ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

Estuscience – Se, 2025, 26 (2) pp. 132-149, DOI: 10.18038/estubtda.1624909

RESEARCH ARTICLE

AUDIO COPY-MOVE FORGERY DETECTION WITH MACHINE LEARNING METHODS

Merve ARSLAN [1, *], Şerif Ali SADIK [2]

[1] Department of Software Engineering, Faculty of Engineering, Dumlupınar University, Kütahya, Türkiye
merve.arslan@dpu.edu.tr - 0000-0002-2867-6198

[2] Department of Software Engineering, Faculty of Engineering, Dumlupınar University, Kütahya, Türkiye
serifali.sadik@dpu.edu.tr - 0000-0003-2883-1431

Abstract

Converting original sounds into fake sounds using various methods and using these sounds for fraud or misinformation purposes poses serious risks and threats. In this study, a classification system using machine learning methods is created and performance analysis is performed in order to detect sounds created with copy-move forgery, which is one of the types of sound forgery. Sound files are treated as raw data. Then, Mel-spectrograms are obtained to visually represent the spectral features of the sound over time. Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN) and XGBoost algorithms are used in the classification phase. As a result of the performance analysis of the created models, the highest success is achieved with the XGBoost algorithm. The performance of the XGBoost algorithm is further improved by performing hyperparameter optimization with the Random Search method. The results of the models are analyzed using various metrics. According to the study results, it is seen that it gives competitive results with the XGBoost algorithm.

Keywords

Audio,
Copy-Move Forgery,
Machine Learning,
XGBoost

Time Scale of Article

Received : 22 January 2025
Accepted : 13 June 2025
Online date :25 June 2025

1. INTRODUCTION

With the rapid development of the modern age, digital voices are used extensively in all areas of our lives. Voice recordings of individuals are strong evidence especially in legal and forensic cases. In such a case, it is of great importance to verify the authenticity of the voice recording. As a result of the rapid increase in technological innovations, it has become much easier to create fake voices. Unfortunately, even non-professionals can easily manipulate sounds and produce fake audio.

Given the growing ease with which audio can be manipulated, detecting forged or tampered audio has become a critical challenge. Traditional methods of authentication often fall short when dealing with subtle manipulations, necessitating the adoption of more advanced solutions. Machine learning techniques have emerged as powerful tools in this regard, capable of identifying patterns and anomalies in audio data that are imperceptible to human hearing. By leveraging these technologies, it is possible to build systems that automatically detect tampering with high accuracy, providing a robust solution in the face of increasingly sophisticated audio forgeries.

Machine learning methods have been used effectively in studies on the classification or detection of sound-based events. In a study from 2024, the classification process for hate speech detection from an audio file was performed using Support Vector Machines, Random Forest, eXtreme Gradient Boosting and multilayer perceptron (MLP). First, various features were extracted from the audio files and then

*Corresponding Author:merve.arslan@dpu.edu.tr

machine learning methods were used to classify hate speech. The voices in the dataset consist of English and Kiswahili languages. While Random Forest gave the most successful result with 95.8% in the classification process using the voices in English language, XGB gave the most successful result with 91.8% in the classification process using the voices in Kiswahili language [1].

In another study carried out in 2024, sound classification was performed using "UrbanSound8K" and "Sound Event Audio Classification dataset" [2-3]. Mel-frequency cepstral coefficients (MFCC) feature extraction was applied to the "UrbanSound8K" dataset and STFT feature extraction process was applied to the "Sound Event Audio Classification dataset". Both datasets were classified using Artificial Neural Network model, Logistic Regression, SVM, KNN, Naive Bayes, Decision Tree and RF algorithms. The datasets for which classification is performed have 10 and 8 different classes respectively. As a result of the study, the artificial neural network model gave the most successful result on both datasets with 91.41% and 91.27% respectively [4].

In 2021, a dataset consisting of real voices and computer-generated voices is used in a study to distinguish between fake and original voices [5]. In this study, two methods, feature-based classification and image-based classification, were created. Under the feature-based classification approach, 20 MFCCs were considered in addition to extracting various audio features in the feature extraction step. These features were then provided as input to machine learning algorithms. Within the scope of the study, five machine learning algorithms including SVM [6], Light Gradient-Boosting Machine (LGBM) [7], XGBoost [8], KNN [9], RF [10] were employed. GridSearchCV was used for parameter optimization. When the test results were analyzed, SVM algorithm gave the most successful result with 67% [11].

The techniques that researchers have developed as a solution to the challenges of digital voice authentication are divided into passive and active methods. The passive method is the detection of forgery through the signal itself and its characteristics. Active is the method of detecting this situation as a result of embedding certain information in the audio with various techniques. For example, active methods such as watermarking involve embedding additional information in the signal. In many cases, watermarks may not be able to detect areas that need to be deleted, and in some cases, counterfeiting can be done without serious damage to the watermark. For such a situation, passive forgery detection would be a more appropriate solution [12].

Copy-move forgery, which is one of the passive methods, is basically based on copying certain parts of a sound recording and moving them to another part within the same recording. By creating a fake audio recording in this way, the meaning of the phrase is completely changed [13]. Due to the methodology used, it is difficult to recognize the production of forged audio as a result of the changes made within the same audio recording. Therefore, an efficient and reliable method for detecting the authenticity of audio recordings is an important need in this field.

Copy-move forgery is a security threat on digital media. Unfortunately, audio files are also subject to such manipulations. One of the detection techniques for copy-move forgery is based on pitch similarity. Pitch is associated with the frequency of a sound and is a feature that allows the ranking of sounds based on their frequency [14]. In one of the studies, after the pitch sequences of the sound were extracted, a detection study was carried out as a result of calculations and comparison with threshold values [15]. In a study conducted to detect such forgeries, Discrete cosine transform (DCT) of audio signals and voice activity detection (VAD) algorithm are used together [16].

In a recent copy-move detection study, a fake audio file was created using the TIMIT dataset [17] [18]. Then, MFCC, delta-MFCC, delta-delta-MFCC and LPC data were obtained by feature extraction. Afterwards, the original and fake data detection process was performed with an artificial neural network. Tests were performed using various epoch numbers and batch sizes. From the results obtained, 76.48% test accuracy was achieved using 1500 epochs and batch size of 8 [19].

Considering all these, detecting copy-move forgery is of great importance for ensuring digital data security. In particular, the use of digital content as evidence in legal, commercial and social fields makes the detection of such forgeries more critical. For this reason, the development of reliable methods that can detect copy-move forgery is essential to preserve the authenticity and integrity of the audio recording. In this study, we use a copy and paste forgery dataset created using audio recordings of 100 people in different environments. The dataset was collected in 2024 using up-to-date technologies and designed to reflect real-life scenarios, allowing for effective analysis against modern copy-move forgery techniques. The 200 texts included in the dataset were either purposefully produced or carefully selected to be suitable for copy-move forgery, ensuring that no semantic or logical inconsistencies occur after manipulation. The texts incorporate various communicative functions such as requests, announcements, and informational messages, with linguistic and expressive features that vary depending on the assumed speaker and the topic. Furthermore, a wide range of sentence types affirmative, negative, interrogative, exclamatory, imperative are exemplified using simple, compound, sequential, and complex sentence structures. This diversity enables realistic testing of forgery detection methods and allows the evaluation of machine learning models in the context of linguistic variability. In addition to the contribution of a newly collected and linguistically diverse Turkish audio dataset for copy-move forgery detection, this study presents a systematic evaluation of several classical machine learning algorithms. While many recent approaches focus computationally expensive deep learning methods, our work offers a reproducible and computationally efficient classical ML baseline, which can be especially useful in settings where access to advanced computational resources is limited. With an optimized XGBoost model achieving reliable accuracy, the results demonstrate that well-tuned classical algorithms remain a viable and interpretable alternative for audio forgery detection tasks in resource-constrained environments. Within the scope of the study, six machine learning methods are used for forgery detection. The performance of machine learning methods on forgery detection is interpreted with the outputs obtained. The rest of the paper is as follows: Copy-move forgery is discussed in Section 2. The machine learning algorithms used in the study are given under Section 3. Examination of the dataset, feature selection and Mel-spectrogram for fake audio detection experimental outputs are analyzed in Section 4. Finally, Section 5 discusses the study with a conclusion.

2. COPY-MOVE FORGERY

Emerging artificial intelligence techniques have made it possible to imitate a person's voice, manipulate their speech, change the content of the speech or, in addition to all these, produce completely fake voices. The privacy of individuals is also threatened by voice forgery. In addition, individuals' speech recordings are used as evidence in courts of law.

Voice forgery is basically the alteration of original voice recordings using various techniques. In some cases, voice forgery is also performed by creating completely fake voices without manipulating the voice. Voice forgery can be performed using digital audio processing methods. Figure 1 shows the grouping scheme of digital audio forgery types.

As shown in Figure 1, audio forgery methods are primarily categorized as Active and Passive. In active audio forgery, certain information is embedded into the original audio using Digital Audio Watermarking, Digital Audio Signatures and Hash Values techniques shown in Figure 1. In this approach, in order to preserve the authenticity of the audio data, it is analyzed to determine whether there is any forgery in the audio recording by analyzing whether the pieces of information actively embedded in the audio are preserved. Passive analysis, on the other hand, focuses on the characteristics of the audio signals and does not require any extra information to be embedded. Audio copy-move, Audio Splicing and Audio Compression are passive audio analysis techniques [19]. Such methods are of great importance for analyzing audio manipulation or detecting audio forgery.

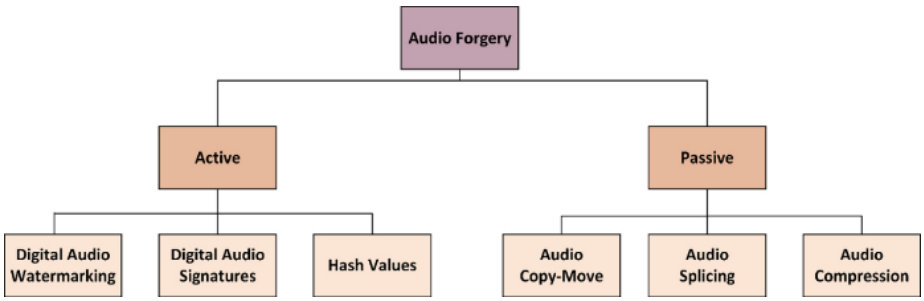


Figure 1. Types of voice forgery [19]

The copy-move method is one of the most common audio forgery techniques. In this method, the attacker copies some parts of a person's voice recording and pastes them into another part of the same voice recording. This type of forged data is usually not recognized as it is derived from the same speech recording. Detecting forged audio recordings only by listening to them without using any technical methods will result in time waste and low accuracy [20]. Figure 2 (a) and (b) show the time domain signals of the original and copy-move generated fake audio files, respectively. In Figure 2(b), the part where a different word is pasted with the copy-move method is marked with a red box. However, it should be noted that the pasted region is only identifiable when prior knowledge about the forgery location is available. Without such information, it is practically impossible to distinguish between the original and fake signals through visual inspection alone. Subtle spectral and temporal inconsistencies, phase discontinuities, and slight changes in background noise patterns introduced by the copy-move operation are not easily perceivable by the human eye in the time domain. Therefore, automatic detection relies on machine learning algorithms capable of analyzing fine-grained signal characteristics that are beyond human perception.

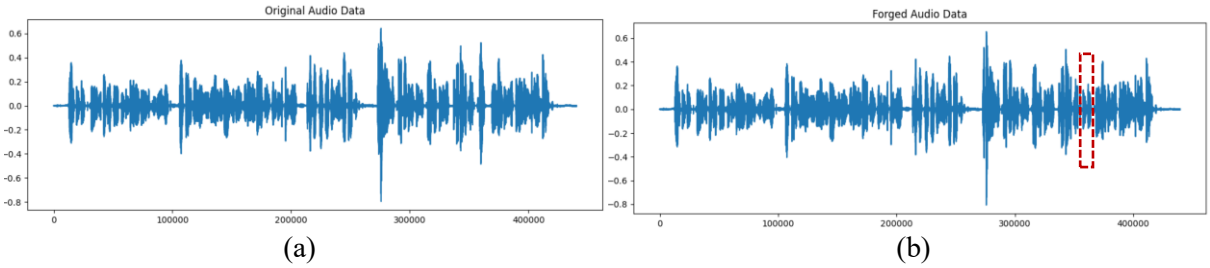


Figure 2. An example to copy-move forgery. Time signal of (a) original sound (b) fake sound.

The text of the data, whose original audio file is given in Figure 2 (a) in Turkish, reads as follows: “Bir isim fiil ile bir çekimli fiilin bir araya geliş ilişkilerinden, ortaya çıkan birimin cümlede ve bağlam içinde kazandığı değere kadar taşıdığı ipuçları bizi çok farklı yaklaşımlara sevk edebilir. Bu bağlamda Trabzon ağızlarının fiil şekilleri ve özellikle zarf fiiller açısından ele alınması önem kazanır.” The data of the fake voice obtained with the copy-move method is given in Figure 2(b) and its text is as follows: Bir isim fiil ile bir çekimli fiilin bir araya geliş ilişkilerinden, ortaya çıkan birimin cümlede ve bağlam içinde kazandığı değere kadar taşıdığı ipuçları bizi çok farklı yaklaşımlara sevk edebilir. Bu bağlamda Trabzon ağızlarının fiil şekilleri ve özellikle ~~zarf~~ isim fiiller açısından ele alınması önem kazanır."

Copy-move audio forgery analysis can detect internal manipulations of an audio recording. Especially in legal investigations, it is very important to determine the reliability of the audio recording. Therefore, such analysis is necessary to ensure the reliability of audio recordings.

3. MACHINE LEARNING ALGORITHMS

3.1. Logistic Regression

Logistic Regression (LR) is characterized as both a regression and a classification algorithm. However, it is generally used for binary classification problems. It is one of the most frequently used supervised machine learning algorithms [21] [22].

This algorithm can be used, for example, to predict whether a person is "sick" or "healthy". Logistic regression estimates the probability and performs the classification process using this probability. In other words, the predicted value is compared with a threshold value and classification is performed. Due to its simple and straightforward structure, it is one of the first preferred algorithms in classification studies. Figure 3 illustrates the decision boundary of the LR classifier. The model estimates the probability that a given input belongs to one of two classes using a sigmoid function. Data points located near the lower end of the curve are classified into one class (orange circles), while those near the upper end are classified into the other (blue circles). The nonlinear S-shaped curve reflects the gradual probability transition across the feature space, with the decision boundary typically set at a probability threshold of 0.5.

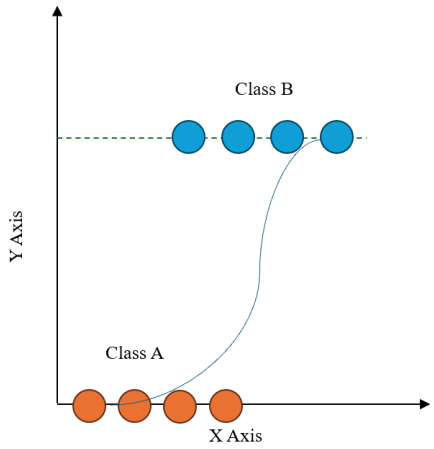


Figure 3. Binary Classification with LR algorithm [23]

3.2. Support Vector Machine (SVM)

SVM is one of the supervised machine learning algorithms developed in the 1990s. It can be used in various tasks such as classification and regression. The main goal of the algorithm is to find the hyperplane that provides the most optimal separation of data points. That is, the distance between the hyperplane and the data points closest to the boundary should be maximum. This system provides a more accurate separation of classes and a better classification of incoming data [6] [24] [25]. The data classification representation of the support vector machines algorithm is given in Figure 4.

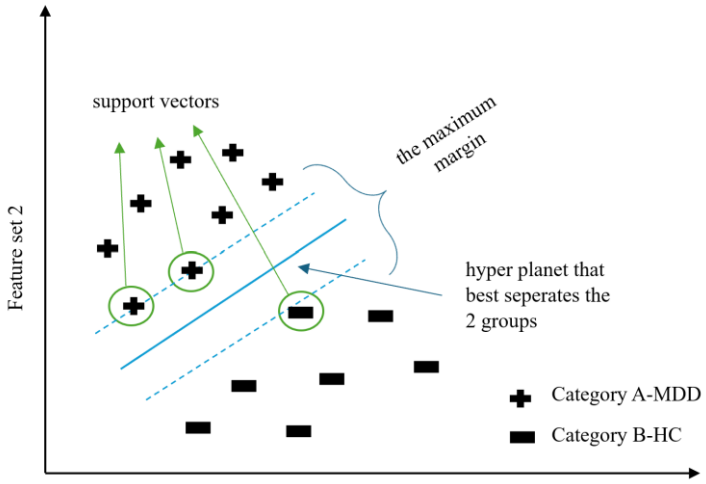


Figure 4. Demonstration of the separation of data into two classes with SVM (where the classes are Major depressive disorder (MDD) and Healthy controls (HC) [26].

When the figure is analyzed, firstly, the two axes of the graph represent certain features belonging to two different classes. Classification is done based on these features. According to the algorithm, a hyperplane is drawn to separate these data in the best way. The maximum margin represents the plane that widens the difference between the groups the most. The data points on this margin are defined as support vectors.

3.3. Random Forest (RF)

RF, proposed in 2001, is one of the most widely used machine learning algorithms for classification and regression problems. RF consists of multiple decision trees. Randomization is used to generate multiple decision trees. In accordance with the type of problem, the output of the trees are combined into a single result using voting for classification and averaging for regression [10] [27] [28]. The basic representation of the RF algorithm is given in Figure 5.

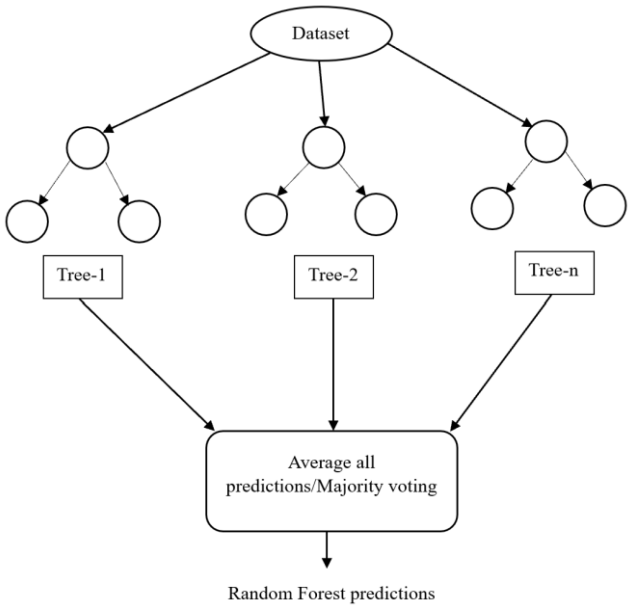


Figure 5. Structure of Random Forest [29] [30].

When Figure 5, which explains the basic structure of Random Forest, is examined, each tree produces a prediction in classification problems. The class that reaches the majority among the trees is considered as the prediction result of the model. If the problem is a regression problem, the model result is obtained by averaging the prediction results. The random forest algorithm provides diversity and improves prediction performance by using multiple trees instead of a single tree.

3.4. K-Nearest Neighbors (KNN)

KNN is one of the supervised learning algorithms used in classification and regression problems. The basic logic of the algorithm is that when making a decision about a data point, other data surrounding it is considered. Therefore, the entire training data set is consulted during classification. The decision to classify each new data point is made by using all the examples in the data set.

When determining which class a new data point belongs to, the class of the *K* closest data points belonging to that data point is considered. The class to which the most data belongs is the class of the new data point. The *K* value is usually chosen as a small integer value such as 3 and 5. In the KNN algorithm, Euclidean distance calculation is usually used to measure the distance between data. Euclidean distance is the distance calculated along a straight line between two points. This means that data with the same class label are close to each other in terms of distance [9] [31]. Figure 6 shows a basic representation of the KNN algorithm.

Figure 6, where the classification process is performed using the KNN algorithm, shows two classes consisting of blue squares and green circles. The data to be classified is indicated by the black plus symbol. The k value of the algorithm is set to 3. Therefore, the three closest neighbors of the data to be classified are examined. Two of the examined neighbors belong to class A and one belongs to class B. For this reason, our data is included in class A.

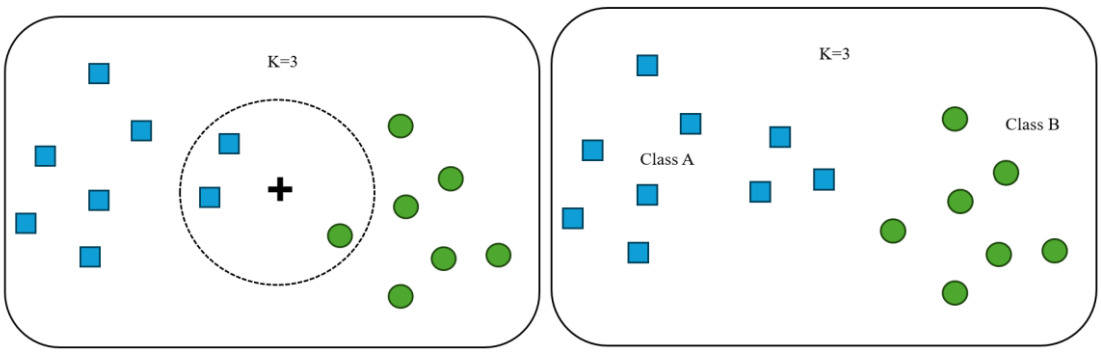


Figure 6. Demonstration of basic classification with KNN algorithm.

3.5. Extreme Gradient Boosting (XGBoost)

XGBoost was developed by Chen et al [8]. This method is a scalable implementation of gradient boosting machines. Boosting is an ensemble method where new models are added to correct the errors of the models. The added models are added recursively until a significant improvement is seen in the result. Gradient boosting is an algorithm in which the errors of previous models are estimated and determined, and new models are developed and combined to form the outcome prediction. A gradient descent algorithm is used to minimize the loss when adding new models.

In order to achieve an optimal result, the parameters of the XGBoost algorithm should be set correctly. This is quite difficult as XGBoost has a large number of parameters. “Grid Search” or 'Random Search' methods are used for the parameter tuning task. In this study, “Random Search” technique is used for hyperparameter tuning of XGBoost algorithm. The random search method usually shows a fast performance [8] [32] since it tries on a certain number of random samples instead of trying all combinations. A basic illustration of the XGBoost algorithm is given in Figure 7.

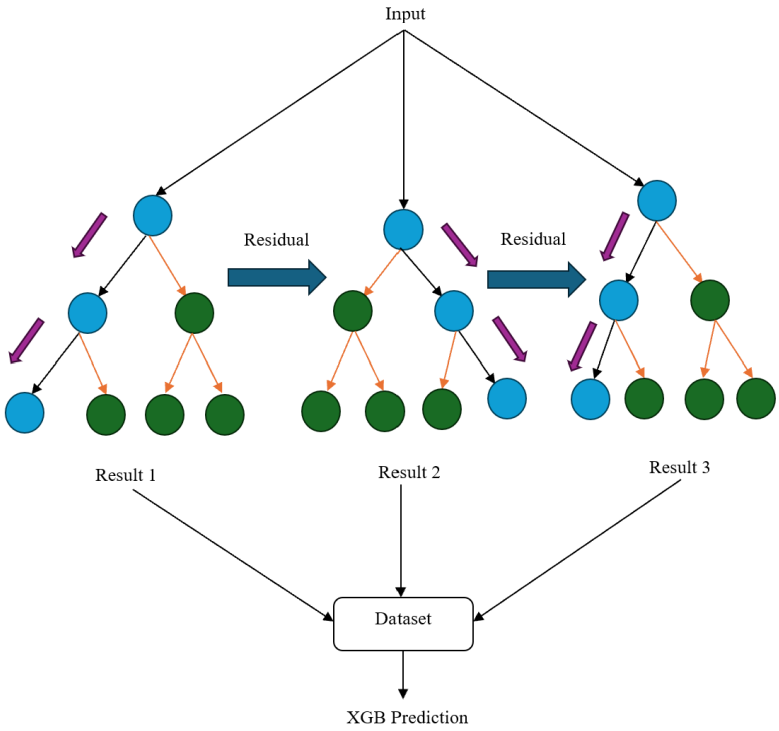


Figure 7. XGBoost [33]

Figure 7, which shows the fundamental operation of XGBoost, shows that a new tree is added to the model to eliminate the error generated by the previous tree. This process improves the performance of the model. This process continues for several cycles until no further improvement is achieved or until the number of trees reaches a specified upper limit.

4.EXPERIMENTAL RESULTS

4.1. Dataset

The data set used in this study consists of 100 different people, 50 women and 50 men, reading 200 different texts. The texts were read in three different environments: office, cafeteria and quiet room. The ages of the speakers are 50 people between 18-25 years old, 30 people between 25-35 years old and 20 people between 35-55 years old. The audio files have a sampling rate of 44.1 kHz and 16-bit coding. The voice files are in wav format.

Within the scope of the dataset, fake voices are created by copy-move method using the original voice recordings. First, Matlab's speech2text tool is used to detect the beginning and end of the words in the original voices. Then, word pairs are determined for the copy and paste process. Thus, fake audio recordings are created by copy-paste forgery. Table 1 gives a detailed representation of the dataset [34].

Table 1. Distribution of the dataset used in the study

Environment	Voice recording counts	
	Original	Fake
Office	200	349
Cafe	200	349
Quiet room	200	349

As can be seen in Table 1, each environment has 200 original voice recordings. A total of 600 original audio recordings were used to generate 1047 fake voices. Since the sounds were taken in three different environments, the noise levels may vary.

Extracting features is an important step to study and analyze audio signals. In this study, the mel-spectrogram feature is considered. The mel spectrogram represents the frequency components of audio signals similar to human hearing. It is effective in time-frequency analysis of audio signals. In order to classify audio signals, audio files are visualized with mel spectrogram. Sample mel-spectrogram representations of the sounds in the dataset are given in Figure 8.

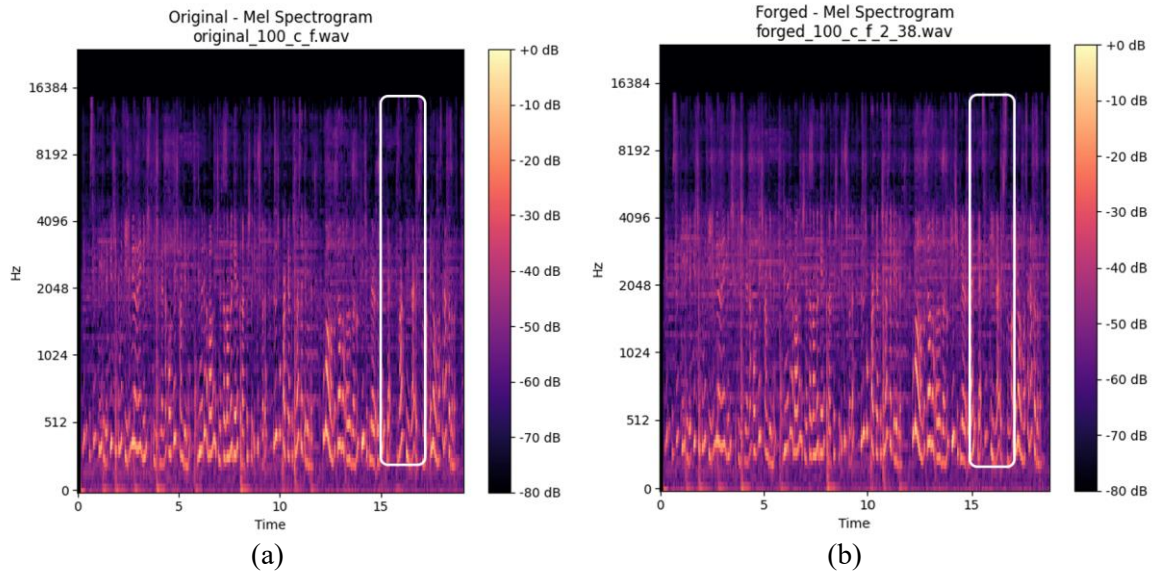


Figure 8. Mel-spectrogram representation a) original sound b) fake sound

In this study, mel-spectrograms were computed using a window size (win_length) of 2048 samples and a hop length of 512 samples, which corresponds to approximately 10.67 milliseconds of temporal resolution at a 48 kHz sampling rate. A 2048-point FFT was applied with a Hann window function to reduce spectral leakage. These settings allow for a detailed time-frequency analysis suitable for capturing subtle manipulations in the audio signals. Mel-spectrograms show the change of the audio signal over time. The horizontal axis shows time and the vertical axis shows frequency components. Figure 8(a) shows the mel-spectrogram representation of the original audio file with the file name 'original_100 c_f', while Figure 8(b) shows the mel-spectrogram representation of the forged audio file named 'forged_100_c_f_2_38' obtained using the same audio file. Mel-spectrograms visually show how the energy of a sound signal is distributed across its frequency components over time, and as a result, such manipulations can sometimes lead to spectral anomalies. In particular, the addition of the copied region can create inconsistencies in frequency components, as well as noticeable repetitions or discontinuities along the time axis. These issues may manifest in the mel-spectrogram as differences in color intensity or spectral patterns. However, factors such as the scale of the manipulation and the similarity of the audio source can affect the visibility of these changes. Small-scale manipulations may

create subtle differences, making it difficult to detect them visually. Therefore, detecting such forgery may not always be possible through traditional visual inspection, which is why more precise and automated machine learning algorithms are needed.

4.2. Fake Voice Detection with Mel-spectrogram Feature

After extracting the mel-spectrogram features of both original and fake audio files in the dataset, machine learning methods are used to classify the original and fake audio. Figure 9 shows a flow diagram of the study.

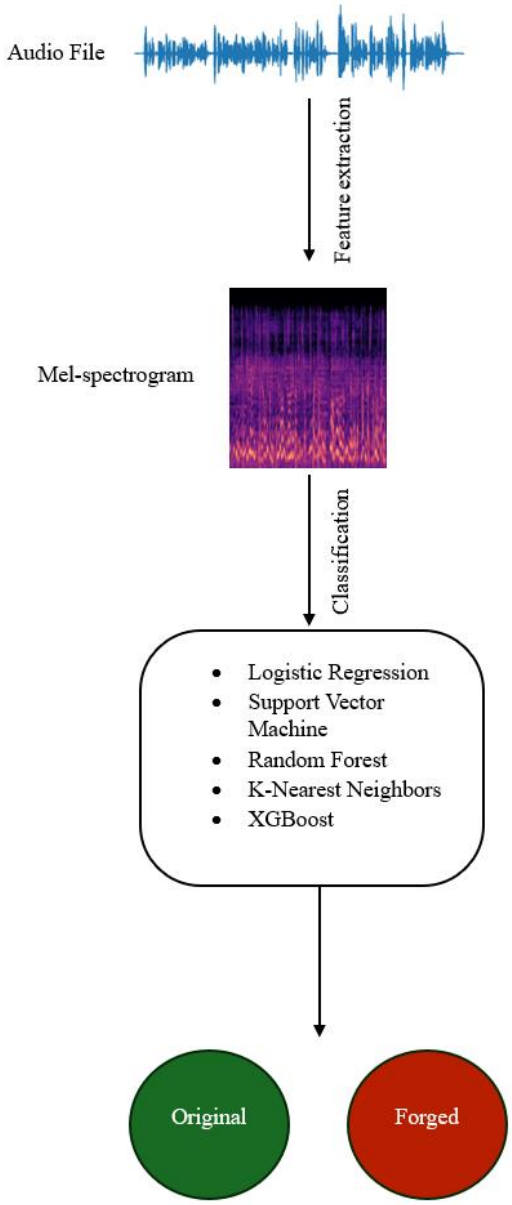


Figure 9. Real and forged audio classification flow diagram

The steps of forgery detection from an audio file are given in Figure 9. First, the audio file is received as input to the system as raw data. Then, the raw audio signal is analyzed with time and frequency components and converted into a Mel-spectrogram. Using the Mel-spectrogram, the changes of the

sound over time are presented visually. This feature is then presented as input to machine learning algorithms. The classification step indicated in the diagram involves different machine algorithms. These are LR, SVM, RF, KNN and XGBoost algorithms. The most successful result among the algorithms was obtained with XGBoost. Parameter tuning is performed with 'Random Search' to improve the performance of the XGBoost algorithm. While LR makes a probabilistic classification, SVM aims to find the hyperplane that best classifies the data. On the other hand, decision tree-based algorithms such as RF and XGBoost provide high accuracy on complex data structures. Finally, KNN performs the classification process based on the nearest neighbors of the data. According to the results of these algorithms, it is finally determined whether the sound is original or fake. Parameters of the algorithms used in the study are shown in Table 2.

Table 2. Hyperparameters of the algorithms used in the study

Methods	Hyperparameters	Definition
Logistic Regression	• max_iter=100 • solver='lbfgs' • tol=1e-4 • class_weight= None	• Maximum number of iterations. • Optimization algorithm used: 'lbfgs' • Stopping criterion of optimization • It is used to determine the weights of the classes. The default value is None and each class has equal weight.
Support Vector Machine	• kernel='rbf' • C=1.0	• The kernel function • The penalty parameter
Random Forest	• ntree=100	• The number of trees
K-Nearest Neighbors	• K=5 • weights: uniform	• Number of neighbors • Weight function used in prediction
XGBoost	• max_depth=6 • colsample_bytree=1 • subsample=1 • learning_rate=0.3 • n_estimators=100	• Maximum depth of a tree • Subsample ratio of columns when constructing each tree • Subsample ratio of the training instance • Control the learning rate • Number of trees
XGBoost (XGBoost with hyperparameters tuned using Random Search)	• n_estimators= [100, 200, 300] • learning_rate= uniform (0.01,0.2) • max_depth= [3,5,7] • subsample=uniform (0.7, 1.0) • colsample_bytree=uniform (0.7, 1.0)	• Number of trees • Control the learning rate • Maximum depth of a tree • Subsample ratio of the training instance • Subsample ratio of columns when constructing each tree

In this study, the XGBoost algorithm used for audio forgery detection was optimized using the Random Search method. This approach involved performing random searches over various hyperparameters to maximize the model's performance. The parameters used include 'n_estimators' (100, 200, 300) to determine the number of trees, 'learning_rate' (random between 0.01 and 0.2) to control the learning rate, 'max_depth' (3, 5, 7) to limit the tree depth, 'subsample' (random between 0.7 and 1.0) to define the sample ratio for each tree, and 'colsample_bytree' (random between 0.7 and 1.0) to define the feature selection ratio for each tree. Each of these hyperparameters was carefully selected and optimized to improve the overall performance of the model.

Figure 10 shows the Confusion Matrix outputs obtained to evaluate the classification performance of the algorithms. Confusion matrix is an evaluation tool used to analyze the performance of a machine learning model and shows the true and false classifications of the model in a quantitative table.

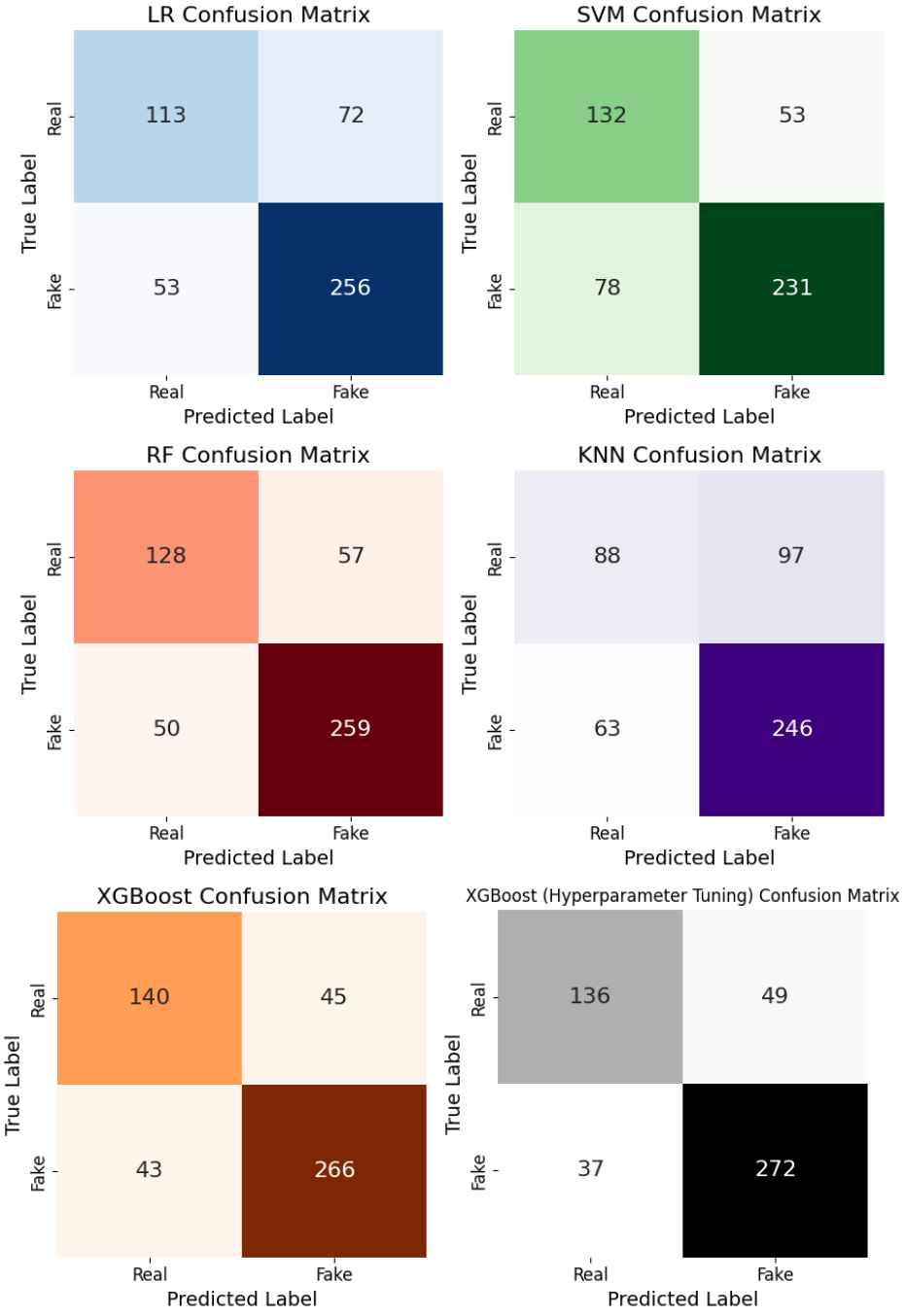
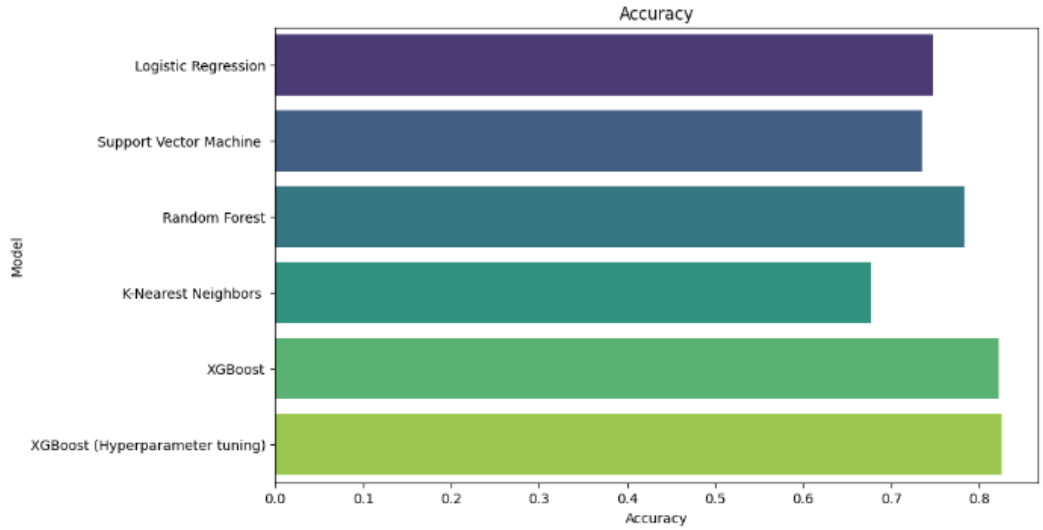


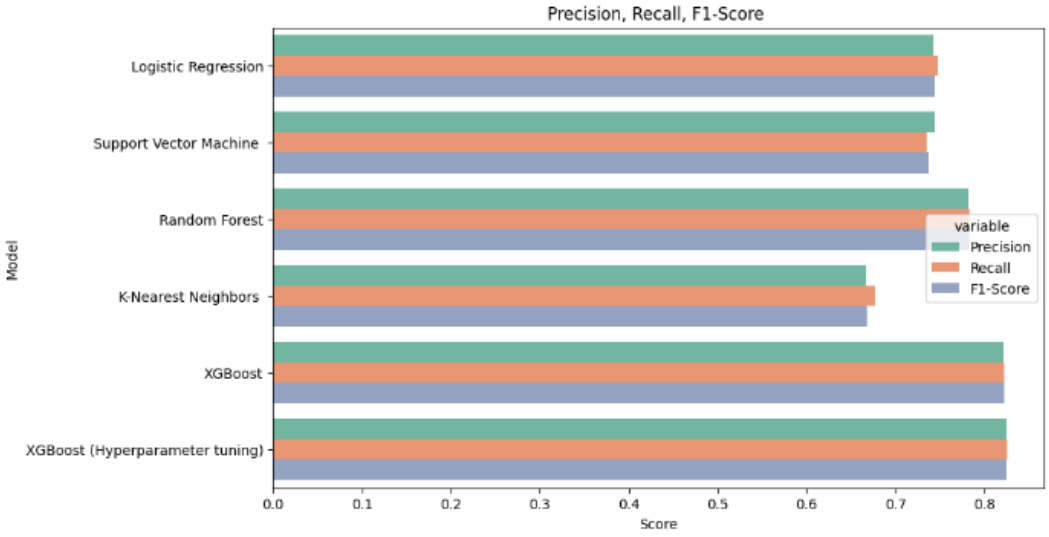
Figure 10. Confusion matrices

A detailed analysis of the confusion matrices reveals that the XGBoost model with hyperparameter optimization achieves the highest number of true positive predictions (272) and the lowest number of false negatives (37) among all evaluated models, indicating an enhanced capability in correctly detecting fake audio samples. Although the default XGBoost model exhibits the lowest false positive count (45),

the overall trade-off between false positives and false negatives appears more balanced in the hyperparameter-tuned model. This balance is particularly important in forgery detection tasks, where minimizing both types of errors contributes significantly to the reliability of the classification system. Therefore, the hyperparameter-tuned XGBoost model demonstrates a comparatively more favorable performance profile, suggesting its potential suitability for practical deployment in audio forgery detection scenarios. The results of other metrics calculated based on the confusion matrices are shown in Figure 11.



(a)



(b)

Figure 11. Comparisons of the models (a) accuracy score (b)precision, recall, f1 score

Figure 11(a) presents the overall accuracy scores of the evaluated models, while Figure 11(b) provides a comparative analysis of their Precision, Recall, and F1-Score metrics. As can be observed, the XGBoost and hyperparameter-tuned XGBoost models consistently outperform the other algorithms across all evaluation metrics. Notably, the hyperparameter-tuned XGBoost model achieves the highest F1-Score (0.8247), indicating a balanced and reliable classification performance between precision and recall. In contrast, the K-Nearest Neighbors (KNN) model demonstrates the lowest performance across all metrics, with an accuracy of 67.61% and a corresponding decline in precision (66.69%), recall

(67.61%), and F1-Score (66.82%). This suggests that KNN may not effectively capture the underlying structure of the audio forgery data in this study. The numerical results, summarized in Table 3, further highlight that while Random Forest (RF) and Support Vector Machine (SVM) models show competitive precision values, they lag slightly behind XGBoost-based methods in terms of recall and F1-Score. Considering the relatively high F1-Score values obtained with XGBoost models, it can be inferred that these methods offer a more robust balance between correctly identifying forged audio samples and minimizing false alarms. Overall, these findings reinforce the suitability of ensemble-based approaches, particularly optimized XGBoost models, for the detection of copy-move forgeries in audio data, especially when computational efficiency and interpretability are also prioritized.

Table 3. Test results of the models.

Model	Accuracy	Precision	Recall	F-1 score
LR	0.7470	0.7431	0.7470	0.7439
SVM	0.7348	0.7442	0.7348	0.7376
RF	0.7733	0.7707	0.7733	0.7713
KNN	0.6761	0.6669	0.6761	0.6682
XGBoost	0.8219	0.8215	0.8219	0.8217
XGBoost (tuned)	0.8259	0.8244	0.8259	0.8247

The Receiver Operating Characteristic (ROC) curve is a widely used tool for evaluating the performance of classification models. It shows the trade-off between the true positive rate and the false positive rate across various thresholds. The closer a model's ROC curve is to the upper left corner of the graph, the better the classification performance. Further, the Area Under the Curve (AUC) value provides a scalar metric that gives a single metric summarizing the overall effectiveness of the model, with higher AUC values indicating better performance.

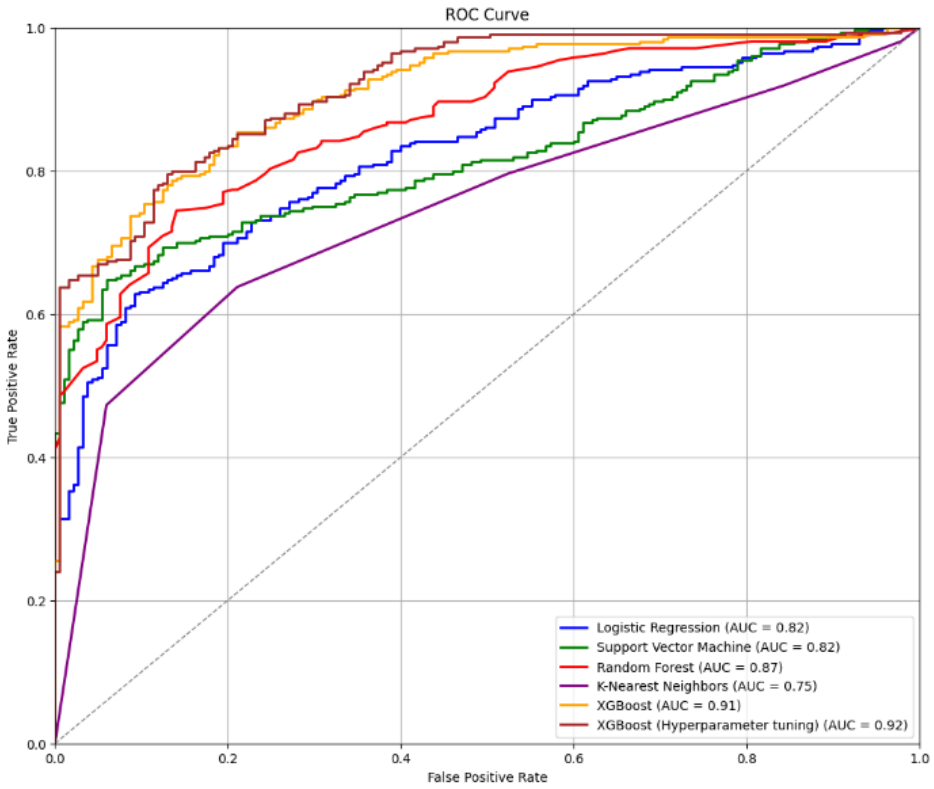


Figure 12. ROC curve

Figure 12 presents the ROC curves for the six classification algorithms used in this study. Analyzing the ROC curves in Figure 12, it can be seen that the XGBoost algorithm with hyperparameter tuning achieves the best performance and its curve is closer to the upper left corner than the others. This observation is supported by the AUC value of 0.92, which is higher than the other models. The standard XGBoost implementation also performs strongly (AUC = 0.91), followed by Random Forest (AUC = 0.87). In contrast, SVM and Logistic Regression models show moderate performance with the same AUC values of 0.82, while KNN shows the lowest AUC value of 0.75, indicating relatively weaker classification ability. The XGBoost (hyperparameter tuning) model stands out as the most successful model, achieving the highest results in all metrics. Random Forest and XGBoost models also show balanced performances. KNN gives the lowest results in all metrics compared to other models. Depending on the application area, the metrics that should be emphasized may change.

These results highlight the effectiveness of hyperparameter tuning in improving model performance, as demonstrated by the significant improvement of XGBoost after tuning. Furthermore, the comparison between the algorithms highlights the importance of choosing not only the right classification method but also the optimal parameter configuration for robust audio forgery detection.

Many studies on audio forgery detection in the literature have been conducted using outdated datasets, which often fail to reflect modern audio manipulation techniques; Table 4 summarizes the information about these studies. The KTUCengAudioForgerySet, with its up-to-date structure, enables more accurate and reliable results in forgery detection. This dataset contains comprehensive and rich examples designed to model contemporary audio forgeries, and its inclusion of both original and forged audio files allows for more reliable analyses.

Table 4. Summary of Studies on Audio Copy-Move Forgery Detection

Study	Methods	Dataset
Akdeniz & Becerikli (2024) [19]	MFCCs, MFCCs, MFCCs, MFCC + MFCC + MFCCs, and LPCs	TIMIT database (1993)
Su et al. (2023) [35]	CQCC, sliding window	Chinese speech and LibriSpeech dataset (2015)
Yan et al. (2019) [36]	Pitch feature and formant feature	Wall Street Journal(WSJ) speech database (1992) TIMIT database (1993)
Imran et al. (2017) [37]	1D LBP	King Saud University Arabic Speech Database (2014)
Our study	**XGBoost (hyperparameters tuned using Random Search) with Mel spectrogram features**	**KTUCengAudioForgerySet (2024)**

5. CONCLUSION

In this study, the results of machine learning based models developed for voice forgery detection are analyzed. Comparisons between different algorithms show that XGBoost provides the highest success rate and has a balanced performance in terms of both precision and recall. In particular, the Random Search method applied for the hyperparameter tuning of XGBoost increased the accuracy of the model. The use of Mel-spectrograms for analyzing audio data within the scope of the study helps to understand

the structure of the voice and to make an accurate classification in forgery detection. Audio forgery is a significant threat in the digital age. In this context, techniques such as "copy-move forgery" can be applied to audio files to easily obtain and use fake voices, creating more risk. The results emphasize that XGBoost is more effective in audio forgery detection than other machine learning models. In addition, this study lays the groundwork for future work and research on analyzing audio data. Integration of different feature extraction methods or model development to improve performance for more complex and real-world problems will provide guidance for future studies. The increasing accessibility of audio manipulation tools raises ethical concerns, particularly regarding the potential misuse of forged audio in malicious or deceptive contexts. Moreover, the risk of false positives in detection systems could lead to unintended consequences, especially in sensitive applications such as legal or forensic investigations. Therefore, while developing detection systems, it is crucial to balance technical performance with considerations of fairness, transparency, and responsible use. In future work, the dataset will be expanded to include a wider range of audio samples, enabling the application of more complex models. This extension will allow for the integration of deep learning architectures such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which are well-suited to capturing intricate temporal and spectral patterns in audio data. These models are expected to improve forgery detection performance, particularly in more diverse and realistic scenarios.

CONFLICT OF INTEREST

The authors stated that there are no conflicts of interest regarding the publication of this article.

CRediT AUTHOR STATEMENT

Merve Arslan: Software, Visualization, Data Curation, Writing – Original Draft, Writing – Review & Editing **Şerif Ali Sadık:** Conceptualization**,** Methodology, Visualization, Writing – Review & Editing.

REFERENCES

[1] Imbwaga JL, Chittaragi NB, Koolagudi SG. Automatic hate speech detection in audio using machine learning algorithms. Int J Speech Technol 2024; 27(2): 447-469.

[2] Salamon J, Jacoby C, Bello JP. A dataset and taxonomy for urban sound research. In: 22nd ACM International Conference on Multimedia; Nov 2014; pp. 1041-1044. Retrieved 14 December 2020 from https://urbansounddataset.weebly.com/urbansound8k.html.

[3] Chathuranga S. Sound Event Dataset. [Online]. Retrieved 14 December 2020 from https://github.com/chathuranga95/SoundEventClassification.

[4] Gourisaria MK, Agrawal R, Sahni M, Singh PK. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. Discov Internet Things 2024; 4(1): 1.

[5] Reimao R, Tzerpos V. For: a dataset for synthetic speech detection. In: 2019 International Conference on Speech Technology and Human–Computer Dialogue (SpeD); 2019. IEEE; pp. 1-10.

[6] Evgeniou T, Pontil M. Support vector machines: theory and applications. In: Advanced Course on Artificial Intelligence. Berlin: Springer; 1999. pp. 249-257.

[7] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017; 30: 3146-3154.

[8] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. pp. 785-794.

[9] Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In: OTM Confederated International Conferences on the Move to Meaningful Internet Systems. Berlin: Springer; 2003. pp. 986-996.

[10] Breiman L. Random forests. Mach Learn 2001; 45(1): 5-32.

[11] Khochare J, Joshi C, Yenarkar B, Suratkar S, Kazi F. A deep learning framework for audio deepfake detection. Arab J Sci Eng 2021; 1-12.

[12] Gupta S, Cho S, Kuo CCJ. Current developments and future trends in audio authentication. IEEE Multimedia 2011; 19(1): 50-59.

[13] Ustubioglu B, Tahaoglu G, Ulutas G. Detection of audio copy-move-forgery with novel feature matching on Mel spectrogram. Expert Syst Appl 2023; 213: 118963.

[14] Klapuri A, Davy M. Signal processing methods for music transcription. Springer; 2006. ISBN 978-0-387-30667-4.

[15] Yan Q, Yang R, Huang J. Copy-move detection of audio recording with pitch similarity. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Apr 2015. IEEE; pp. 1782-1786.

[16] Wang F, Li C, Tian L. An algorithm of detecting audio copy-move forgery based on DCT and SVD. In: 2017 IEEE 17th International Conference on Communication Technology (ICCT); Oct 2017. IEEE; pp. 1652-1657.

[17] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST Speech Disc 1-1.1. NASA STI/Recon Tech Rep N 1993; 93: 27403.

[18] Akdeniz F, Becerikli Y. Detection of copy-move forgery in audio signal with mel frequency and delta-mel frequency kepstrum coefficients. In: 2021 Innovations in Intelligent Systems and Applications Conference (ASYU); 2021. IEEE; pp. 1-6.

[19] Akdeniz F, Becerikli Y. Detecting audio copy-move forgery with an artificial neural network. Signal Image Video Process 2024; 18(3): 2117-2133.

[20] Yan Q, Yang R, Huang J. Robust copy-move detection of speech recording using similarities of pitch and formant. IEEE Trans Inf Forensics Secur 2019; 14(9): 2331-2341.

[21] Shah K, Patel H, Sanghvi D, Shah M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augment Hum Res 2020; 5(1): 12.

[22] Yang Z, Li D. Application of logistic regression with filter in data classification. In: 2019 Chinese Control Conference (CCC); Jul 2019. IEEE; pp. 3755-3759.

[23] Mondal PK, Foysal KH, Norman BA, Gittner LS. Predicting childhood obesity based on single and multiple well-child visit data using machine learning classifiers. Sensors 2023; 23(2): 759.

[24] Vapnik VN. The Nature of Statistical Learning Theory. 2nd ed. Springer Verlag; 1995. pp. 1-20.

[25] Kotsiantis SB. Supervised machine learning: a review of classification techniques. Informatica 2007; 31: 249-268.

[26] Schnyer DM, Clasen PC, Gonzalez C, Beevers CG. Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder. Psychiatry Res Neuroimaging 2017; 264: 1-9.

[27] Biau G, Scornet E. A random forest guided tour. Test 2016; 25: 197-227.

[28] Rigatti SJ. Random forest. J Insur Med 2017; 47(1): 31-39.

[29] Sahour H, Gholami V, Torkaman J, Vazifedan M, Saeedi S. Random forest and extreme gradient boosting algorithms for streamflow modeling using vessel features and tree-rings. Environ Earth Sci 2021; 80: 1-14.

[30] Kiranmai SA, Laxmi AJ. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. Prot Control Mod Power Syst 2018; 3(3): 1-12.

[31] Bramer M. Principles of data mining. Springer; 2007.

[32] Ogunleye A, Wang QG. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform 2019; 17(6): 2131-2140.

[33] Wang CC, Kuo PH, Chen GY. Machine learning prediction of turning precision using optimized XGBoost model. Appl Sci 2022; 12(15): 7739.

[34] Ustubioglu B, Tahaoglu G, Ayaz GO, Ustubioglu A, Ulutas G, Cosar M, Kılıc M. KTUCengAudioForgerySet: a new audio copy-move forgery dataset. In: 2024 47th International Conference on Telecommunications and Signal Processing (TSP); Jul 2024. IEEE; pp. 123-129.

[35] Su Z, Li M, Zhang G, Wu Q, Wang Y. Robust audio copymove forgery detection on short forged slices using sliding window. Journal of Information Security and Applications, 2023; 75, 103507.

[36] Yan Q, Yang R, Huang J. Robust copy-move detection of speech recording using similarities of pitch and formant. IEEE Trans. Inform. Forensics Secur., 2019;4(9), 2331–2341.

[37] Imran M, Al Z, Bakhsh ST, Akram S. Blind detection of copy-move forgery in digital audio forensics. IEEE Access, 2017; 5, 12843–12855.