EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Investigation of Imbalanced Sentiment Analysis in Voice Data: A Comparative Study of Machine Learning Algorithms

Viraj Nishchal Shah[1], Deep Rahul Shah[1], Mayank Umesh Shetty[1], Deepa Krishnan[1], Vinayakumar Ravi[3*], and Swapnil Singh[2]

[1]Department of Computer Engineering, Mukesh Patel School of Technology Management and Engineering, SVKM's NMIMS University, Mumbai, Maharashtra, India
[2]Computer Science Department, Virginia Tech, Blacksburg, VA, USA
[3]Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar 34754, Saudi Arabia

Abstract

INTRODUCTION: Language serves as the primary conduit for human expression, extending its reach into various communication mediums like email and text messaging, where emoticons are frequently employed to convey nuanced emotions. In the digital landscape of long-distance communication, the detection and analysis of emotions assume paramount importance. However, this task is inherently challenging due to the subjectivity inherent in emotions, lacking a universal consensus for quantification or categorization.
OBJECTIVES: This research proposes a novel speech recognition model for emotion analysis, leveraging diverse machine learning techniques along with a three-layer feature extraction approach. This research will also through light on the robustness of models on balanced and imbalanced datasets.
METHODS: The proposed three-layered feature extractor uses chroma, MFCC, and Mel method, and passes these features to classifiers like K-Nearest Neighbour, Gradient Boosting, Multi-Layer Perceptron, and Random Forest.
RESULTS: Among the classifiers in the framework, Multi-Layer Perceptron (MLP) emerges as the top-performing model, showcasing remarkable accuracies of 99.64%, 99.43%, and 99.31% in the Balanced TESS Dataset, Imbalanced TESS (Half) Dataset, and Imbalanced TESS (Quarter) Dataset, respectively. K-Nearest Neighbour (KNN) follows closely as the second-best classifier, surpassing MLP's accuracy only in the Imbalanced TESS (Half) Dataset at 99.52%.
CONCLUSION: This research contributes valuable insights into effective emotion recognition through speech, shedding light on the nuances of classification in imbalanced datasets.

Keywords: Audio Feature Extraction, Emotion Detection, Gradient Boosting, K Nearest Neighbours, Multi-Layer Perceptron, Speech Emotion Recognition

Received on 10 January 2024, accepted on 22 April 2024, published on 22 April 2024

doi: 10.4108/eetsis.4805

*Corresponding Author: vravi@pmu.edu.sa

1. Introduction

Understanding the sentiments embedded within an individual's psychological state holds profound significance, fuelling a compelling research domain dedicated to discerning desired emotions. Historically, researchers explored diverse methodologies encompassing physiological signals [1], facial expressions [2], and speech [3] to unravel a speaker's emotional landscape. Notably, speech-based approaches gained prominence owing to the abundance of available data. The surge in voice recognition technology has facilitated the extraction of emotional cues from speech signals, propelling emotion recognition in speech into a rapidly expanding frontier of machine learning research [4]. This trend is not only indicative of technological advancements but also highlights the multifaceted applications of emotion identification in various domains such as human-computer interfaces, clinical psychology, and market research.

Sentiment analysis from voice finds diverse applications across industries. It enhances Human-Computer Interaction (HCI) by allowing voice-activated assistants to understand and respond to user emotions, creating personalized experiences, especially in smart homes. In customer service, real-time sentiment analysis prioritizes high-emotion calls, improving overall satisfaction, while market research benefits from data-driven decisions derived from understanding consumer sentiments in various contexts [5]. Additionally, healthcare, education, entertainment, security, and internal organizational dynamics all leverage voice-based sentiment analysis for purposes ranging from early detection of emotional disorders to enhancing workplace satisfaction and productivity [6].

This study aims to examine the application of machine learning algorithms to identify emotions from speech and evaluate their performance in terms of precision, robustness, and generalizability. In this study, we want to find the impact of various feature extraction approaches, classifiers, and data preparation techniques on the overall performance of an emotion detection system. The findings of this work will lead to the development of more precise and effective emotion identification systems that may be utilized in various fields and aid in improving human-machine interaction. The proposed methodology incorporates a three-layer feature extraction approach, emphasizing a thorough analysis of speech characteristics. This innovative feature extraction process goes beyond traditional methods, enhancing the model's ability to discern subtle nuances in emotional expression within the speech data.

The study evaluates the proposed model's performance across different datasets, including Balanced TESS, Imbalanced TESS (Half), and Imbalanced TESS (Quarter) datasets. This approach accounts for real-world scenarios where imbalanced datasets are common, providing insights into the model's adaptability and robustness across diverse data distributions. By focusing on imbalanced datasets, the study sheds light on the challenges associated with emotion recognition in scenarios where certain emotions are underrepresented. This insight is crucial for refining models and addressing biases, contributing to the development of more inclusive and accurate emotion recognition systems.

Our research confronts the challenge of accurately identifying emotions from voice data amid the complexities of imbalanced datasets. Emotion recognition from voice, while critical in enhancing human-computer interaction, is hindered by subjective interpretations and skewed data representations. This imbalance significantly affects the performance of machine learning algorithms, leading to potential biases and inaccuracies in emotion detection. Our study aims to explore the effectiveness of various machine learning approaches in addressing these issues, with a focus on improving model robustness and reliability in the face of dataset imbalances.

The major contributions of the proposed research can be summarized as follows:

- Adoption of a three-layer feature extraction approach, enhancing the model's ability to discern subtle nuances in emotional expression within speech data.
- Evaluation of the model's performance on Balanced TESS, Imbalanced TESS (Half), and Imbalanced TESS (Quarter) datasets, providing insights into adaptability and robustness across varied data distributions.
- Evaluation of diverse machine learning techniques, including K-Nearest Neighbour, Gradient Boosting, Multi-Layer Perceptron, and Random Forest.

This research work is organized as follows: Section 2 describes the detailed literature review, and Section 3 illustrates the proposed methodology. Section 4 presents our results on balanced and imbalanced datasets. Finally, in Section 5, we have summarised our proposed work with an overview of the future scope of the work.

2. Literature Review

Several new articles and studies have been published in recent years regarding sentiment analysis using speech [7]. This section thoroughly analyses and evaluates the research methodology mentioned in the latest academic materials concerning Speech Emotion Recognition (SER). It was observed that several machine learning algorithms were utilized for classification purposes, mainly feature extraction over multiple datasets covered by the studies. The omission of deep learning techniques such as neural networks was deemed a mistake by the authors. Khalil et al.'s [8] review on deep learning for Speech Emotion Recognition (SER) highlights their exploration of various deep learning systems in the context of emotion detection from speech. The paper briefly touches upon deep neural

networks (DNNs) and convolutional neural networks (CNNs) as part of the broader discussion. Additionally, recurrent neural networks (RNNs) and auto encoders are mentioned in the study, indicating a comprehensive overview of different deep learning architectures commonly employed in SER.

Si, Anjali et al. [9] conducted a study aimed at identifying the most effective machine learning approach for Speech Emotion Recognition (SER). In their research, they comprehensively reviewed and analyzed previous studies in SER spanning the years 2009 to 2018. The analysis encompassed the detailed examination of feature extraction methodologies and machine learning approaches utilized in the reviewed research. While the study acknowledges its limitation in terms of the breadth of analysis, it nevertheless presents itself as a valuable reference for future enhancements in research methodologies within the field of SER.

Basu et al. [10] presented a comprehensive examination of speech emotion datasets, features, and noise reduction methods in the year 2020. Their study delves into the application of various classification techniques, including Support Vector Machine (SVM) and Hidden Markov Model (HMM), for the categorization of speech emotions. The research offers valuable insights into several factors associated with Speech Emotion Recognition (SER). However, its limitation lies in the omission of a thorough evaluation of more recent methodologies, providing only a brief overview of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) as deep learning techniques.

In [11] authors have conducted experiments on the MEMOS database and provide the baseline result for temporal emotional state detection research, with average mean-average-precision (mAP) score as 8.109% on detecting the five emotions (happiness, sadness, anger, surprise, other negative emotions) in videos. It is higher than 5.47% where the emotions are detected by averaging the frame-level confidence scores (obtained by Face++ emotion recognition API) in the segments from a sliding window.

Akcay et al. [12] comprehensively assess speech emotion datasets, their features, as well as existing machine learning models and classifiers actively used in sentiment analysis. The authors have delineated the components constituting a speech emotion recognition system. These systems rely on training data sourced from speech databases, which can be generated through acted, elicited, or natural means. Subsequently, the signals undergo preprocessing to ensure suitability for feature extraction. Speech Emotion Recognition (SER) systems predominantly employ prosodic and spectral features, as they accommodate a broader spectrum of emotions and yield superior outcomes. Enhancements in results can be achieved by incorporating features from other modalities, including those dependent on visual or linguistic aspects.

A.S. Nasim et al. [13] conducted an analysis of speech emotions by utilizing a merged dataset comprising the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS). The study employed diverse classifiers, including Gradient Boosting and MLP, with Gradient Boosting attaining the highest accuracy at 84.96%. The classification considered seven fundamental human emotions, encompassing happiness, anger, sadness, neutrality, fear, disgust, and surprise. Extracting 180 speech features from the audio files using Mel-Frequency Cepstral Coefficient (MFCC), Chroma, and Mel Spectrogram techniques, we applied various traditional classifiers to both the combined dataset and the RAVDESS and TESS datasets independently. Comparative analysis revealed that Gradient Boosting surpassed other classifiers on the combined dataset, achieving an accuracy of 84.96%. Additionally, the MLP classifier demonstrated superior performance across all three datasets compared to other classifiers.

X. Yuan et .al [14] utilized the Berlin Database to train an MLP Model to recognize speech emotion. They used MFCC and openSmile to extract features. In [15] the authors merged the spatial and temporal feature representations using parallel Convolutional Neural Networks and a transformer encoder. Two parallel CNNs capture spatial features, while the Transformer encoder handles temporal features, optimizing filter depth and feature map reduction for a more expressive hierarchical representation at a lower computational cost. Applied to the RAVDESS dataset, the SER model, incorporating spatial and sequential features, achieves an impressive 82.31% accuracy for eight emotions on a hold-out dataset. Further evaluation on the IEMOCAP dataset yielded a recognition accuracy of 79.42% for five emotions.

R. Arya et al. [16] introduced an approach for emotion recognition based on speech, employing diverse machine learning algorithms. Features were extracted using Chroma, Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram Frequency, Mel Tonnetz, and Spectral Contrast. The study conducted a performance comparison among various classification algorithms, including Support Vector Machine (SVM), Recurrent Neural Network (RNN), Multi-Layer Perceptron (MLP), K-Nearest Neighbors (k-NN), Adaboost, Gradient Boosting Classifier, and Random Forest, using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. Notably, MLP outperformed other algorithms, achieving the highest accuracy at 89.5%.

H. Chen et al. [17] systematically explored voice features extracted from the CASIA Emotional Corpus to recognize speech emotion, employing Support Vector Machine, Neural Network, K Nearest Neighbors, and Random Forest. The results demonstrated 55.56% accuracy for K Nearest Neighbors, 58.89% for Random Forest, 80.56% for Neural Network, and 81.11% for Support Vector Machine.

In [18], a novel approach to emotion classification in speech signals by combining Capsule Networks (CapsuleNets) with Time Distributed 2D-Convolution layers is introduced. While CapsuleNets excel at capturing spatial cues, they fall short in considering temporal cues in

time series data like speech. To address this limitation, Time Distributed 2D-Convolution layers are integrated before CapsuleNets to capture both spatial and temporal cues. The proposed architecture is experimented with using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Interactive Emotional Dyadic Motion Capture (IEMOCAP) speech datasets. The log-mel spectrogram of speech samples is extracted for model training and testing. The combined CapsuleNets with Time Distributed 2D-Convolution layers achieve an impressive classification accuracy of 92.6% on the RAVDESS dataset and 93.2% on the IEMOCAP dataset.

While sentiment analysis provides valuable insights in many applications, there are instances where it may fall short, necessitating the need for emotion detection. Emotion detection goes beyond sentiment analysis by precisely determining an individual's emotional or mental state. In [19] authors have given a detailed review of the techniques used in sentiment analysis. It also outlines the challenges involved in emotion detection. The usage of speech-based intelligent applications is one of the potential applications of emotion detection. This is explored in [20] where a study comprising semi structured interviews with fourteen older adults, investigates their interactions with speech-based intelligent personal assistants (sIPAs) and pinpoints usability barriers hindering their widespread adoption.

Another big leap in speech synthesis is done in [21]. This paper introduces a novel deep learning model, leveraging a recurrent neural network (RNN), to address text normalization challenges in speech synthesis. Traditional rule-based models face limitations in handling contextual information and exceptions outside predefined rules, prompting the adoption of a seq2seq neural network framework. The proposed model, employing gated recurrent units (GRU) and a local attention mechanism, demonstrates effectiveness in achieving meaningful results for specific applications, showcasing its superiority over existing attention-based and non-attention models in text normalization for speech synthesis.

Another interesting use case of emotion analysis is done in [22]. This article explores the impact of emotions on team performance by collecting audio recordings and game logs from players during an eSports tournament. Utilizing machine learning models, the study demonstrates a 92.7% accuracy in classifying six common emotional states and background sounds, revealing a strong correlation between team performance, player communication, and emotional sentiment during gameplay, highlighting the significance of internal team conversations for achieving better result. An incremental graph convolution network (I-GCN) for emotion detection in conversations, employing a graph structure to represent semantic correlations and an incremental graph structure to capture temporal changes in dynamic conversations is proposed in [23]. The proposed approach includes utterance-level GCN (U-GCN) and speaker-level GCN (S-GCN) in the initial step, enhancing feature learning for emotion detection. U-GCN emphasizes correlations among utterances using multi-head attention, while S-GCN focuses on speaker-utterance relationships, providing a novel perspective for guiding feature learning in emotion detection from conversations.

While there are many research works that focus on sentiment analaysis from voice, the accuracy, precision and recall measures are quite low in many works. In [24], the author addresses the lack of systematic analysis in sentiment analysis (SA) methods, particularly in the context of multimodal sentiment analysis (MSA). It introduces a novel framework that focuses on individual modalities, providing an extensive review of single-modal SA workflows, recent advancements, and datasets. The article then proposes a new taxonomy for MSA, delving into multimodal representation learning and data fusion, and explores advanced studies on topics like multimodal alignment and the application of Chat Generative Pre-Trained Transformer (ChatGPT) in sentiment analysis, concluding with discussions on open research challenges and potential avenues for improvement in future MSA works.

[25] introduces a machine learning-based model designed to improve resource matching in intelligent education systems, addressing the common challenges of cold starts and data sparsity. By leveraging K-means clustering to calculate similarities between users and resources, the model can predict and match the most suitable educational resources to users with high precision. Test results show recall and coverage rates exceeding 98% and 96%, respectively, highlighting the model's effectiveness in optimizing resource allocation and enhancing the learning experience by accurately catering to individual needs.

Audio along with machine learning is also used to make a training assistant as discussed in [26]. This paper introduces music training assistant system designed to enhance music training by improving input accuracy and response speed through artificial intelligence. The system architecture includes infrastructure, data, application, and presentation layers, integrating ARM and digital signal processors (DSP) for data analysis and human-interface interaction. It employs cepstrum algorithms for feature extraction, linear smoothing for filtering, dynamic time warping for pattern matching, and a radial basis function algorithm for output. The system demonstrates high operability and effectiveness in music assistant training, with signal-to-noise ratios above 14dB, input accuracy over 99.5%, and the capacity to serve 240 users with a response time of only 1 second, showcasing its strong performance in supporting music training.

Employing machine learning and deep learning techniques, [27] successfully identifies antisocial behavior on Twitter. It leverages natural language processing to proactively detect such behavior, using a dataset labeled for this purpose. The study found the Convolutional Neural Network (CNN) and Support Vector Machine (SVM) to be the most effective models, with CNN reaching an accuracy of 99.86%. These findings demonstrate the potential of computational approaches to enhance the safety of online

communities by monitoring and addressing harmful behaviors.

Leveraging a cutting-edge evolutionary algorithm with specialized local search methods, Chuan Wang et al. [28] target the intricate NP-hard problem posed by Sudoku puzzles. The developed algorithm, dubbed LSGA, introduces a matrix coding scheme alongside innovative crossover and mutation techniques, significantly optimizing the search process for solving Sudoku puzzles. In comparative tests with existing algorithms, LSGA outperformed others, demonstrating impressive recall and coverage rates of over 98% and 96%, respectively. These results not only highlight LSGA's superior search efficiency and faster convergence but also its potential for practical applications beyond gaming, particularly in enhancing intelligent education systems through advanced resource matching.

The literature studied shows the diverse use of trivial machine learning algorithms including K Nearest Neighbours, Support Vector Machines, and Multi-Layer Perceptron, along with an emphasis on the importance of deep learning techniques like GRU and RNN. The review studies provide a comprehensive examination of the speech emotion datasets used along with the effective machine learning approaches employed by researchers. This review also highlights innovative approaches like the use of the combination of spatial and temporal features and the MSA framework to address gaps in systematic analysis and proposing a new taxonomy.

The current landscape of sentiment analysis and emotion recognition research presents several notable challenges that hinder a comprehensive understanding of the field. The performance of the classifiers in the case of imbalanced nature of the dataset is not explored by any of the research discussed above. This research significantly contributes to the previously published work and analyses the performance of various machine learning algorithms for sentiment analysis from voice in the context of imbalanced classification.

3. Proposed Methodology

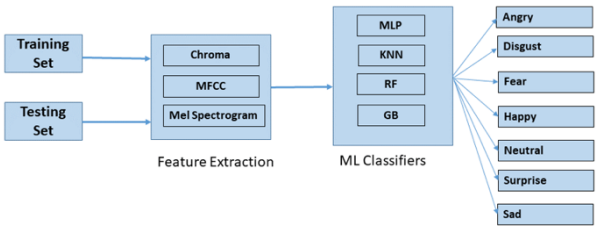

Figure 1. Proposed Methodology for Sentiment Analysis using Voice

Figure 1 illustrates the proposed methodology for sentiment analysis. The Toronto emotional speech dataset consisting of the voice of two actresses for different emotions, was employed for feature extraction. The feature extraction step used the Librosa library of python to extract Chroma features, Mel-frequency cepstral coefficients, and Mel spectrogram. Subsequently, the features were divided into training and testing set keeping a 70-30 train-test ratio. This split was then passed to four distinct machine learning models, namely, Multi-Layer Perceptron, K-Nearest Neighbours, Random Forest, and Gradient Boosting for emotion recognition. These models were selected based on the merits highlighted by various authors in the literature review. Model training and evaluation involved a rigorous process of hyper-parameter tuning and performance metrics assessment. At the end we compared the efficacy of the algorithms for identifying emotions based on the extracted features from the audio data.

3.1. Dataset Used

The Toronto emotional speech dataset [29] used in the research consist of audio samples where two actresses (ages 26 and 64) recited a set of 200 keywords in the predeceased phrase "Say the word" and samples were created of the set to represent seven different moods which were anger, disgust, fear, happiness, neutral, pleasant surprise, and sadness. The total number of stimuli is 28,000 which were divided into a train-test split of 70:30. The train data is fed directly to the feature extraction method for pre-processing into a data frame.

3.2. Feature Extraction

We have used Librosa library of python to extract features from the wave files. The feature extraction process has been divided into 3 parts, namely, Chroma feature extraction, Mel-frequency cepstral coefficients and Mel spectrogram.

(i) Chroma Features: These features are extracted from the short time fourier transform (STFT) of the audio sample extracting frequency-based data over time. The STFT data is assigned to the 12 pitch classes and an aggregate of each of the classes is computed for each audio frame. This data is averaged and added to the analysis dataframe. Mohammed Jawad et.al [30] have used a similar fourier transform to represent a speech signal (n) x divided into L frames using the equation (1):

$$X(n) = \sum_{K=1}^{M} \left(H_K^1(n) \left(cos\left(2\pi \frac{f_N^2}{Fs} n \right) \right) + \varphi_k^1 \right) \quad (1)$$

(ii) Mel-frequency cepstral coefficients (MFCCs): These features deal with the spectral shape of the audio signal. These coefficients are computed by transforming the Mel-Spectrogram over the time

frame of the audio sample. Ali Bou et.al [31] has concluded during survey that MFCC features result in better results with the deep neural netowrk than the traditional practices creating a suitable feature for extraction.

(iii) Mel Spectrogram: Mel spectrogram is logarithmically scaled version of the spectrogram. The Mel-Spectrogram has frequency bins placed on the frequency vs pitch scale, corresponding to human auditory perceptions providing and intuitive representation of the spectral shape of the audio signal.

Figure 2 illustrates the waveform and spectrogram for sad and neutral emotions. It can be observed that the waveform for sad and neutral are distinct.

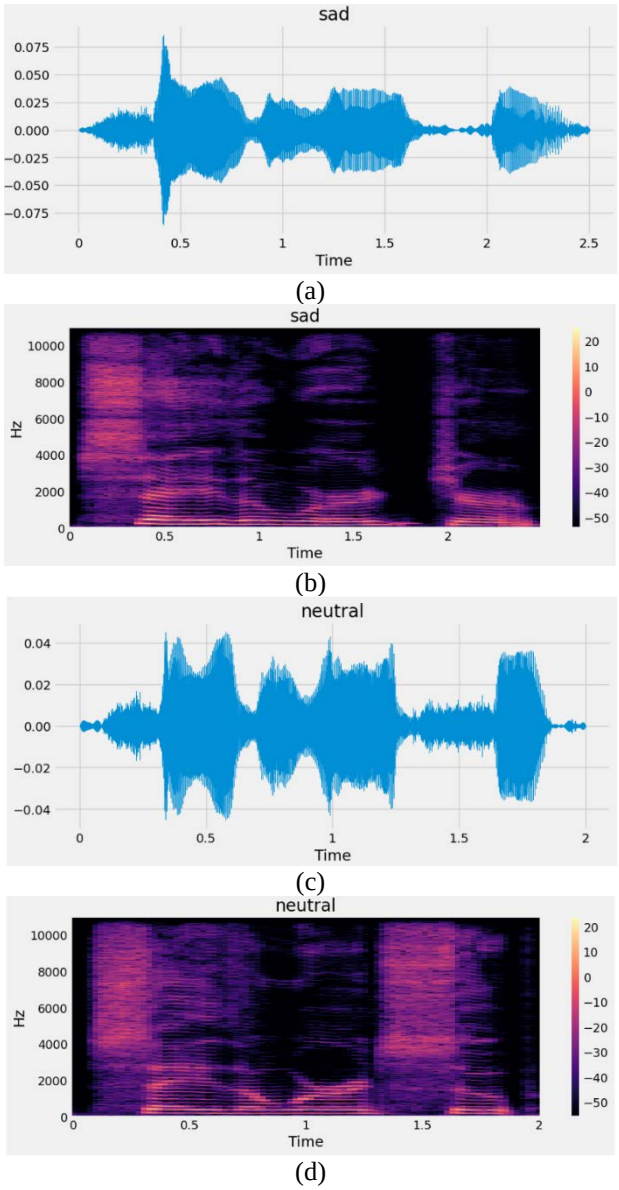

Figure 2. (a) waveform for sad emotion (b) spectrogram for sad emotion (c) waveform for neutral emotion (d) spectrogram for neutral emotion

3.3. Algorithms Used

To get the most accurate findings, we as researchers in the area of emotion recognition from audio are always on the lookout for new and improved algorithms. Multi-layer Perceptron (MLP), K-Nearest Neighbours (KNN), Random Forest, and Gradient Boosting are just few of the powerful and flexible algorithms we've used for this purpose. We want to improve our odds of generating reliable and accurate answers by combining the benefits of many algorithms, each of which has its own strengths and shortcomings.

(i) Multi-Layer Perceptron (MLP): The acronym MLP refers to the feedforward artificial neural network, Multi-Layer Perceptron. MLP is a popular deep learning technique used in several fields, of audio analysis, and Natural Language Processing (NLP). A MLP has many artificial neurons (perceptron) buried in its various levels. In an MLP, the input is sent through a series of hidden layers, each of which performs some kind of non-linear activation function on the data before passing it on. An MLP's last layer, the output layer, makes a final prediction or call based on the information it has received. Ferras et.al [32] have used the following formula for an input vector v and the factors x which are written by equation (2)

$$z = f_0(x) + \epsilon \qquad (2)$$

The hyperparameters of the MLP were configures by setting regularization parameter as 0.01, batch size as 256, with an adaptive learning rate, and epochs to 600. The model has an input layer, a hidden layer with 320 neurons and an output layer with 7 neurons.

(ii) K-Nearest Neighbour (KNN): One easy and uncomplicated machine learning approach that may be used for emotion identification in audio is called k-Nearest Neighbours (KNN). Here, KNN is useful for identifying the underlying emotions in audio files and categorising them accordingly. KNN for emotion detection works on the premise that characteristics relevant to the emotion classification task may be extracted from the audio signals. Some examples of such characteristics include spectral features, prosodic character traits, and rhythmic attributes. A feature vector, which is used to represent the audio signal, is then computed using the features. As discussed by Taunk et.al [33], the Euclidian distance between the training dataset is calculated using the equation (3).

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (3)$$

KNN was trained with 5 neighbours and Euclidean distance as distance matric.

(iii) Random Forest (RF): By generating several separate decision trees and then averaging their results, Random Forest is an ensemble learning approach for classification and regression. It's a method used in machine learning and has many practical uses, such

as in image recognition, audio processing, and NLP. Several decision trees in a Random Forest model are educated using shuffled samples of the whole training dataset. Predictions are produced for an input by each tree in the forest, and the final forecast is based on voting (in classification problems) or averaging all the trees' predictions (in regression problems). The prediction time for Random Forest scales linearly with the number of trees in the forest, which means it may be too costly for use in real-time applications. Since Random Forest tends to favour features with many possible values, the data may need to be pre-processed before the method can be used. Thambi et.al [34] have also discussed how random forest has received the maximum accuracy on creating a speech/non-speech detection machine, an inference of our current problem statement, and recommend the use of random forest as it effectively estimates missing data and continues to maintain accuracy even if a part of data is missing.

Random forest used in this experiment was trained with 100 estimators.

(iv) Gradient Boosting (GB): Gradient Boosting, is used for both regression and classification tasks. One of the key components of audio-based emotion detection is predicting a speaker's emotional state using elements derived from an audio stream. Each successive decision tree in Gradient Boosting is taught to improve upon its predecessor by attempting to fix the shortcomings of its predecessor's training. An initial basic model is fitted to the training data, and then the method repeatedly improves the model by fitting new trees to the residuals (the comparison between the real label and the predicted label). The ultimate forecast is arrived at by summing the forecasts of all preceding trees.

In our experiment we used Gradient Boosting classifier with a learning rate of 0.1 and 100 estimators.

4. Results and Analysis

The experiments are performed on a desktop with AMD Ryzen 5 5600X CPU and 16GB RAM. Python 3.9 serves as the development environment, comprising packages such as Librosa for key feature extraction and analysis, Scikit-learn 1.0.2 for implementing machine learning models and the computation of evaluation metrics, and additional supplementary libraries such as OS, Math, Matplotlib, NumPy, and Pandas. The evaluation metrics for the proposed framework include the Accuracy, Precision, Recall, and F1-score for each classifier.

The classifiers are evaluated over the TESS dataset (balanced and imbalanced). The proposed model classifies the audio recording recited by two actors into one of the seven classes (angry, disgust, fear, happy, neutral, surprised, and sad). The balanced dataset consists of 200 records evenly distributed per class per actor, totalling 2800 recordings. In contrast, the imbalanced dataset in this experiment is created by unevenly halving the records of the balanced dataset, thereby obtaining half and a quarter of the records found in the balanced dataset for certain classes. Table I shows the class-wise dataset record split in detail.

Tables II, III, and IV highlight the four metrics used for assessing the performance of the proposed framework over the balanced and imbalanced datasets, namely the Accuracy, Precision, Recall, and F1-score for Multi-Layer Perceptron (MLP), K-Nearest Neighbour (KNN), Random Forest, and Gradient Boosting. A comparison of all four classifiers is also provided in Figure 3.

Table 1. Class-wise samples in dataset

Emotion Class	Toronto Emotion Speech Set (TESS) Dataset		
	Original	Half-Class	Quarter-Class
ANGRY	400	400	100
DISGUST	400	400	100
FEAR	400	200	400
HAPPY	400	200	400
NEUTRAL	400	200	100
SURPRISED	400	400	100
SAD	400	200	400

Table 2. Results for Balanced TESS Dataset

Classifier	Accuracy	Precision	Recall	F1-Score
MLP	0.9964	0.9964	0.9964	0.9964
KNN	0.9957	0.9957	0.9957	0.9957
RF	0.9957	0.9957	0.9957	0.9957
GB	0.9921	0.9921	0.9921	0.9921

Table 3. Results for Imbalanced TESS Dataset (Half Class Records)

Classifier	Accuracy	Precision	Recall	F1-Score
MLP	0.9943	0.9943	0.9943	0.9943
KNN	0.9952	0.9953	0.9952	0.9952
RF	0.9923	0.9925	0.9924	0.9924
GB	0.9838	0.9841	0.9838	0.9838

Table 4. Results for Imbalanced TESS Dataset (Quarter Class Records)

Classifier	Accuracy	Precision	Recall	F1-Score
MLP	0.9931	0.9931	0.9931	0.9931
KNN	0.9920	0.9920	0.9920	0.9920
RF	0.9783	0.9790	0.9783	0.9779
GB	0.9748	0.9748	0.9749	0.9743

From the table results, it is evident that MLP is the best-performing classifier among the other classifiers in the proposed framework, with accuracies of 99.64% in the Balanced TESS Dataset, 99.43% in the Imbalanced TESS (Half) Dataset and 99.31% in the Imbalanced TESS (Quarter) Dataset. KNN is the second-best classifier, surpassing MLP (99.43%) in accuracy only in the Imbalanced TESS (Half) Dataset (99.52%).Furthermore, the accuracies of all classifiers gradually drop as the dataset records are reduced to half and then one-fourth of the original size. RF and GB exhibit the most pronounced accuracy declines compared to the other classifiers in the proposed framework. The accuracy of RF decreases from 99.57 percent to 99.23 percent and subsequently to 97.8 percent. GB is the least accurate classifier in the proposed framework, with accuracy decreasing from 99.21% to 98.38% and subsequently to 97.48%.

Therefore, the findings indicate that MLP and KNN are the most accurate classifiers for predicting an individual's emotion based on their speech. This suggested framework also includes other classifiers that may be deemed to fulfil the same purpose optimally, compared to classifiers that are not a part of this proposed framework. The high accuracy across all scenarios suggests that the classifiers perform well in correctly classifying instances. However, when dealing with imbalanced datasets, especially with a quarter of class records, the models' performance is noticeably affected. Gradient Boosting shows sensitivity to imbalanced datasets, experiencing a more pronounced decrease in metrics. Precision, Recall, and F1-Score remain consistent across balanced and moderately imbalanced datasets, indicating the classifiers' ability to maintain a balance between correctly identifying positive instances and minimizing false positives and negatives. However, as the imbalance increases, there is a noticeable decline in these metrics, especially for Gradient Boosting which is evident in Figure 3.

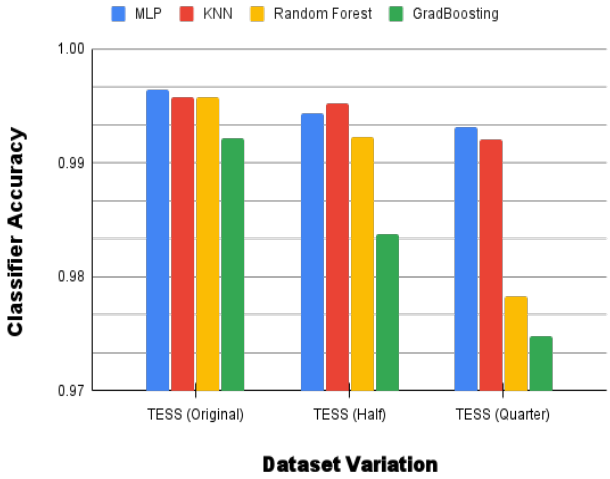

Figure 3. Comparison of all four classifiers in Balanced and Imbalanced TESS datasets

4.1. Confusion Matrices for Balanced Dataset

The Confusion matrices given in Figure 4 shows the misclassifications when MLP, K Nearest Neighbour, Random Forest and Gradient Boosting is used with the balanced dataset. Notably, both KNN and MLP exhibited impeccable accuracy, achieving zero misclassifications across the emotional spectrum, including categories such as angry, disgust, fear, neutral, and sad. This outcome underscores the efficacy of KNN's proximity-based approach and MLP's capacity to discern intricate non-linear patterns within the feature space. However, the ensemble models, Random Forest, and Gradient Boosting, displayed misclassifications in the realms of disgust, happy, and surprise, indicative of potential challenges in capturing nuanced decision boundaries or susceptibility to noise.

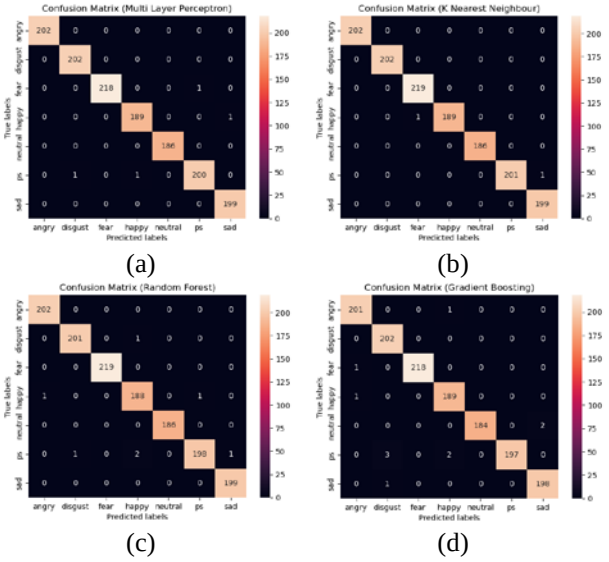

Figure 4. (a) Confusion Matrix for MLP for balanced TESS dataset (b) Confusion Matrix for KNN for balanced TESS dataset (c) Confusion Matrix for RF for balanced TESS dataset (d) Confusion Matrix for GB for balanced TESS dataset

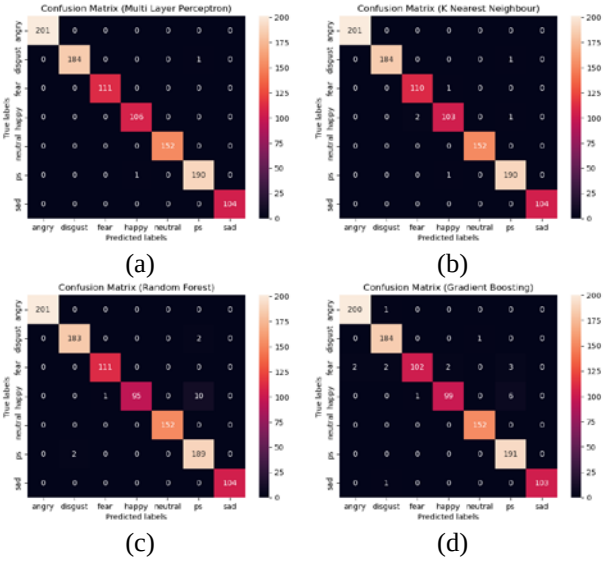

Figure 5. (a) Confusion Matrix for MLP for imbalanced TESS dataset - half (b) Confusion Matrix for KNN for imbalanced TESS dataset - half (c) Confusion Matrix for RF for imbalanced TESS dataset - half (d) Confusion Matrix for GB for imbalanced TESS dataset - half

4.2. Confusion Matrices for Imbalanced Dataset - Half

The experimental repetition with reduced samples of fear, happy, neutral, and sad, as illustrated in Figure 5, has yielded noteworthy results with respect to the performance of classification algorithms, particularly MLP (Multi-Layer Perceptron) and KNN (K-Nearest Neighbours). The reduced misclassifications across all classes, except for a single misclassification in the surprise class by MLP, underscore the resilience of these models to variations in sample sizes. MLP's continued accuracy suggests its ability to generalize well even under reduced data, emphasizing its robustness. Similarly, KNN's consistent performance attests to its capacity to maintain proximity-based distinctions effectively. Notably, Random Forest and Gradient Boosting algorithms, while still exhibiting strong performance, displayed a few misclassifications, particularly between the happy and surprise classes. These misclassifications may be attributed to the inherent complexity of decision boundaries created by ensemble models and potential sensitivity to the reduced sample sizes.

4.3. Confusion Matrices for Imbalanced Dataset - Quarter

In the course of the iterative experimental phase marked by a quarter reduction in samples for fear, happy, neutral, and sad, as illustrated in Figure 6, the performance of classification algorithms, notably MLP (Multi-Layer Perceptron) and KNN (K-Nearest Neighbors), persists in demonstrating robustness as seen in Figure 5.The absence of misclassifications across all classes for MLP, except for instances involving the happy and surprise categories, underscores the model's ability to uphold elevated accuracy levels even amidst substantially diminished data. In parallel, the sustained outstanding performance of KNN suggests a commendable resilience to diminishing sample sizes, accentuating its suitability for scenarios characterized by constrained training instances. Conversely, Random Forest and Gradient Boosting, while retaining competitiveness, exhibited some misclassifications, particularly within the angry, disgust, and surprise classes. These misclassifications likely emanate from the heightened complexity associated with capturing nuanced decision boundaries and patterns within the dataset when confronted with severely reduced samples. The observed competitive scores illuminate the adaptability of these models to the diminished dataset, yet concurrently emphasize the critical necessity for meticulous considerations when contending with limited training examples.

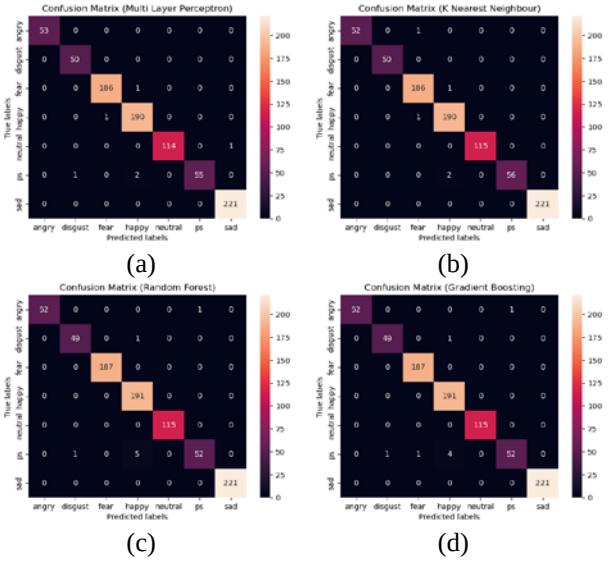

Figure 6. (a) Confusion Matrix for MLP for imbalanced TESS dataset - quarter (b) Confusion Matrix for KNN for imbalanced TESS dataset - quarter (c) Confusion Matrix for RF for imbalanced TESS dataset - quarter (d) Confusion Matrix for GB for imbalanced TESS dataset – quarter

5. Discussion

Based on the results discussed above we can see that MLP or Multi-Layer Perceptron is the best performance model for all the experiments. MLP has performed better than other models based on its inherent ability to learn complex relationships withing the data due to its deep architecture and non-linear activation functions. The performance of MLP is also influenced by the quantity and quality of training data, pre-processing techniques, and the tuning of hyperparameter. We can say that the size and quality of our training set, the feature extraction and pre-processing steps and the hyperparameters tuned have helped in the better performance of MLP. By delving deeper into these aspects, we aim to elucidate the underlying mechanisms driving MLP's superior performance and provide valuable insights into optimizing model design and training strategies for enhanced emotion recognition from audio data across different contexts and datasets. We find these facts to support our model performance based on our experiments, the literature review we undertook, and the review undertaken in [35].

Furthermore, an emphasis needs to be given on the significance of model interpretability in the domain of sentiment analysis, particularly for applications that require a high degree of trust and credibility. In our study, while leveraging machine learning algorithms such as Gradient Boosting, Random Forest, and Multi-Layer Perceptron (MLP), we recognize that these models vary in their inherent interpretability. Specifically, ensemble methods like Random Forest and Gradient Boosting provide insights into decision-making through feature importance scores, highlighting which voice data features (e.g., pitch, tone) significantly influence emotional predictions. Although MLPs are complex and perceived as less interpretable, the transparency in our methodology, from feature extraction to model evaluation, aims to mitigate this by clarifying how voice data are transformed into emotional insights.

Driving deeper, to enhance the interpretability of our findings, we delve into the analysis of feature importance across models, offering a glimpse into the predictive power of various audio features for different emotions. This approach not only aids in understanding how models discern emotions from voice data but also reinforces the credibility of our application by elucidating the judgment basis of our models. Acknowledging the critical role of interpretability, we are committed to exploring more transparent and interpretable modeling techniques in our future work. This endeavor will focus on further elucidating the connection between voice data features and emotional states, thereby advancing the field of sentiment analysis in a direction that fosters user trust and model accountability.

We emphasize the practical significance of our research by defining specific application scenarios where our findings can be employed to enhance emotional intelligence in machine-human interactions. For instance, our models' ability to accurately recognize emotions from voice data holds profound implications for improving customer service in call centers, where understanding a caller's emotional state can enable more empathetic and efficient responses. Additionally, our research can be pivotal in developing more intuitive and responsive voice-activated assistants for smart homes and personal devices, creating interactions that feel more natural and understanding to the user. In the realm of healthcare, our findings could support mental health monitoring by detecting emotional distress signals in patients' speech, facilitating timely intervention. By outlining these scenarios, we aim to bridge the gap between theoretical research and tangible benefits, showcasing how advancements in sentiment analysis from voice data can contribute to societal well-being and technological advancement.

6. Conclusion

In summary, this research aims to assess the efficacy of diverse machine learning algorithms in the domain of speech-based emotion detection. The outcomes revealed that the MLP (Multi-Layer Perceptron) and KNN (K-Nearest Neighbours) algorithms excelled in terms of F1 Score, accuracy, recall, and precision when applied to the TESS dataset. These findings suggest the practical viability of employing MLP and KNN for effective emotion recognition in speech-related applications. Nevertheless, it is imperative to conduct additional research to enhance the robustness and generalizability of these algorithms, extending the assessment to encompass a more extensive

array of datasets. In our forthcoming investigations, we aspire to explore diverse applications of speech emotion recognition, such as mood disorder detection for psychological conditions or emotion recognition within call center systems. Despite this, the present study contributes significant insights into the utilization of machine learning methodologies for emotion detection from speech, illuminating the prospects for continued advancements in this burgeoning field.

References

[1] Ragot M, Martin N, Em S, Pallamin N, Diverrez JM. Emotion recognition using physiological signals: laboratory vs. wearable sensors. InAdvances in Human Factors in Wearable Technologies and Game Design: Proceedings of the AHFE 2017 International Conference on Advances in Human Factors and Wearable Technologies, July 17-21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8 2018 (pp. 15-22). Springer International Publishing.

[2] Ali H, Hariharan M, Yaacob S, Adom AH. Facial emotion recognition using empirical mode decomposition. Expert Systems with Applications. 2015 Feb 15;42(3):1261-77.

[3] Liu ZT, Wu M, Cao WH, Mao JW, Xu JP, Tan GZ. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing. 2018 Jan 17;273:271-80.

[4] Acheampong FA, Wenyu C, Nunoo-Mensah H. Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports. 2020 Jul;2(7):e12189.

[5] Sahoo C, Wankhade M, Singh BK. Sentiment analysis using deep learning techniques: a comprehensive review. International Journal of Multimedia Information Retrieval. 2023 Dec;12(2):41.

[6] Atmaja BT, Sasou A. Sentiment analysis and emotion recognition from speech using universal speech representations. Sensors. 2022 Aug 24;22(17):6369.

[7] Abbaschian BJ, Sierra-Sosa D, Elmaghraby A. Deep learning techniques for speech emotion recognition, from databases to models. Sensors. 2021 Feb 10;21(4):1249.

[8] Khalil RA, Jones E, Babar MI, Jan T, Zafar MH, Alhussain T. Speech emotion recognition using deep learning techniques: A review. IEEE access. 2019 Aug 19;7:117327-45.

[9] Tripathi A, Singh U, Bansal G, Gupta R, Singh AK. A review on emotion detection and classification using speech. InProceedings of the international conference on innovative computing & communications (ICICC) 2020 May 15.

[10] Basu S, Chakraborty J, Bag A, Aftabuddin M. A review on emotion recognition using speech. In2017 International conference on inventive communication and computational technologies (ICICCT) 2017 Mar 10 (pp. 109-114). IEEE.

[11] Li Y, Xia X, Jiang D, Sahli H, Jain R. MEMOS: A Multi-modal Emotion Stream Database for Temporal Spontaneous Emotional State Detection. InCompanion Publication of the 2020 International Conference on Multimodal Interaction 2020 Oct 25 (pp. 370-378).

[12] Akçay MB, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication. 2020 Jan 1;116:56-76.

[13] Nasim AS, Chowdory RH, Dey A, Das A. Recognizing Speech Emotion Based on Acoustic Features Using Machine Learning. In2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS) 2021 Oct 23 (pp. 1-7). IEEE.

[14] Yuan X, Wong WP, Lam CT. Speech emotion recognition using multi-layer perceptron classifier. In2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN) 2022 Aug 23 (pp. 644-648). IEEE.

[15] Ullah R, Asif M, Shah WA, Anjam F, Ullah I, Khurshaid T, Wuttisittikulkij L, Shah S, Ali SM, Alibakhshikenari M. Speech emotion recognition using convolution neural networks and multi-head convolutional transformer. Sensors. 2023 Jul 7;23(13):6212.

[16] Arya R, Pandey D, Kalia A, Zachariah BJ, Sandhu I, Abrol D. Speech based emotion recognition using machine learning. In2021 IEEE Mysore Sub Section International Conference (MysuruCon) 2021 Oct 24 (pp. 613-617). IEEE.

[17] Chen H, Liu Z, Kang X, Nishide S, Ren F. Investigating voice features for Speech emotion recognition based on four kinds of machine learning methods. In2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS) 2019 Dec 19 (pp. 195-199). IEEE.

[18] Yalamanchili B, Anne KR, Samayamantula SK. Speech emotion recognition using time distributed 2D-Convolution layers for CAPSULENETS. Multimedia Tools and Applications. 2022 May;81(12):16945-66.

[19] Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Social network analysis and mining. 2021 Dec;11(1):81.

[20] Islam MU, Chaudhry BM. A framework to enhance user experience of older adults with speech-based intelligent personal assistants. IEEE Access. 2022 Dec 22;11:16683-99.

[21] Huang L, Zhuang S, Wang K. A text normalization method for speech synthesis based on local attention mechanism. IEEE Access. 2020 Feb 18;8:36202-9.

[22] Abramov S, Korotin A, Somov A, Burnaev E, Stepanov A, Nikolaev D, Titova MA. Analysis of video game players' emotions and team performance: An esports tournament case study. IEEE Journal of Biomedical and Health Informatics. 2021 Oct 11;26(8):3597-606.

[23] Nie W, Chang R, Ren M, Su Y, Liu A. I-GCN: Incremental graph convolution network for conversation emotion detection. IEEE Transactions on Multimedia. 2021 Oct 8;24:4471-81.

[24] Lu Q, Sun X, Long Y, Gao Z, Feng J, Sun T. Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. IEEE Transactions on Neural Networks and Learning Systems. 2023 Jul 21.

[25] Xiang CZ, Fu NX, Gadekallu TR. Design of resource matching model of intelligent education system based on machine learning. EAI Endorsed Transactions on Scalable Information Systems. 2022 Feb 10;9(6):e1-.

[26] Zhihan H, Yuan L, Jin T. Design of music training assistant system based on artificial intelligence. EAI Endorsed Transactions on Scalable Information Systems. 2022 Feb 11;9(6):e2-.

[27] Singh R, Subramani S, Du J, Zhang Y, Wang H, Miao Y, Ahmed K. Antisocial Behavior Identification from Twitter Feeds Using Traditional Machine Learning Algorithms and Deep Learning. EAI Endorsed Transactions on Scalable Information Systems. 2023 May 12;10(4).

[28] Wang C, Sun B, Du KJ, Li JY, Zhan ZH, Jeon SW, Wang H, Zhang J. A novel evolutionary algorithm with column

and sub-block local search for sudoku puzzles. IEEE Transactions on Games. 2023 Jan 12.
[29] Pichora-Fuller MK, Dupuis K. Toronto emotional speech set (TESS). Scholars Portal Dataverse. 2020 Feb 13;1:2020.
[30] Al Dujaili MJ, Ebrahimi-Moghadam A, Fatlawi A. Speech emotion recognition based on SVM and KNN classifications fusion. International Journal of Electrical and Computer Engineering. 2021 Apr 1;11(2):1259.
[31] Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. Speech recognition using deep neural networks: A systematic review. IEEE access. 2019 Feb 1;7:19143-65.
[32] Ferras M, Bourlard H. MLP-based factor analysis for tandem speech recognition. In2013 IEEE International Conference on Acoustics, Speech and Signal Processing 2013 May 26 (pp. 6719-6723). IEEE.
[33] Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. In2019 international conference on intelligent computing and control systems (ICCS) 2019 May 15 (pp. 1255-1260). IEEE.
[34] Thambi SV, Sreekumar KT, Kumar CS, Raj PR. Random forest algorithm for improving the performance of speech/non-speech detection. In2014 First International Conference on Computational Systems and Communications (ICCSC) 2014 Dec 17 (pp. 28-32). IEEE.
[35] SÖNMEZ YÜ, VAROL A. In-depth investigation of speech emotion recognition studies from past to present The importance of emotion recognition from speech signal for AI. Intelligent Systems with Applications. 2024 Mar 11:200351.