Archives available at journals.mriindia.com

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319-2526

Volume 15 Issue 01, 2026

AI-Generated Voice Detection Using Machine Learning and Deep Learning

[1]G. G. Desai, [2]Aniket K. Pawar, [3]Onkar N. Upase, [4]Aditya S. Sid, [5]Soham S. Magdum
[1]Assistant professor, Artificial Intelligence and Data Science Dr. J. J. Magdum College of Engineering Jaysingpur, India
Artificial Intelligence and Data Science Dr. J. J. Magdum College of Engineering Jaysingpur, India
Email: [1]gousiya.desai@jjmcoe.ac.in, [2]ap1660392@gmail.com, [3]onkarupase332@gmail.com, [4]adityasid2512@gmail.com, [5]magdumsoham20@gmail.com

Peer Review Information

Submission: 18 March 2026

Revision: 05 April 2026

Acceptance: 27 April 2026

Keywords

AI Voice Detection, Deepfake Audio, Sentiment Analysis, Machine Learning, LibROSA, Speech Processing

Abstract

Recent advancements in artificial intelligence have enabled the generation of highly realistic synthetic voices using text-to-speech and voice cloning technologies. While these innovations have numerous applications, they also introduce serious security threats such as impersonation, fraud, and misinformation. This paper proposes an offline AI-based system for detecting whether an audio sample is human or AI-generated. The system utilizes audio preprocessing techniques, feature extraction using LibROSA, and classification through machine learning and deep learning models including Random Forest and Convolutional Neural Networks. Additionally, sentiment analysis is incorporated to analyze emotional tone in user inputs, enhancing system intelligence. Experimental results demonstrate high classification accuracy and robustness, making the system suitable for forensic and cybersecurity applications.

Introduction

Artificial Intelligence has significantly transformed the field of speech processing, enabling the generation of highly realistic synthetic voices through advanced text-to-speech and voice cloning technologies. while these developments have improved applications such as virtual assistants, accessibility tools, and media production, they have also introduced serious security challenges, including impersonation, fraud, and the spread of misinformation. Distinguishing between human and AI-generated voices has become increasingly difficult due to their similar acoustic characteristics, making manual detection unreliable.to address this issue, the research proposes an offline audio based classification system that analyzes recorded speech using advanced preprocessing techniques and feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral features, and chroma representations. Machine learning and deep learning models, including Random Forest and Convolutional Neural Networks (CNNs), are employed to accurately classify audio samples as human or synthetic. In addition, the system integrates sentiments analysis on user input to determine emotional context, thereby, enhancing the interpretability and applicability of the model. By combining robust audio analysis with natural language processing, the proposed system aims to provide a reliable and scalable solution for voice authentication in cybersecurity, digital forensics, and real-world applications.

Literature Review

Mahine Learning-Based Audio Classification (Alexsoft, 2022) :
This study explored intelligent sound detection systems using machine learning algorithms such as Random Forest, Support Vector Machines (SVM), and environmental sounds. Although it achieved good accuracy, it was not specially designed for distinguishing AI-generated voices from human speech

Real-Time Deepfake Audio Detection (Bird & Lotfi, 2023) :
This research proposed a deep learning framework combining CNN and BilSTM architectures along with wav2vec 2.0 embeddings for detecting AI-generated speech. The system achieved accuracy above 90%, demonstrating strong performance in cybersecurity applications. However, it required high computational resources and needed further validation in noisy real-world environments

Neural Network-Based Speech Recognition (Al Smadi & Al Issa, 2015) :
This work focused on speech recognition using deep neural networks such as CNNs, RNNs and Transformer-based architectures. The system showed improved adaptability to different accents and noisy conditions, reducing error rates. However, the primary focus was speech-to-text conversion rather than detecting deepfake audio.

Deep Learning for Forensic Audio Detection (Mucuba etal.,2023) :
This study compared multiple deep learning models including CNN. CNN-LSTM, and Transformer-based approaches for deepfake audio detection. Results indicated that CNN-LSTM models provided better accuracy and interpretability for forensic applications. The limitation of this work was its evaluation on controlled datasets, lacking real-world testing.

Explainable AI in Deepfake Detection (Govindu etal.,2023):
This research introduced explainability techniques such as SHAP and Grad-CAM combined with CNN-LSTM models to improve transparency in deepfake detection systems. The model provided both high accuracy and interpretability, which is crucial in security and investigation scenarios. However, the system required optimization for real-time deployment and resistance to adversarial attacks.

Spectrogram- Based Deepfake Detection with Ensemble Models(LamPham 2024) :
This study utilized advanced spectrogram techniques such as Short-Time Fourier Transform (STFT), Constant-Q Transform (CQT), and Wavelet Transform along with ensemble deep learning models. The system achieved very low error rates, demonstrating high effectiveness. However, the approach was computationally intensive and needed further optimization for practical deployment.

Problem Statement

The advancement of artificial intelligence has made it possible to generate highly realistic synthetic voices that are difficult to distinguish from human speech, leading to increased risks of fraud, impersonation, and misinformation. Traditional detection methods and human perception are no longer reliable due to the similarity in acoustic characteristics between real and AI-generated voices. Existing systems also face limitations such as poor performance in noisy environments and lack of generalization to new voice generation techniques. Therefore, there is a need for an efficient system that can accurately classify recorded audio as human or AI-generated, while also incorporating sentiment analysis to enhance contextual understanding and real-world applicability.

Proposed System

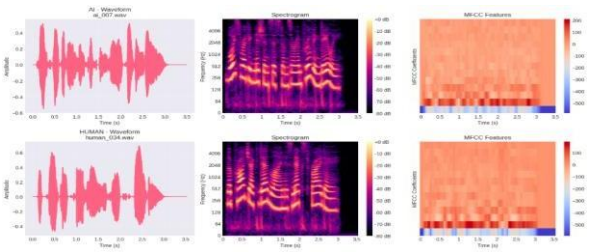
The proposed system is designed to accurately distinguish between human and AI-generated voices using an offline audio analysis approach. The system accepts recorded audio files as input, which are first preprocessed through steps such as resampling, noise reduction, silence removal, and normalization to ensure consistency. Relevant acoustic features, including MFCCs, chroma features, spectral properties, and energy-based attributes, are then extracted using audio processing techniques. These features are fed into machine learning and deep learning models such as Random Forest and Convolutional Neural Networks for classification. To enhance accuracy and robustness, an ensemble approach may be used to combine predictions from multiple models. Additionally, the system integrates a sentiment analysis module that analyzes user input text or speech to identify emotional context, classifying it as positive, negative, or neutral. The final output provides both voice authenticity (human or AI-generated) and sentiment insights, making the system useful for applications in cybersecurity, digital forensics, and intelligent user interaction systems.

Methodology

The proposed system is developed as a complete

pipeline for detecting whether an audio sample is human or AI- generated, along with performing sentiment analysis on user input. The methodology is divided into several stages, where each stage plays a crucial role in ensuring the accuracy, robustness, and usability of the system.

- **Data Collection**

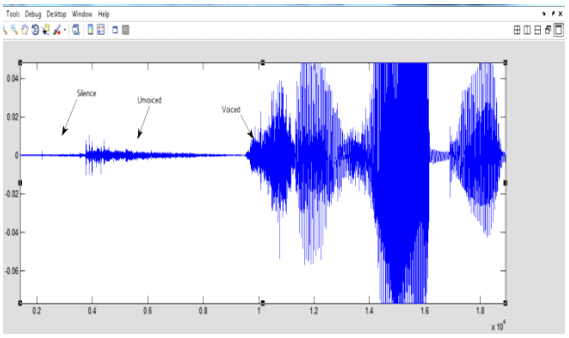


The first stage involves collecting a diverse dataset containing both human voice recordings and AI-generated speech samples. The audio data is obtained from multiple sources, including publicly available datasets and AI voice generation tools. The dataset consists of audio files in commonly used formats such as WAV and MP3 to ensure compatibility with real-world applications.

To improve the diversity and robustness of the dataset, data augmentation techniques are applied. These include time shifting, pitch variation, and speed modification, which simulate different speaking conditions and recording environments. This helps the model generalize better and reduces the chances of overfitting. A balanced dataset is maintained to ensure that both classes (human and AI-generated) are equally represented, preventing bias during training.

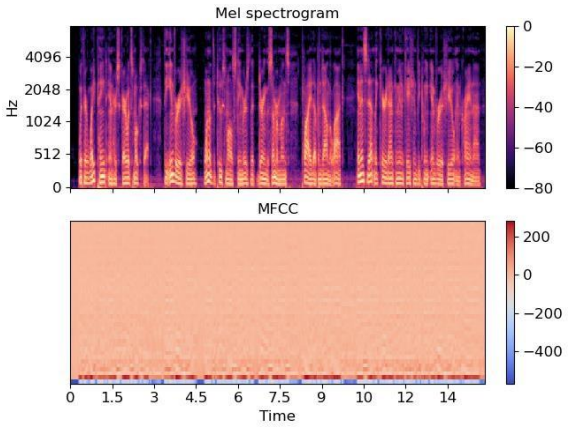
- **Audio Processing**

Raw audio signals often contain noise, silence, and inconsistencies that can negatively affect model performance. Therefore, preprocessing is performed to standardize the input data.
Initially, all audio files are resampled to a fixed sampling rate (e.g., 44.1 kHz) to ensure uniformity across the dataset. Silence at the beginning and end of the recordings is removed so that only relevant speech segments are analyzed. The amplitude of the signal is then normalized to maintain consistent loudness



across different samples, preventing bias caused by varying recording volumes.
Additionally, noise reduction techniques are applied to eliminate background disturbances and enhance the clarity of the audio signal. These preprocessing steps ensure that the input data is clean, consistent, and suitable for feature extraction.

- **Feature Extraction**



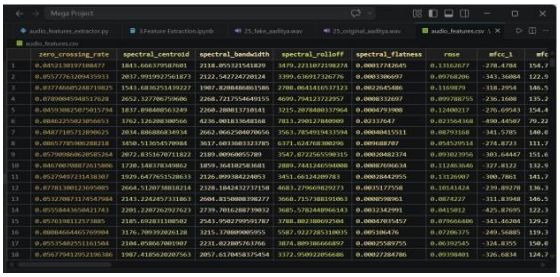
Feature extraction is a crucial step in which raw audio signals are transformed into meaningful numerical representations. These features capture important characteristics of speech that help distinguish between human and AI-generated voices.
The system extracts Mel-Frequency Cepstral Coefficients (MFCC), which represent the perceptual aspects of sound and are widely used in speech recognition tasks. Chroma features are used to capture pitch-related information, while Zero-Crossing Rate measures the rate of signal sign changes, indicating the noisiness of the signal. Spectral features such as centroid, bandwidth, and rolloff describe the distribution of frequencies, and Root Mean Square Energy represents the loudness of the signal.

MFCC Equation :

$$MFCC = \sum_{k=1}^{N} \log(S_k)\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{N}\right]$$

These extracted features collectively provide a detailed representation of audio signals, enabling effective classification.

- **Feature Representation**



After feature extraction, the data is converted into structured formats suitable for model training. Numerical features are stored in tabular form for machine learning algorithms, while audio signals are also transformed into spectrogram images for deep learning models such as CNN.
The dataset is then divided into training and testing sets, typically in an 80:20 ratio. The training set is used to train the model, while the testing set is used to evaluate its performance on unseen data. This separation ensures that the model does not memorize the data and can generalize effectively.

- **Model Training:**

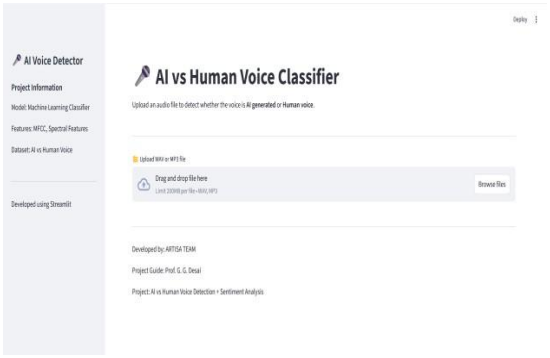
In this stage, classification models are trained using the prepared dataset. Machine learning models such as Random Forest are applied to tabular feature data due to their ability to handle complex feature interactions. Deep learning models such as Convolutional Neural Networks are used for analyzing spectrogram images, as they can capture spatial patterns in frequency-time representations.
During training, the models learn patterns and relationships that differentiate human voices from AI-generated voices. Hyperparameters may be adjusted to improve model performance. In some cases, predictions from multiple models are combined using ensemble techniques to enhance accuracy and stability.

- **Model Testing and Prediction Pipeline :**

Once training is completed, the model is tested using unseen data to evaluate its generalization capability. The testing process follows the same steps as training, including preprocessing and feature extraction.

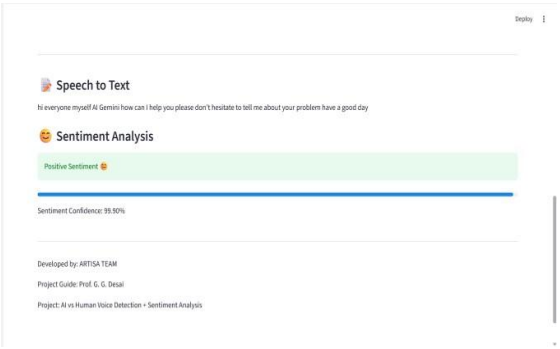
When a new audio input is provided, it is first preprocessed and converted into features. These features are then passed to the trained model, which predicts whether the input is a human voice or AI-generated. This ensures consistency between training and testing processes.

- **Streamlit-Based User Interface :**



To make the system accessible, it is deployed using a Streamlit-based web application. This interface allows users to interact with the model easily without requiring technical knowledge.
The user uploads an audio file through the interface, and the system automatically processes it using the trained pipeline. The interface provides real-time feedback, displaying the classification result along with a confidence score. This makes the system practical for real-world applications.

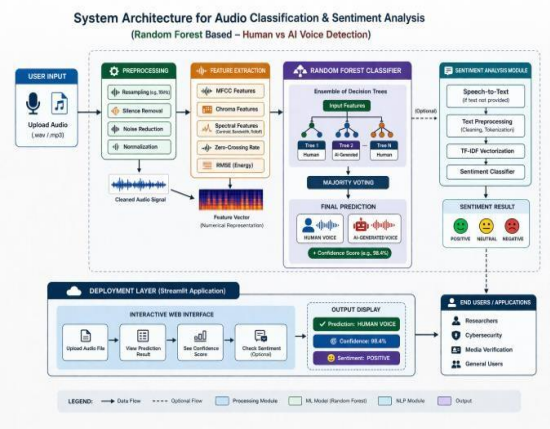
- **Sentiment Analysis Module:**



In addition to voice classification, the system includes a sentiment analysis module to analyze user input text or speech converted to text.
The process begins with text preprocessing, where unwanted characters are removed, and the text is tokenized into meaningful units. Feature extraction is then performed using techniques such as TF-IDF, which converts text into numerical form. A classification model is used to determine the sentiment of the input as positive, negative, or neutral.
This module adds an additional layer of

intelligence to the system by providing contextual understanding of user behavior and emotional tone.

System Architecture



The proposed system architecture is designed as an end-to-end pipeline that processes audio input, performs classification using a Random Forest model, and provides results through a user-friendly Streamlit interface along with sentiment analysis. The architecture consists of multiple interconnected modules, each responsible for a specific task.

Overall System Design

The proposed system is structured as an end-to-end pipeline that processes user-provided audio input and generates classification results along with sentiment analysis. Each module in the architecture is interconnected and performs a specific function, ensuring smooth data flow and accurate prediction.

- **Input Layer:**

The system begins with user input in the form of an audio file uploaded through the Streamlit interface. Supported formats include WAV and MP3, making the system flexible for real-world usage. Before processing, the input file is validated to ensure it meets format and quality requirements. This step prevents errors and ensures that only suitable audio data is passed to the next stage.

- **Preprocessing Module:**

Once the audio is received, it undergoes preprocessing to improve quality and consistency. The audio is first resampled to a fixed sampling rate to maintain uniformity across all samples. Silence present at the beginning and end of the audio is removed to focus only on meaningful speech content. Noise reduction techniques are applied to eliminate background disturbances, and normalization is performed to standardize the amplitude. These operations ensure that the input signal is clean and suitable for accurate feature extraction.

- **Feature Extraction Layer:**

The preprocessed audio is then transformed into numerical representations through feature extraction techniques. Important acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), chroma features, zero-crossing rate, spectral centroid, spectral bandwidth, spectral rolloff, and root mean square energy are extracted. These features capture both time-domain and frequency-domain properties of the audio signal, which are essential for distinguishing between human and AI-generated voices.

- **Feature Representation:**

The extracted features are organized into a structured feature vector that represents each audio sample numerically. This feature vector serves as the input to the classification model. Proper representation of features ensures that the model can effectively learn patterns and relationships within the data, leading to improved prediction accuracy.

- **Random Forest Classification Module:**

The core of the system is the Random Forest classifier, which is an ensemble machine learning algorithm composed of multiple decision trees. Each tree is trained on different subsets of the dataset and independently predicts the class of the input. The final classification is determined using a majority voting mechanism, where the most frequently predicted class among all trees is selected. This approach improves robustness, reduces overfitting, and enhances overall accuracy. The model classifies the audio input as either human voice or AI-generated voice.

- **Prediction and Confidence Layer:**

After classification, the system generates the final prediction along with a confidence score. The confidence score indicates how certain the model is about its decision, providing additional insight into the reliability of the result. This helps users interpret the output more effectively.

- **Sentiment Analysis Module:**

In addition to voice classification, the system incorporates a sentiment analysis module to analyze user input text or speech converted into text. The input text undergoes preprocessing steps such as cleaning, tokenization, and

removal of unnecessary elements. Feature extraction is performed using techniques like TF-IDF, which convert textual data into numerical form. A classification model is then used to determine the sentiment of the input as positive, negative, or neutral. This module adds contextual understanding to the system.

- **Streamlit User Interface:**

The entire system is deployed using a Streamlit-based web application, which provides a simple and interactive interface for users. Through this interface, users can upload audio files, initiate analysis, and view results in real time. The interface ensures ease of use, even for non-technical users, and makes the system accessible for practical applications.

- **Output Layer:**

The final output is displayed through the user interface in a clear and structured format. It includes the voice classification result (human or AI-generated), the confidence score, and the sentiment analysis result. The output presentation is designed to be easily understandable, allowing users to quickly interpret the system's decision.

Result Analysis

The performance of the proposed system is evaluated based on its capability to accurately classify audio samples as either human or AI-generated using the Random Forest classifier, along with the effectiveness of the integrated sentiment analysis module. The model is trained on a diverse dataset consisting of both real and synthetic voice samples, ensuring balanced representation. To validate the robustness of the system, testing is performed on previously unseen data, allowing the evaluation to reflect real-world performance rather than memorized patterns.

During experimentation, the system demonstrated strong classification performance, achieving a high level of accuracy in distinguishing between human and AI-generated voices. This indicates that the selected feature set and model architecture are effective for the given task. The ensemble nature of the Random Forest algorithm, which combines predictions from multiple decision trees, contributed significantly to improving prediction stability and reducing overfitting. As a result, the system was able to maintain consistent performance across different test samples.

A key factor influencing the system's performance is the quality of feature extraction. Acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), chroma features, spectral characteristics, and energy-based measures provided a comprehensive representation of the audio signals. These features helped the model capture subtle variations between natural human speech and AI-generated voices, which are often difficult to distinguish through manual observation. The combination of multiple feature types enhanced the discriminative capability of the model.

The preprocessing stage also played a critical role in improving system performance. Techniques such as noise reduction, silence trimming, and amplitude normalization ensured that the input data was clean and standardized. By removing irrelevant components and minimizing distortions, the system was able to focus on meaningful speech characteristics. This resulted in improved feature consistency and allowed the model to learn more accurate patterns during training.

The system was further evaluated using real-time inputs through the Streamlit-based user interface. Users were able to upload audio files, and the system successfully processed these inputs through the complete pipeline, including preprocessing, feature extraction, and classification. The model generated predictions along with confidence scores, providing users with both the result and an indication of reliability. This real-time evaluation demonstrated the practical applicability of the system in user-facing environments.

In addition to voice classification, the sentiment analysis module was tested on user-provided textual inputs. The module effectively categorized the input into positive, negative, and neutral sentiments, indicating its ability to interpret basic emotional context. The inclusion of sentiment analysis enhances the overall functionality of the system, making it more versatile for applications such as user interaction analysis, feedback evaluation, and fraud detection scenarios where emotional tone can be significant.

Despite achieving promising results, certain limitations were observed during testing. The system's performance tends to decline when the input audio contains excessive background noise, distortions, or poor recording quality. Such conditions can negatively affect feature extraction and lead to incorrect predictions. Additionally, the model may face challenges when dealing with highly advanced or previously unseen AI-generated voices, as these may exhibit patterns not captured during training.

Another limitation is related to the evaluation methodology. The system currently relies primarily on accuracy as the performance

metric, which provides an overall measure of correctness but may not fully represent the model's behavior in all scenarios. Metrics such as precision, recall, and F1-score could provide deeper insights into performance, especially in cases involving imbalanced datasets or edge conditions.

Overall, the experimental results indicate that the proposed system is effective, reliable, and capable of handling real-world audio inputs. The integration of audio classification and sentiment analysis provides a comprehensive solution that can be applied in various domains, including cybersecurity, digital forensics, and intelligent user interaction systems. The system demonstrates strong potential for practical deployment, with scope for further improvements in robustness and evaluation.

Challenges And Limitations

The proposed system faces several challenges related to the quality and variability of input audio data. Since the dataset consists of recordings collected from different sources, there can be significant differences in recording devices, environments, and formats. These variations may introduce inconsistencies that affect the reliability of preprocessing and feature extraction. In real-world scenarios, audio inputs often contain background noise, echoes, or compression artifacts, which can distort important acoustic features such as MFCC and spectral properties. As a result, the system's performance may decrease when handling low-quality or noisy audio samples, limiting its effectiveness in uncontrolled environments.

Another important limitation lies in the model's generalization capability and feature dependency. The Random Forest classifier is trained on a specific dataset, and while it performs well on known data, it may struggle to accurately classify voices generated by newly developed or unseen AI models. As synthetic voice technologies continue to evolve, the system may require frequent updates and retraining to maintain its accuracy. Additionally, the model relies on extracted features such as MFCC, chroma, and spectral attributes, which may not fully capture complex temporal and contextual patterns present in speech. Unlike deep learning models, Random Forest has limited ability to model sequential dependencies, which can restrict its performance in capturing subtle variations in audio signals.

Furthermore, there are limitations related to evaluation methods and additional system components. The system is primarily evaluated using accuracy as the performance metric, which provides a general measure of correctness but may not fully reflect performance under all conditions, especially in edge cases or imbalanced datasets. The sentiment analysis module also has certain constraints, as it may not accurately interpret complex language patterns such as sarcasm, irony, or mixed emotions due to its reliance on basic text processing techniques. In terms of deployment, the system depends on properly formatted audio inputs, and any corrupted or unsupported files may lead to processing errors. Environmental factors, input variability, and system dependencies collectively influence the real-world performance of the proposed solution.

References

A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed., Upper Saddle River: Prentice Hall, 2010, pp. 45–78.

T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Upper Saddle River: Prentice Hall, 2002, pp. 120–156.

D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed., Upper Saddle River: Pearson, 2009, pp. 89–135.

S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, vol. 87, no. 4,

pp. 1738–1752, Apr. 1990.

L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs: Prentice Hall, 1993, pp. 200–245.

F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., New York: Springer, 2009,

pp. 587–604.

L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

Y. LeCun, Y. Bengio, and G. Hinton, "Deep

learning," *Nature*, vol. 521, pp. 436–444, May 2015.

A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

J. Brownlee, *Machine Learning Mastery With Python*, Melbourne: Machine Learning Mastery, 2016, pp. 150–200.

B. McFee et al., "librosa: Audio and music signal analysis in Python," in *Proc. SciPy Conf.*, 2015, pp. 18–25.

F. Chollet, *Deep Learning with Python*, 2nd ed., Shelter Island: Manning, 2021, pp. 210–260.

T. Mikolov et al., "Efficient estimation of word representations in vector space," in *Proc. ICLR*, 2013, pp. 1–12.

J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

A. Graves, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE ICASSP*, 2013, pp. 6645–6649.

K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge: MIT Press, 2016, pp. 321–350.

N. J. Nilsson, *Introduction to Machine Learning*, Stanford: Stanford University Press, 1998, pp. 75–110.