Voice Disorder Classification Using Wav2vec 2.0 Feature Extraction

[*,a]Jie Cai, [*,†,a]Yuliang Song, [*,a]Jianghao Wu, and [*]Xiong Chen, *Wuhan, China, and* †*Vandoeuvre-les-Nancy, France*

Summary: Objectives. The study aims to classify normal and pathological voices by leveraging the wav2vec 2.0 model as a feature extraction method in conjunction with machine learning classifiers.

Methods. Voice recordings were sourced from the publicly accessible VOICED database. The data underwent preprocessing, including normalization and data augmentation, before being input into the wav2vec 2.0 model for feature extraction. The extracted features were then used to train four machine learning models—Support Vector Machine (SVM), K-Nearest Neighbors, Decision Tree (DT), and Random Forest (RF)—which were evaluated using Stratified K-Fold cross-validation. Performance metrics such as accuracy, precision, recall, F1-score, macro average, micro average, receiver-operating characteristic (ROC) curve, and confusion matrix were utilized to assess model performance.

Results. The RF model achieved the highest accuracy (0.98 ± 0.02), alongside strong recall (0.97 ± 0.04), F1-score (0.95 ± 0.05), and consistently high area under the curve (AUC) values approaching 1.00, indicating superior classification performance. The DT model also demonstrated excellent performance, particularly in precision (0.97 ± 0.02) and F1-score (0.96 ± 0.02), with AUC values ranging from 0.86 to 1.00. Macro-averaged and micro-averaged analyses showed that the DT model provided the most balanced and consistent performance across all classes, while RF model exhibited robust performance across multiple metrics. Additionally, data augmentation significantly enhanced the performance of all models, with marked improvements in accuracy, recall, F1-score, and AUC values, especially notable in the RF and DT models. ROC curve analysis further confirms the consistency and reliability of the RF and SVM models across different folds, while confusion matrix analysis revealed that RF and SVM models had the fewest misclassifications in distinguishing "Normal" and "Pathological" samples. Consequently, RF and DT models emerged as the most robust performers, making them particularly well-suited for the voice classification task in this study.

Conclusions. The method of wav2vec 2.0 combining machine learning models proved highly effective in classifying normal and pathological voices, achieving exceptional accuracy and robustness across various machine evaluation metrics.

Key Words: Voice disorder–Wav2vec 2.0–Machine learning.

INTRODUCTION

Voice disorders encompass a diverse array of conditions that affect the vocal folds, leading to alterations in pitch, volume, or overall voice quality.[1] These disorders can arise from various etiologies, including structural abnormalities, neurological conditions, vocal misuse, or environmental factors, presenting significant challenges to individuals across different age groups and professions.[2,3] Such challenges can impair daily activities, social interactions, and professional endeavors.[4,5] Moreover, untreated voice disorders can lead to serious long-term complications, such as vocal fold nodules, polyps, or even laryngeal cancer.[6,7] This highlights the critical importance of early detection and intervention in managing voice disorders effectively.

Current assessments of voice quality predominantly rely on subjective perceptual evaluations and objective acoustic analysis.[8] Subjective evaluations, which involve clinicians' judgments of a patient's voice, are often imprecise due to inherent variability and potential biases.[8,9] On the other hand, objective assessments, including acoustic analysis and stroboscopy, require a controlled environment and detailed evaluation by skilled operators.[10] These stringent requirements make the assessment process time-consuming and resource-intensive, posing substantial challenges for implementation in general health care settings.

In recent years, machine learning methods have emerged as powerful tools in medical diagnostics, including the classification and diagnosis of voice disorders.[11,12] These techniques utilize advanced computational methods to analyze large datasets of voice recordings, extract meaningful features, and accurately predict the presence and severity of voice abnormalities.[12,13] By leveraging machine learning, clinicians can potentially improve the early detection, diagnosis, and management of voice disorders.[14,15]

Traditional machine learning approaches for voice classification commonly have typically relied on features

Accepted for publication September 3, 2024.

This study was supported by the National Natural Science Foundation of China (82071033).

From the *Department of Otorhinolaryngology, Head and Neck Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China; and the †Department of Ear, Nose, and Throat–Head and Neck Surgery, University Hospital of Nancy, Hospital of Brabois, Vandoeuvre-les-Nancy, France

[a] Jie Cai, Yuliang Song, and Jianghao Wu contributed equally to the work.

Address correspondence and reprint requests to Xiong Chen, Department of Otorhinolaryngology, Head and Neck Surgery, Zhongnan Hospital of Wuhan University, #169, East Lake Road, Wuchang District, Wuhan 430000, China. E-mail: zn_chenxiong@whu.edu.cn

Journal of Voice, Vol xx, No xx, pp. xxx–xxx

0892-1997

https://doi.org/10.1016/j.jvoice.2024.09.002

such as mel-frequency cepstral coefficients (MFCC)[16] and cepstral peak prominence.[17] These features have shown promise in previous studies for the classification of voice disorders across various datasets. For instance, Chen and Chen utilized a Deep Neural Network (DNN) to extract 12 MFCC features from each voice sample in the VOICED dataset, achieving impressive results with high sensitivity (97.8%), specificity (99.4%), precision (99.4%), accuracy (98.6%), and F1-scores (98.4%).[18] Similarly, Verde et al compared the performance of several machine learning techniques for voice pathology detection using the Saarbruecken Voice Database. They employed parameters such as fundamental frequency (F0), jitter, shimmer, and harmonic-to-noise ratio (HNR), finding that Support Vector Machine (SVM) and Decision Tree (DT) algorithms yielded the highest accuracy depending on selected features.[19]

Despite these researches, traditional features like F0, jitter, shimmer, and HNR capture only specific aspects of speech signals, resulting in low-dimensional features with limited representational capacity.[17] While MFCC are widely adopted in speech processing tasks, their extraction process is intricate and often falls short in capturing nonlinear relationships, long-term dependencies, and in effectively handling noise.[16] Furthermore, the evolving landscape of voice signals classification has seen the emergence of new methods. For example, Fonseca et al utilized a discriminative paraconsistent machine for an acoustic investigation of speech pathologies, achieving an accuracy of 95%.[20] Contreras et al proposed a novel framework based on multiple projections of cepstral coefficients to enhance the detection of dysphonic alterations through machine learning techniques.[21] Although these methods demonstrate good performance, they involve relatively complex processing and remain relatively underexplored in the research.

Wav2vec 2.0, a model architecture designed for self-supervised learning of speech representations, offers a promising solution to many of these challenges. By leveraging raw audio data to learn meaningful features without the need for extensive labeled datasets, wav2vec 2.0 significantly reduces the dependency on annotated data.[22,23] The model employs a contrastive loss function to better manage variability in speech, achieving state-of-the-art results on various speech recognition benchmarks.[23] Additionally, the use of a transformer encoder within wav2vec 2.0 enables the model to capture long-distance dependencies in the input sequence, which facilitates the processing of speech signals and the generation of high-quality feature representations suitable for diverse speech processing tasks.[23]

Wav2vec 2.0 has demonstrated exceptional performance in numerous speech recognition tasks.[23] Wang et al fine-tuned wav2vec/Hubert benchmarks for tasks such as speech emotion recognition, speaker verification, and spoken language understanding, showcasing its versatility.[24] Klempíř et al evaluated wav2vec embeddings in detecting Parkinson's disease, underscoring the model's potential generalizability and effectiveness across different databases.[25] Despite its impressive results in speech recognition, the application of wav2vec 2.0 in classifying normal and pathological voices remains relatively unexplored.

This study seeks to address this gap by employing wav2vec 2.0 as a feature extraction method for the classification of normal and pathological voices, in conjunction with four machine learning models. The objective is to demonstrate the efficacy of wav2vec 2.0 in achieving high performance in supervised classification tasks without the need for extensive manual labeling or complex feature extraction process, even when working with small datasets. By harnessing the advanced capabilities of wav2vec 2.0, this research aims to develop a novel approach for diagnosing voice disorders.

MATERIALS AND METHODS

Database

In this study, we utilized sound data from the publicly accessible VOICED PhysioNet database, available since May 2018.[26] This database comprises a total of 208 voice recordings, including 150 representing pathological conditions and 58 recordings from healthy individuals. The age range of the volunteers contributing to these recordings spans from 18 to 70 years. The pathological voice recordings are categorized into three specific conditions: hyperkinetic dysphonia (72 recordings), hypokinetic dysphonia (40 recordings), and reflux laryngitis (38 recordings). Within the hyperkinetic dysphonia category, specific conditions include vocal fold nodules, Reinke's edema, chorditis, rigid vocal fold, polyps, and prolapse. The hypokinetic dysphonia category encompasses conditions such as dysphonia of the chordal groove, adduction deficit, presbyphonia, glottic insufficiency, vocal fold paralysis, conversion dysphonia, laryngitis, and extraglottic air leak.

All recordings were acquired in a professionally equipped environment designed to maintain a noise level below 30 dB, with humidity levels controlled between 30% and 40%. The voice signals were captured using a Samsung Galaxy S4 smartphone equipped with the Vox4Health management system. The voice sensor was positioned 20 cm away from each volunteer's mouth at a 45-degree downward angle to ensure consistent recording conditions. All the recordings were of 32-bit resolution and were sampled at 8000 Hz. Furthermore, dedicated sound filters were applied to each recording to eliminate any unintended noise, thereby preserving the integrity and quality of the dataset.

Speech preprocessing

Prior to feature extraction, all audio data underwent a preprocessing stage to ensure compatibility with the wav2vec 2.0 model. The audio data were initially loaded in waveform format and resampled to 16 kHz to meet the input requirements of wav2vec 2.0. To enhance the diversity of training data and improve the model's generalization ability, several data augmentation techniques were applied. These techniques included reducing the volume to half its original level to simulate variations in loudness, randomly masking specific frequencies in the spectrum (with a parameter set to 30) to mimic the loss of certain frequency bands, and randomly masking segments on the time axis (with a masking parameter set to 30) to simulate the loss of information during specific time intervals. Following the data augmentation process, the augmented audio data were normalized using "layer_norm" function, which standardizes audio features to zero mean and unit variance. The purpose of normalization was to eliminate magnitude differences between different audio samples, ensuring that all data input to the wav2vec 2.0 model were comparable and processed within the same dimensional space.

Feature extraction using wav2vec 2.0

The wav2vec 2.0 model was utilized for feature extraction from audio data (Figure 1). This pretrained model, based on the transformer architecture, has proven effective extracting high-quality feature representations from raw audio waveforms.[22] We utilized transformers library from Hugging Face to load both the pretrained wav2vec 2.0 processor and model. The "facebook/wav2vec2-large-960h" model, trained on 960 hours of labeled audio data,[27] was selected for its robust feature extraction capabilities.

The processed audio data were input into the wav2vec 2.0 model for inference, during which the model encoded the audio data through multiple transformer layers, each responsible for extracting distinct features from the audio. The feature encoder in wav2vec 2.0 comprises seven layers of convolutional neural networks, where the input feature dimension corresponds to the original audio waveform. The extracted low-level features were then passed to a transformer encoder, the core component of wav2vec 2.0, consisting of multiple self-attention layers and feedforward neural networks. This architecture enables the model to capture long-range dependencies and contextual information, resulting in high-quality feature representations.

After feature extraction, the resulting features were standardized again using the StandardScaler to ensure they were on a comparable scale before further processing steps, such as principal component analysis (PCA) and classifier training. PCA was subsequently applied to the normalized features to reduce dimensionality, retaining the most informative components while minimizing computational complexity for the downstream classification tasks.

Training and validation stage

Four machine learning models were selected for comparison and evaluation: SVM, K-Nearest Neighbors (KNN), DT, and Random Forest (RF). These models were chosen for their diverse approaches to classification, each offering unique strengths: SVM is known for its effectiveness in high-dimensional spaces,[28] KNN is valued for its simplicity and ability to adapt to different distributions,[29] DT provides interpretability through decision rules,[30] and RF combines the benefits of multiple DTs to enhance robustness and reduce overfitting.[31] To optimize model performance, hyperparameters for each model were systematically tuned using Grid Search Cross-Validation (CV), a comprehensive method that explores all possible parameter combinations. For the SVM model, we optimized the regularization parameter C across the values $\{0.01, 0.1, 1, 10, 100\}$ and tested different kernel functions, including linear and radial basis functions. The optimal configuration was determined based on CV performance. For the KNN model, we explored various values of k (3, 5, 7, 9) and compared uniform versus distance-based weighting schemes. For the DT model, complexity was managed by optimizing parameters such as maximum tree depth (none, 10, 20, 30), minimum samples required to split an internal node (2, 5, 10), and minimum samples required at a leaf node (1, 2, 4). The RF model's hyperparameters were fine-tuned by adjusting the number of trees (50, 100, 200), maximum tree depth, and criteria for node splitting and leaf formation. Grid Search CV ensured that the selected hyperparameters maximized performance across multiple folds.

To evaluate the performance of the classification models, we employed the Stratified K-Fold CV method, which ensures that the proportion of class labels in each fold matches that of the original dataset.[32] This technique is particularly valuable for datasets with imbalanced classes, as it maintains the distribution consistency, thereby enhancing the reliability of model evaluation. By using Stratified K-Fold CV, we gained a comprehensive understanding of the model's performance across different data partitions, addressing class imbalance issues effectively.

The dataset was divided into five folds, with each fold used sequentially for training and testing. In each iteration, the model was trained on the training set and evaluated using the test set. Performance metrics such as accuracy, precision, recall, F1-score, macro average, micro average, receiver-operating characteristic (ROC) curve, and confusion matrix were calculated.

Precision (P), Recall (R), and the F1-score were computed using Equations (1), (2), and (3).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

where TP denotes true positives, FP represents false positives, and FN indicates false negatives.

Additionally, the performance was evaluated through micro and macro averages, with the specific formulas provided in Equations (4)-(9).

$$Micro\ Precision = \frac{Total\ TP}{Total\ TP + Total\ FP} \quad (4)$$

$$Micro\ Recall = \frac{Total\ TP}{Total\ TP + Total\ FN} \quad (5)$$

$$MicroF1Score = \frac{2 \times (Micro\ Precision \times Micro\ Recall)}{Micro\ Precision \times Micro\ Recall} \quad (6)$$

$$Macro\ Precision = \frac{1}{N}\sum_{i=1}^{N} Precision_{\mathrm{i}} \quad (7)$$

$$Macro\ Recall = \frac{1}{N}\sum_{i=1}^{N} Recall_{\mathrm{I}} \quad (8)$$

$$MacroF1Score = \frac{1}{N}\sum_{i=1}^{N} Score_{\mathrm{I}} \quad (9)$$

where N represents the number of classes.

Statistical analysis

GraphPad Prism 9.0 was employed for graph preparation and data analysis. The results are presented as mean ± standard deviation. Statistical significance was determined using Student's t test or one-way analysis of variance (ANOVA), with $P < 0.05$ considered statistically significant.

RESULTS

The performance evaluation of classifier models

Table 1 presents the performance metrics of four machine learning models—RF, SVM, KNN, and DT—evaluated across accuracy, precision, recall, and F1-score. The results are reported as mean values accompanied by standard deviation (mean ± SD), providing a comprehensive assessment of each model's consistency and reliability. The RF exhibits the highest accuracy (0.98 ± 0.02), underscoring its superior overall performance in the classification task. The model's high recall (0.97 ± 0.04) further highlights its proficiency in correctly identifying true-positive cases, while its F1-score (0.95 ± 0.05) indicates a well-balanced trade-off between precision and recall, affirming RF's robustness across various evaluation metrics. In contrast, the SVM model records the lowest accuracy (0.91 ± 0.06) among the four classifiers. However, it maintains a recall comparable to that of RF (0.97 ± 0.04), suggesting that while SVM may be less accurate overall, it is still effective in detecting positive instances. The SVM's F1-score (0.94 ± 0.04) reflects a strong, though slightly less consistent, performance relative to RF. The KNN model shows performance metrics closely aligned with SVM, achieving an accuracy of 0.92 ± 0.04 and an F1-score of 0.95 ± 0.03. Notably, KNN's precision (0.93 ± 0.04) surpasses that of SVM, despite their comparable recall values, indicating KNN's particular strength in reducing false positives. The DT model demonstrates a high accuracy (0.95 ± 0.03) and some of the highest precision (0.97 ± 0.02) and recall (0.96 ± 0.04) scores observed among the models. The DT's F1-score (0.96 ± 0.02) further underscores its effectiveness in achieving a balanced performance between precision and recall. Overall, RF and DT emerge as the top-performing models, with RF excelling in accuracy and DT

TABLE 1.
The Performance of Machine Learning Models

Models	Accuracy (Mean ± SD)	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)
RF	0.98 ± 0.02	0.92 ± 0.05	0.97 ± 0.04	0.95 ± 0.05
SVM	0.91 ± 0.06	0.92 ± 0.05	0.97 ± 0.04	0.94 ± 0.04
KNN	0.92 ± 0.04	0.93 ± 0.04	0.97 ± 0.03	0.95 ± 0.03
DT	0.95 ± 0.03	0.97 ± 0.02	0.96 ± 0.04	0.96 ± 0.02

Note: RF (Random Forest), SVM (Support Vector Machine), KNN (K-Nearest Neighbors), and DT (Decision Tree).

TABLE 2.
The Macro Average and Micro Average of Performance in Machine Learning Models

Models	Precision (Mean ± SD)		Recall (Mean ± SD)		F1-Score (Mean ± SD)	
	Macro Avg	Micro Avg	Macro Avg	Micro Avg	Macro Avg	Micro Avg
RF	0.92 ± 0.09	0.92 ± 0	0.88 ± 0.09	0.92 ± 0	0.90 ± 0.09	0.92 ± 0
SVM	0.91 ± 0.08	0.91 ± 0	0.87 ± 0.08	0.91 ± 0	0.89 ± 0.08	0.91 ± 0
KNN	0.92 ± 0.05	0.92 ± 0	0.88 ± 0.06	0.92 ± 0	0.89 ± 0.05	0.92 ± 0
DT	0.94 ± 0.05	0.95 ± 0	0.94 ± 0.03	0.95 ± 0	0.93 ± 0.04	0.95 ± 0

Abbreviations: macro avg, macro averaged; micro avg, micro averaged.

leading in precision, making those models particularly well-suited for the voice classification task.

Table 2 provides the macro average and micro average performance metrics—precision, recall, and F1-score—across the four models. The DT model achieves the highest macro average precision (0.94 ± 0.05), recall (0.94 ± 0.03), and F1-score (0.93 ± 0.04), indicating consistent and balanced performance across all classes. RF and KNN exhibit similar macro average precision values (0.92 ± 0.09 and 0.92 ± 0.05, respectively). However, RF slightly outperforms KNN in macro average recall (0.88 ± 0.09 vs 0.88 ± 0.06) and F1-score (0.90 ± 0.09 vs 0.89 ± 0.05), indicating a marginally better balance between precision and recall. Although the SVM model lags behind the other models, it still demonstrates adequate macro average precision (0.91 ± 0.08), recall (0.87 ± 0.08), and F1-score (0.89 ± 0.08), reflecting reasonable performance across different classes. In terms of micro-averaged metrics, all models exhibit identical precision, recall, and F1-score values, with RF, SVM, and KNN each achieving 0.92, while DT leads with a score of 0.95. This consistency suggests that all models perform uniformly well across the entire dataset, with DT showing a slight edge. In summary, the DT model emerges as the most robust performer, excelling in both macro- and micro-average metrics, particularly in terms of precision and recall. RF and KNN also demonstrate strong performance, especially in micro average metrics, while SVM remains competitive, despite its slightly lower overall performance. These results indicate that the DT model is particularly well-suited for the classification tasks in this study, given its superior performance across various evaluation metrics.

Figure 2 illustrates the performance of different classifiers across four evaluation metrics, comparing scenarios with and without data augmentation. Each box plot reflects the variability and consistency across CV folds. For DT model, data augmentation significantly improves accuracy and recall, with reduced variability. Precision remains stable, while the F1-score shows a notable increase. The KNN model also benefits from augmentation, with noticeable improvement in accuracy, precision, and F1-score. The RF model experiences significant enhancements across all metrics, particularly in accuracy and recall. Similarly, the SVM model sees improvements in accuracy, precision, and F1-score with data augmentation. The *P* values indicate that the DT model shows a statistically significant improvement ($P = 0.0263$) with augmentation, while the other models, although improved, do not reach such improvements. Overall, data augmentation positively impacts the performance of all classifiers, especially the RF and DT models, where the improvements are most substantial.

Receiver-operating characteristic curve

Figure 3 displays the ROC curves for four machine learning models across five CV folds, with corresponding area under the curve (AUC) values provided for each fold. These AUC values offer a quantitative assessment of the models' ability to distinguish between classes. The DT exhibits variability across folds, with AUC values ranging from 0.86 to 1.00. Notably, folds 2, 3, and 4 demonstrate strong classification capabilities, achieving AUCs of 0.96, 1.00, and 0.89, respectively. However, fold 5, with an AUC of 0.86, indicates a decline in performance, suggesting that the DT model's effectiveness may be sensitive to the specific data split, thus highlighting potential inconsistencies. The KNN model shows greater variability in its AUC values across folds, ranging from 0.50 to 0.83. The model's weakest performance is observed in fold 3, where the AUC drops to 0.50, indicating poor discriminative ability. Other folds demonstrate moderate performance, with AUCs between 0.73 and 0.83. This variability implies that KNN may be less consistent in its performance compared with the other models evaluated. The RF model emerges as the most consistent and robust performer, with AUC values near or equal to 1.00 across all folds. Folds 2, 3, 4, and 5 achieve perfect AUC scores of 1.00, signifying excellent classification performance. Even in fold 1, the model achieves a near-perfect AUC of 0.98, underscoring its reliability and effectiveness across different data splits. Similarly, the SVM model demonstrates strong and consistent performance, with AUC values close to 1.00 across all folds. Folds 2, 3, 4, and 5 achieve near-perfect AUC scores of 1.00, paralleling the performance of the RF

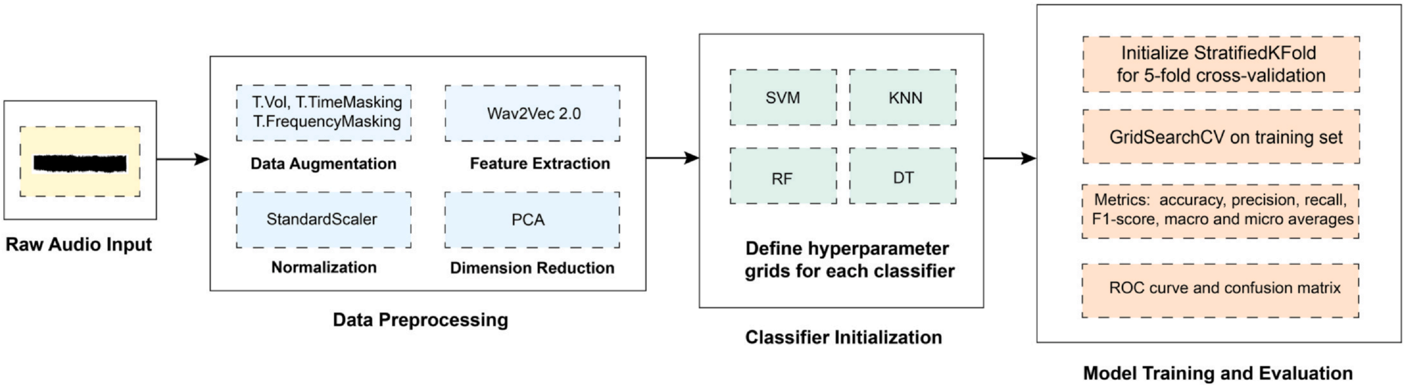
Raw Audio Input
T.Vol, T.TimeMasking
T.FrequencyMasking
Data Augmentation
Wav2Vec 2.0
Feature Extraction
StandardScaler
Normalization
PCA
Dimension Reduction
Data Preprocessing
SVM
KNN
RF
DT
Define hyperparameter grids for each classifier
Classifier Initialization
Initialize StratifiedKFold for 5-fold cross-validation
GridSearchCV on training set
Metrics: accuracy, precision, recall, F1-score, macro and micro averages
ROC curve and confusion matrix
Model Training and Evaluation

FIGURE 1. Workflow for voice classification using wav2vec 2.0. Audio is preprocessed with wav2vec 2.0 to extract features. Data augmentation is selectively applied to the training set. Features are normalized and PCA is applied to reduce dimensionality. KNN, SVM, RF, and DT classifiers are trained using five-fold cross-validation, with Grid Search CV for tuning. Performance is evaluated with metrics like accuracy, precision, recall, and AUC.

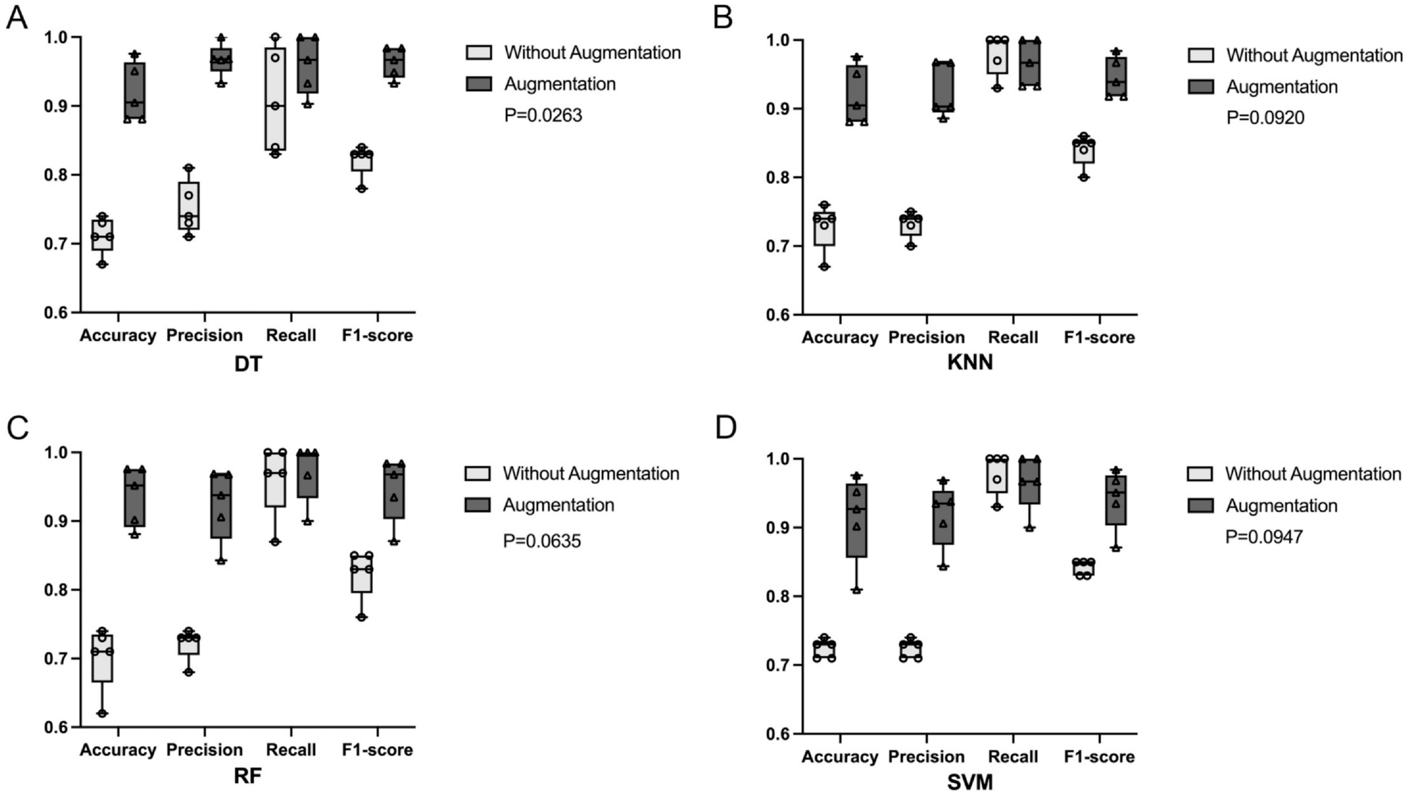

FIGURE 2. Box plots showing performance of DT, KNN, RF, and SVM models with and without data augmentation across accuracy, precision, recall, and F1-score. Each point represents a fold from cross-validation. A *P* value < 0.05 indicates a statistically significant difference between augmented and nonaugmented models, with no significant difference observed between folds.

model. Although fold 1 shows a slightly lower AUC of 0.99, it still indicates excellent discriminative ability. The consistency of the SVM model across folds further establishes it as a robust classifier for this task.

Confusion matrix

Figure 4 presents confusion matrices for four machine learning models across five CV folds, detailing the true positives, true negatives, false positives, and false negatives for two classes: “Normal” and “Pathological.” The DT model demonstrates strong performance across all folds, with most predictions accurately classified. In fold 1, the model correctly classifies 29 pathological samples and 11 normal samples, with minimal misclassifications (1 false positive and 1 false negative). Similar trends are observed in other folds, though occasional misclassifications occur, such as 3 false positives in fold 2, indicating some variability in performance. The KNN model exhibits more variability across folds. For instance, in fold 1, the model misclassifies seven normal samples, and in fold 5, it misclassifies five normal samples, suggesting less-robust performance relative to the other models. Nevertheless, KNN consistently classifies all pathological samples correctly across all folds, highlighting its strength in recognizing pathological cases. The RF model consistently delivers high performance, with most confusion matrices showing no or minimal misclassifications. In folds 3 and 4, the model achieves perfect classification with no false positives or false negatives. Only minor misclassifications are observed in folds 1, 2, and 5, with the worst cases involving 2 false positives and 1 false negative, emphasizing the model's reliability. The SVM model also displays strong and consistent performance across the folds, similar to RF. Most folds show few-to-no misclassifications, with fold 1 achieving perfect classification, while fold 2 shows a few misclassifications (4 false positives and 4 false negatives). Overall, the SVM model maintains high accuracy, particularly in correctly identifying pathological samples. Across all models, RF and SVM exhibit the most robust and consistent performance, with minimal misclassifications across all folds. The DT also performs well but shows slightly more variability in misclassifications. In contrast, KNN shows greater variability and higher misclassification rates, particularly in distinguishing normal samples. These confusion matrices further corroborate the superior performance of RF and SVM in this classification task.

DISCUSSION

This study underscores the efficacy of wav2vec 2.0 as a feature extractor for the classification of voice disorders. By harnessing its advanced feature extraction capabilities, we successfully trained four machine learning models that exhibited high performance in differentiating between normal and pathological voices.

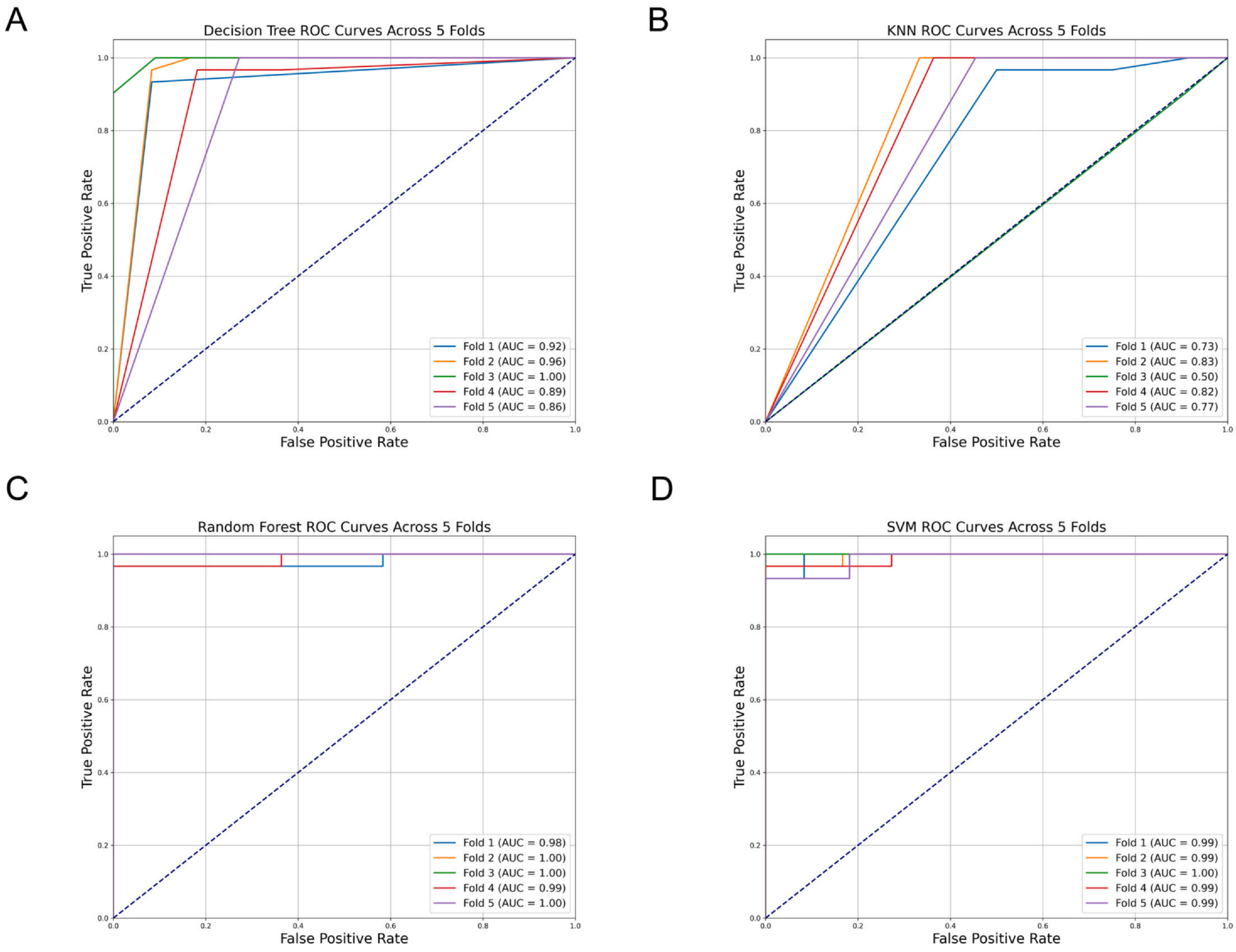

FIGURE 3. The ROC curve of the machine learning models. The AUC value, nearing 1.0, corroborates the model's superior performance in distinguishing between normal and pathological voices.

A notable strength of wav2vec 2.0 lies in its ability to automatically extract features directly from raw audio data, thereby obviating the need for traditional handcrafted features such as MFCC. While MFCC, derived from conventional signal processing techniques, effectively capture fundamental acoustic characteristics like the spectral envelope, but their static nature limits adaptability across different tasks.[33] In contrast, wav2vec 2.0 leverages self-supervised pretraining on large-scale, unlabeled data to learn generalizable representations, thereby enhancing the applicability of these features to a wide array of speech-related tasks. This approach not only reduces the reliance on manually engineered features but also maximizes the utilization of inherent information within the audio data.[23,34]

Recent studies have begun to explore the application of wav2vec 2.0 in voice-related research. Zhang et al used transfer learning with wav2vec 2.0 for speech depression detection in low-source environments, achieving commendable F1-scores of 79% on the DAIC-WOZ dataset and 90.53% on the CMDC dataset.[35] Getman et al demonstrated the efficacy of wav2vec 2.0 in predicting pronunciation levels in children with speech sound disorders, further supporting its potential for speech-related applications.[36]

In this study, wav2vec 2.0 features were utilized across four distinct machine learning models. When compared to other studies using the VOICED dataset, such as the work by Chen et al, which reported accuracies of 97.8%, 91.6%, and 90.3% with DNN, SVM, and RF models, respectively,[37] our results also reflect strong performance. Notably, the RF model consistently outperformed the others across various evaluation metrics, achieving an accuracy of 0.98 ± 0.04, recall of 0.97 ± 0.04, and F1-score of 0.95 ± 0.05, thereby demonstrating its robustness and reliability in classifying "Normal" and "Pathological" voices. To address potential class imbalance, we further assessed each model using macro and micro average values, all of which exceeded 0.9, as reported in Table 2. These findings

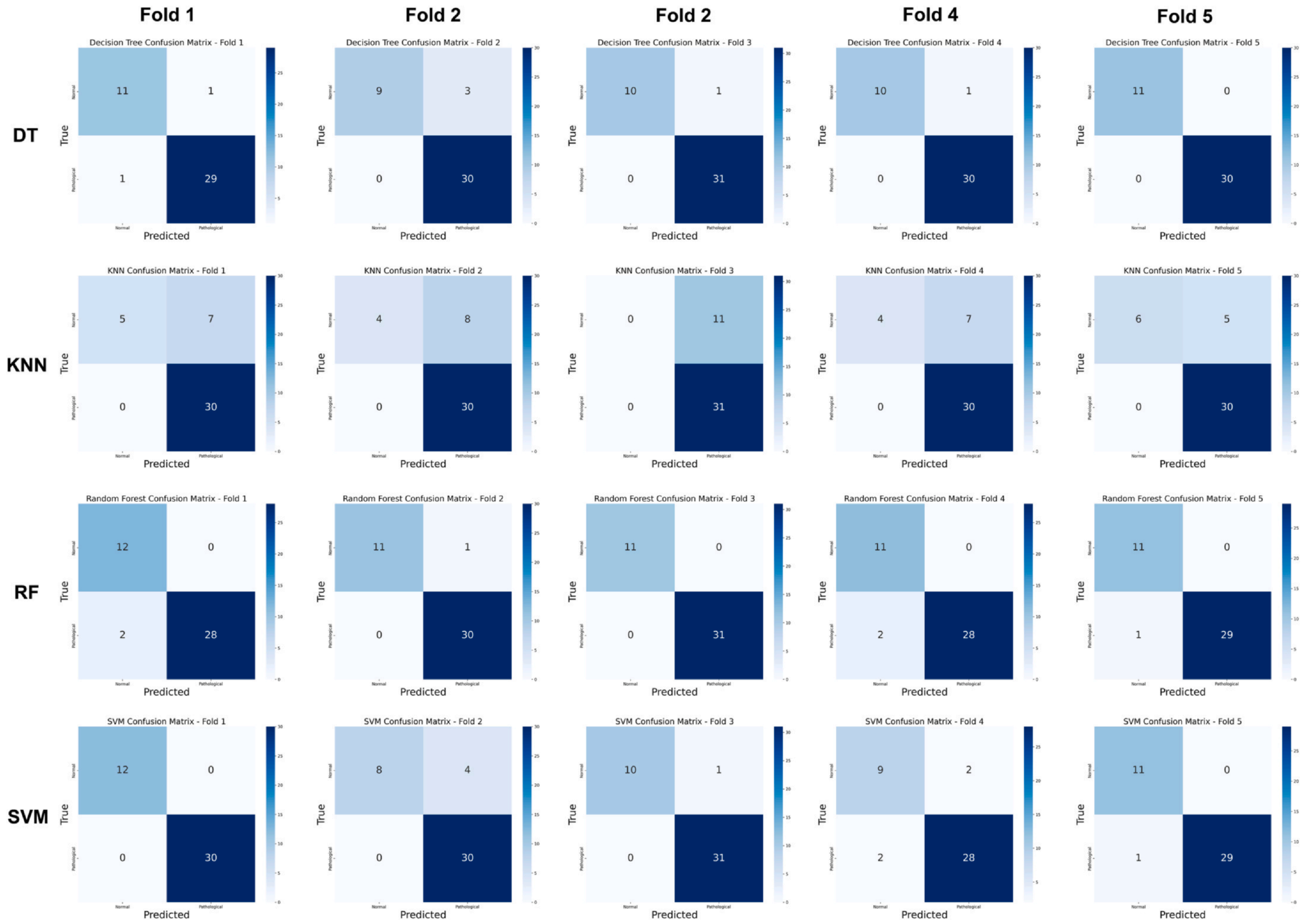

FIGURE 4. The confusion matrix of the machine models' performance in different folds. The x-axis was denoted by the predicted categories, while the y-axis was represented by the true categories. Each cell contained the count of samples where the actual category was aligned with the predicted category. Color variations were used to emphasize different sample counts, with darker shades typically indicating higher sample counts and lighter shades indicating lower sample counts.

underscore the high classification performance achieved through the integration of wav2vec 2.0 and machine learning techniques.

To further evaluate the robustness of these models, we analyzed their performance across different CV folds, with and without data augmentation. As depicted in Figure 2, data augmentation had a generally positive impact on model performance, particularly in terms of accuracy and recall. The RF and DT models exhibited the most significant improvements, highlighting their sensitivity to augmentation techniques. Additionally, ROC curve analysis (Figure 3) revealed consistently high AUC values across most models, with the RF model consistently outperforming the others. Furthermore, confusion matrix analysis (Figure 4) confirmed the robust recognition capabilities of the DT, KNN, RF, and SVM models, as evidenced by minimal misclassifications and stable performance across folds. This comprehensive evaluation suggests that the models are not overfitting to specific data folds but rather exhibit consistent performance across the entire dataset.

Despite these promising results, the study has several limitations. First, the generalizability of the acoustic features extracted by wav2vec 2.0 to other datasets remains unvalidated, necessitating further research to assess its performance across diverse data sources. Additionally, the relatively small dataset used in this study limits the generalizability of the findings. While CV and data augmentation techniques were employed to mitigate the effects of limited data, these strategies cannot fully compensate for the absence of a larger, more diverse dataset. Future studies should prioritize the collection of a more extensive dataset encompassing a broader range of pathologies to enhance the robustness and applicability of the models. Last, while wav2vec 2.0 effectively reduces the dependence on handcrafted features, the challenge of interpretability in deep learning models persists. Future work should explore methods to improve the transparency and interpretability of models utilizing wav2vec 2.0 features, especially in clinical settings where understanding the decision-making process is critical.

CONCLUSION

In conclusion, this study demonstrates the effectiveness of using wav2vec 2.0 embeddings combined with machine learning models to classify normal and pathological voices within the VOICED dataset. The results achieved high accuracy and robust evaluation metrics, underscoring the potential of wav2vec 2.0 with deep learning techniques in advancing the diagnosis of voice disorders.

CRediT Authorship Contribution Statement

Jie Cai: Conceptualization, formal analysis, methodology, and writing—original draft. **Jianghao Wu:** Data curation, formal analysis, and conceptualization. **Yuliang Song**: Data curation, formal analysis, and conceptualization. **Xiong Chen**: Conceptualization, review, editing, and supervision.

Data Availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Martins RHG, do Amaral HA, Tavares ELM, et al. Voice disorders: etiology and diagnosis. *J Voice.* 2016;30:761.e1–761.e9. https://doi.org/10.1016/j.jvoice.2015.09.017.
2. Sataloff RT, Spiegel JR, Hawkshaw M. Voice disorders. *Med Clin North Am.* 1993;77:551–570. https://doi.org/10.1016/S0025-7125(16)30239-5.
3. Cohen SM, Kim J, Roy N, et al. Prevalence and causes of dysphonia in a large treatment-seeking population. *Laryngoscope.* 2012;122: 343–348. https://doi.org/10.1002/lary.22426.
4. Martins RHG, Pereira ERBN, Hidalgo CB, et al. Voice disorders in teachers. A review. *J Voice.* 2014;28:716–724. https://doi.org/10.1016/j.jvoice.2014.02.008.
5. Oliveira P, Ribeiro VV, Constantini AC, et al. Prevalence of work-related voice disorders in voice professionals: systematic review and meta-analysis. *J Voice.* 2022. https://doi.org/10.1016/j.jvoice.2022.07.030.S0892-1997(22)00232-6.
6. Lee SH, Yu JF, Fang TJ, et al. Vocal fold nodules: a disorder of phonation organs or auditory feedback? *Clin Otolaryngol.* 2019;44:975–982. https://doi.org/10.1111/coa.13417.
7. Hinerman RW, Mendenhall WM, Amdur RJ, et al. Early laryngeal cancer. *Curr Treat Options Oncol.* 2002;3:3–9. https://doi.org/10.1007/s11864-002-0036-x.
8. Costello D. Acoustic assessment. *Adv Otorhinolaryngol.* 2020;85:55–58. https://doi.org/10.1159/000456683.
9. Reghunathan S, Bryson PC. Components of voice evaluation. *Otolaryngol Clin North Am.* 2019;52:589–595. https://doi.org/10.1016/j.otc.2019.03.002.
10. Sachdeva K, Mittal N, Sachdeva N. Role of video laryngostroboscopy in benign disease of larynx. *Indian J Otolaryngol Head Neck Surg.* 2020;72:267–273. https://doi.org/10.1007/s12070-020-01827-8.
11. Idrisoglu A, Dallora AL, Anderberg P, et al. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: systematic literature review. *J Med Internet Res.* 2023;25:e46105. https://doi.org/10.2196/46105.
12. Al-Hussain G, Shuweihdi F, Alali H, et al. The effectiveness of supervised machine learning in screening and diagnosing voice disorders: systematic review and meta-analysis. *J Med Internet Res.* 2022;24:e38472. https://doi.org/10.2196/38472.
13. Syed SA, Rashid M, Hussain S. Meta-analysis of voice disorders databases and applied machine learning techniques. *Math Biosci Eng.* 2020;17:7958–7979. https://doi.org/10.3934/mbe.2020404.
14. Reid J, Parmar P, Lund T, et al. Development of a machine-learning based voice disorder screening tool. *Am J Otolaryngol.* 2022;43: 103327. https://doi.org/10.1016/j.amjoto.2021.103327.
15. Dhief F.T.A., Latiff N.M.A., Malik N.N.N.A., et al., Voice Pathology Detection Using Machine Learning Technique, 2020 IEEE 5th International Symposium on Telecommunication Technologies (ISTT), 2020, pp. 99-104. doi: 10.1109/ISTT50966.2020.9279346.
16. Sidhu MS, Latib NAA, Sidhu KK. MFCC in audio signal processing for voice disorder: a review. *Multimed Tools Appl.* 2024;83:1–21. https://doi.org/10.1007/s11042-024-19253-1.2024/04/27.
17. Heman-Ackah YD, Heuer RJ, Michael DD, et al. Cepstral peak prominence: a more reliable measure of dysphonia. *Ann Otol Rhinol Laryngol.* 2003;112:324–333. https://doi.org/10.1177/000348940311200406.
18. Chen L, Chen J. Deep neural network for automatic classification of pathological voice signals. *J Voice.* 2022;36:288.e15–288.e24. https://doi.org/10.1016/j.jvoice.2020.05.029.
19. Verde L, Pietro GD, Sannino G. Voice disorder identification by using machine learning techniques. *IEEE Access.* 2018;6:16246–16255. https://doi.org/10.1109/ACCESS.2018.2816338.
20. Fonseca ES, Guido RC, Junior SB, et al. Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM). *Biomed Signal Process Control.* 2020;55:101615. https://doi.org/10.1016/j.bspc.2019.101615.
21. Contreras RC, Viana MS, Fonseca ES, et al. An experimental analysis on multicepstral projection representation strategies for dysphonia detection. *Sensors.* 2023;23:5196. https://doi.org/10.3390/s23115196.
22. Schneider S, Baevski A, Collobert R, Auli M. wav2vec: unsupervised pre-training for speech recognition. *Interspeech.* 2019:3465–3469. https://doi.org/10.21437/Interspeech.2019-1873.
23. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. *arXiv.* 2020;2006:11477. https://doi.org/10.48550/arXiv.2006.11477.
24. Wang Y, Boumadane A, Heba A. A fine-tuned Wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv.* 2021:2111.02735. https://doi.org/10.48550/arXiv.2111.02735.
25. Klempíř O, Příhoda D, Krupička R. Evaluating the performance of wav2vec embedding for Parkinson's disease detection. *Meas Sci Rev.* 2023;23:260–267. https://doi.org/10.2478/msr-2023-0033.
26. Cesari U, De Pietro G, Marciano E, et al. A new database of healthy and pathological voices. *Comput Electr Eng.* 2018;68:310–321. https://doi.org/10.1016/j.compeleceng.2018.04.008.
27. wav2vec large.pt. Accessed May 21, 2024. Available at: https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_large.pt.
28. Pisner DA, Schnyer DM. Chapter 6 - Support vector machine. In: Mechelli A, Vieira S, eds. *Machine Learning.* London, UK: Academic Press; 2020:101–121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7.
29. Zhang Z. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med.* 2016;4:218. https://doi.org/10.21037/atm.2016.03.37.
30. Yang Y, Morillo I, Hospedales T. Deep neural decision trees. *arXiv.* 2018:1806.06988. https://doi.org/10.48550/arXiv.1806.06988.
31. Rigatti SJ. Random forest. *J Insur Med.* 2017;47:31–39. https://doi.org/10.17849/insm-47-01-31-39.1.
32. Yadav S., Shukla S., Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification, 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016, pp. 78-83. doi:10.1109/IACC.2016.25.

33. Abdul ZK, Al-Talabani AK. Mel frequency cepstral coefficient and its applications: a review. *IEEE Access.* 2022;10:122136–122158. https://doi.org/10.1109/ACCESS.2022.3223444.
34. Jain R, Barcovschi A, Yiwere MY, et al. A WAV2VEC2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access.* 2023;11:46938–46948. https://doi.org/10.1109/ACCESS.2023.3275106.
35. Zhang X, Zhang X, Chen W, et al. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Sci Rep.* 2024;14:9543. https://doi.org/10.1038/s41598-024-60278-1.
36. Getman Y, Al-Ghezi R, Voskoboinik E, et al. Wav2vec2-based speech rating system for children with speech sound disorder. *Int Speech Commun Assoc (ISCA).* 2022;23:3618–3622. https://doi.org/10.21437/Interspeech.2022-10103.
37. Chen L, Wang C, Chen J, et al. Voice disorder identification by using Hilbert-Huang transform (HHT) and K nearest neighbor (KNN). *J Voice.* 2021;35:932.e1–932.e11. https://doi.org/10.1016/j.jvoice.2020.03.009.