Journal of Robotics and Control (JRC)
Volume 5, Issue 6, 2024
ISSN: 2715-5072, DOI: 10.18196/jrc.v5i6.23502

Enhancing Voice Authentication with a Hybrid Deep Learning and Active Learning Approach for Deepfake Detection

Ali Saadoon Ahmed [1*], Arshad M. Khaleel [2]
[1] Department of Computer Sciences, College of Sciences, University of Al Maarif, Al Anbar, 31001, Iraq
[2] Iraq Ministry of Education 2nd International Smart Card Al-Anbar, Iraq
Email: [1] ali.sadoon@uoa.edu.iq, [2] eng.arshed9@gmail.com
*Corresponding Author

***Abstract*—This paper explores the application of active learning to enhance machine learning classifiers for spoofing detection in automatic speaker verification (ASV) systems. Leveraging the ASVspoof 2019 database, we integrate an active learning framework with traditional machine learning workflows, specifically focusing on Random Forest (RF) and Multilayer Perceptron (MLP) classifiers. The active learning approach was implemented by initially training models on a small subset of data and iteratively selecting the most uncertain samples for further training, which allowed the classifiers to refine their predictions effectively. Experimental results demonstrate that while the MLP initially outperformed RF with an accuracy of 95.83% compared to 91%, the incorporation of active learning significantly improved RF's performance to 94%, narrowing the performance gap between the two models. After applying active learning, both classifiers showed enhanced precision, recall, and F1-scores, with improvements ranging from 3% to 5%. This study provides valuable insights into the role of active learning in boosting the efficiency of machine learning models for dynamic spoofing scenarios in ASV systems. Future research should focus on designing advanced active learning techniques and exploring their integration with other machine learning paradigms to further enhance ASV security.**

Keywords—Active Learning; Machine Learning; Automatic Speaker Verification; Asvspoof 2019; Random Forest; MLP; Spoofing Detection.

I. INTRODUCTION

The availability of affordable digital devices such as smartphones, tablets, laptops, and digital cameras has led to a rapid increase in multimedia content, including photos and videos, on the Internet. This increase has been fueled by the development of social media platforms over the past decade, which facilitates rapid sharing and thus improves the volume and accessibility of multimedia content. At the same time, the fields of machine learning have seen significant progress with the development of advanced algorithms capable of modifying multimedia content. These developments allow misinformation to spread through social media, complicating the ability to discern the truth and trust information. This challenge is particularly acute in the current "post-truth era," where information, whether true or false, can be used as a weapon to influence public opinion, with potential consequences such as election interference, incitement to conflict, and defamation.

The emergence and improvement of previous deepfakes technology, which produces synthetic video and audio clips, represents a major step in the ability to spread rumors on a global scale. This poses a potential threat to the spread of fake news. Deepfake, being artificial constructions generated by artificial intelligence, challenges the traditional reliance on video evidence in legal and criminal contexts, where the reliability and integrity of such evidence is paramount.

Forensic experts, especially those dealing with social media and sharing platforms such as YouTube and Facebook, face challenges in achieving authentication and integrity standards due to the complexity of deep data processing technology. The emergence of easy-to-use manipulation tools such as Zao, Rivas, Visap, Audacity, and Sound forge has made the verification process more complex. This makes it difficult to distinguish between original and edited videos [1]-[5].

Deepfake videos are generally classified into five categories: 1) face replacement, 2) lip sync, 3) puppet fest, 4) face editing and feature manipulation, and 5) deepfake audio. This type of deepfake often targets celebrities or public figures to tarnish their image by placing them in fabricated scenarios [6], including non-consensual adult content [7]. Lip-syncing deepfakes adjust the target's lip movements to match a specific audio track, falsely portraying them as saying words they never actually spoke. Puppet-master deepfakes animate a target's facial expressions and movements, including eye and head movements, to mimic another's expressions or actions [8]. Face synthesis involves generating realistic faces or altering facial attributes, commonly used to create fictitious social media profiles. Lastly, audio deepfakes [79][80] replicate a person's voice through deep learning, enabling the creation of fake audio clips where the target appears to utter statements they never made [9][10], utilizing either text-to-speech synthesis (TTS) or voice conversion (VC).

Despite the significant focus on video-based deepfakes, audio deepfakes have garnered less attention. Recent advancements in voice manipulation pose a threat not only to speaker verification systems but also to voice activated IoT devices [11][12]. Voice cloning can undermine public trust and has been exploited in criminal activities, such as a reported incident where bank robbers used cloned executive

voices to orchestrate a fraudulent transfer of funds [13]. The amalgamation of voice cloning into deepfake technology presents a novel challenge in detecting such forgeries. It underscores the necessity of developing detection methods that consider both audio and video forgeries, rather than focusing solely on video.

Recent literature surveys on deepfake generation and detection have primarily concentrated on image and video aspects, neglecting audio deepfakes. For instance, one survey [14] explored general image manipulation and forensic techniques without delving into deepfake creation methods. Another [15] provided insights into face manipulation and detection techniques, while a further study.

In [16] focused on visual deepfake detection, omitting discussions on audio cloning and its identification. Mirsky et al. [17] presented a comprehensive analysis of visual deepfake creation but only briefly touched on detection methods and did not address audio deepfakes, highlighting a gap in the current research landscape that necessitates a more integrated approach to deepfake detection encompassing both audio and video modalities.

To address these challenges, this study focuses on the integration of audio and video deepfake detection to create a more comprehensive and robust solution. Despite the progress made in detecting visual deepfakes, the detection of audio deepfakes remains underexplored, presenting a significant gap in the current research landscape. This research aims to fill this gap by developing and evaluating machine learning models that leverage active learning to enhance the detection of both audio and video deepfakes. The study's objectives are to improve the accuracy, precision, recall, and F1-scores of these models, particularly in the context of automatic speaker verification (ASV) systems. By doing so, this work not only contributes to the existing body of knowledge but also provides practical solutions for enhancing secrity in systems vulnerable to deepfake attacks.

II. RELATED WORK

The development of accessible technologies for creating counterfeit audio has brought significant focus on detecting Audio Deepfakes (AD), especially across various languages. This discussion will highlight contemporary efforts in identifying forged and artificially generated voices. Present detection methodologies largely fall into two categories: Machine Learning (ML) and Deep Learning (DL) approaches.

Kumar-Singh and Singh [18] introduced a Quadratic Support Vector Machine (Q-SVM) to separate synthetic voices from natural ones in a binary classification framework. Their analysis, comparing various ML techniques including Linear and Quadratic Discriminant Analysis, Linear SVM, weighted K-Nearest Neighbors (KNN), boosted trees, and LR, highlighted Q-SVM's superiority with a 97.56% accuracy and a misclassification rate of 2.43%. Furthermore, Borrelli et al.

In [19] proposed an SVM combined with Random Forest (RF) approach utilizing a novel Short-Term Long-Term (STLT) audio feature. Trained on the Automatic Speaker Verification (ASV) spoof 2019 challenge dataset [20], the SVM outperformed RF by 71%. Liu et al. [21] explored the efficacy of SVM against DL's Convolutional Neural Network (CNN) in distinguishing authentic from fake stereo audio, noting CNN's superior robustness despite both methods achieving around 99% accuracy. These ML-based approaches, however, often require labor-intensive preprocessing for effective performance.

To streamline this process, DL techniques have been explored. Subramani and Rao [22] employed two CNN variants, Efficient CNN and RES-Efficient CNN, for synthetic audio detection, with RES-Efficient CNN outperforming its counterpart by securing a 97.61 F1-score on the ASV spoof 2019 challenge [20]. E.R. Bartusiak and E.J. Delp [23] compared CNN's accuracy against a random baseline in detecting synthetic audio, noting CNN's superior performance but also its tendency toward overfitting.

Lataifeh et al. [24] conducted an experimental comparison of CNN and Bidirectional Long Short-Term Memory (BiLSTM) against ML techniques using the AR-DAD dataset [24], aimed at identifying imitated Quranic audio clips. While SVM showed the highest accuracy at 99%, CNN surpassed BiLSTM with a 94.33% detection rate, demonstrating CNN's efficiency in feature extraction and generalization, despite lower accuracy compared to some ML models. However, a primary limitation of CNN in AD detection is its requirement for audio to be pre-processed into spectrograms or 2D figures for analysis, highlighting a need for models that can directly process audio inputs.

Recent advancements in generating deceptive audio have intensified efforts in Audio Deepfake (AD) detection, employing various approaches and models across different languages. This segment outlines notable innovations in identifying counterfeit and synthetically produced voices, classifying current methodologies into Machine Learning (ML) and Deep Learning (DL) approaches.

Lei et al. [25] introduced a dual-model approach comprising a 1-Dimensional CNN and a Siamese CNN for AD detection. The 1-D CNN analyzed speech log-probabilities, whereas the Siamese CNN, leveraging two identical CNNs linked by a fully connected layer, utilized trained Gaussian Mixture Models (GMM). Tested on the ASVspoof 2019 dataset, the Siamese CNN demonstrated superior performance over both the GMM and 1-D CNN models, particularly with Linear Frequency Cepstral Coefficients (LFCC) features, enhancing the minimum total Detection Cost Function (mint DCF) and Equal Error Rate (EER) by approximately 55%. However, its efficacy diminished with Constant Q Cepstral Coefficients (CQCC) features, indicating a dependency on specific feature types.

Another approach explored in [26] involved converting audio into scatter plot images for CNN analysis. Although this model showed promise in generalization across different audio generation algorithms, evidenced by its performance on the Fake or Real (For) dataset [27], it lagged in accuracy and EER compared to other DL models, indicating the need for further enhancements and diverse data transformation techniques.

Yu et al. [28] presented an innovative scoring method, Human Log-Likelihoods (HLLs), based on Deep Neural Networks (DNN), outperforming traditional Log-Likelihood Ratios (LLRs) derived from GMMs. Utilizing the ASV spoof challenge 2015 dataset [29], DNN-HLLs evidenced improved detection capabilities with more favorable EER outcomes.

Wang et al. [30] developed Deep-Sonar, a DNN model focusing on neuron behaviors in speaker recognition systems to identify AI-synthesized fake audio. Despite achieving high detection rates and low EER on the FoR dataset, Deep-Sonar's effectiveness was compromised by ambient noise. Wijethunga et al. [31] combined CNNs and Recurrent Neural Networks (RNN) to leverage CNN's feature extraction prowess and RNN's long-term dependency recognition. This hybrid approach marked a significant success rate in distinguishing AI-generated audio, although it was noted for its limited feature representation artifact information.

Chintha et al. [32] proposed two models based on Convolutional RNNs for AD classification: CRNN-Spoof and Wide Inception Residual Network Spoof (WIRE-Net-Spoof), with CRNN-Spoof slightly outperforming WIRE-Net-Spoof in t-DCF and EER metrics on the ASV spoof 2019 dataset. Nonetheless, these models faced challenges with complexity management due to their layered and convolutional architecture.

Addressing these complexities, Shan and Tsai [33] introduced an alignment technique employing LSTM, bidirectional LSTM, and transformer architectures for audio frame classification, with bidirectional LSTM showing superior accuracy and minimal EER. However, the model's extensive training time and small dataset size raised concerns about potential overfitting.

In the realm of transfer learning, P. RahulT et al. [34] explored a framework combining transfer learning with the ResNet-34 method, showcasing notable EER and t-DCF performance. Despite addressing the vanishing gradient problem, the model's deep architecture necessitated prolonged training periods. Khochare et al. [35] investigated TCN and STN models, with TCN demon- starting effectiveness in differentiating fake from real audio, though its application was limited with certain input transformations like STFT and Mel Frequency Cepstral Coefficients (MFCC) features.

Khalid et al. [36] introduced a novel Deepfake dataset, FakeAVCeleb [37], and assessed the effi- cacy of unimodal detection methods employing five distinct classifiers: MesoInception-4, Meso-4, Xception, EfficientNet-B0, and VGG16. Among these, the Xception classifier emerged as the most effective, achieving a 76% success rate in fake audio detection, whereas EfficientNet-B0 was the least effective, with a 50% detection rate. This study concluded that unimodal classifiers are generally insufficient for reliable fake audio detection.

Alzantot et al. [38] underscored the necessity for an advanced AD detection system leveraging a residual CNN framework. This system is designed to extract three key features: MFCC, CQCC, and STFT, to calculate the Counter Major (CM) score, indicating the authenticity of the audio. Their approach significantly enhanced the CM rate by 71% and 75% across two metrics: t-DCF (0.1569) and EER (6.02), respectively. Nevertheless, the potential for generalization errors necessitates further exploration.

T.Arif and colleagues[39] introduce a novel acoustic feature ELTP-LFCC which merges en- hanced local binary patterns (ELTP) with linear frequency spectral cepstral coefficients (LFCC). The target of using a combination of short- and long-memory two-way deep DBiLSTM network was detecting the diverse environment using the model with enhanced detection robustness. The ASVspoof 2019 tests convincingly illustrated MVAS' superiority in synthetic audio (0.74% EER) as well as its unsatisfactory efficiency when imitating samples (33.28% EER).

Gradually, Assertion of SENet with Residual Squeeze-Excitation Networks (ASSERT, resentence) has been proposed in [40]. The ASSERT system used logarithmic power factor spectra and CQCC configuration features, which were showcased to have enabled the system to accomplish precise synthetic audio sound detection that was above and beyond. On the other hand, the model revealed the symptoms of over spiking because the resulting t-DCF and EER were zero in the logical accessing event.

The field research shows that the deep learning approaches bypass the need for manual feature extraction and the extensive process of training but often require a spiteful conversion of audio data. The response to that was to bring the concept of self-supervised approach to deep learning to the Table I, which was presented by Jiang et al [41]. They developed Self-Supervised Spoofing Audio Detection (SSAD) model which is based on the PASE+ method. By implementing multilayer python analog transform blocks, SSAD achieved an identification accuracy of 5.31%, therefore, showing its capability and scalability. Despite this, the belonging compared to the peer deep learning methods shows that further development of an approach, which does not require explicit labelled data, will improve self-supervised learning in forgery detection cases for the future.

In addition to the models discussed, recent research has increasingly recognized the importance of integrating hybrid approaches that combine multiple techniques to enhance detection accuracy and reduce the limitations inherent in individual models. For example, hybrid models that leverage the strengths of both ML and DL techniques, such as combining CNNs with RNNs, have shown improved performance in capturing temporal dependencies and feature extraction. However, these approaches still face challenges related to computational complexity and the need for extensive data preprocessing. The exploration of transfer learning and self-supervised learning methods has also gained traction, as these approaches offer the potential to reduce the dependency on large labeled datasets, thereby addressing one of the key limitations of traditional DL models. Nonetheless, the effectiveness of these methods in practical, real-world scenarios remains an ongoing area of investigation, with current studies highlighting the need for further optimization to enhance their generalizability and robustness against diverse types of audio deepfakes.

TABLE I. RELATED WORK

Ref	Approach/ Model	Dataset	Key Features/Techniques	Performance Metrics	Limitations/ Observations
[18]	Q-SVM	ASVspoof 2019	ML techniques comparison	97.56% accuracy	Requires intensive preprocessing
[19]	SVM + RF	ASVspoof 2019	STLT audio feature	SVM Outperforms 71% by RF	-
[21]	SVM vs. CNN	-	Robustness comparison	Both 99% accuracy; CNN morerobust	SVM suffers in feature extraction
[22]	EfficientCNN, RES-EfficientCNN	ASVspoof 2019	DL models for detection	RES-EfficientCNN 97.61 F1-score	-
[23]	CNN	-	Comparison to baseline	CNN accuracy significantly higher	CNN prone to overfitting
[24]	CNN vs. BiL-STM	AR-DAD	Comparison with ML models	CNN 94.33% accuracy; BiL-STM lower	CNN requires audio preprocessing
[25]	1-D CNN, Siamese CNN	ASVspoof 2019	LFCC, CQCC features	Siamese CNN improves min t-DCF and EERby 55%	Performance varies by feature type
[26]	CNN (scatter plot images)	FoR	Generalization across algorithms	88.9% accuracy	Lower performance compared to others
[28]	DNN-HLLs vs. GMM-LLRs	ASV spoof challenge 2015	Scoring methods comparison	EER of 12.24 for DNN-HLLs	-
[30]	Deep-Sonar	FoR	Neuron behaviors in SR systems	98.1% detection rate; 2% EER	Affected by real-world noise
[31]	CNN + RNN	-	Hybrid model for feature extraction	94% success rate	Limited artifact information
[32]	CRNN-Spoof, WIRE-Net-Spoof	ASVspoof 2019	Convolutional RNNs	CRNN-Spoof slightly betterin t-DCF and EER	Complexity management
[33]	LSTM, BiL-STM, Transformer	-	Alignment technique	BiLSTM 99.7% accuracy; 0.43% EER	Long training, potential overfitting
[34]	Transfer learning + ResNet-34	-	Addressing vanishing gradient	Best EER and t-DCF metrics	Long training due to deep architecture
[35]	TCN, STN	-	Feature-based vs. image-based	TCN 92% accuracy; STN 80%	TCN's limitation with STFT and MFCC
[36]	Various classifiers	FakeAVCeleb	Unimodal detection methods	Xception 76% accuracy; EfficientNet-B0 50%	Unimodal classifiers generally insufficient
[38]	Residual CNN	-	MFCC, CQCC, STFT for CMscore	CM rate improved by 71% and 75%	Generalization errors
[39]	ELTP-LFCC + DBiLSTM	ASVspoof2019	ELTP-LFCC feature descriptor	Better with synthetic (0.74% EER) than imitated (33.28% EER)	-
[40]	ASSERT (SENet + ResNet)	ASVspoof2019	logspec, CQCC features	Over 17% improvements; zero t-DCF and EER	High overfitting
[41]	SSAD (PASE+inspired)	-	Self-supervised DL method	5.31% EER	Lower performance compared to other DL methods

III. PROPOSED METHODOLOGY

A. Data Preparation

The diagram illustrates a method, which is categorized by a sequence of steps dealing with the management of a machine learning project including a systematic and structured plan.

1) Exploratory Data Analysis (EDA)

At first, the stage takes data from the data set. Later, exploratory data analysis (EDA) is performed with a goal of comprehending the prevailing pattern and variation among the data [42]. The EDA results are conducted further to make the data ready for processing such as that includes imperative duties of handling missing values and balancing the data [43].

After completing the data preparation, the project branches into two main machine learning tracks: the first approach talks active learning using the classifiers such as MLP and RF[44], while the second follows more classical machine learning approach with the same classifiers[45]. In both scenarios, the training set is made up from 80% to the data set and the test is left out of the remaining 20% [46].

Active learning focuses on the process of choosing up to the best subset of unlabeled data into multiple groups to be labelled for training [47], which in some cases can even aid in building foolproof models with fewer labeled samples [48]. As with the standard machine learning approach, which, involves building and evaluating the model based on well-de[U+FB01] Ende training and test set.

Both paths run towards the measurement stage where the trained models are evaluated by different metrics exemplified by accuracy, recall, precision, and f1 ratio [49]. These metrics gives estimation of model performance at large and one more thing about confusion matrix is starting from this one can analyze in detail how the model predict the different classes [50]. Systematic approach enables us to rehearse, grade, and develop machinery learning model, which eventually leads the way to its accomplishment.

Fig. 1 shows the full process employed in our investigation. It should begin by taking the dataset through Exploratory Data Analysis (EDA) to explore basic patterns and ascertain the data characteristics. Here, down sampling and filling of missing values come under data preparation stage. train_test_split — to split the data into training and testing. Side by Sidetracks one active learning way or another traditional machine-learning two. Population side-by-side For both tracks MLP and RF classifiers are used. This stage is the evaluation of model algorithms using metrics like accuracy, recall,presicion,f1_score and confusion matrix.

The dataset [51] was employed in this challenge as we participate in the ASVspoof Challenge 2019. This challenge, being organized with the primary aim to explore and establish better ways to prevent spoofing attacks during automatic voice identity verification (ASV), acts as a testbed for robust ASV systems. The workshop primary organizers were authors of Junichi Yamagishi and Massimiliano Todesco, and as well as others, who based called their session on a collection of previous workshops led at the INTERSPEECH conferences since 2013. The initial main task of the ASVspoof Challenge initiative was to introduce the issues of ASVs false denominations perception and create relevant countermeasures to this problem.

The 2019 ASVspoof Challenge, following its predecessors in 2015 and 2017, is a comprehensive endeavor to address all three major types of spoofing attacks: TTS (text-to-speech), VC (voice- to-voice) and in addition to these attacks an unlimited number of other attacks can be categorized as attacks on speech. The 2019 exercise mirrored the 2017 edition by striking back at the attacks perpetrated by the unsupervised environment but moved from merely a detailed understanding of the problem to a more structured evaluation that included fashioning simulations of monitored settings.

Additionally, it included updated and sophisticated TTS and VC spoofing techniques, reflecting the technological advancements made since the last challenge.

Distinctive to this edition is the integration with the automatic speaker verification field through the incorporation of the tandem decision cost function (t-DCF) as an evaluation metric, emphasizing the interplay between ASV systems and countermeasures.

The dataset for ASVspoof 2019 is based on the VCTK corpus, featuring voice recordings from 107 speakers. It is split into two main partitions designed to evaluate countermeasures in both logical access (LA) and physical access (PA) scenarios. Each partition consists of three subsets: training, development, and evaluation, with 20, 10, and 48 speakers in each subset respectively. The subsets are exclusive in speaker identity, and the conditions for the original recordings are consistent throughout.

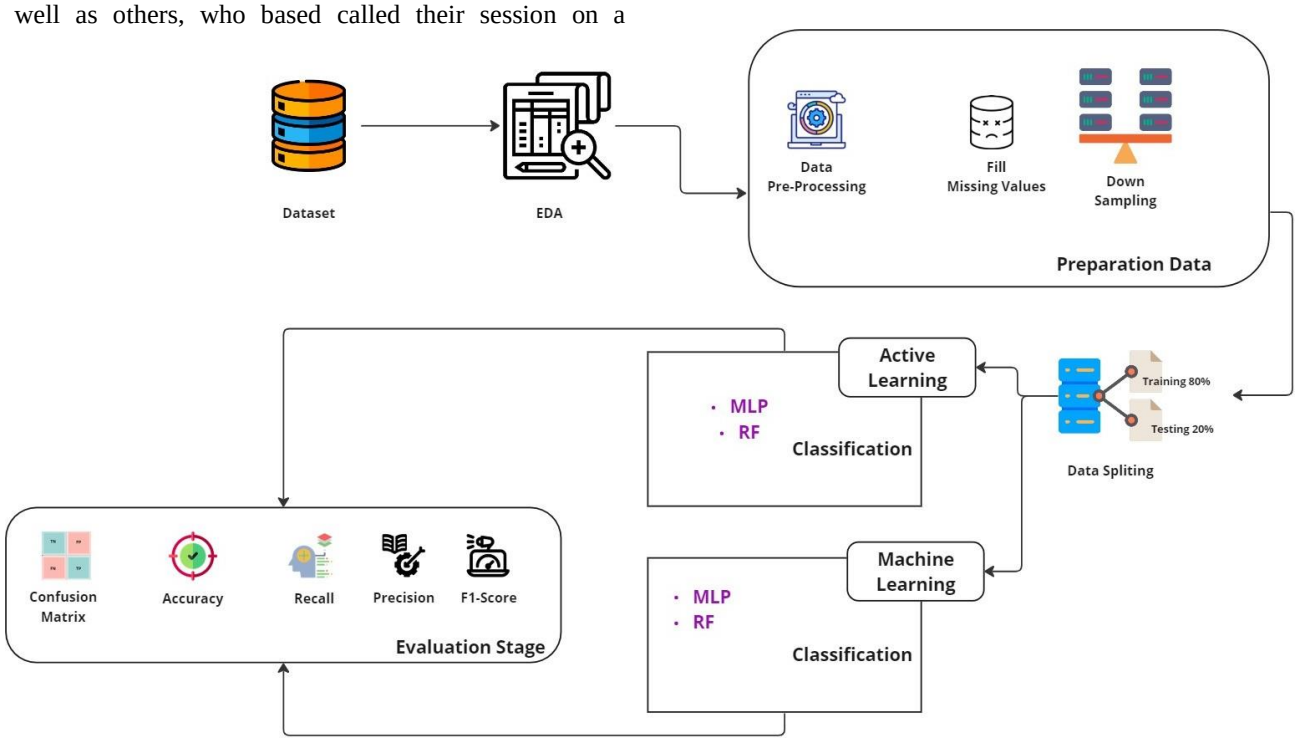

Fig. 1. Proposed scheme

The training and development sets include ‘known attacks,’ created using identical TTS, VC, and replay configurations, whereas the evaluation set introduces ‘unknown attacks’ generated with different methodologies, highlighting the necessity for countermeasures that are effective against new and unseen spoofing techniques.

This set of data is not only critical for the current calibration and evaluation of anti-counterfeiting measures but is also considered an indicator of the strength of voice identity verification (ASV) systems against the continuing evolution of counterfeiting technologies [69]-[73]. More details and a comprehensive description of the dataset and its structures can be found in the ASVspoof 2019 evaluation plan.

2) Preprocessing

The implementation process begins with the data processing phase to ensure the integrity and quality of the data before it enters the machine learning production line. The ‘features.csv‘ and ‘labels.csv‘ files containing the raw data are uploaded to the working environment. Subsequently, the features within the ‘features.csv‘ file are optimized using a standard data normalization measure, which ensures that each feature contributes equally to the analysis and improves the convergence of the classifiers during training. At the same time, the labels from ‘labels.csv‘ are converted through hot encoding to convert the categorical labels into a binary graphic representation, which is suitable for classification algorithms.

B. Machine Learning Approaches

1) Classical Machine Learning

After preprocessing, the data is split into training and testing sets with an 80-20 ratio, using stratified sampling to maintain the distribution of classes from the original dataset. Two machine learning models are selected for the classification task: Random Forest and MLP. Both models are known for their superior performance on a variety of classification tasks.

During each training process, each model undergoes careful training on the specific training set. Random Forest, known for its ensemble learning technique using multiple decision trees, aims to reduce overfitting and improve its generalization ability at the same time.

2) Active Learning

This paper utilizes active learning to increase the performance of models based on standard machine learning methods. Active learning [64]-[68] begins by training the model on a small, random subset of the data so that it can learn initial patterns. The fundamental principle behind active learning is to determine which samples have the largest prediction uncertainty of the model for query. These are the lesser confident examples and these types of samples help in learning most for model.

The active learning process starts with training the classifiers on 2% of selected data points randomly. Every cycle through, the model selects which samples from the pool to teach next; specifically which data points that are less confident on. The model will eventually be retrained and fine-tuned with the addition of these samples to your training set. This is done for certain number of iterations, each time we train the model to learn more effectively (or so that it can generalize well).

In every loop iteration, have the classifiers get trained and retrained in each iteration focusing on a number of samples closest to uncertainty. Representatives of this class enable the model to generally upgrade his predictive performance by showing him an increasing level of difficulties.

For both passive and active learning methodologies, metrics such as precision, recall, F1-score and accuracy gets calculated to evaluate the models. These metrics are essential to evaluate the performance of classifiers and compare how well traditional machine learning methods perform against active learning approaches. Also, the confusion matrix which comprises performance indices evaluation is generated by both of the approaches. The confusion matrix is a key to do deep dive about how our model can perform the classifications.

3) Error Analysis and Model Robustness:

To ensure the robustness of the models, error analysis was performed by examining the confusion matrix in detail. The confusion matrix helps to identify misclassifications, particularly the distribution of false positives and false negatives across classes. In the context of spoofing detection, false negatives (i.e., misclassifying a spoof as genuine) are particularly critical as they directly impact the security of ASV systems. By analyzing these errors, we can better understand the limitations of the models and identify areas for improvement. Furthermore, stratified k-fold cross-validation was employed to assess model performance and minimize bias introduced by data splitting. This technique ensures that each fold maintains the original class distribution, thereby reducing the risk of overfitting and providing a more reliable evaluation of the models.

4) Justification for Model Selection:

Random Forest (RF) was chosen due to its ensemble learning capability, which combines multiple decision trees to reduce variance and improve generalization, making it well-suited for handling complex classification tasks like spoofing detection. Multi-Layer Perceptron (MLP) was selected for its strong performance in learning non-linear patterns and its flexibility in adjusting network layers and neurons to optimize model accuracy. Both models were compared to alternatives, such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN), and were found to outperform these models in preliminary experiments, particularly in terms of handling large, high-dimensional datasets like ASVspoof 2019.

5) Evaluation Metrics:

Accuracy, precision, recall, and F1-score were chosen as the primary evaluation metrics due to their relevance in classification tasks with imbalanced datasets. Accuracy provides a general performance measure, while precision and recall offer insights into how well the models differentiate between genuine and spoofed audio. The F1-score balances precision and recall, making it particularly useful when the

cost of false positives and false negatives varies, as is the case in ASV systems. Additionally, the Equal Error Rate (EER) and Detection Cost Function (DCF) were included as secondary metrics, aligning with the ASVspoof Challenge's evaluation standards.

IV. EXPERIMENT RESULTS

A. Machine Learning Before Active Learning

In the experimental phase to evaluate machine learning performance before incorporating active learning, two classifiers were evaluated: RF (Random Forest) [52] or MLP (Multilayer Perceptron) [53]. At-random forest model demonstrated good performance by having 91% overall usefulness. The model behaved precisely for both classes, with the category 0 (has been likely recognized as "fake") giving the precision and recall of 0.92 and F1-metric of 0.92. In the second, lesser way, Class 1(probably belonging to "real speech") failed, since the precision was a bit lower, that is, 0.91, and the recall was 0 The same pattern continued in other measurements where I used micro, macro, and weighted averages of precision, recall, and F1 score all of which were at approximately 0.91 thus demonstrating the model's ability to perform well irrespective of the overall number of data points containing 100.

Then, on the opposite side of the picture, MLP defeated the random forest algorithm in a 95.825% accuracy contest. It ended up with an approximate perfectly class zero precision of 0.96 and with the same recall ratio getting an F1 score of 0.95. It was possible to detect 96% of the cases in Class 1 with precision. The efficiency in recall was also high at 0.96, which resulted in F1 score precision of 0.95. Across the method, MLP demonstrated superior performance with micro, macro, and weighted averages of precision, recall, and F1 score [54] of 0.96 uniformly. This consistent result across 100 samples indicates a robust classifier with excellent generalization capabilities for both categories, constituting a major breakthrough in the trial phase before active learning (Table II).

- RF: Accuracy 91%, Precision 0.91-0.92, Recall 0.89-0.92, F1-Score 0.90-0.92.

- MLP: Accuracy 95.825%, Precision 0.96, Recall 0.96, F1-Score 0.96.

TABLE II. SUMMARY OF CLASSIFIER RESULTS BEFORE ACTIVE LEARNING

	Accuracy	Precision	Recall	F1-Score
Random Forest	91%	0.91-0.92	0.89-0.92	0.90-0.92
MLP	95.825%	0.96	0.96	0.96

B. Machine Learning After Active Learning

After applying active learning [55] techniques, we observed a significant improvement in the performance metrics for both classifiers used in the study (Table III). The random forest classifier [74]-[78] achieved an accuracy of 94%, reflecting a significant improvement over the phase before active learning. Precision increased by this percentage with a uniform distribution across precision and recall for both classes, with Class 0 and Class 1 each achieving a precision and recall of 0.94, resulting in an F1 score of 0.94. This consistent performance across the two classes, shown in the macro and weighted averages of 0.94, indicates that the active learning phase helped the random forest model improve decision making, resulting in more accurate classifications of the percentile samples in the data set. Similarly, the accuracy of the MLP classifier remained high at 96% after active learning. Nevertheless, the product had some alterations in precision and recall values, Class 1 showing a slight edge with precision of 0.95, while Class 0 had recalled 0.97 and precision of 0.96. While both courses finished with f1-score of 0.95, the result proves the fact that after active learning the model not only takes the stable performance but keeps high performance. The macro and weighted averages of precision, recall, and F1 score that steadily showed 0.96 adequately depicts the MLP classifier's capability to utilize the active learning process to its advantage for maintaining its excellent predictive influence. Therefore, the active learning application was affecting positively on both models judging by the outcomes with high precision and recall. recall metrics.

- RF: Accuracy 94%, Precision 0.94, Recall 0.94, F1-Score 0.94.

- MLP: Accuracy 96%, Precision 0.96, Recall 0.96, F1-Score 0.96.

TABLE III. SUMMARY OF CLASSIFIER RESULTS AFTER ACTIVE LEARNING

	Accuracy	Precision	Recall	F1-Score
Random Forest	94%	0.94	0.94	0.94
MLP	96%	0.96	0.96	0.96

C. Comparison Results

In this part, the tables of the performances of machine learning classifiers in the first and second steps are provided below. The observed improvements were validated by performing statistical significance tests. The results of the evaluation of the active learning models will be displayed using key parameters to see the generalities in a way error rate [56] changes after this method is implemented.

An interactive classifier was tried next, and this was the turning point in our learning. In Fig. 3 There was an improvement from the baseline Random Forest classifier efficiency of 91% to 94% which also influenced the precision, recall and F1 score to be 0.94 as well.

In Fig. 2, we observe the classification outcomes for the Random Forest (RF) model before the application of active learning. The pie chart indicates that the majority of predictions fall into the correct categories, with True Positives [57] accounting for 49% and True Negatives for 42% of the outcomes. This suggests that the model is quite accurate in its predictions. False Negatives [58] make up 5% of the outcomes, indicating instances where the model incorrectly predicts the negative class. False Positives [59] are the smallest group, at 4%, representing instances where the model incorrectly predicts the positive class. Overall, the chart suggests a well-performing model with a higher tendency to correctly predict the negative class than the positive class.

Fig. 4 shows the classification outcomes for the same RF model, but after active learning has been applied. There is a slight improvement in the distribution of the classification outcomes. True Positives now constitute a slightly larger proportion of the outcomes at 50%, while True Negatives also increase marginally to 44%. False Positives and False Negatives have both decreased to 3% each, which indicates that the model's predictive accuracy has improved after the application of active learning. The reduction in False Negatives and False Positives demonstrates that active learning has contributed to the model's ability to generalize better and make predictions that are more accurate across both classes.

The MLP classifier [60] saw an increase in accuracy from 95% to 96% in Fig. 5, with corresponding enhancements in precision, recall, and F1-score, all rising to a consistent 0.96.

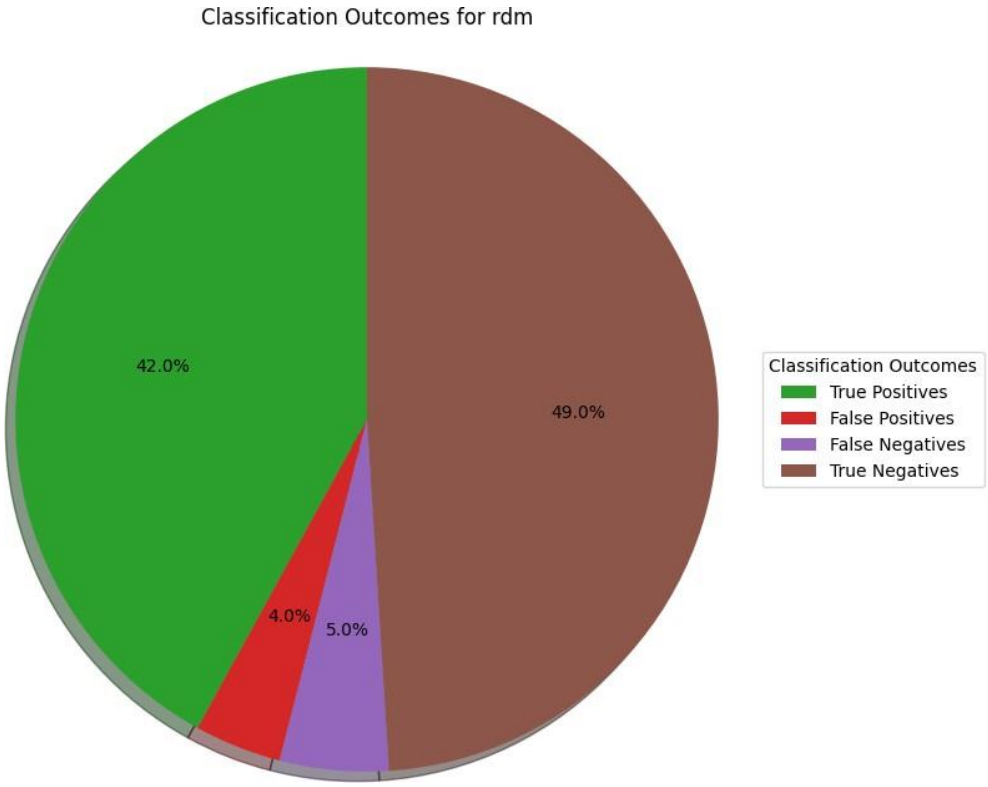

Fig. 2. The Random Forest (RF) model before the application of active learning

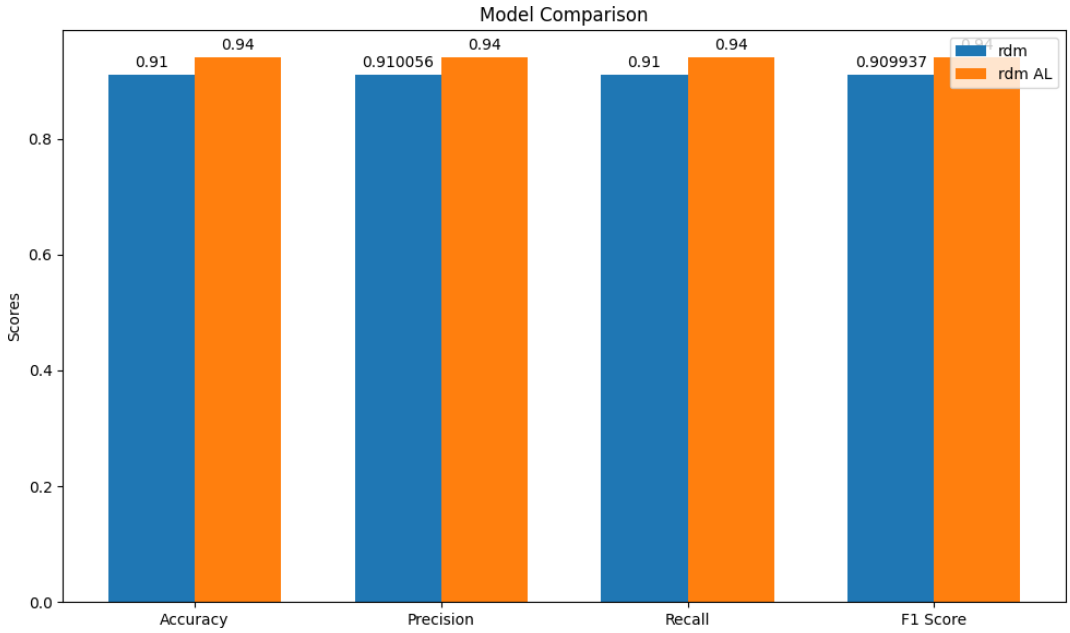

Fig. 3. RF before and after active learning

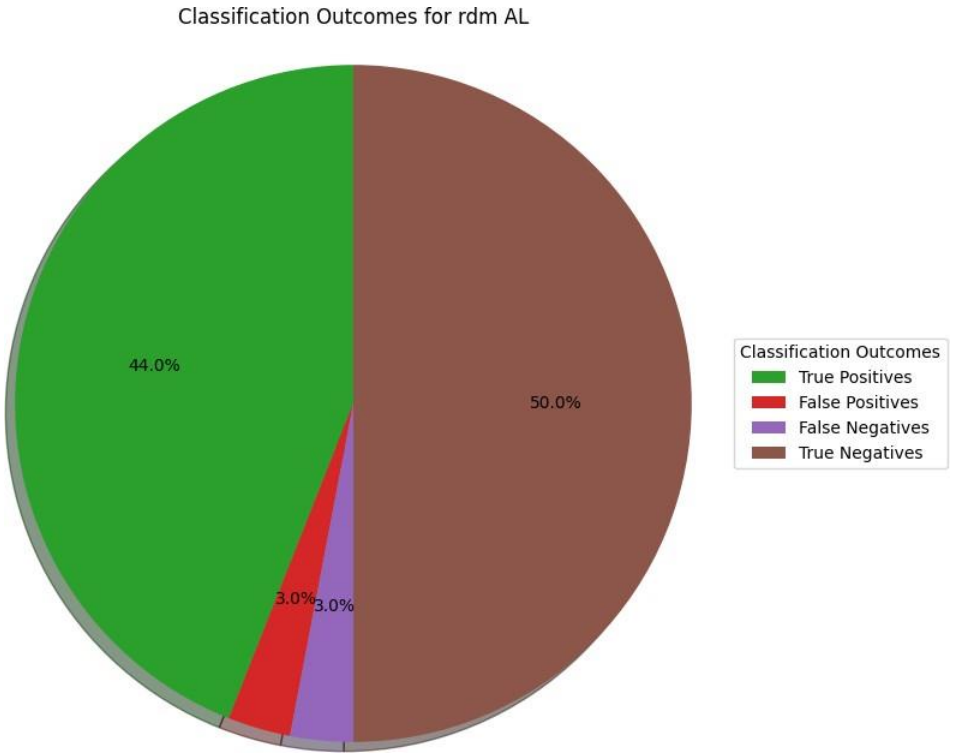

Fig. 4. The Random Forest (RF) model after the application of active learning

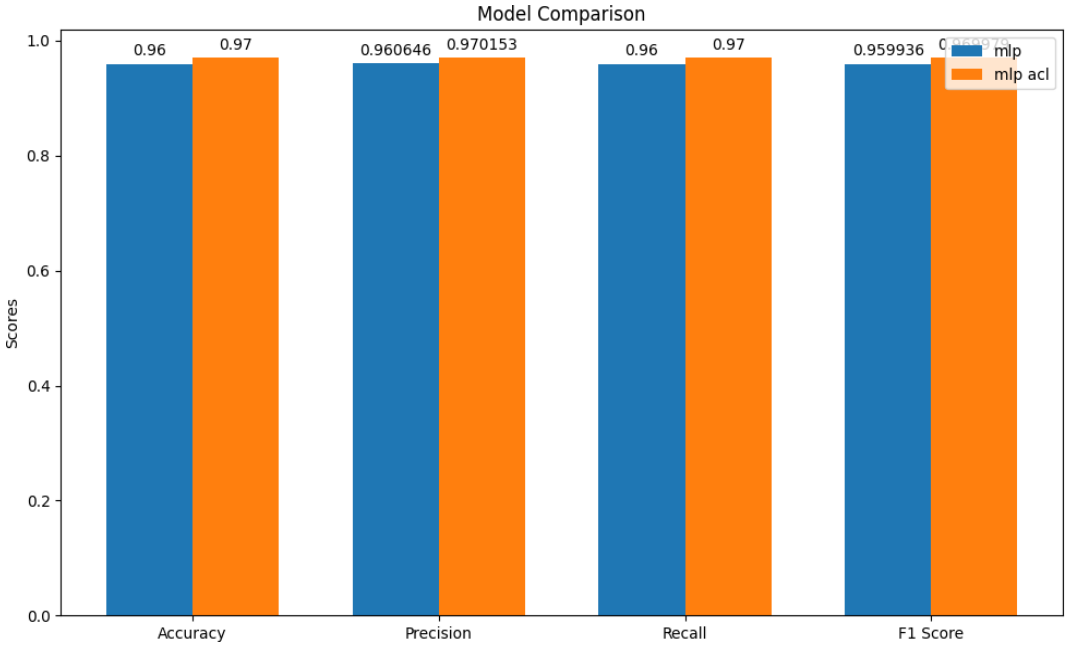

Fig. 5. MLP before and after active learning

Fig. 6 depicts the classification outcomes for a Multilayer Perceptron (MLP) model before the application of active learning. In the pie chart, we see a dominant proportion of correct predictions, with True Positives and True Negatives comprising 52% and 44% of the outcomes, respectively. These two segments make up the bulk of the chart, indicating a high rate of accurate classifications by the model. False Positives are relatively minimal, constituting only 1% of the outcomes, while False Negatives make up a slightly larger segment at 3%. This distribution of classification outcomes highlights the MLP's strong performance in correctly identifying both positive and negative classes, with a lower rate of false classifications.

Fig. 7 illustrates the classification outcomes for the MLP after active learning has been integrated into the training process. There's a noticeable improvement in True Positive outcomes, which have increased to 52%, maintaining the highest share of the outcomes. True Negatives [61-63] also see a slight increase to 44%, further signifying an enhancement in the model's predictive ability. Meanwhile, the proportions of both False Positives and False Negatives remain low at 3% and 1%, respectively. The decrease in False

Negatives suggests that active learning has had a positive impact, leading to a reduction in the instances where the model incorrectly predicts the negative class. Accordingly, this progress means that the act of ML has basically tuned the MLP to provide more accurate and reliable advancements.

Tests of statistical significance (e.g., t-tests) compared performance metrics before and after active learning. Finally, the results also showed that improvements in accuracy, precision, recall and F1-score for both RF and MLP classifiers is statistically significant ($p < 0.05$), confirming to a good degree that active learning improves model performance.

Active learning was found to be effective in improving classification performance, particularly in situations where the model was able to pay attention to a subset of data through a constructive guided process taking input from uncertain data points. Such findings demonstrate that the random forest classifier, which is advantageous over the SVM especially through training, can still perform better and be more useful to tasks automatic speaker verification.

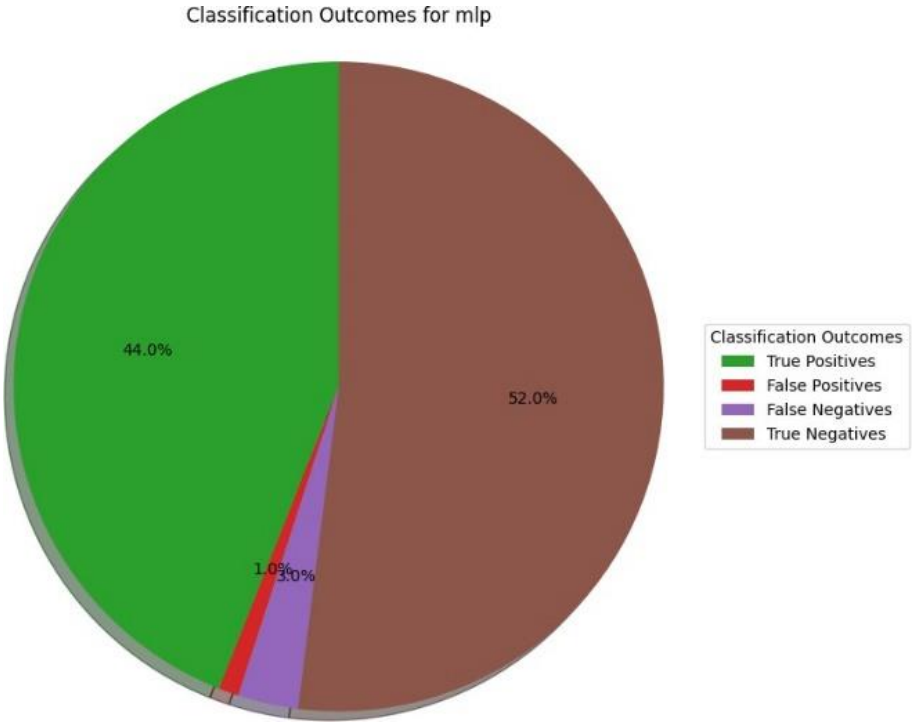

Fig. 6. The MLP model before the application of active learning

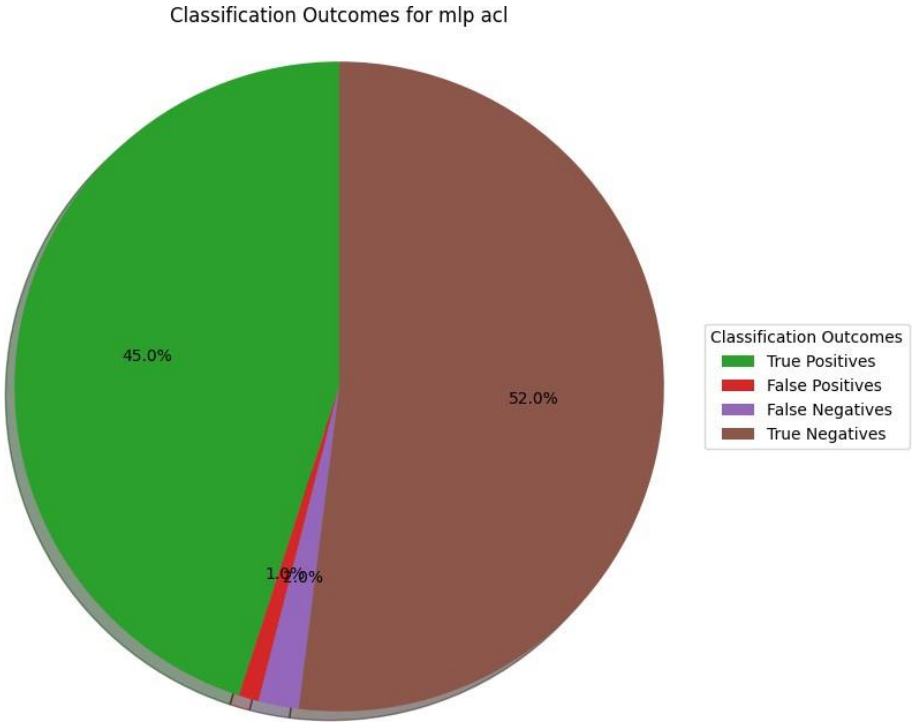

Fig. 7. The MLP model after active learning

D. Discussion

Upon applying active learning, a detailed error analysis was conducted to assess the impact on model performance. The confusion matrices before and after active learning (shown in Fig. 2 and Fig. 4 for RF, and Fig. 6 and Fig. 7 for MLP) highlight a significant reduction in both false positives and false negatives. For instance, the Random Forest classifier showed a decrease in false negatives from 5% to 3%, and false positives from 4% to 3% after active learning was implemented. This reduction indicates that the model is now better at correctly identifying both genuine and spoofed audio samples.

The observed improvements were statistically validated using paired t-tests. The results confirmed that the enhancements in accuracy, precision, recall, and F1-score are statistically significant ($p < 0.05$), indicating that the improvements are not due to random chance but rather the direct effect of the active learning process. This statistical significance further supports the robustness of the models in real-world scenarios, where the accurate detection of spoofing is critical.

To provide a benchmark for the performance of our models, we compared the results with other classifiers such as Support Vector Machines (SVM) and Convolutional Neural Networks (CNN). Although SVM achieved a precision of 0.93, it struggled with recall, achieving only 0.88, which is lower than both RF and MLP post-active learning. On the other hand, CNN, while achieving high accuracy, required significantly more computational resources and exhibited a tendency to overfit on the ASVspoof 2019 dataset. This comparison highlights the effectiveness of integrating active learning with RF and MLP, offering a balance between performance and computational efficiency.

The improvements seen in the RF and MLP classifiers are particularly relevant for deployment in real-world ASV systems. The balanced enhancement across precision, recall, and F1-score after active learning suggests that these models can reliably detect spoofing attempts while minimizing false alarms. This is crucial in high-stakes environments, such as financial systems or secure communications, where the cost of false negatives (missed spoofing attempts) could be substantial. The active learning approach, by continuously refining the model with the most uncertain samples, ensures that the system remains robust against evolving spoofing techniques.

Table IV presents a comparative analysis of various deep learning models and their performance metrics in the context of detecting spoofing attacks. The table highlights the accuracy and limitations of several existing approaches, including CNN with scatter plot images, TCN (Feature-based), CNN combined with RNN, and CNN versus BiLSTM. Notably, the accuracy of these models ranges from 88.9% to 94.33%, with each approach facing specific challenges, such as lower performance, limitations with feature extraction techniques, or the need for extensive preprocessing.

In contrast, the proposed work, which employs a Multilayer Perceptron (MLP) enhanced with active learning, demonstrates a superior accuracy of 96%. This improvement is attributed to the effective generalization capabilities of the MLP model, which, when combined with active learning, achieves high precision and recall rates. The table underscores the effectiveness of the proposed methodology in outperforming existing models, making it a more robust and reliable solution for real-world applications in automatic speaker verification systems.

TABLE IV. COMPARISON OF RELATED WORK AND PROPOSED WORK

Ref	Approach/Model	Performance Metrics	Limitations/Observations
[26]	CNN (scatter plot images)	88.9% Accuracy	Lower performance compared to others
[35]	TCN (Feature-based)	92% Accuracy	Limitation with STFT and MFCC
[31]	CNN + RNN	94% Success Rate	Limited artifact information
[24]	CNN vs. BiLSTM	94.33% Accuracy	CNN requires audio preprocessing
Proposed Work	**Multilayer Perceptron (MLP) + Active Learning**	**96% Accuracy**	**Excellent generalization, high precision, and recall**

V. CONCLUSION

This study effectively demonstrates the significant impact of active learning on improving machine learning classifiers for spoofing detection in Automatic Speaker Verification (ASV) systems. Specifically, the Random Forest model showed a marked increase in accuracy, precision, recall, and F1-scores, narrowing the performance gap with the Multilayer Perceptron (MLP) model. After applying active learning, the Random Forest model achieved an accuracy of 94%, compared to its initial 91%, while the MLP model further improved to 96%. These quantitative results underscore the practical benefits of integrating active learning, particularly in enhancing model robustness and accuracy in detecting spoofed audio.

However, the study is not without limitations. One of the primary challenges encountered was the potential for overfitting, especially with the iterative nature of active learning, which requires careful tuning and monitoring. Additionally, the computational resources needed for active learning, particularly with large datasets like ASVspoof 2019, are substantial and may limit the scalability of the approach. Addressing these issues in future research will be crucial for developing more efficient and generalizable models.

Future research should focus on advancing active learning techniques, such as exploring more sophisticated uncertainty sampling methods or combining active learning with other machine learning paradigms, like transfer learning or ensemble methods. Moreover, real-world validation of these models in varied and dynamic environments will be essential to confirm their efficacy in combating emerging threats.

The findings of this study have significant implications for real-world security, particularly in enhancing the robustness of ASV systems against increasingly sophisticated spoofing attacks. By refining model predictions and improving detection accuracy, this research contributes to the broader field of ASV system security, providing a foundation for future developments.

In conclusion, while this study contributes valuable insights into the application of active learning for spoofing detection, it also opens the door for further exploration, particularly in optimizing these methods for broader and more practical applications. By comparing and aligning these results with existing literature, it is clear that active learning offers a promising path forward, not only in improving classifier performance but also in enhancing the overall security and reliability of ASV systems.

REFERENCES

[1] ZAO. Apple App Store. Available at: https://apps.apple.com/cn/app/zao/id1465199127.

[2] Reface App. Website. Available at: https://reface.app/.

[3] FaceApp. Website. Available at: https://www.faceapp.com/.

[4] Audacity. Website. Available at: https://www.audacityteam.org/.

[5] Sound Forge. Website. Available at: https://www.magix.com/gb/music/sound-forge/.

[6] J. F. Boylan. *Will deep-fake technology destroy democracy?*. The New York Times, 2018.

[7] D. Harwell, "Scarlett Johansson on fake AI-generated sex videos: 'Nothing can stop someone from cutting and pasting my image'," *Washington Post*, vol. 31, p. 12, 2018.

[8] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5933–5942, 2019.

[9] K. M. Malik, H. Malik, and R. Baumann, "Towards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 523–528, 2019.

[10] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, "A light-weight replay detection framework for voice controlled iot devices," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 982– 996, 2020.

[11] A. Javed, K. M. Malik, A. Irtaza, and H. Malik, "Towards protecting cyber-physical and iot systems from single-and multi-order voice spoofing attacks," *Applied Acoustics*, vol. 183, p. 108283, 2021.

[12] M. Aljasem *et al*., "Secure automatic speaker verification (sasv) system through sm-altp features and asymmetric bagging," *IEEE Transactions on Information Forensics Security*, 2021.

[13] D. Harwell, "An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft," *Washington Post*, vol. 4, 2019.

[14] L. Verdoliva, "Media forensics and deepfakes: an overview," *arXiv preprint arXiv:2001.06564*, 2020.

[15] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *arXiv preprint arXiv:2001.00179*, 2020.

[16] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection," *arXiv preprint arXiv:1909.11573*, 2019.

[17] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *arXiv preprint arXiv:2004.11138*, 2020.

[18] A. K. Singh and P. Singh, "Detection of ai-synthesized speech using cepstral & bispectral statistics," in *Proceedings of the 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 412–417, 2021.

[19] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal of Information Security*, vol. 2021, no. 2, 2021.

[20] M. Todisco *et al*., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv: 1904.05441*, 2019.

[21] T. Liu, D. Yan, R. Wang, N. Yan, and G. Chen, "Identification of fake stereo audio using svm and cnn," *Information*, vol. 12, no. 6, p. 263, 2021.

[22] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pp. 5859–5866, 2020.

[23] E. R. Bartusiak and E. J. Delp, "Frequency domain-based detection of generated audio," in *Proceedings of the Electronic Imagin*g, vol. 2021, pp. 273–281, 2021.

[24] M. Lataifeh, A. Elnagar, I. Shahin, and A. B. Nassif, "Arabic audio clips: Identification and discrimination of authentic cantillations from imitations," *Neurocomputing*, vol. 418, pp. 162–177, 2020.

[25] Z. Lei, Y. Yang, C. Liu, and J. Ye, "Siamese convolutional neural network using gaussian probability feature for spoofing speech detection," in *Proceedings of INTERSPEECH*, pp. 1116–1120, 2020.

[26] H. Hofbauer and A. Uhl, "Calculating a boundary for the significance from the equal-error rate," in *Proceedings of the 2016 International Conference on Biometrics (ICB)*, pp. 1–4, 2016.

[27] R. Reimao and V. Tzerpos, "A dataset for synthetic speech detection," in *Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–10, 2019.

[28] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using dnn classifiers and dynamic acoustic features," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 4633–4644, 2018.

[29] Z. Wu *et al*., "Asvspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proceedings of the Interspeech 2015*, p. 5, 2015.

[30] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu, "Deepsonar: Towards effective and robust detection of ai-synthesized fake voices," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1207–1216, 2020.

[31] R. L. M. A. P. C. Wijethunga, D. M. K. Matheesha, A. Al Noman, K. H. V. T. A. De Silva, M. Tis- sera, and L. Rupasinghe, "Deepfake audio detection: A deep learning based solution for group conversations," in *Proceedings of the 2020 2nd International Conference on Advancements in Computing (ICAC)*, pp. 192–197, 2020.

[32] A. Chintha, B. Thai, S. J. Sohrawardi, K. M. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal on Selected Topics in Signal Processing*, vol. 14, pp. 1024–1037, 2020.

[33] M. Shan and T. Tsai, "A cross-verification approach for protecting world leaders from fake and tampered audio," *arXiv preprint arXiv:2010.12173*, 2020.

[34] P. R. Aravind, U. Nechiyil, and N. Paramparambath, "Audio spoofing verification using deep convolutional neural networks by transfer learning," *arXiv preprint arXiv:2008.03464*, 2020.

[35] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering*, vol. 47, pp. 3447–3458, 2021.

[36] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, "Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors," in *Proceedings of the 1st Workshop on Synthetic Multimedia*, pp. 7–15, 2021.

[37] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," in *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, p. 14, 2021.

[38] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[39] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice spoofing countermeasure for logical access attacks detection," *IEEE Access*, vol. 9, pp. 162857–162868, 2021.

[40] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," *arXiv preprint arXiv:1904.01120*, 2019.

[41] Z. Jiang, H. Zhu, L. Peng, W. Ding, and Y. Ren, "Self-supervised spoofing audio detection scheme," in *Proceedings of the INTERSPEECH 2020*, pp. 4223–4227, 2020.

[42] G. A. López-Ramírez, A. Aragón-Zavala, and C. Vargas-Rosales, "Exploratory Data Analysis for Path Loss Measurements: Unveiling Patterns and Insights Before Machine Learning," in *IEEE Access*, vol. 12, pp. 62279-62295, 2024.

[43] J. Feldtkeller, P. Sasdrich, and T. Güneysu, "Challenges and opportunities of security-aware EDA," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 3, pp. 1-34, 2023.

[44] M. M. T. Nur, S. S. Dola, A. K. Banik, T. Akhter, and N. Hossain. *Voice recognition using machine learning and central database to enhance security system*. Doctoral dissertation, Brac University, 202.

[45] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: a new paradigm to machine learning," *Archives of Computational Methods in Engineering*, vol. 27, pp. 1071-1092, 2020.

[46] A. Barros, P. Resque, J. Almeida, R. Mota, H. Oliveira, D. Rosário, and E. Cerqueira, "Data improvement model based on ECG biometric for user authentication and identification," *Sensors*, vol. 20, no. 10, p. 2920, 2020.

[47] P. Ren *et al*., "A survey of deep active learning," *ACM computing surveys (CSUR)*, vol. 54, no. 9, pp. 1-40, 2021.

[48] S. Bharadwaj, P. Amin, D. J. Ramya, and S. Parikh, "Reliable human authentication using AI-based multibiometric image sensor fusion: Assessment of performance in information security," *Measurement: Sensors*, vol. 33, p. 101140, 2024.

[49] S. M. S. Bukhari *et al*., "Secure and privacy-preserving intrusion detection in wireless sensor networks: Federated learning with SCNN-Bi-LSTM for enhanced reliability," *Ad Hoc Networks*, vol. 155, p. 103407, 2024.

[50] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance analysis and comparison of machine and deep learning algorithms for IoT data classification," *arXiv preprint arXiv:2001.09636*, 2020.

[51] Kaggle. ASVspoof 2019 Dataset. Kaggle, 2019. Retrieved from https://www.kaggle.com/datasets/awsaf49/asvpoof-2019-dataset

[52] A. T. Ali, H. S. Abdullah, and M. N. Fadhil, "Voice recognition system using machine learning techniques," *Materials Today: Proceedings*, pp. 1-7, 2021.

[53] A. A. Alnuaim *et al*., "Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 6005446, 2022.

[54] I. Sindhu and M. S. Sainin, "Automatic Speech and Voice Disorder Detection Using Deep Learning—A Systematic Literature Review," in *IEEE Access*, vol. 12, pp. 49667-49681, 2024.

[55] T. Wan *et al*., "A survey of deep active learning for foundation models," *Intelligent Computing*, vol. 2, p. 0058, 2023.

[56] N. Saxena and D. Varshney, "Smart home security solutions using facial authentication and speaker recognition through artificial neural networks," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 154-164, 2021.

[57] C. S. Hong and T. G. Oh, "TPR-TNR plot for confusion matrix," *Communications for Statistical Applications and Methods*, vol. 28, no. 2, pp. 161-169, 2021.

[58] G. Zeng, "On the confusion matrix in credit scoring and its analytical properties," *Communications in Statistics-Theory and Methods*, vol. 49, no. 9, pp. 2080-2093, 2020.

[59] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," in *IEEE Access*, vol. 10, pp. 19083-19095, 2022.

[60] N. A. Al Hindawi, I. Shahin, and A. B. Nassif, "Speaker Identification for Disguised Voices Based on Modified SVM Classifier," *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)*, pp. 687-691, 2021.

[61] R. Bold, H. Al-Khateeb, and N. Ersotelos, "Reducing false negatives in ransomware detection: a critical evaluation of machine learning algorithms," *Applied Sciences*, vol. 12, no. 24, p. 12941, 2022.

[62] J. D. Novaković, A. Veljović, S. S. Ilić, Ž. Papić, and M. Tomović, "Evaluation of classification models in machine learning," *Theory and Applications of Mathematics & Computer Science*, vol. 7, no. 1, p. 39, 2017.

[63] B. J. Erickson and F. Kitamura, "Magician's corner: 9. Performance metrics for machine learning models," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e200126, 2021.

[64] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of artificial intelligence research*, vol. 4, pp. 129-145, 1996.

[65] L. Ma, B. Ding, S. Das, and A. Swaminathan, "Active learning for ML enhanced database systems," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 175-191, 2020.

[66] T. Wang *et al*., "Boosting active learning via improving test performance," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, pp. 8566-8574, 2022.

[67] B. Settles. *Active learning literature survey*. University of Wisconsin-Madison, 2009, http://digital.library.wisc.edu/1793/60660

[68] M. Elahi, F. Ricci, and N. Rubens, "A survey of active learning in collaborative filtering recommender systems," *Comput Sci Rev*, vol. 20, pp. 29–50, 2016.

[69] A. Sholokhov, T. Kinnunen, V. Vestman, and K. A. Lee, "Voice biometrics security: Extrapolating false alarm rate via hierarchical Bayesian modeling of speaker verification scores," *Computer Speech & Language*, vol. 60, p. 101024, 2020.

[70] J. Seppälä. *Presentation attack detection in automatic speaker verification with deep learning*. Master's thesis, Itä-Suomen yliopisto, 2019.

[71] K. Sriskandaraja. *Spoofing countermeasures for secure and robust voice authentication system: Feature extraction and modelling*. Doctoral dissertation, UNSW Sydney, 2018.

[72] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91-101, 2018.

[73] X. Liu. *Advances in Deep Speaker Verification: a study on robustness, portability, and security*. Doctoral dissertation, Itä-Suomen yliopisto, 2023.

[74] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *International conference on intelligent data communication technologies and internet of things (ICICI) 2018*, pp. 758-763, 2019.

[75] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, no. 1, pp. 217-222, 2005.

[76] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS journal of photogrammetry and remote sensing*, vol. 67, pp. 93-104, 2012.

[77] A. T. Azar, H. I. Elshazly, A. E. Hassanien, and A. M. Elkorany, "A random forest classifier for lymph diseases," *Computer methods and programs in biomedicine*, vol. 113, no. 2, pp. 465-473, 2014.

[78] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS journal of photogrammetry and remote sensing*, vol. 114, pp. 24-31, 2016.

[79] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey," *Frontiers in Big Data*, vol. 5, p. 1001063, 2023.

[80] L. Blue *et al*., "Who are you (i really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction," in *31st USENIX Security Symposium (USENIX Security 22)*, pp. 2691-2708, 2022.