Safeguarding Audio Integrity: Ensemble Architectures for Real-World Fake Audio Identification

Soundarya Lahari Kasturi1, Nivetha S1, Dr. Selva Rani B1, Dr. S Vairamuthu2

Sourabh Tiwari3, Devendra Yadav3, Rashmi T Shankarappa3

1School of Computer Science Engineering and Information Systems, VIT University, Vellore, India

2School of Computer Science and Engineering, VIT University, Vellore, India

3Voice Intelligence Group, Samsung R&D Institute, Bengaluru, India

{kasturi.srisoundarya2020, nivetha.s2020a}@vitstudent.ac.in, {bselvarani, svairamuthu}@vit.ac.in, {sourabh.t, devendra.yd, rashmi.ts}@samsung.com

*Abstract*— **In this research, we introduce a unique Audio Spoof Detection System for distinguishing between actual, fraudulent, and produced audio recordings. Strong detection algorithms are required because the spread of audio modification techniques poses a serious danger to the integrity of audio-based systems.**

**Combination of 7 key voice features from cepstral coefficients and spectrograms are used to cover all possible spoofing. We are proposing ensembling methods to enhance the accuracy of conventional machine learning models such as Random Forest, State Vector Classifier (SVC), K-Nearest Neighbors (KNN) etc. and deep learning models such as Recurrent Neural Networks (RNN), Time-Delay Neural Networks (TDNN), SqueezeNet etc. The proposed model is evaluated using the ASVspoof 2021 logical access (LA) and Deep Voice dataset. Experimental results show that our proposed model significantly elevates performance compared to the baseline and state-of-the-art models.**

Keywords—Audio Spoof Detection, Machine learning, Deep Learning, Ensemble learning, CNN, Random Forest, TDNN

# Introduction

Automatic speaker verification (ASV) is one of the widely used technologies to automate tasks and biometric recognition. Unfortunately, when ASV is used for security purposes, it faces a major threat due to spoofing attacks, including replay attacks, speech synthesis, voice conversion etc. These spoofing attacks include both generated and replayed audio. With the advancement of technologies like artificial intelligence, a lot of criminal activities are also rapidly increasing by generating fake audio to surpass edge devices. This has also become one of the growing concerns in safeguarding our privacy with devices such as cell phones and intelligent speakers. Existing voice spoofing attacks can be classified into physical access attacks, i.e., replay [1][2] or logical-access attacks, i.e., speech synthesis [3][4], VC [5][6].

In this paper, we analysed the existing works and have presented our study using popular datasets including ASV Spoof 2021 LA (Logical Access) and DEEP-VOICE audio dataset [7][8]. The Logical Access dataset of the ASV Spoof Challenge consists of spoofed and bona fide utterances generated using text-to-speech (TTS) and voice conversion (VC) algorithms are communicated across telephony and VoIP networks with various coding and transmission effects. DEEP-VOICE is comprised of real human speech from eight well-known figures and their speech is converted to one another using Retrieval-based Voice Conversion.

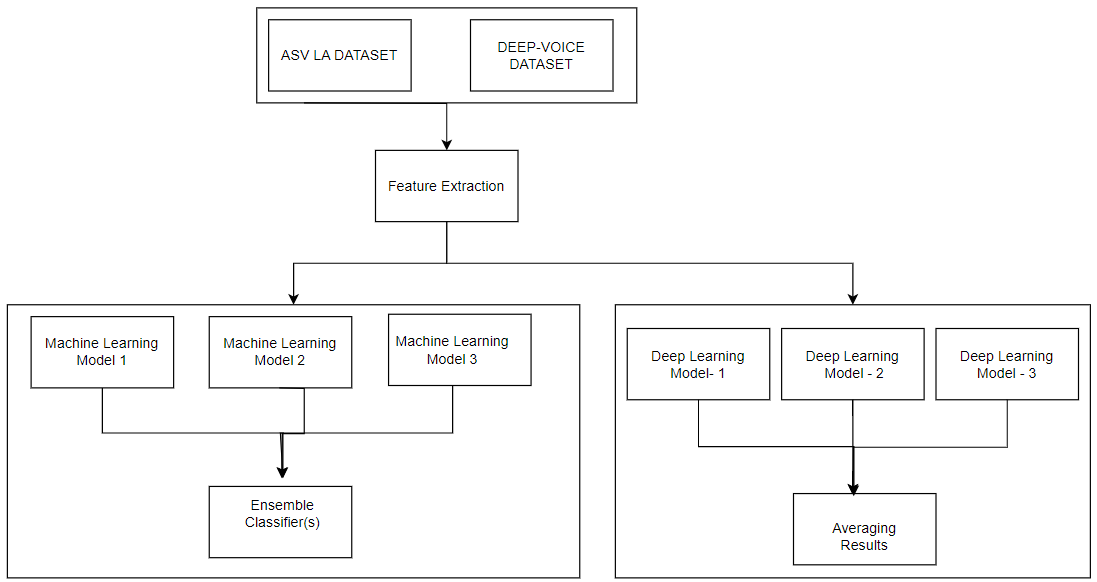


Fig 1. General Architecture of the Audio spoof detection system

The rest of the paper is organized as follows: Section II. Related Works in the field of spoof detection. Section III. Methodology: introduces and explain the important concepts that have been implemented in this work. Section IV. Experimental Results describes the results obtained measured using various standard metrics. Section V. Conclusion and Future Works concludes the paper while highlighting the future possibilities in the same direction and scope. Section VI. References used for the study.

# Related Work

To avoid the potential misuse of the fake audio with the ASV systems a lot of research is going on with focus on robust voice deep fake detection. One such instance is described through the ASV Spoof initiative, which focuses on automatic speaker verification spoofing and countermeasures [9]. Similarly, the FAD dataset was introduced for fake audio detection, to address the need for more generalized detection methods [10].

In view of detection methods, researchers have used the techniques of deep learning, machine learning, statically novel methods in feature engineering. Ensemble architectures have also been studied for exploring spoof detection with spectrograms.

Zheng Gu, Xinzhou Xu et al. proposed a novel approach that jointly uses Synthesized Classifiers (SynC) and pre-trained models for the zero-shot audio classification task, addressing the deficiencies by considering synthesized classifiers from discriminative learning on seen audio classes and the usage of pre-trained models learnt from image processing. This method consisted of three major steps, feature extraction (Spectrogram, ResNet Model), class-label encoding and zero-shot learning. For class-label encoding word2vec was employed which pre-trained on the Google News corpus. The Environmental Sound Classification (ESC-50) dataset was used for evaluation purposes. [11]

Hussian Dawood, Sajid Saleem et al. proposed a novel feature descriptor Center Lop-Sided Binary Patterns (CLS-LBP) for audio representation. The proposed CLS-LBP features are used to train the long short-term memory (LSTM) network for detection of both the physical- (replay) and logical-access attacks (speech synthesis, voice conversion). They have used ASV Spoof 2019 corpus to evaluate the performance of their work. They employed the LSTM network comprising of 10 LSTM layers with 100 hidden units in each layer, fully connected layer followed by a SoftMax layer and a classification layer. The approach obtained an EER of 0.58% and 2.91%, and min-tDCF of 0.16 and 0.072 on the evaluation and development datasets. [12]

Juan M. Martin-Donas and Aitor Alvarez employed a W2V2 feature extractor for audio deepfake detection system, which obtains encoded speech representations, and a classification model that scores the input audio as genuine or spoof. They evaluated their system using ASVSpoof2021 and 2022 ADD challenge database. The models were trained using the Adam optimizer with a default learning rate. At each of the 10 epochs the model was evaluated using the corresponding dev set. It was mainly developed for tracks 1& 2 of the ADD challenge and the training was carried out as per the challenge’s conditions. [13]

Piotr Kawa et al. worked towards decreasing the computational requirements and inference time of audio DF detection. As a part of their study, they introduced SpecRNet — a novel spectrogram–based model inspired by RawNet2 backbone. The testing of the model was done using WakeFake dataset. Each of the two architectures (RawNet2 and SpecRNet) were trained and tested using 3 different random seeds. Their model gave an improvement of 40% in inference speed. [14]

Kishor Barasu Bhangale et al. surveyed various deep learning architectures with respect to speech processing. Major applications such as speech enhancement, speech separation, speech recognition, speaker recognition was focused while analyzing the structures and applications of Auto-Encoders (AE), Generative Adversarial Network (GAN), Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Deep Reinforcement Learning (DRL). Different datasets were used for each of the application such as TIMIT, LibriSpeech, VoxCeleb etc. [15]

Janavi Khochare, Chaitali Joshi et al. focused on comparing two approaches for fake audio detection namely, a feature-based approach using machine learning algorithms and an image-based approach using deep learning algorithms. After their extensive comparison, they proposed an architecture that is inclined towards deep learning. They used datasets namely FoR Dataset and Generated Audios. Accuracies of the models were in the range of 0.60–0.70 for feature based and TCN, STN models gave an accuracy of 0.80-0.92 for image-based classification. [16]

Abderrahim Fathan, Jahangir Alam, Woo Hyun Kang explored the use of Mel-spectrogram image features and proposed a selection of multiple audio augmentations specially designed to resemble audio spoofing attack. They adopted WaveletCNN and VGG16 architectures and Mel-spectrogram image features in their study. They experimentally achieved a performance of 2.8% EER using ASVSpoof 2019 & 2021 challenge datasets. [17]

Akmalbek Bobomirzaevich Abdusalomov, Furkat Safarov et al. improved feature parameter extraction from speech signals using machine learning algorithms. They aimed to reduce the computation time. The dataset used contained 120hours of audio with over 90,650 utterances, 415,780 words and 65,810 unique words. As a part of their experiments, they evaluated their method of feature extraction with various recently published methods. The proposed model reduces the processing time and improves the feature extraction accuracy by 98.4% by effectively using MFCCs. [18]

Zhongjie Ba et al. employs transfer learning to enhance audio deepfake detection across languages. It utilizes a diverse multilingual dataset and adapts a pre-trained model. Effective cross-lingual detection is achieved, demonstrating improved accuracy and generalization in identifying audio deepfakes, showcasing the model's robustness and practical applicability. They have used RawNet2, AASIST, and Res-TTDSNet as the architectures of the model. [19]

Jun Xue et al., presents a novel approach for detecting audio deepfakes by combining fundamental frequency (F0) information with real and imaginary spectrogram features. The proposed method aims to improve the accuracy of deepfake detection using these combined audio characteristics. [20]

Rui Yan et al., adopts data augmentation as a primary strategy to improve the robustness of our systems based on known-unknown information as per the challenge evaluation plan. In addition, they proposed an innovative network stitching method, which improves the robustness of the model on different distributed test sets, and it shows good performance on all test sets of Track 3.2. Finally, the system outperforms all other proposed systems and won competition with an EER of 10.1%. [21]

Emanuele Conti et al. have presented a novel method for synthetic speech detection based on high-level semantic feature extraction. We focused on detecting deepfake speech tracks generated with TTS algorithms exploiting their emotional voice content. The system is composed of two main components. The first one is a SER network trained on a speech dataset annotated with the emotion expressed by the speaker and used as emotional feature extractor. By applying a transfer learning approach to this network, we can create an embedding space that is meaningful not only for the original task, i.e., SER, but also for the task at hand, i.e., synthetic speech detection. The second component is a supervised classifier that takes as input the emotional features and predicts if the given speech track is real or deepfake. [22]

# Proposed Methodology

The framework for robust audio spoof detection involves various stages, data pipelining, feature engineering, model training and evaluation. As we proceed in this section, let us look at the steps at each phase.

## Dataset: ASV Spoof 2021(Logical Access)

Designed to evaluate the robustness of spoofing countermeasures against text-to-speech (TTS) and voice conversion (VC) attacks that have been transmitted over real-world telephony and VoIP networks. Aims to bridge the gap between ideal laboratory conditions and more realistic scenarios encountered in practical applications. The Bona fide speech consists of recordings from 48 speakers (24 male, 24 female), each speaking short phrases in English. The Spoofed speech was generated using 17 different TTS and VC systems. The dataset is publicly available for research purposes.[23]

## Dataset: DEEP-VOICE

This dataset was produced from the study "Real-time Detection of AI-Generated Speech for Deep Fake Voice Conversion"[24]. It comprises of real human speech of eight speakers and their speech converted to one another using Retrieval-based Voice Conversion. The dataset consisted of 11778 rows with features extracted from one-second windows of audio and are balanced through random sampling.

## Feature Extraction

7 kinds of features were considered for both the datasets while that of ASV Spoof 2021 were manually extracted the dataset of DEEP-VOICE already had features extracted. The following features were extracted due to their varying importance in understanding the audio clip and developing the model.

Popularly known as fingerprint of audio, cepstral coefficients are the numerical values derived from transforming a signal’s frequency distribution in a specific way. These are less vulnerable to background noises or channel variations and hence best fit for real-world situations. We have used:

MFCC [25] which employs the Mel scale to mirror the non-linear way humans perceive pitch.

RFCC [26] is a variant of MFCC designed to capture relative spectral information less sensitive to channel variations. This resilience makes it effective for spoof detection.

LFCC [27] which employs a linear frequency scale instead of Mel scale. It is useful for tasks involving the need for high spectral resolution i.e., high quality recording.

Zero Crossing Rate (ZCR) [28] is the measure of the rate at which an audio signal crosses the zero-amplitude level. Thus, it distinguishes spoof by identifying the lack of natural variations in air pressure within the vocal tract.

Root Mean Square [29] measures the average over of an audio signal over a given time window. Thus, it reveals spoofed audio by examining inconsistent energy levels and manipulated background noises.

Spectrogram Centroid [30] represents the centre of mass of the spectrum. It provides insights into the characteristics of audio signals, aiding in distinguishing genuine and fake audio.

Spectrogram Roll Off [31] determines the frequency below which a certain percentage of the total spectral energy is concentrated. In context of spoof detection, it helps in identifying anomalies in the frequency distribution of audio signals.

The Chromagram [32] represents the energy distribution of different musical notes in an audio signal. It aids in identification of anomalies that may indicate spoofing with musical context.

Together all these features combating each other’s strengths and weakness would aid in robust and wholesome training of the model.

## Model Training

This phase involved various experiments with deep learning and machine learning algorithms.

1. Machine Learning Models:

Machine Learning algorithms have always been playing a crucial role in audio spoof detection from past works. Machine learning for audio classifier systems is highly effective when ensemble approaches are used. They can be methods such as bagging, boosting, stacking, and voting which combine the predictions of multiple base classifiers to improve overall accuracy and robustness.

The study involved training and evaluating various ensemble models such as XGBoost (Extreme Gradient Boosting), CatBoost (Categorical Boosting), LightGBM (Light Gradient Boosting Machine), Gradient Boosting, Stacking, Voting, and Bagging Classifiers, in addition to the Random Forest Classifier[33][34][35][36][37].

Stratified Sampling method was used to divided the huge datasets into homogeneous subgroups of 2000 samples each. Label Encoder has been used to convert the category labels into numerical values appropriate for machine learning models, which facilitated the training process. This method allowed for efficient training on the datasets and guaranteed compatibility with the algorithms used.

1. Machine Learning Ensemble Models:

After training the individual boosting ensemble models voting classifier, stacking classifier and bagging classifiers were used to combine the predictions of Gradient Boosting, Categorical Boosting and Random Forest and the results were compared with Extreme Gradient Boosting and Light Gradient Boosting.

In context of voting classifier ensembles, the selection of the right kind of voting method is crucial for achieving high classification accuracy in ensemble systems. The voting classifier employed "soft" voting strategies in all the experiments. A soft voting strategy involves weighing the predictions of each model based on their confidence scores. It aims to capitalize on the strengths of each model while mitigating their individual weaknesses and thus makes more accurate and robust ensemble predictions [38].

In context to stacking and bagging classifiers logistic regression and random forest were used respectively for combining the predictions of the base models.

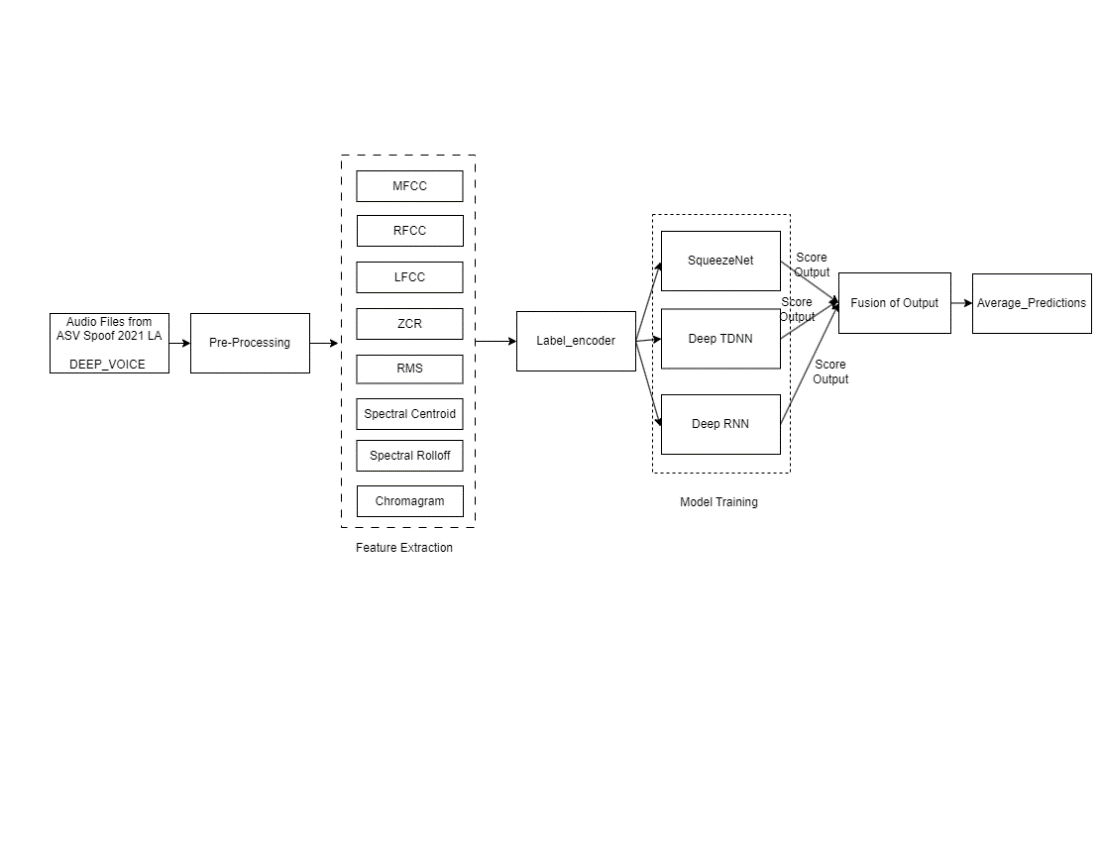
The training and construction of an ensemble of various neural network models for audio spoof detection use the individual strengths of TDNN, RNN, CNN, SqueezeNet. These models are leveraged to enhance the overall predictive performance, resulting in a more accurate and robust ensemble model. The ensemble is made up of three Deep Learning models best suited for audio spoof detection i.e. TDNN (Time-Delayed neural network), RNN (Recurrent Neural Network), SqueezeNet. The Keras API is used to define and compile each model separately.

1. Deep Learning Model-1:

The training and construction of an ensemble of various neural network models for audio spoof detection use the individual strengths of TDNN, RNN, CNN, SqueezeNet. These models are leveraged to enhance the overall predictive performance, resulting in a more accurate and robust ensemble model. The ensemble is made up of three Deep Learning models best suited for audio spoof detection i.e. TDNN (Time-Delayed neural network), RNN (Recurrent Neural Network), SqueezeNet. The Keras API is used to define and compile each model separately.

Similarly, RNN is used to capture sequential dependencies, which is helpful for time series data. The architecture consists of a simple Recurrent layer, Dense layers with Dropout and Output layer as similar in TDNN. The third model in the ensemble architecture is the CNN Model. The CNN model is proficient at learning heirarchial features in spatial data. The architecture is as follows: Convolution layer with MaxPooling, Flatten layer, Dense layer with Dropout, Output Layer. The two convolutional layers with 64 filters and kernel size 3 is followed by a maxpooling layer, the dense layers with 256 neurons and relu activation are introduced and each followed by a dropout layer with rate of 0.5. For the LA dataset, the RNN and CNN models attain competitive accuracies of 89.81%, and for the DeepVoice dataset, 84.76%. Reduced generalization may result from overfitting CNN and RNN architectures to the training set. We need to investigate stronger architectures to mitigate overfitting.

1. Deep Learning Model-2:

The purpose of the TDNN model is to extract temporal dependencies from sequential data, like audio characteristics. The architecture includes the following layers: Convolutional layers, MaxPooling Layers, Flatten Layer, Dense layer with Dropout and the Output Layer. The Convolutional layer has 64 filters and a kernel size of 3 which is used to convolve over the input sequence. In this way, local patterns and feature representations can be learned by the model. To add non-linearity, the relu activation function is utilized. The MaxPooling layer with pool size 2 are applied to down sample the spatial dimensions. Similarly, the Flatten layer is used to convert the 2D spatial output into a 1D vector, the Dense layer wih Dropout is made up of 2 dense layers, each with 256 neurons and relu activation with a dropout rate of 0.5 to prevent overfitting. The Output Layer with Sigmoid activation function produces binary classification results. For the LA and DeepVoice datasets, this change led to notable accuracy gains of 89.81% and 89.98%, respectively. One of the primary challenges is the need for robust feature extraction. Fake audio detection requires robust features, and there is ongoing research on methods for extracting features that are resilient to spoofing attacks [39]. TDNN models may require significant hyperparameter adjustment and are vulnerable to computational complexity issues. To overcome these obstacles, careful model selection and fine-tuning were required to achieve the best possible balance between computational efficiency and model performance.

1. Deep Learning Model-3:

SqueezeNet's architecture was selected to solve issues with earlier models. This architecture, which achieves a noteworthy accuracy of 97.66%, is well-known for its effective yet strong design. SqueezeNet, which consists of global average pooling, fire modules, and convolutional layers, is an efficient feature extraction algorithm that requires little processing power. In order to decrease model parameters and speed up training and inference without compromising expressive capability, the convolutional layers use 1x1 filters. Fire modules allow the network to capture complex patterns at various scales by effectively combining 1x1 and 3x3 convolutions. By substituting fully connected layers with global average pooling, overfitting is decreased, and model generalization is improved. The potential of this simplified design to attain excellent accuracy with fewer parameters, so addressing issues related to overfitting and computational complexity, makes it stand out. SqueezeNet's novel design, featuring fire modules and global average pooling, leads to its success in audio spoof detection, making it a valuable complement to the ensemble model.

1. Deep Learning Ensemble Architecture:

The ensemble model has used SquuezeNet, Deep TDNN, Deep RNN.

SqueezeNet architecture, Deep Time-Delay Neural Network (TDNN), and Deep Recurrent Neural Network (RNN) can be combined to create an ensemble model for audio spoof detection. SqueezeNet is well-known for its effective performance and lightweight architecture. SqueezeNet is a lightweight deep learning architecture. Global average pooling, fire modules, and convolutional layers make up this system. To improve feature extraction and capture temporal dependencies in the audio data, it can be combined with Deep TDNN and Deep RNN. SqueezeNet's architecture prioritizes accuracy preservation over model sizing reduction, which helps with effective audio data feature extraction—a critical component of spoof detection. Furthermore, the ability of the TDNN and RNN components to recognize long-term relationships and temporal patterns in the audio signals is crucial for differentiating between real and fake sounds. By utilizing the advantages of each design, the ensemble model can enhance audio spoof detection performance overall (Fig 2.).

Fig 2. Ensemble Deep Learning Architecture

The ensemble model then combines the result of all three models trained using a simple averaging approach. Each model's predictions from the ensemble architecture are averaged, and a threshold of 0.5 is set to obtain the final ensemble prediction. Using an averaging method, the ensemble model incorporates predictions from SqueezeNet, TDNN, and RNN. A balanced and reliable conclusion is produced by averaging the individual model outputs to arrive at the final prediction.

Ensemble Prediction = (Deep TDNN\_Predcition + Deep RNN\_Prediction + SqueezeNet\_Prediction)/3

The ensemble model containing SqueezeNet architecture, TDNN, and RNN can be a strong tool for audio spoof detection, harnessing the capabilities of each architecture to boost feature extraction and capture temporal relationships. By combining the advantages of SqueezeNet, TDNN, and RNN, the ensemble model improves overall performance and tackles specific problems such as overfitting. The ensemble model's robustness and accuracy can be further increased by using the averaging strategy for output prediction, which makes it a good choice for identifying audio spoofing attempts.

## Evaluation Metrics

The results of the models were evaluated using four primary metrics namely accuracy, f1-score, equal error rate(EER) and tandem detection cost function(t-DCF) [40]. Accuracy is one the intuitive measure for any classification model as it gives the measure of the model predicting the correct classes. F1-score is the harmonic mean of recall and precision thus helps in balanced assessment of the model in identifying both the classes. Tandem detection cost function (t-DCF) is the primary evaluation metric of ASV Spoof 2021. Equal Error Rate (EER) is a point at which the miss rate is equivalent to the false alarm rate in a countermeasure error detection system. A lower EER value corresponds to higher accuracy.

# Experimental Results

This section describes the results obtained from the experiments conducted as described in the previous sections for audio deep fake detection. Table 1 represents the results obtained using machine learning models in ensemble architecture with voting classifier. Table 2&3 represents the results obtained using deep learning models for feature-based audio spoof detection task.

1. Machine Learning Models (DEEP-VOICE)

| Model | Evaluation Metrics | | | |
| --- | --- | --- | --- | --- |
| Accuracy | F1-Score | EER | t-DCF |
| Random Forest | 98 | 98 | 0.00 | 52 |
| Gradient Boosting | 96 | 96 | 0.00 | 79 |
| Categorical Boosting | 99 | 99 | 0.00 | 32 |
| XGBoosting | 98 | 98 | 0.00 | 53 |
| Light GBM | 98 | 98 | 0.00 | 52 |
| Voting  (RF+GB+CB) | 98 | 99 | 0.00 | 42 |
| Stacking  (RF+GB+CB) | 99 | 99 | 0.00 | 31 |
| Bagging (RF+GB+CB) | 97 | 99 | 0.00 | 67 |

1. Machine Learning Models (LA DATASET)

| Model | Evaluation Metrics | | | |
| --- | --- | --- | --- | --- |
| Accuracy | F1-Score | EER | t-DCF |
| Random Forest | 90 | 95 | 0.00 | 401 |
| Gradient Boosting | 89 | 94 | 0.00 | 413 |
| Categorical Boosting | 90 | 95 | 0.00 | 410 |
| Light GBM | 90 | 95 | 0.00 | 410 |
| XGBoosting | 90 | 95 | 0.00 | 410 |
| Voting  (RF+GB+CB) | 90 | 95 | 0.00 | 410 |
| Stacking  (RF+GB+CB) | 90 | 95 | 0.00 | 410 |
| Bagging (RF+GB+CB) | 90 | 95 | 00 | 410 |

1. Deep Learning Models (la dataset)

| Model | Evaluation Metrics | | | |
| --- | --- | --- | --- | --- |
| Accuracy | F1-Score | EER | t-DCF |
| TDNN | 89.81 | 94.63 | 0 | 0.01 |
| RNN | 89.81 | 94.63 | 0 | 0.01 |
| Ensemble(TDNN+RNN+CNN) | 89.81 | 94.63 | 0 | 0.01 |
| SqueezeNet | 89.81 | 94.63 | 0 | 0.89 |
| Ensemble(SqueezeNet+Deep TDNN+Deep RNN) | 89.81 | 94.63 | 0 | 0.01 |

1. Deep Learning Models (deep-voice dataset)

| Model | Evaluation Metrics | | | |
| --- | --- | --- | --- | --- |
| Accuracy | F1-Score | EER | t-DCF |
| TDNN | 89.98 | 90.55 | 0 | 0.01 |
| RNN | 84.76 | 83.97 | 0 | 0.01 |
| Ensemble(TDNN+RNN+CNN) | 98.47 | 97.44 | 0 | 0.01 |
| SqueezeNet | 96.98 | 97.01 | 0 | 0.49 |
| Ensemble(SqueezeNet+Deep TDNN+Deep RNN) | 98.30 | 0.98 | 0 | 0.00033 |

# Conclusion

In this work, we have understood the necessity of developing a robust spoofed audio detection system. Our main contribution to the task was to provide an ensemble architecture to combat the misuse of generated and forged audio. This work focuses mainly on exploring the various machine learning and deep learning algorithms with a variety of feature sets extracted from diverse and huge datasets. Through extensive study of different approaches our study found that the Deep Learning Ensembles show more promising results in comparison to traditional Machine Learning Ensembles. The ensemble model with machine learning algorithms gave an accuracy of 70.50% to 98% for DEEP-VOICE dataset and 89% to 90% for LA Dataset. The proposed deep learning approach gave 98.30% and 89.81% accuracy for DEEP-VOICE & LA datasets respectively proving equally competing and promising.

These findings will have an impact on the future works in the field of securing ASVs from spoofed and fake audio. Keeping in view the real-world scenarios with huge noise distortions and poor channel quality, the future models should focus on incorporating deep learning with other hybrid techniques. We aim to further carry out our research in combining memory-based algorithms with standard traditional algorithms to combat each other's weaknesses.

##### References

1. Alegre, F., Janicki, A., Evans, N. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. In: 2014 International Conference of the Biometrics Special Interest Group (BIOSIG). 2014. IEEE.
2. Rosenberg, A.E., 1976. Automatic speaker verification: a review. Proc. IEEE 64 (4), 475–487.
3. Yamagishi, J. et al., 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Language Process 17 (1), 66–83.
4. Lindberg, J., Blomberg, M., 1999. Vulnerability in speaker verification-a study of technical impostor techniques. Sixth European Conference on Speech Communication and Technology.
5. Evans, N. et al., 2009. Anti-spoofing: voice conversion. Encycl Biometr, 1–10.
6. Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064.
7. Delgado, H., Evans, N., Kinnunen, T., Lee, K. A., Liu, X., Nautsch, A., Patino, J., Sahidullah, M., Todisco, M., Wang, X., & Yamagishi, J. (2021, May 28). ASVspoof 2021 Challenge - Logical Access Database (1.1). Zenodo. https://doi.org/10.5281/zenodo.4837263
8. <https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition>
9. Z. Wu et al., "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 588-604, June 2017, doi: 10.1109/JSTSP.2017.2671435.
10. Ma, H., Yi, J., Wang, C., Yan, X., Tao, J., Wang, T., … & Fu, R. (2022). Fad: a chinese dataset for fake audio detection. <https://doi.org/10.48550/arxiv.2207.12308>
11. Z. Gu, X. Xu, S. Liu and B. Schuller, "Zero-Shot Audio Classification Using Synthesised Classifiers and Pre-Trained Models," 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 2022, pp. 1-6, doi: 10.1109/CISP-BMEI56279.2022.9980329.
12. Hussain Dawood, Sajid Saleem, Farman Hassan, Ali Javed, A robust voice spoofing detection system using novel CLS-LBP features and LSTM, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 9,2022, Pages 7300-7312, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2022.02.024.
13. Martín-Doñas, J. M., & Álvarez, A. (2022). The Vicomtech Audio Deepfake Detection System based on Wav2Vec2 for the 2022 ADD Challenge. arXiv [Eess.AS]. Retrieved from http://arxiv.org/abs/2203.01573
14. Kawa, P., Plata, M., & Syga, P. (2022). SpecRNet: Towards Faster and MoreAudio DeepFake Detection. arXiv [Cs.SD]. Retrieved from http://arxiv.org/abs/2210.06105
15. Bhangale, K.B., Kothandaraman, M. Survey of Deep Learning Paradigms for Speech Processing. Wireless Pers Commun 125, 1913–1949 (2022). https://doi.org/10.1007/s11277-022-09640-y
16. Khochare, J., Joshi, C., Yenarkar, B. et al. A Deep Learning Framework for Audio Deepfake Detection. Arab J Sci Eng 47, 3447–3458 (2022). https://doi.org/10.1007/s13369-021-06297-w
17. A. Fathan, J. Alam and W. H. Kang, "Mel-Spectrogram Image-Based End-to-End Audio Deepfake Detection Under Channel-Mismatched Conditions," 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 2022, pp. 1-6, doi: 10.1109/ICME52920.2022.9859621.
18. Abdusalomov, A.B.; Safarov, F.; Rakhimov, M.; Turaev, B.; Whangbo, T.K. Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm. Sensors 2022, 22, 8122. https://doi.org/10.3390/s22218122
19. Ba, Z., Wen, Q., Cheng, P., Wang, Y., Lin, F., Lu, L., & Liu, Z. (2023, April). Transferring Audio Deepfake Detection Capability across Languages. In Proceedings of the ACM Web Conference 2023 (pp. 2033-2044).
20. Xue, J., Fan, C., Lv, Z., Tao, J., Yi, J., Zheng, C., ... & Shao, S. (2022, October). Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features. In Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia (pp. 19-26).
21. Yan, R., Wen, C., Zhou, S., Guo, T., Zou, W., & Li, X. (2022, May). Audio deepfake detection system with neural stitching for add 2022. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 9226-9230). IEEE.
22. Conti, E. (2020). Synthetic speech detection through emotion recognition: a semantic approach.
23. Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., ... & Delgado, H. (2021). ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. arXiv preprint arXiv:2109.00537.
24. Bird, J. J., & Lotfi, A. (2023). Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion. *arXiv [Cs.SD]*. Retrieved from http://arxiv.org/abs/2308.12734
25. Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012, July). Speech recognition using MFCC. In *International conference on computer graphics, simulation and modeling* (Vol. 9).
26. [https://s311354.github.io/Louis.github.io/2019/09/10/Introduction\_of\_feature\_representations\_of\_the\_speech\_signal/#RASTA](#RASTA)
27. Dewi, S. P., Prasasti, A. L., & Irawan, B. (2019, July). The study of baby crying analysis using MFCC and LFCC in different classification methods. In *2019 IEEE International Conference on Signals and Systems (ICSigSys)* (pp. 18-23). IEEE.
28. Ito, M., & Donaldson, R. (1971). Zero-crossing measurements for analysis and recognition of speech sounds. *IEEE Transactions on Audio and Electroacoustics*, *19*(3), 235-242.
29. Phinyomark, A., Limsakul, C., & Phukpattaranont, P. (2011). Application of wavelet analysis in EMG feature extraction for pattern classification. *Measurement Science Review*, *11*(2), 45-52.
30. Kua, J. M. K., Thiruvaran, T., Nosratighods, M., Ambikairajah, E., & Epps, J. (2010, June). Investigation of spectral centroid magnitude and frequency for speaker recognition. In *Odyssey* (p. 7).
31. Kos, M., Kačič, Z., & Vlaj, D. (2013). Acoustic classification and segmentation using modified spectral roll-off and variance-based features. *Digital Signal Processing*, *23*(2), 659-674.
32. Birajdar, G. K., & Patil, M. D. (2020). Speech/music classification using visual and spectral chromagram features. *Journal of Ambient Intelligence and Humanized Computing*, *11*, 329-347.
33. Strobl, C., Malley, J. D., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.. Psychological Methods, 14(4), 323-348. <https://doi.org/10.1037/a0016973>
34. Naranjo-Alcazar, J., Perez-Castanos, S., Martín-Morató, I., Zuccarello, P., Ferri, F. J., & Cobos, M. (2020). A comparative analysis of residual block alternatives for end-to-end audio classification. IEEE Access, 8, 188875-188882. <https://doi.org/10.1109/access.2020.3031685>
35. Iqbal, F., Abbasi, A., Javed, A. R., Jalil, Z., & Al-Karaki, J. (2022). Deepfake Audio Detection via Feature Engineering and Machine Learning. In Woodstock’22: Symposium on the irreproducible science. IEEE (pp. 70-75).
36. X. Sun, J. Fu, B. Wei, Z. Li, Y. Li and N. Wang, "A Self-Attentional ResNet-LightGBM Model for IoT-Enabled Voice Liveness Detection," in IEEE Internet of Things Journal, vol. 10, no. 9, pp. 8257-8270, 1 May1, 2023, doi: 10.1109/JIOT.2022.3230992.
37. Jin W, Wang X, Zhan Y. Environmental Sound Classification Algorithm Based on Region Joint Signal Analysis Feature and Boosting Ensemble Learning. Electronics. 2022; 11(22):3743. https://doi.org/10.3390/electronics11223743
38. Anagnostopoulos, Theodoros & Skourlas, Christos. (2014). Ensemble Majority Voting Classifier for Speech Emotion Recognition and Prediction. Journal of Systems and Information Technology. 16. 10.1108/JSIT-01-2014-0009.
39. Lou, J., Xu, Z., Zuo, D., & Liu, H. (2021). Feature extraction method for hidden information in audio streams based on hm-emd. Security and Communication Networks, 2021, 1-12. https://doi.org/10.1155/2021/5566347
40. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, “t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification,” in Speaker Odyssey 2018 The Speaker and Language Recognition Workshop, (Les Sables d’Olonne, France), June 2018.