, fail to capture these long-term dependencies. As a result, their detection accuracy decreases when exposed to high-quality synthetic speech.

Another drawback of existing approaches is their limited generalization capability. Many systems are trained and evaluated on specific datasets and perform poorly when tested on unseen data or audio generated using different synthesis techniques. Additionally, most existing detection systems function in offline or research-oriented environments and lack real-time processing capabilities. They often do not provide user-friendly interfaces, making them unsuitable for practical deployment.

Furthermore, existing systems usually rely on a single detection model, making them vulnerable to evolving deepfake generation methods. Manual feature engineering and model tuning are often required, increasing system complexity and reducing adaptability. These limitations highlight the need for more advanced, flexible, and robust deepfake audio detection frameworks.

3.2Proposed System

The proposed system introduces a robust and flexible deepfake audio detection framework that integrates traditional machine learning and advanced deep learning models to improve detection accuracy and reliability. The system utilizes Mel-Frequency Cepstral Coefficients (MFCCs) as the primary feature representation, as they effectively capture perceptually relevant spectral characteristics of speech signals. The extracted features are processed using three complementary models: Support Vector Machine (SVM), Long Short-Term Memory (LSTM) networks, and Transformer-based models.

a. Support Vector Machine (SVM) Model

The Support Vector Machine (SVM) model serves as a baseline classifier in the proposed system. It operates on statistical representations of MFCC features, specifically the mean MFCC vector extracted from each audio file. The SVM constructs an optimal decision boundary that separates real and deepfake audio samples in a high-dimensional feature space. Due to its strong generalization capability and low computational complexity, the SVM model provides fast inference and serves as a reference point for evaluating the performance of deep learning models. However, since it does not explicitly model temporal dependencies, its effectiveness is limited when detecting subtle dynamic inconsistencies in deepfake audio.

b. Long Short-Term Memory (LSTM) Model

The Long Short-Term Memory (LSTM) model is a type of recurrent neural network designed to process sequential data. In the proposed system, the LSTM model receives MFCC feature sequences rather than aggregated statistics, enabling it to capture temporal dependencies and speech dynamics over time. This allows the model to detect inconsistencies in rhythm, intonation, and transitions between speech frames that are often introduced during deepfake audio generation. The LSTM model significantly improves detection accuracy by learning long-term patterns in speech signals that traditional classifiers fail to capture.

c. Transformer Model

The Transformer-based model represents the most advanced detection component in the proposed system. Unlike recurrent models, the Transformer relies on self-attention mechanisms to analyze relationships between all frames in an audio sequence simultaneously. This allows the model to capture long-range dependencies and global patterns in

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

speech signals more effectively than LSTM networks. By focusing on the most informative regions of the audio sequence, the Transformer model excels at identifying subtle artifacts and inconsistencies present in high-quality deepfake audio. Its ability to model complex temporal relationships makes it particularly robust against modern neural speech synthesis techniques.

System Integration and Deployment

All three models are integrated into a unified Flask-based web application that allows users to upload audio files and select the desired detection model dynamically. The system processes the input audio, extracts MFCC features, applies the selected model, and outputs a classification result along with a confidence score. Designed to operate efficiently on CPU-based environments, the proposed system offers a practical, scalable, and user-friendly solution for real-world deepfake audio detection.

IV. METHODOLOGY

4.1 Data Collection

The first step in the methodology involves collecting a labeled dataset containing both real human speech and AI-generated (deepfake) audio samples. Publicly available datasets such as DEEP-VOICE, ASVspoof, and other synthetic speech corpora are used to ensure diversity in speakers, accents, recording conditions, and synthesis techniques.

The dataset is organized into two main classes:

- Real Audio – Natural human speech recordings
- Fake Audio – Speech generated using voice cloning or text-to-speech models

This labeled dataset forms the foundation for supervised learning, enabling the models to learn discriminative patterns between genuine and manipulated audio.

4.2 Data Preprocessing

Raw audio files often contain noise, silence, and variations in sampling rate that can negatively affect model performance. Therefore, preprocessing is applied to standardize the input data.

Key preprocessing steps include:

- Conversion of all audio files to WAV format
- Resampling audio to a uniform sampling rate
- Removal of silence and irrelevant segments
- Normalization of audio amplitude
- Trimming or padding audio signals to a fixed duration

These steps ensure consistency across the dataset and improve feature extraction quality.

4.3 Feature Extraction using MFCC

Mel-Frequency Cepstral Coefficients (MFCCs) are used as the primary handcrafted features for the LSTM and Transformer models. MFCCs are widely adopted in speech processing because they closely represent how the human auditory system perceives sound.

The MFCC extraction process includes:

1. Framing the audio signal into short overlapping frames
2. Applying a Fast Fourier Transform (FFT)
3. Mapping frequencies onto the Mel scale
4. Taking the logarithm of Mel energies
5. Applying Discrete Cosine Transform (DCT)

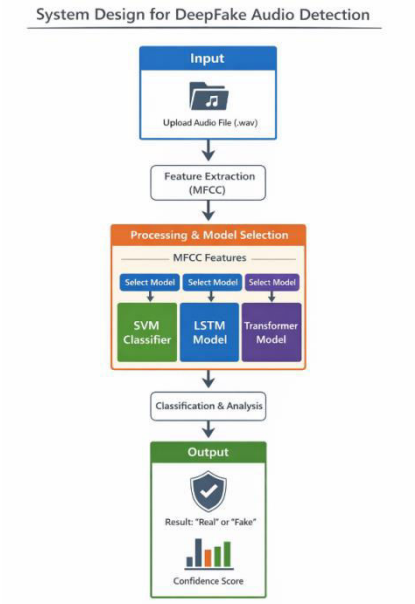
The below descries the architecture of the Deepfake Audio Detection:

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI:10.15680/IJMRSET.2026.0903088

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



4.4 Model Architecture and Training

To improve detection accuracy and compare performance, three different deep learning models are implemented and trained independently.

4.4.1 LSTM Model :

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning long-term dependencies in sequential data. In this system, the LSTM model processes MFCC sequences to capture temporal patterns such as unnatural transitions and rhythm inconsistencies present in deepfake audio.

The LSTM architecture consists of:

- Input layer receiving MFCC sequences
- One or more LSTM layers
- Fully connected (dense) layer
- Sigmoid output layer for binary classification

LSTM is particularly effective for modeling speech dynamics over time.

4.4.2 Transformer Model:

The Transformer model leverages self-attention mechanisms to analyze relationships between different time frames in the MFCC sequence simultaneously. Unlike LSTMs, Transformers do not rely on sequential processing, making them efficient and powerful for long audio sequences.

Key components include:

- Positional encoding
- Multi-head self-attention layers
- Feed-forward neural networks
- Output classification layer

The Transformer model excels at identifying subtle spectral inconsistencies introduced by voice synthesis models

4.4.3 Support Vector Machine (SVM):

The Support Vector Machine (SVM) model is trained using statistical MFCC features extracted from audio signals. For each audio file, the mean MFCC vector is computed and normalized using a standard scaler. The SVM classifier learns an optimal decision boundary that separates real and deepfake audio samples. Due to its low computational cost and strong generalization ability, SVM serves as a baseline model for comparison.

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI:10.15680/IJMRSET.2026.0903088

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. IMPLEMENTATION AND TOOLS

The implementation of the proposed deepfake audio detection system is carried out using Python as the core programming language, following a modular and scalable design approach. The system integrates audio preprocessing, feature extraction, model inference, and result visualization within a unified framework. Audio samples in WAV format are processed using the Librosa library, which enables efficient loading and extraction of Mel-Frequency Cepstral Coefficients (MFCCs). These features capture the essential spectral and temporal characteristics of human speech and serve as the primary input to the classification models. For traditional machine learning, a Support Vector Machine (SVM) model is trained using statistical MFCC features, providing a computationally efficient baseline for deepfake detection. In addition, deep learning models including Long Short-Term Memory (LSTM) and Transformer architectures are implemented using the PyTorch framework to analyze MFCC feature sequences and learn temporal dependencies and long-range patterns present in synthetic speech. Trained models are saved and loaded dynamically during inference to enable real-time prediction. A Flask-based web application is developed to provide an interactive interface through which users can upload audio files, select the desired detection model, and view classification results along with confidence scores. The frontend of the application is built using HTML5, CSS3, and JavaScript to ensure usability and responsiveness. The system is designed to operate entirely on a CPU-based environment, making it suitable for low-resource systems and academic deployment. Supporting libraries such as NumPy and Scikit-learn are used for numerical computation, feature scaling, and model evaluation, while Joblib facilitates model persistence. File handling and validation are managed using standard OS utilities and FNMatch to ensure secure and consistent processing. Development and testing are conducted within a Python virtual environment using Visual Studio Code, ensuring dependency isolation and reproducibility. Overall, the chosen implementation strategy and tools enable the development of an efficient, reliable, and extensible deepfake audio detection system suitable for real-world applications.

VI. RESULTS AND DISCUSSION

The performance of the proposed deepfake audio detection system was evaluated using real and AI-generated speech samples, with experiments conducted on three different classification models: Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Transformer. The evaluation focused on measuring the system's ability to accurately distinguish between genuine and deepfake audio samples using MFCC-based feature representations. Standard performance metrics such as accuracy, precision, recall, and F1-score were used to assess model effectiveness. The experimental results demonstrate that the SVM model provides fast and computationally efficient predictions, making it suitable for baseline detection; however, its performance is limited when handling complex temporal patterns present in sophisticated deepfake audio. In contrast, the LSTM model shows improved detection accuracy by effectively modeling temporal dependencies in speech signals, enabling it to capture inconsistencies introduced during synthetic audio generation. The Transformer model achieves the highest overall performance among the evaluated approaches due to its self-attention mechanism, which allows it to identify long-range dependencies and subtle spectral anomalies in MFCC sequences. Despite its higher computational cost compared to SVM and LSTM, the Transformer model maintains acceptable inference time in a CPU-based environment. Comparative analysis indicates that deep learning models significantly outperform traditional machine learning techniques in detecting advanced deepfake audio. The integration of these models into a Flask-based web application validates the system's real-time applicability and usability. The results confirm that MFCC features remain effective for deepfake audio detection, particularly when combined with deep learning architectures capable of learning complex speech patterns. Overall, the experimental findings highlight the robustness, scalability, and practical relevance of the proposed system while emphasizing the trade-off between computational efficiency and detection accuracy.

VII. CONCLUSION

This project successfully presents a robust and comparative framework for deepfake audio detection by integrating traditional machine learning and advanced deep learning techniques. The system employs Mel-Frequency Cepstral Coefficients (MFCCs) for effective feature extraction and utilizes three classification models—Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Transformer networks—to distinguish between real and synthetic speech. By comparing these models, the study demonstrates the advantages of deep learning approaches in capturing temporal and spectral inconsistencies commonly introduced during deepfake audio generation. The implementation of a

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Flask-based web application enables real-time deepfake audio detection through an intuitive user interface. The system supports dynamic model selection, allowing users to analyze audio files using different classifiers and observe comparative results. Designed to operate efficiently on CPU-based environments, the system is accessible, cost-effective, and suitable for real-world deployment in academic and low-resource settings. Experimental evaluation shows that while the SVM model provides fast and computationally efficient baseline performance, LSTM and Transformer models achieve higher detection accuracy by effectively modeling temporal dependencies in speech signals. Overall, the proposed system enhances trust and security in audio-based digital communication by providing an accurate, scalable, and user-friendly deepfake detection solution.

REFERENCES

[1] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015). Spoofing and countermeasures for speaker verification: A survey. Speech Communication, 66, 130–153.
[2] Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech Communication, 52(1), 12–40.
[3] Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., & Shchemelinin, V. (2017). Audio replay attack detection with deep learning frameworks. Interspeech 2017, 82–86.
[4] Wang, X., Yamagishi, J., & King, S. (2020). Neural source-filter-based waveform model for statistical parametric speech synthesis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 402–415.
[5] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008
[6] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems (NeurIPS).
[7] Todisco, M., Delgado, H., & Evans, N. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. Interspeech 2019, 1008–1012.
[8] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. Advances in Neural Information Processing Systems, 577–585.
[9] Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2), 236–243.
[10] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com