IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 June 2025, accepted 24 June 2025, date of publication 30 June 2025, date of current version 8 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3584534

RESEARCH ARTICLE

Performance Evaluation of Different Speech-Based Emotional Stress Level Detection Approaches

JÁN STAŠ, STANISLAV ONDÁŠ, AND JOZEF JUHÁR

Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, 042 00 Košice, Slovakia

Corresponding author: Ján Staš (jan.stas@tuke.sk)

This work was supported in part by the Ministry of Education, Research, Development, and Youth of the Slovak Republic under Project KEGA 049TUKE-4/2024 and Project KEGA 041TUKE-4/2025; and in part by the Slovak Research and Development Agency under Project APVV-22-0261 and Project APVV-22-0414.

ABSTRACT This study presents a comprehensive evaluation and comparison of selected conventional and advanced approaches for detecting and classifying emotional stress levels from speech. Three representative and publicly available stress datasets–CRISIS, StressDat, and SUSAS–were used. A detailed comparative analysis of their speech characteristics, including fundamental frequency (F_0), formant frequencies (F_1, F_2, F_3), voice intensity, harmonic-to-noise ratio, jitter, and shimmer, was conducted to investigate how these features change with increasing stress and contribute to its detection. Both conventional feature-based methods and deep learning techniques, including transfer and self-supervised learning, are explored in the experimental part of this research. The best classification results among the conventional methods were achieved using a combination of features extracted via the Discrete Wavelet Transform and GammaTone Cepstral Coefficients (DWT+GTCC), paired with a Subspace k-Nearest Neighbors classifier, which yielded F1-scores of up to 91.1% on CRISIS and 82.4% on StressDat. Among the deep learning approaches, we fine-tuned self-supervised models such as Wav2Vec 2.0 and BYOL-S (Bootstrap Your Own Latent for Speech), as well as transfer learning models including VGGish and YAMNet. Wav2Vec 2.0 consistently outperformed the others, achieving an F1-score of 94% on CRISIS and over 87% on StressDat. On the SUSAS dataset, which contains speech under simulated and real stress conditions, an F1-score of 77% was achieved using both the fine-tuned Wav2Vec 2.0 model as well as a pipeline combining BYOL-S feature extraction with a Support Vector Machine (SVM) classifier. After merging all stress-related datasets into a single balanced dataset, Wav2Vec 2.0 achieved the highest F1-score of 85%. Incorporating additional emotional speech data improved robustness and led to a slight increase in the accuracy of classifying stress into three distinct levels. These results highlight the importance of robust feature selection, appropriate classifier design, and the integration of emotional speech data to enhance generalization performance in real-world applications.

INDEX TERMS Deep learning, emotional speech, feature selection, fundamental frequency, generalization, levels of stress, self-supervised learning, speech under stress, transfer learning.

I. INTRODUCTION

Stress is the natural response of the human body to external pressures or increased demand. Although the physical effects of stress vary from person to person, it is a common mechanism that triggers a chain reaction that prepares the body to cope with a perceived threat. These reactions include the release of adrenaline, noradrenaline, and cortisol, which can increase heart rate, breathing rate, blood pressure, muscle tension, sweating, alertness, or a rush of energy. While

The associate editor coordinating the review of this manuscript and approving it for publication was Alexander Kocian.

this response is beneficial in certain situations and can motivate us to perform better, experiencing it too frequently or in situations where it is unnecessary can lead to chronic stress and, often, health problems. In cases of chronic stress, the body remains in a state of alertness even when there is no immediate danger, which can contribute to a range of psychological and physical health issues.

In today's hectic world, stress is slowly but certainly becoming an everyday part of our lives. Few people can say that they do not know what stress is, or that they do not experience symptoms in daily life. According to the Global Organization of Stress[1] and other research, stress affects billions of people worldwide. It is estimated that more than 75% of adults worldwide report experiencing stress at least once per month. Approximately 35–40% of people experience excessive or chronic stress. Around 20–25% of the population suffers from chronic (long-term) stress. Stress levels vary by region, age group, and lifestyle. In the workplace, stress affects 60–80% of employees.

The determining factor (or trigger) of stress is the stressor, which causes the corresponding stress reaction and exhibits significant individual variability. A stressor can have a physical, chemical, biological, or psychosocial origin, but it is usually a combination of several factors. The most common stressors include excessive workload (33%), health problems (28%), and disturbed relationships within the family or with a partner (25%).[2] A high level of stress is experienced primarily by people aged 18–34, university graduates, mental workers, employees working with people, managers, etc.

Kirchhubel et al. [1] defined four levels of stressors. *Zero-order stressors* occur in everyday situations, for example, due to vibrations in the surrounding environment or discomfort caused by wearing various personal equipment (e.g., an oxygen mask). These stressors lead to psychological and articulatory changes. *First-order stressors* include factors such as sleep deprivation, illness, fatigue, or withdrawal from nicotine, alcohol, or drugs. This type of stressor causes changes in breathing rate and muscle tension. *Second-order stressors* arise during communication when it is necessary to increase voice intensity due to environmental noise or a poor communication channel (e.g., during a phone call). *Third-order stressors*, the highest level, occur under excessive psychological or cognitive load, leading to changes in articulation and phonation.

An optimal level of stress can act as a creative and motivational force, leading a person to increase productivity. Stress acting in this way is referred to as *eustress*. It is associated with positive, pleasant feelings and helps to find balance and stability. Eustress releases the so-called happiness hormones enkephalins and endorphins. Another type of stress is *acute stress*, which is a very short-term type of stress that can either be positive or negative. We most often encounter this type of stress in everyday life. It results from sudden events that require an immediate response. Acute stress triggers the body's stress response, while the triggers can be different (e.g. verbal argument). Acute stress itself does not significantly affect a person's health if we find ways to relax quickly. The opposite of eustress is *distress*, which is chronic traumatic stress that is potentially destructive. It damages mental and physical health. It is associated with negative, unpleasant feelings and disturbing physical states. If the amount of acting stressors exceeds the body's capacity to cope over time or in intensity, it can have a negative impact on our health. In this situation, we experience *chronic stress*. Chronic stress can cause various symptoms and affect the overall well-being of the individual. It affects the functioning of the endocrine, cardiovascular, digestive and immune systems.

Various methods for detecting emotional stress using physiological signals were reviewed in [2], including those based on body gestures, facial expressions, human voice, blood volume pulse, electrocardiogram, electroencephalogram, electromyogram, galvanic skin resistance, heart rate variability, respiratory parameters, skin conductance and so on.

Speech-based emotional stress detection relies on key acoustic features such as mean fundamental frequency (mean F_0), vocal intensity (the acoustic energy of speech), formant frequencies (F_1, F_2, and F_3), jitter, shimmer, speaking rate, the number of pauses and hesitations, as well as spectral and non-linear attributes. Stress affects speech quality through changes in voice level, pitch, and muscle tension. Emphasizing spectral and prosodic features enhances the accuracy of speech recognition systems, making them crucial for identifying stress-related emotions.

The detection of emotional stress in speech has applications in various fields, including mental health care, child psychology, customer service, e-learning, driver behavior monitoring, and human-computer interaction. For instance, Perera et al. [3] introduced an artificial intelligence-based framework that detects emotional states and stress levels, monitors speech fluency, and identifies speakers through their voice in real time.

As part of our research, we focus on several topics related to natural language and speech processing, where speech stress detection plays an important role. This motivates us to create a robust tool capable of classifying stress levels in speech in real time. For example, in the project "*The Role of Early Diagnostic and Therapeutic Supportive Tools for Children with Hearing Disabilities and Speech Disorders*", we are collaborating with specialists to develop various speech therapy exercises integrated into a game, which reduce stress and increase motivation in children. Similarly, in the project "*Multimodal Detection of Toxic Behavior in Social Media*", the emotional state or stress level can serve as one of the input modalities monitored for detecting toxic behavior in audiovisual content published on social networks. In our latest project, entitled "*Impact of Driver's Emotions and*

[1] https://gostress.com/stress-facts

[2] https://www.medante.sk/clanky/viac-ako-stvrtina-slovakov-trpi-chronickym-stresom/

Behavior on Driving Style", we investigate the relationship between emotional state, stress level, and specific driving behaviors, with the goal of evaluating a driver's style and its impact on road safety.

The objectives and contributions of the article can be summarized as follows:

- to provide an overview of existing databases of speech under stress for the purpose of training and testing stress classification algorithms;
- to perform a detailed analysis of available databases of speech under stress from the perspective of selected speech characteristics, such as fundamental frequency, formant frequencies, voice intensity, jitter, shimmer, etc., in order to understand how these features change with increasing stress;
- to compare both conventional and modern approaches to classifying the presence of stress in human speech, including feature extraction, or embeddings combined with suitable classification algorithms, transfer learning using pre-trained speech models, and self-supervised learning of speech representations;
- to extend the stressed speech database with emotional data in order to increase the size and diversity of the training set, and to monitor whether this improves stress classification accuracy. This helps ensure that the resulting algorithm can be deployed under a variety of acoustic conditions.

The rest of the paper is organized as follows. In Section II, we summarize almost 20 different research studies, either from the perspective of extracted features and embeddings, or from the perspective of classification algorithms and the application of deep learning techniques. Section IV focuses on the description and comparison of the algorithms and methods used in this research. Section III describes and analyzes the available datasets of stressed and emotional speech used in the experimental part of the study. Section V presents the experiments conducted with the proposed stress classification approaches and summarizes their results. Section VI discusses the limitations and potential improvements of the applied methods. Finally, Section VII summarizes the contributions of our work, and Section VIII concludes the paper with suggestions for future research.

II. LITERATURE REVIEW

This section provides a critical overview of current research on voice stress analysis and the detection of emotional stress levels from speech, employing various feature extraction techniques, speech embedding methods, and classification approaches, some of which are used in this study.

A. VOICE STRESS ANALYSIS

Sigmund [4] investigated psychological stress detection in Czech speech using the ExamStress corpus [5], which contains recordings of neutral and stressed speech from 31 speakers during final oral exams. The research focused on a detailed analysis of fundamental frequency (F_0), vowel formants, and spectral characteristics to capture both long-term and short-term changes in speech under stress. A key advantage of the study is its realistic stress induction setting and the quantitative analysis of F_0, which revealed speaker-specific stress markers. However, the corpus is limited in speaker diversity and remains publicly inaccessible, reducing its utility for broader research. Additionally, the analysis relied on simplified acoustic features and did not incorporate automatic classification methods, limiting its applicability.

The paper [6] reviews psychological stress analysis in speech, focusing on differences between stressed and neutral speech signals. Stress can be cognitive or noise-induced, affecting the short-time spectrum of vowel phonemes. The study explores two spectral analysis methods – Short-Time Fourier Transform (STFT) and Short-Time Chirp Transform (STChT) – comparing their effectiveness in detecting stress-related changes in speech. As part of this research, a new "Exam Stress" dataset was created, consisting of speech recordings collected during exams. STFT was applied to segment speech into short utterances, capturing time-frequency features indicative of stress-induced variations in pitch and energy. Results indicate that stressed speech exhibits higher frequencies in the chirp spectrum due to enhanced pitch modulation. The study employs real stress indicators, simple feature extraction methods, and a comparative approach that highlights the advantages of STChT over STFT. However, its small and inaccessible dataset limits reproducibility, and the absence of automatic classification and statistical validation reduces its applicability in real-world systems.

B. STRESS DETECTION

He et al. [7] focused on the detection of stress in speech with applications in behavioral sciences, human-computer interaction, and mental health monitoring. This research employs non-linear speech analysis using the Teager Energy Operator (TEO) to capture additional excitation sources present in stressed speech. The study introduces a novel feature extraction approach that combines TEO with Discrete Wavelet Transform (DWT), Wavelet Packet (WP), and Perceptual Wavelet Packet (PWP) transforms. Wavelet transforms were applied to decompose speech signals into frequency components, enabling the extraction of time-frequency features sensitive to stress-induced variations. Stress classification was performed using a Gaussian Mixture Model (GMM) under speaker-independent conditions. The method was evaluated on the SUSAS dataset [8], which contains both simulated and actual stress speech recordings. Experimental results demonstrate that the TEO-PWP method achieved the highest Accuracy (94–96%) by leveraging non-linear energy-based analysis and perceptual critical band information, outperforming conventional approaches. However, the method involves higher computational complexity, sensitivity to the choice of wavelet basis, and limited generalization beyond the SUSAS dataset. Still, it shows

strong potential for robust stress recognition in controlled settings.

Han et al. [9] presented a deep learning algorithm for automatic stress detection from speech signals, addressing the need for non-invasive psychological stress monitoring. The proposed method extracts mel-filter bank coefficients from speech data and utilizes a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) to classify speech as stressed or unstressed. The dataset consists of speech recordings from 25 subjects, selected from a larger multimodal dataset (56 participants), where stress levels were validated using salivary cortisol measurements. The best-performing model, an LSTM with a Support Vector Machine (SVM) classifier, achieved 66.4% Accuracy. This approach benefits from context-aware modeling through LSTM, objective stress validation via cortisol, and shows potential for real-world use. However, moderate accuracy, a small dataset, and unclear class boundaries limit its robustness and generalizability. Despite this, the study points to promising directions for combining physiological validation with deep learning in speech-based stress detection.

The authors in [10] proposed a Hilbert-Huang MFCC (HHT-MFCC) method for speech-based stress recognition, addressing the limitations of traditional MFCC-based feature extraction. The key innovation lies in integrating the HHT into the MFCC framework, allowing for adaptive time-frequency analysis suitable for nonlinear and non-stationary speech signals. The model was trained and tested on the SUSAS dataset [8], containing recordings from six speakers categorized into neutral, low-stress, and high-stress conditions. An Artificial Neural Network (ANN) was used for stress classification. Experimental results show that HHT-MFCC outperforms MFCC, achieving 96.05% Accuracy, demonstrating its effectiveness in enhancing stress recognition performance. The proposed method captures stress-induced speech variations well, uses a simple neural network, and is validated on real stress data. However, it is limited by a small speaker pool, potentially inflated performance due to extensive cross-validation, and higher computational complexity compared to standard MFCC extraction. Despite these limitations, the method offers a promising direction for more robust stress detection in speech.

Duvvuri et al. [11] explored stress level detection in continuous speech using deep learning models and optimal feature selection. The study employs the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ) dataset [12], processed to extract stress-related features. The research investigates Log Filter Bank (LogFBank), MFCC, GFCC, chroma, and Linear Predictive Coding (LPC) for feature extraction. Five deep learning models were evaluated: LSTM, Bidirectional LSTM (Bi-LSTM), CNN, k-fold CNN ($k = 5$), and a hybrid CNN-LSTM-attention model. The k-fold CNN model, trained on MFCC, GFCC, and LogFBank features, achieved the highest Accuracy of 80%. The study offers a comprehensive comparison of features and models, uses realistic data, and applies augmentation to boost robustness. However, its relatively small dataset, lower accuracy for high stress levels, and risk of overfitting due to augmentation limit generalizability. Despite these challenges, the work demonstrates the potential of deep learning combined with careful feature selection for speech-based stress detection.

The paper [13] introduces a new experimental paradigm to account for individual differences in stress perception. A key innovation is the use of reference speech, allowing for more accurate stress level assessments. The study presents the IFSO dataset, the first public Mandarin speech stress dataset. The proposed Stress Siamese ResNet (SSResNet), a siamese-based deep learning model, improves stress detection by leveraging reference speech as prior knowledge. Experimental validation shows that SSResNet outperforms state-of-the-art models like eGeMAPS, ResNet-101, Wav2Vec 2.0, and HuBERT, demonstrating higher accuracy and better generalization. The work uses a clearly labeled dataset, a strong baseline model, and reference speech to improve generalization across individuals. Challenges include language-specific speech patterns, self-assessed stress labels, and evaluation in controlled lab settings. Nonetheless, the study highlights the value of individualized stress baselines and provides a solid foundation for future mental health monitoring.

The study [14] explores voice stress detection by incorporating speaker individuality into speech analysis. Unlike previous research [15], which evaluates stress detection separately per dataset, this work integrates data from more than 100 speakers in 9 languages and five different types of stress to create a more generalized model. The innovation lies in combining self-supervised BYOL-S (Bootstrap Your Own Latent for Speech) and Hybrid BYOL-S/CvT audio embeddings with speaker embeddings (ECAPA, Resemblyzer) to enhance stress recognition [16]. The datasets come from previous cognitive and physical load studies, totaling 12 hours of speech data. Classification models include SVM and MLP, with experiments performed on 3–5 second audio chunks. Results show that integrating speaker embeddings significantly improves performance, achieving an Unweighted Average Recall (UAR) of up to 80.21% on short-duration speech samples. The study effectively incorporates speaker embeddings, improves cross-data generalization across languages and stress types, and supports short audio inputs suitable for real-time use. However, performance drops when combining heterogeneous datasets, and the approach relies on high-quality speaker embeddings for optimal results.

Ghose et al. [17] focused on multimodal stress detection using audio-visual data, addressing the challenge of non-invasive stress recognition in real-world settings. Unlike traditional physiological approaches, this study integrates four affective modalities: facial expressions, vocal prosody, speech sentiment, and physical fidgeting. A key innovation is the introduction of the MultiAffectStress dataset, consisting of 353 labeled video clips of individuals discussing personal experiences. For stress detection in speech, the study employs Wav2Vec 2.0 for vocal prosody analysis and DistilBERT for

sentiment classification. These features are combined with visual indicators using three fusion techniques: intermediate fusion, voting-based late fusion, and learning-based late fusion. Experimental results demonstrate that learning-based late fusion with a Random Forest model achieves the highest F1 score of 85%, showing that multimodal integration enhances stress detection accuracy. The study uses multimodal integration, a custom well-annotated dataset, and compares multiple fusion strategies. However, it relies heavily on sentiment analysis, uses a relatively small dataset, and shows limited generalization to diverse real-world scenarios. Despite these challenges, it highlights the benefits of combining complementary modalities for stress recognition.

The study by Rituerto-González et al. [18] examines the impact of stress on speaker recognition performance and proposes methods to mitigate its effects. The research introduces a stress-robust speaker identification approach by leveraging data selection and augmentation. Feature extraction methods include MFCC, pitch, and formant analysis. The study uses the VOCE Corpus [19], which contains speech samples recorded under both neutral and stress conditions. A Multi-Layer Perceptron (MLP) is used as the primary classification model, and its performance is compared with GMM and SVM. A novel approach involving data augmentation is introduced, generating synthetic stress-like speech by modifying pitch and speech rate. Experimental results show that training with both natural and synthetic stressed speech improves speaker recognition accuracy, achieving up to 99.45%. The method benefits from low computational complexity, effective data augmentation, and extensive evaluation across different classifiers. However, the dataset size is limited, the augmentation approach is relatively simple, and the focus remains solely on speaker identification rather than stress detection itself. Despite these limitations, the study demonstrates that targeted stress modeling can significantly enhance speaker recognition performance under adverse conditions. The authors also mention using Linear Frequency Cepstral Coefficients (LFCC) as one of the important features for speaker recognition under stress conditions. Along with LFCC, the research also considers MFCC and Linear Prediction Cepstral Coefficients (LPCC) as relevant features.

The study, led by Baird et al. [20], focused on evaluating the ability of acoustic features in speech to predict physiological markers of stress, such as cortisol levels, heart rate, and respiration. The research utilizes three German-language corpora – FAU-TSST, REG-TSST, and ULM-TSST [21] – involving over 100 participants subjected to the Trier Social Stress Test (TSST). The innovation lies in combining traditional handcrafted acoustic features (e.g., ComParE and eGeMAPS) with deep learning-based spectrogram features using the DeepSpectrum toolkit and VGGish embeddings. VGGish, pre-trained on AudioSet [22], extracts 128-dimensional embeddings from log-scale mel spectrograms, contributing to stress prediction models. The study employs machine learning models such as Support Vector Regression (SVR) and LSTM-RNN. Results demonstrate that audio features alone can effectively predict cortisol levels, achieving a Spearman correlation coefficient (ρ) of 0.770, with further gains through multimodal fusion. The work benefits from the use of validated physiological markers and a hybrid deep learning architecture. However, limitations include inconsistencies in cortisol scales across datasets, the use of only one cortisol label per session, and limited generalization to non-controlled environments.

The paper [23] reviews 24 research works covering different datasets, algorithms, and approaches related to stress detection through speech analysis using machine learning methods. These studies employ techniques such as CNN, RNN, SVM, and Radial Basis Function (RBF) networks for emotion and stress detection. The datasets mentioned include RAVDESS [24], EMO-DB [25], and DAIC-WOZ [12], among others. Commonly used feature extraction methods include MFCC, spectrograms, and Low-Level Descriptors (LLD). Some studies also explore non-speech-based stress detection, using image processing, physiological signals, and smartphone sensor data. The effectiveness of deep learning models is highlighted, with varying accuracies reported across studies. The review concludes that speech-based stress detection is a promising field, but further refinement of models and feature selection is necessary to improve accuracy.

C. DETECTION OF STRESSED EMOTIONS

Vaikole et al. [26] explored stress detection through speech analysis using deep learning. The study introduces a novel model based on Convolutional Neural Networks (CNNs) to classify speech as either stressed or unstressed. MFCC are extracted from preprocessed speech signals to capture essential acoustic features. The system is trained and evaluated using the RAVDESS dataset [24], a multimodal database of emotional speech and song. Results indicate that the MFCC-based CNN model significantly outperforms traditional pitch- and sample-rate-based methods, achieving an Accuracy of 94.3%. Among the advantages of the approach are its high classification accuracy, use of a validated dataset, and ability to infer stress through emotional cues. Limitations include a narrow range of modeled emotions, sensitivity to background noise, limited generalizability to real-world settings, and a potential risk of overfitting due to the complexity of the CNN model.

Nordin et al. [27] presented a speech-based emotional stress detection approach using a combination of TEO and MFCC, termed T-MFCC, for feature extraction. The study aims to improve emotional stress detection accuracy by enhancing the representation of non-linear speech components. A CNN classifier is employed to recognize stressed emotions (anger, disgust, sadness, and fear) in comparison to neutral speech. The dataset used is RAVDESS [24], which contains 1,440 speech samples. Results show that

the T-MFCC method outperforms traditional MFCC-based models, achieving 95.83% Accuracy for male speakers and 95.37% for female speakers, compared to 84.26% (male) and 93.98% (female) using MFCC alone. These findings highlight the effectiveness of TEO-enhanced speech features, improved feature representation, and robust performance across genders, along with efficient CNN-based classification. Limitations include the small, acted dataset, the focus on negative emotions, and limited generalizability to real-world stress.

Sugan et al. [28] proposed an innovative speech emotion recognition approach by introducing novel triangular filter banks based on the Bark and Equivalent Rectangular Bandwidth (ERB) frequency scales, resulting in Human Factor Cepstral Coefficients (HFCC). The study explores how these filter banks improve feature extraction, comparing them with conventional MFCC. Two emotional speech datasets are used: the Berlin Emo-DB [25] and SAVEE [29]. The extracted features undergo dimensionality reduction using the ReliefF algorithm, and classification is performed using an SVM for both speaker-dependent and speaker-independent recognition. Results indicate that the proposed HFCC features achieve competitive Accuracy, up to 86.96% for speaker-dependent and 77.08% for speaker-independent scenarios on Emo-DB. The study presents a novel filter bank design, effective dimensionality reduction, and thorough evaluation across datasets and conditions. Limitations include reliance on acted data, use of a single SVM classifier, and lower performance in speaker-independent settings.

Rahali et al. [30] investigated the effectiveness of different cepstral feature extraction techniques for automatic speech recognition (ASR) in noisy environments. They introduced Modified Human Factor Cepstral Coefficients (MHFCC), an improved version of HFCC that incorporates a gammachirp filter bank and prosodic features such as jitter and shimmer. The study compares MHFCC with MFCC and HFCC using the AURORA database [31], which contains speech data with various levels of impulsive noise. Hidden Markov Models (HMM) and GMM are employed for classification. Results indicate that MHFCC outperforms traditional MFCC and HFCC, achieving higher recognition accuracy, especially in low Signal-to-Noise Ratio (SNR) conditions. These findings demonstrate the robustness of the proposed MHFCC method and highlight its potential to improve speech-based technologies in real-world noisy environments.

Yerigeri et al. [32] introduced a speech-based stress recognition system using Semi-Eager (SemiE) learning, which balances the efficiency of eager learning with the adaptability of lazy learning. The key innovation lies in the application of perceptual speech features, including MFCC, Inverted MFCC (IMFCC), Revised Perceptual Linear Prediction (RPLP), and Bark and Gammatone Frequency Cepstral Coefficients (BFCC, GFCC). The system was trained and evaluated on the SUSAS database [8] and a newly created Marathi speech dataset. Experimental results show that the SemiE classifier achieved 90.66% Accuracy on SUSAS and 88.08% on the Marathi dataset, outperforming traditional classifiers. The study leverages rich perceptual features, benefits from low training and prediction costs via the SemiE approach, and introduces a regional language dataset to enhance applicability. Limitations include the use of acted speech, absence of deep learning comparisons, and restricted access to the small, non-public Marathi dataset.

Elbanna et al. [15] introduced BYOL-S, a self-supervised learning method for detecting emotional stress and other audio tasks by learning speech representations. It extends BYOL-A with training on speech-only subsets from large datasets like AudioSet [22] and LibriSpeech [33]. Evaluated on emotional stress detection (e.g., CREMA-D [34]), BYOL-S outperformed models like BYOL-A and Wav2Vec 2.0. A hybrid version using openSMILE[3] features improved accuracy and robustness further. Tested on the HEAR 2021 benchmark[4] across 16 tasks, it showed strong results in speech emotion recognition. BYOL-S generalizes well and benefits from hybrid training but can overfit with large encoders and is sensitive to hyperparameters. With a Convolutional vision Transformers (CvT) encoder, it offers a competitive, efficient framework for speech-based emotional stress detection.

The article by Hosain et al. [35] focuses on detecting emotions in speech using machine learning algorithms. It utilizes the CREMA-D [34] and TESS [36] datasets, which contain over 10,000 audio samples labeled with seven emotions: anger, disgust, fear, happiness, neutrality, pleasant surprise, and sadness. The research applies Empirical Mode Decomposition (EMD) for signal decomposition and extracts key features such as MFCC, GammaTone Cepstral Coefficients (GTCC), spectral descriptors, and harmonicity. The machine learning models used include SVM, ANN, Ensemble methods, and k-Nearest Neighbors (kNN). Among them, SVM achieved the highest test Accuracy of 67.7%, followed by ANN (63.3%), Ensemble (61.6%), and kNN (59.0%). The study offers comprehensive feature extraction, broad dataset coverage, and comparison across several machine learning models. However, its moderate accuracy, reliance on acted speech, and absence of deep learning benchmarks limit insight into current state-of-the-art performance.

D. SUMMARY

Table 1 provides a detailed comparison of individual research studies and approaches for detecting stress and stress-related emotions, as summarized in the previous sections.

Besides Prasetio et al. [10] and Duvvuri et al. [11], other authors also employ binary classification to distinguish between neutral and stressed recordings. Overall, reported Accuracy (Acc) values are high, often exceeding 95%. However, it should be noted that Accuracy is a reliable metric only for balanced datasets. For imbalanced datasets, the

[3]https://audeering.github.io/opensmile/
[4]https://hearbenchmark.com/

TABLE 1. Comparison of selected approaches for emotional stress detection.

Authors	Topic	Dataset	Classes	Best Approach	Performance	Strengths	Weaknesses
Sigmund, M. [4]	stress analysis	ExamStress [5]	neutral, stressed	monitoring changes in F_0, formants, and spectral coefficients	–	realistic stress induction; quantitative F_0 analysis; speaker-specific stress markers	limited speaker diversity; inaccessible dataset; simplified features; no automatic classification
Jena & Singh [6]	stress analysis	Exam Stress [6]	normal, stressed	STFT, STChT	–	real stress indicators; simple feature extraction methods; comparative analysis approach	small, inaccessible dataset; no automatic classification; no statistical validation
He et al. [7]	stress detection	SUSAS [8]	neutral, stressed	TEO-PWP + GMM	Acc = 95.95%	high accuracy; nonlinear speech analysis; critical band consideration	computational complexity; sensitivity to wavelet choice; limited generalization
Han et al. [9]	stress detection	Custom Dataset [9]	unstressed, stressed	LSTM-RNN + SVM	Acc = 66.40%	context-aware modeling; validated with cortisol levels; potential for real use	moderate accuracy; small and limited dataset; unclear boundary between classes
Prasetio et al. [10]	stress detection	SUSAS [8]	neutral, low, high	HHT-MFCC + ANN	Acc = 96.05%	better capture of stress-induced variations in speech; simple neural network architecture; validated on real stress data	limited speaker diversity; higher number of folds in validation; increased computational complexity
Duvvuri et al. [11]	stress detection	DAIC-WOZ [12]	mild, moderate, high	MFCC + k-fold CNN	Acc = 79.38% F1 = 64.69%	comprehensive feature and model comparison; realistic data; data augmentation for robustness	limited dataset size; low accuracy for high stress; risk of overfitting due to augmented data
Chen et al. [13]	stress detection	IFSO [13]	eustress, distress	mel-spectrogram + Stress Siamese ResNet	Acc = 75.10%	clearly labeled dataset; strong baseline model; reference speech enhances generalization	language-specific speech patterns; stress labels are self-assessed; tested in controlled lab conditions
Wu et al. [14]	stress detection	CREMA-D [34]	unstressed, stressed	BYOL-S/CvT + MLP	UAR = 82.43%	incorporation of speaker embeddings; cross-data generalization; short audio inputs (3-5 seconds)	performance drop when combining datasets; dependency on high-quality speaker embeddings
Ghose et al. [17]	multimodal stress detection	MultiAffect-Stress [17]	not stressed, stressed	multimodal fusion with Random Forest	F1 = 85.00%	multimodal integration; custom well-annotated dataset; comparison of multiple fusion strategies	strong reliance on sentiment analysis; relative small dataset; limited generalization
R.-González et al. [18]	speaker recognition under stress	VOCE Corpus [19]	neutral, stressed	MFCC + MLP	Acc = 99.45%	effective data augmentation; low computational complexity; extensive experimental evaluation	limited data size; simplified data augmentation; speaker identification focus only
Baird et al. [20]	cortisol prediction	TSST [21]	no stress, stress	eGeMAPS + LSTM-RNN	ρ = 0.77	multimodal approach; validated physiological markers; deep learning architecture	one cortisol label per session; incompatible cortisol scales across datasets; limited generalization
Vaikole et al. [26]	detection of stressed emotions	RAVDESS [24]	unstressed, stressed	MFCC + CNN	Acc = 94.33%	high accuracy; detection of stress through emotions; validated dataset	limited range of emotions; noise sensitivity; limited generalization; risk of overfitting
Nordin et al. [27]	detection of stressed emotions	RAVDESS [24]	neutral, sad, anger, disgust, fearful	TEO-MFCC + CNN	Acc = 95.60%	high accuracy; improved feature representation; gender-robust performance; efficient use of CNN	small and limited dataset; acted and negative emotions only; limited generalization
Sugan et al. [28]	emotion detection	Emo-DB [25]; SAVEE [29]	neutral, happy, boredom, anger, fearful, disgust, sad	TFBCC(MFCC-HFCC) + SVM	Acc = 86.96%	novel filter bank design; robust dimensionality reduction; comprehensive evaluation	only acted datasets; performance drops in speaker-independent settings; only SVM classifier
Yerigeri et al. [32]	emotion detection	SUSAS [8]; Marathi [32]	neutral, sad, anger, happy, surprise	spectral features + Semi-Eager Learning	Acc = 90.66% Acc = 88.08%	low training and prediction cost; rich perceptual feature set; regional language dataset created	acted speech used; no comparison with deep learning models; small, non-public Marathi dataset
Elbanna et al. [15]	emotion detection	CREMA-D [34]	neutral, happy, anger, disgust, fearful, sad	BYOL-S/CvT + MLP	Acc = 67.20%	general-purpose representations; high benchmark performance; efficient learning without labels	overfitting risk with complex encoders; window size sensitivity; high resource requirements
Hosain et al. [35]	emotion detection	CREMA-D [34]; TESS [36]	neutral, happy, surprise, anger, fearful, disgust, sad	MFCC-GTCC-spectral features + SVM	Acc = 61.01% Acc = 83.31%	comprehensive feature extraction; combination of datasets; multi-model evaluation	moderate accuracy; acted speech used; no comparison with deep learning models

F1-score is more appropriate. For instance, in the multimodal approach by Ghose et al. [17], which combines facial and vocal emotion recognition with sentiment analysis for stress detection, F1-scores reach approximately 85%. In the three-class stress classification presented by Duvvuri et al. [11], a case most similar to ours, the reported F1-score is only 65%.

Other approaches that focus on detecting specific emotions or stress-related emotional states generally achieve lower accuracy due to the increased complexity of multi-class classification. While researchers naturally aim to maximize performance through novel methods, results largely depend on the acoustic environment in which the dataset was recorded. The more controlled the environment, the higher the reported accuracy, yet the lower the generalizability of the approach in real-world deployment scenarios.

For these reasons, we chose to evaluate the selected approaches in our experimental section using datasets recorded in diverse acoustic environments, simulating a wide range of real-world conditions.

III. DATASETS

It is strongly recommended to use real-world (actual) data when training speech stress level detection systems. However, the legislative framework in many countries does not permit the use of such data, even for research purposes. Legal and privacy concerns regarding data leakage are significantly higher compared to other types of data. On the other hand, simulated databases tend to be highly exaggerated compared to real emotional stress. These databases are typically created by actors attempting to replicate stressful situations as accurately as possible. To mitigate the issue of exaggeration, part of the database is often recorded by non-native speakers.

Real emotional stress databases can be recorded either in a naturally stressful environment (e.g., emergency calls, high-pressure workplaces) or generated in laboratory conditions (e.g., cognitive load tasks, public speaking). The goal is to capture authentic physiological and psychological responses and extract natural changes in speech patterns under stress. In the case of simulated databases, ethical concerns are not an issue, and various speech parameters can be better controlled, allowing for the creation of more balanced datasets. However, the main drawbacks are lower authenticity, limited speaker variability, and the introduction of biases (such as stereotypical or exaggerated patterns).

A. STRESS DATASETS

In the following sections, we provide a detailed description of the speech-under-stress datasets used in the experimental part of this research. The statistics of the number of speech utterances for each dataset are summarized in Table 2.

1) CRISIS

The expressive speech database CRISIS [37] was created to study expressive and emotional speech under different levels of tense arousal. It contains recordings of phonetically rich, semantically neutral, and emotionally loaded sentences.

TABLE 2. The number of speech utterances in stress datasets.

Stress level	Stress datasets		
	CRISIS	StressDat	SUSAS
low	2,765	5,435	7,002
medium	2,765	3,722	3,621
high	2,765	3,680	4,033

The speech material includes warning messages and soothing texts, designed to represent different levels of emotional intensity. The dataset categorizes speech into six levels of tense arousal, ranging from extremely urgent (high arousal, level 3) to very calming (low arousal, level -3). The recordings were made in a studio environment using high-quality audio equipment (RODE K2 microphone, 48 kHz, 16-bit resolution). The speech material consists of short phrases to multi-sentence messages, with arousal levels systematically controlled. The recording method involves a three-step expressive variation, where each sentence is spoken in neutral, heightened, and extreme arousal levels. Similarly, for lower arousal, sentences are spoken in neutral, lower activation, and extremely calming tones. The dataset is emotionally structured but unbalanced, as it focuses on arousal rather than valence. The recorded sentences were manually segmented and designed to maintain expressivity consistency across different arousal levels. This dataset is valuable for speech emotion recognition, human-computer interaction, and emergency communication studies.

2) StressDat

The StressDat [38] is a Slovak speech database designed to study speech under stress. It was created to help improve automatic speech recognition and emotion detection in stressful situations. The dataset includes recordings from 30 professional actors (16 females, 14 males). Each speaker recorded speech under three stress levels: neutral, low, and high stress. The database contains 12 stress-inducing scenarios and four neutral scenarios for balance. Examples of stress scenarios include emergency landings, medical emergencies, and interpersonal conflicts. The recordings were made using actors' own smartphones in home environments. The dataset was annotated by five independent raters using a stress evaluation scale. Initial observations show a reasonable inter-annotator agreement and successful differentiation of stress levels. StressDat will be available for research purposes to advance recognition of speech under stress and related applications.

3) SUSAS

The Speech under Simulated and Actual Stress (SUSAS) dataset [8] was created for research in stress detection and speech recognition under stressful conditions. It contains both simulated and real-world stress speech recordings, making it unique for studying speech variability under pressure. The dataset includes over 16,000 utterances collected from 32 speakers (13 female, 19 male). The speech samples cover a

TABLE 3. A collection of available databases with speech under stress.

Name of dataset	Year	Language	Subjects	Classes		Character	Ref.
SUSAS	1997	English	32	4	neutral, angry, soft, fast	5 different tasks (actual/simulated)	[8]
Soldier of the Quarter	2002	English	6	2	no stress, stress	6 militarily-relevant questions	[40]
Driver's Speech u/Stress	2003	English	4	4	fast/slow driving/response	driver's behavior	[41]
Emo-DB	2005	German	10	4	no stress, normal, strong, emphatic	sentences with different emotions	[25]
ExamStress	2006	Czech	31	3	low, medium, high	final state examinations	[5]
UTDrive	2009	English	68	2	neutral, negative	driver's behavior	[42]
CRISIS	2014	Slovak	4	6	low, medium, high (low/high arousal)	dangerous and critical situations	[37]
DAIC-WOZ	2014	English	140	2	no distress, distress	distress analysis interviews	[12]
Flight Programme	2014	German	8	2	no stress, stress	cockpit communication	[43]
VOCE Corpus	2014	Portugal	45	2	no stress, stress	stressful public speaking situations	[19]
SUSSC	2016	Hindi	15	4	neutral, angry, sad, Lombard	119 word vocabulary	[44]
Multimod. Korean Stress	2018	Korean	91	2	no stress, stress	interviews in English	[9]
MuSE	2020	English	28	2	no stress, stress	final exams	[45]
StressDat	2021	Slovak	30	3	low, medium, high	12 different stressful situations	[38]
ULM-TSST	2021	German	110	2	valence, arousal	highly stress-induced speech task	[21]
I Feel Stressed Out	2022	Mandarin	87	2	eustress, distress	retelling the movie plot	[13]
StressID	2023	English	65	3	relaxed, neutral, stressed	11 different cognitive tasks	[46]
BESST	2024	Czech	79	3	low, moderate, high (PSS14)	hand immersion/reading span tasks	[47]
MultiAffectStress	2024	English	100	2	no stress, stress	interviews - recalling stressful life	[17]
A Multiling. Database of Natur. Stress Emotion	2012	English Mandarin Cantonese	61 42 69	2	no stress, stress	common questionnaire	[48]

range of stress conditions, such as slow, fast, soft, loud, clear, angry, and psychiatric analysis speech data (speech under depression, fear, and anxiety). The recordings were gathered from various real and simulated environments, including military and aviation settings. The dataset is partially balanced in terms of speakers but varies across different stress conditions. The utterances are primarily short phrases and isolated words, making them suitable for real-time stress detection applications. The recordings are available in 16-bit, 8 kHz WAV format, optimized for speech recognition systems. SUSAS is widely used in aviation safety, military communication, and emotion recognition research.

Furthermore, in Table 3, we summarize 19 monolingual datasets and one multilingual dataset of speech under stress that we identified in the literature. The table includes information on the language, content, focus of the dataset, and the number and names of the stress level categories it contains. Most of the listed datasets are either unavailable or accessible only upon request or after meeting the conditions specified in a license agreement. Additionally, a brief summary of six other stress datasets is provided in Reddy et al. [39].

B. EMOTION DATASETS

A wide range of emotional speech datasets is available, a detailed overview of which can be found, for example, in [49], [50], and [51]. These resources can be used to expand the current collection of stress-related speech datasets and help create large-scale emotional stress datasets by effectively mapping emotions to corresponding stress levels. For this research, we used three well-known English datasets, RAVDESS, SAVEE, and TESS, which were also mentioned in the literature review in the context of other research, and the SUS dataset spoken in Slovak, which contained a mixture of different, mostly negative emotions. In the following paragraphs, we will briefly describe these datasets.

1) RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [24] was created for research in speech and music-based emotion recognition. It contains both speech and singing recordings, making it unique among emotional speech datasets. The dataset includes recordings from 24 actors, ensuring gender balance. Each actor performed two spoken statements and two sung phrases in eight emotional categories. The emotions included are neutral, happy, sad, angry, fearful, disgusted, surprised, and calm. The dataset consists of 7,356 recordings, with variations in intensity (normal vs. strong). The files are provided in 16-bit, 48 kHz WAV format, ensuring high audio quality. The duration of each recording varies, with speech clips typically lasting a few seconds and singing samples being slightly longer. RAVDESS is widely used in machine learning and multimodal emotion recognition research. Due to its balanced speaker distribution and inclusion of both speech and song, it is a valuable resource for emotion analysis.

2) SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset [29] was created for research in speech-based and facial emotion recognition. It contains both audio and visual recordings of emotional speech expressions. The dataset includes 480 utterances, spoken by four male actors, all with British English accents. Each speaker performed sentences in seven emotional categories: neutral, happy, sad, angry, fearful, disgusted, and surprised. The sentences were selected from TIMIT [52]. The recordings are available in WAV format (audio) and AVI format (video), ensuring multimodal

emotion analysis. Since it contains only male speakers, the dataset is not fully balanced in terms of gender diversity. The duration of each recording is typically a few seconds, depending on the spoken sentence. SAVEE is commonly used in machine learning and human-computer interaction studies. It is particularly valuable for developing and testing multimodal emotion recognition systems.

3) SUS

The SUS dataset [53] was created for speaker and emotion recognition in Slovak. It consists of emotional audio recordings extracted from a popular Courtroom TV series (''Súdna sieň'' in Slovak). The dataset contains approximately 2,000 utterances from 7 speakers (3 male, 4 female). Since the speech is performed by non-professional actors, it is categorized as an induced emotion dataset. The emotions included are neutral, curiosity, anger, sadness, aggressiveness, and disgust, with few positive emotions. The audio was downsampled to 16 kHz WAV format from the original 48 kHz MPEG stream. Each utterance was manually labeled based on the dominant emotion, with segmentation for mixed-emotion sentences. The average duration of each recording is 5 to 6 seconds. The data set is unbalanced, as it contains mainly neutral and negative emotions. SUS is valuable for speech-based emotion recognition and speaker identification in forensic and linguistic research.

4) TESS

The Toronto Emotional Speech Set (TESS) [36] was designed to study emotional speech recognition. It was created to examine how different emotions affect speech perception, particularly in aging and cognitive research. The dataset consists of 2800 audio speech samples. These recordings were produced by two female speakers, aged 26 and 64. Each speaker read 200 unique sentences, based on the phrase ''*Say the word*...'', followed by a target word. The dataset includes seven emotions: neutral, happy, sad, angry, fearful, disgusted, and surprised. The recordings are available in WAV format, with high-quality 16-bit, 44.1 kHz sampling. Since it features only two speakers, the dataset is not fully balanced in terms of speaker diversity. The duration of individual recordings varies but is typically a few seconds per sample. TESS is widely used in research on speech processing, emotion detection, and machine learning applications.

C. ANALYSIS OF SPEECH IN STRESS DATASETS

For a better understanding of the nature of speech under stress and its effect on the acoustic characteristics of speech, we decided to analyze these characteristics in more detail in the available datasets for individual stress levels.

In previous studies [1], [54], [55], some acoustic correlates of stress in speech were identified, from which the following conclusions emerged:

- A large number of studies confirm an increase in fundamental frequency (F_0) due to heightened stress or cognitive and physical load. In real-life, particularly in dangerous or emergency situations, a significantly greater increase in F_0 is observed compared to laboratory conditions, where the increase is typically lower.
- An increase in voice intensity (amplitude or energy) with increasing stress was also observed, especially in the higher frequency range above 1000 Hz.
- Significant increase in speech rate (speech tempo) in case of stressed speech was observed.
- The number and duration of pauses decreased with stress, but hesitations increased.
- Certain studies have observed a reduction in jitter under different stress conditions, while others found that jitter increased, even in cognitive stress situations and real-life stress scenarios. Some research, however, noted no significant correlation between jitter and stress.
- A moderate decrease in amplitude variations (shimmer) was observed with stress.
- Relatively little attention has been focused on formant values, with minor changes observed only in formants F_1 and F_2. The formant F_3 was not significantly affected by stress. For men, the change in formant values as a function of stress was insignificant.
- In terms of voice quality, the following observations were recorded: increased noise in speech (affrication and aspiration); increased tension in speech (tense voice); presence of low frequency vibration during stress conditions; modifications in articulation of vowels and consonants; voiced segments were devoiced; significant modifications in pitch, rising slope, and closing slope of the glottal pulse were observed; irregular respiration patterns (irregularity and variability of voicing) was present; and, in addition, female voices were more breathy and strained under stress.

In the following, we present typical values of acoustic parameters for a neutral voice [56]:

- *fundamental frequency* (F_0): 85–180 Hz for males and 165–255 Hz for females;
- *formant frequencies*:
  - 300–1000 Hz for F_1;
  - 850–3000 Hz for F_2;
  - and 1700–3700 Hz for F_3.
- *voice intensity*: 60–70 dB SPL (Sound Pressure Level);
- *harmonic-to-noise ratio* (*HNR*): 15–20 dB.
- *jitter*: $\approx$ 0.2–1%;
- *shimmer*: $\approx$ 2–5%.

Due to increased stress, the acoustic characteristics can then change values according to the literature [1] as follows:

- *fundamental frequency* (F_0): may increase by 30–60 Hz or more, depending on the intensity of the stress;
- *formant frequencies*:
  - F_1 may increase by 20–100 Hz;
  - F_2 may increase or decrease by 50–200 Hz;
  - and F_3 may slightly increase or decrease, typically by no more than 50 Hz.
- *voice intensity*: typically increases to 75–85 dB SPL or more, however, when anxious or afraid, the voice may be quieter (around 50–60 dB SPL);

TABLE 4. Extracted speech characteristics across stress datasets (averaged values).

Speech characteristic	Stress level	Stress datasets			
		CRISIS	StressDat	SUSAS actual	SUSAS simulated
mean F_0 [Hz]	low	169.77	161.65	147.06	118.47
	medium	197.23	183.68	147.17	148.20
	high	217.80	198.44	165.84	148.92
pitch range [Hz]	low	135.91	135.36	48.51	44.49
	medium	147.06	146.83	55.63	83.29
	high	164.34	166.44	70.82	70.26
formant F_1 [Hz]	low	740.56	628.57	631.19	492.03
	medium	741.70	635.57	638.54	508.82
	high	771.04	646.86	610.13	522.57
formant F_2 [Hz]	low	1785.52	1709.13	1577.52	1378.37
	medium	1774.71	1708.21	1550.34	1388.79
	high	1781.46	1704.81	1523.65	1403.83
formant F_3 [Hz]	low	2804.49	2717.17	2421.72	2171.04
	medium	2785.65	2699.82	2384.35	2178.79
	high	2777.82	2698.04	2336.19	2187.61
average vocal intensity [dB]	low	55.69	67.98	73.31	77.50
	medium	64.29	72.57	75.18	78.98
	high	70.31	74.93	76.97	79.63
HNR [dB]	low	11.64	10.46	11.06	11.20
	medium	11.52	9.79	9.97	10.28
	high	12.09	9.38	8.89	9.65
jitter [%]	low	2.27	2.48	1.92	2.24
	medium	1.94	2.47	1.91	2.31
	high	1.77	2.53	2.15	2.47
shimmer [%]	low	9.23	10.69	6.95	7.90
	medium	9.14	11.36	6.79	7.51
	high	9.19	11.90	7.88	7.49

- *harmonic-to-noise ratio* (*HNR*): decrease, often below 15 dB, depending on the intensity of the stress;
- *jitter*: may exceed 1 %;
- *shimmer*: may exceed 5 %.

We performed an analysis of basic speech characteristics across stress levels in the CRISIS, StressDat, and SUSAS datasets using the Praat tool.[5] When we examine the measured values of mean F_0, voice intensity, formant frequencies (F_1, F_2, and F_3), jitter, shimmer, and HNR (see Table 4), we can draw the following conclusions: The fundamental frequency F_0 and voice intensity are good predictors of stress levels, with both increasing as the stress level rises. With the exception of the simulated part of the SUSAS dataset, individual stress levels are clearly distinguished in the other datasets. In contrast, parameters such as formant frequencies, jitter, shimmer, and HNR show contradictory results, which also supports the findings of the aforementioned research.

IV. METHODS

A. FEATURE-BASED APPROACHES

The process of stress level classification using speech-based feature extraction typically involves three main steps (see Figure 1). First, the input speech is preprocessed and transformed into meaningful acoustic features. Preprocessing includes downsampling to 16 kHz and loudness normalization. The extracted features are then fed into a machine learning model, such as kNN, SVM, or a deep learning architecture, which learns patterns associated with different stress levels. Finally, the trained model classifies the speech input into predefined stress categories (*low*, *medium*, and *high*), based on statistical and temporal variations in the extracted features.

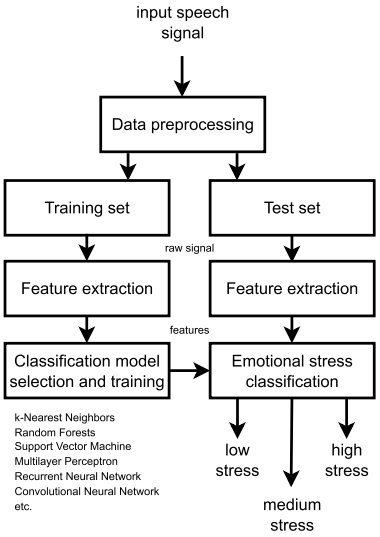

FIGURE 1. The block diagram of feature extraction, classifier training, and evaluation.

Feature extraction is crucial for detecting emotional stress in speech. Different methods emphasize various spectral and temporal characteristics of the speech signal. In this study,

[5]https://www.fon.hum.uva.nl/praat/

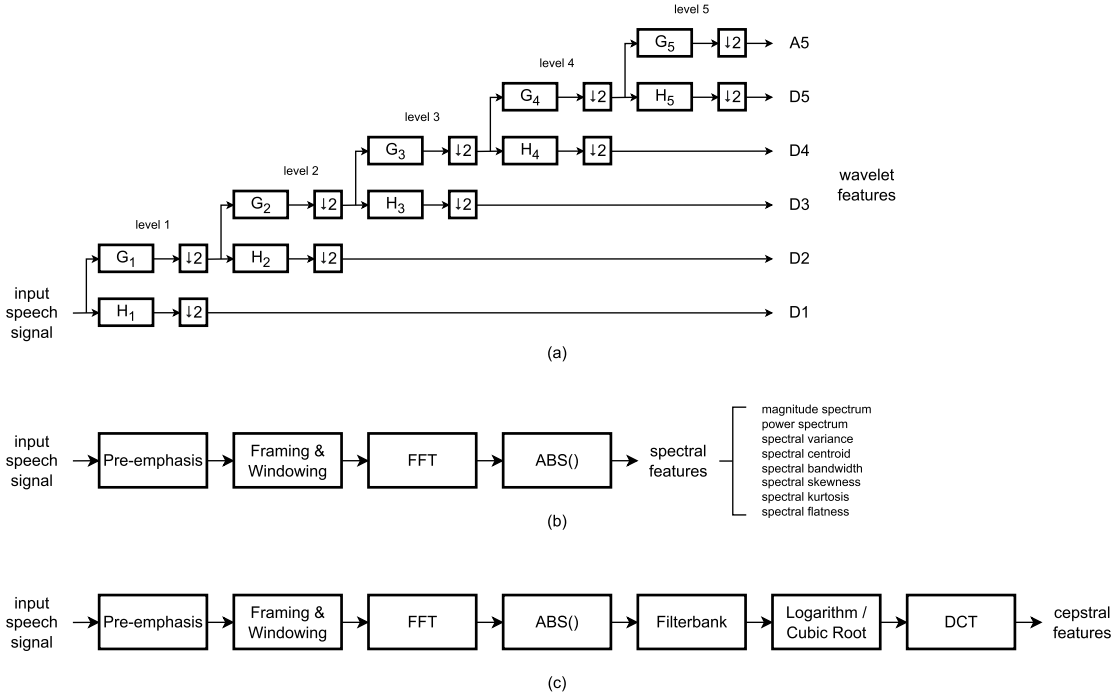

FIGURE 2. Block diagrams illustrating the extraction process of different features: (a) wavelet features; (b) spectral features; (c) cepstral features.

we compare several approaches based on wavelet, spectral, and cepstral feature extraction. Key differences among these techniques are outlined in the following sections.

1) SIGNAL DECOMPOSITION WITH DISCRETE WAVELET TRANSFORM

The Discrete Wavelet Transform (DWT) decomposes signals into multiple frequency bands using wavelet functions, providing better time resolution for high frequencies and better frequency resolution for low frequencies. It emphasizes transient features such as sudden energy changes, rapid pitch fluctuations, and variations in speech rate, which are key indicators of stress. DWT suppresses stable harmonic structures and stationary components, making it well-suited for detecting stress-induced changes in speech. DWT captures both short-term variations and longer-term trends such as vocal tension or breathiness. It is robust to noise in higher frequencies but sensitive to the choice of the wavelet basis [8].

When decomposing the speech signal using DWT (see Fig. 2a), standard low-pass and high-pass filtering, followed by decimation by a factor of 2, was applied to the input speech signal. This was followed by a five-level decomposition into approximation (A5) and detail coefficients (D1–D5). The Daubechies wavelet of order 4 (db4) was used as the mother wavelet, as it provides better performance than other wavelet types (such as Haar, Morlet, etc.) by capturing short-term variations in speech, sudden changes in energy, and offering improved spectral resolution compared to lower-order Daubechies wavelets (e.g., db2). The result is a vector containing six wavelet coefficients (D1–D5 and A5).

2) TIME-FREQUENCY ANALYSIS USING SHORT-TIME FOURIER TRANSFORM

The Short-Time Fourier Transform (STFT) analyzes speech by decomposing it into time-frequency components using a sliding window and the FFT (see Fig. 2b). It provides broad spectral information, emphasizing harmonic structure and fundamental frequency, but its fixed window size imposes a trade-off between time and frequency resolution. STFT is effective for detecting stress-induced variations such as increased energy in high-frequency bands and greater fluctuations in F_0. However, it is limited in capturing fine temporal details and transient aspects of emotional stress. Despite this limitation, STFT reliably tracks changes in pitch, intensity, and speaking rate associated with vocal tension [6]. In addition to standard spectral features such as the magnitude and power spectra, numerous stress-related low-level descriptors can also be extracted from the spectrum.

During the data preprocessing step, pre-emphasis was applied to the input speech signal. To calculate the spectral coefficients, the speech recordings were first divided into short segments of 30 ms duration. A windowing function – a Hamming window – was then applied to these segments with 50% overlap. Finally, the signal was transformed from the time domain to the frequency domain using the FFT, and the magnitude spectrum for each speech segment was obtained by taking the absolute value. As mentioned earlier,

in addition to the magnitude spectrum, we also extracted the power spectrum and six stress-related spectral descriptors: spectral variance, spectral centroid, spectral bandwidth, spectral skewness, spectral kurtosis, and spectral flatness.

3) FILTER BANK-BASED CEPSTRAL PARAMETRIZATION

Cepstral features are widely used in speech processing because they provide a compact representation of the spectral envelope, capturing vocal tract characteristics that are important for detecting stress in speech. The cepstrum is derived from the spectrum of a signal and is typically defined as the inverse magnitude spectrum of the log-magnitude spectrum. Rather than using the raw spectrum directly, a set of band-pass filters is applied, modeling the frequency sensitivity of the Human Auditory System (HAS) by emphasizing perceptually important frequency bands. A DCT is then applied to the logarithmic filter bank energies to decorrelate the features and produce the cepstral coefficients (see Fig. 2c). The cubic root function may also be used instead of the logarithm. Depending on the filter bank, different types of cepstral coefficients can be distinguished, each with specific properties:

- *Mel-Frequency Cepstral Coefficients* (MFCC) extract speech features by applying a mel-scale filter bank to the power spectrum, followed by a DCT to obtain compact feature vectors (13 coefficients). This method simulates human auditory perception by prioritizing perceptually relevant frequency components while discarding phase information. MFCC emphasizes formants, vocal tract features, and low-frequency components, making it useful for capturing stress-related pitch variations, shifts in formants, and changes in energy distribution [18], [26]. However, it suppresses high-frequency details and transient information, which may limit its ability to detect finer stress-induced variations. While widely used for its robustness, MFCC is sensitive to noise and may degrade in noisy environments.
- *Inverted MFCC* (IMFCC) is similar to MFCC but reverses the mel filter bank to emphasize high-frequency details rather than low-frequency components. It preserves high-frequency speech features, making it useful for detecting stress-induced changes in pitch, articulation, and spectral tilt. IMFCC captures sibilant and fricative sounds, which can be affected by stress, while suppressing low-frequency harmonics and formants. This makes it effective for highlighting voice roughness and other high-frequency stress cues that standard MFCC might overlook [57]. As a result, IMFCC provides complementary information to MFCC in stress detection.
- *Linear Frequency Cepstral Coefficients* (LFCC) are similar to MFCC but use a linear frequency scale instead of the mel scale, ensuring equal emphasis across all frequency bands. This method preserves more detailed frequency information, particularly at higher frequencies, making it useful for detecting stress-induced spectral shifts. Unlike MFCC, which prioritizes lower frequencies, LFCC provides a more uniform spectral representation while suppressing psychoacoustic perception effects. It is effective in capturing high-frequency stress-related features such as pitch variations, breathiness, and turbulence in the voice [58].
- *Bark Frequency Cepstral Coefficients* (BFCC) use the Bark scale, which aligns more closely with human auditory perception than the mel scale, to filter the spectrum before applying the DCT. The Bark filter bank divides the frequency spectrum into critical bands, matching the ear's frequency resolution and emphasizing speech components that contribute most to intelligibility. BFCC is effective for detecting stress-induced speech changes compared to MFCC, such as variations in pitch, speech tempo, formant shifts, and vocal tension, particularly in the lower- and middle-frequency ranges [32]. It offers better stress discrimination than MFCC and is more robust to noise than both MFCC and LFCC.
- *Human Factor Cepstral Coefficients* (HFCC) are a variation of MFCC that incorporate auditory-inspired filtering models based on human frequency perception, such as the Bark and Equivalent Rectangular Bandwidth (ERB) scales. By using customizable triangular filter banks, HFCC provides a more accurate representation of frequency regions crucial for human hearing. This method enhances the detection of stress-induced articulatory changes, capturing emotional cues related to pitch, intensity, formants, and timbre [28]. HFCC offers a balance between MFCC and LFCC, making it effective for emotional stress recognition, and is generally more robust to noise than MFCC. Its psychoacoustic principles make it particularly suited for capturing perception-based stress features.
- *Modified HFCC* (MHFCC) are an improved version of HFCC, incorporating a gammachirp filter bank and prosodic features such as jitter and shimmer. This enhancement provides a better dynamic response to speech variations, capturing both low and high frequencies with varying frequency resolution, particularly in stress-relevant regions. MHFCC emphasizes subtle spectral variations linked to stress while suppressing stable, non-stress-sensitive frequencies. It is effective in detecting rapid changes in pitch, tone, and modulation associated with emotional speech [30], offering finer tracking of frequency modulations caused by stress. Additionally, MHFCC is more robust to noise than standard HFCC, improving its accuracy in real-world applications.
- *Gammatone Frequency Cepstral Coefficients* (GFCC) use a gammatone filter bank, which models the HAS more accurately than mel or Bark filters. This filter bank mimics the frequency analysis performed by the human ear, with frequencies distributed non-uniformly—denser at lower frequencies and sparser at higher ones. GFCC enhances perceptual accuracy by better representing

cochlear filtering effects and emphasizing frequencies critical for human hearing. It effectively captures stress-related acoustic shifts, including formants, harmonics, and intonation, and is particularly resilient to noise [11], making it suitable for stress detection in noisy environments. GFCC outperforms MFCC in detecting subtle speech variations induced by stress.

- *GammaTone Cepstral Coefficients* (GTCC) ae an advanced version of GFCC, using gammatone filters and applying a cubic root function instead of a logarithmic function to the filtered coefficients for improved frequency selectivity. This method focuses on the critical-band structure of hearing, enhancing the perceptual relevance of the extracted features. GTCC emphasizes fine-grained auditory spectral details while suppressing harmonics and steady-state components [59]. It is more sensitive to stress-induced vocal changes and is commonly used in noise-robust speech processing, making it highly effective for real-world stress detection scenarios. GTCC captures subtle emotional vrariations in the speaker's voice and is more robust to noise than MFCC or GFCC [35]. Unlike the previous approaches, the GTCC algorithm outputs 20 cepstral coefficients.

In summary, the choice of the feature-based extraction method depends on which stress-related speech properties need to be analyzed. MFCC and HFCC target formant structure, while IMFCC and LFCC emphasize high-frequency details. DWT captures transient stress-related fluctuations, and spectral features reveal key changes in energy distribution. GTCC and GFCC simulate auditory perception. Selecting the appropriate method is therefore crucial for effectively capturing the specific acoustic cues associated with stress.

B. TRANSFER LEARNING

Transfer learning with pre-trained models based on CNNs, such as VGGish or YAMNet, utilizes the already acquired knowledge of these models to perform new tasks without the need to train them from scratch.

In this case, the input audio signal is first converted into a logarithmic mel-spectrogram. The preprocessing step involves resampling the audio to 16 kHz, applying a one-sided STFT with a Hann window, extracting magnitude spectra, mapping them to a 64-band mel scale, converting them to a logarithmic scale, and buffering them into overlapping 96×64 log-scale mel-spectrogram segments. These segments serve as the input for fine-tuning one of the models or for directly recognizing predefined audio events.

The fine-tuned model can then be used for feature extraction, where vector representations (embeddings) are obtained and subsequently used as input for a classifier (as in the case of VGGish). Alternatively, it can be used directly for classification (as with YAMNet). Fine-tuning is performed by partially or fully adapting the model to new data, updating its weights to better suit a specific task or domain. This approach is less computationally demanding and allows for high accuracy even with a smaller amount of labeled data, compared to training a model from scratch.

The following settings were used during the fine-tuning process of both the VGGish and YAMNet models: the *Adam optimizer* with an *initial learning rate* of 1e-4, a *mini-batch size* of 128, *shuffling of the training data* at each epoch, and a *maximum of 5 epochs*. It should be noted that a higher number of epochs did not bring further improvement.

In the following section, we briefly introduce both the VGGish and YAMNet models. As mentioned earlier, while VGGish is designed for extracting vector representations from input audio, the YAMNet model is suitable for the direct recognition of sound/audio events.

1) VGGish

VGGish represents a CNN model based on the VGG (Visual Geometry Group) architecture, designed for the extraction of general-purpose audio features [60]. It transforms audio waveforms into logarithmic mel spectrograms and processes them through convolutional layers to generate embeddings. These embeddings capture high-level acoustic features, which can be used to analyze speech patterns and detect stress-related vocal characteristics. In emotional stress detection, VGGish can serve as a feature extractor for machine learning models that classify stress levels based on voice modulation, pitch, and intensity. Its pre-trained embeddings enable transfer learning without requiring large labeled datasets.

The VGGish network consists of 24 layers: an input layer (identical to that of YAMNet), four convolutional blocks (each comprising a convolutional layer, ReLU activation, and MaxPooling), and three fully connected blocks (each comprising a fully connected layer followed by a ReLU activation). There are nine layers with learnable weights – six convolutional and three fully connected. As a result, VGGish transforms audio input features into a semantically meaningful, high-level 128-dimensional embedding, which can then be used as input to a classification layer. As mentioned earlier, VGGish serves solely as a feature extractor. When fine-tuning the network, we incorporated a Softmax layer as the classification layer by adding the final three layers: a fully connected layer, a Softmax layer, and an output layer.

2) YAMNet

Yet Another Mobile Network (YAMNet) is a pre-trained audio event detection model based on the efficient CNN MobileNetV1 [61], trained to recognize 521 audio categories from AudioSet [22]. It processes speech recordings and identifies relevant acoustic events, such as trembling voices, heavy breathing, or changes in tone, which are indicative of emotional stress. By analyzing the scores of the output class, researchers can correlate stress-related audio patterns with emotional states. The lightweight architecture of YAMNet allows for real-time analysis, making it suitable for applications such as stress monitoring in call centers or mental

health assessments. It can be combined with other models to improve detection accuracy.

The YAMNet network consists of 86 layers: an input layer (96 × 64 log-scale mel spectrograms), a convolutional block (convolution, batch normalization, ReLU activation[6]), followed by a series of 13 depthwise separable convolutional blocks. Each of these blocks includes six layers: grouped convolution, batch normalization, ReLU activation, convolution, batch normalization, and ReLU activation. This results in a 3 × 2 array of activations for 1,024 kernels, which are then averaged to produce a 1, 024-dimensional embedding. This embedding is passed through a Softmax layer serving as the classification layer, followed by the output layer.

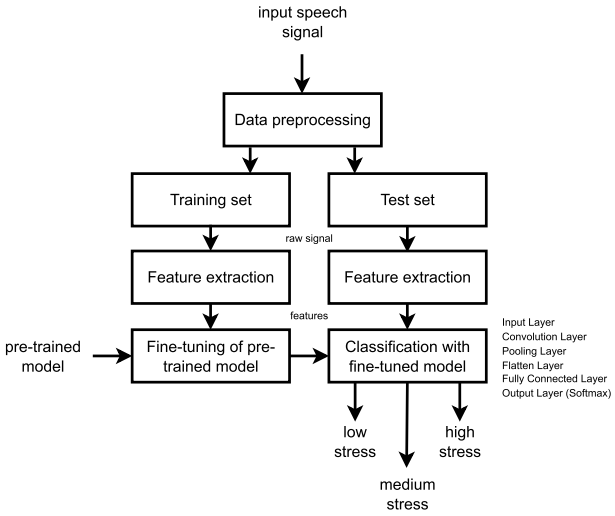

FIGURE 3. The block diagram of transfer learning with pre-trained models.

C. SELF-SUPERVISED LEARNING

Self-supervised learning (SSL) is a machine learning technique where a model learns to extract useful representations of data without the need for manually labeled examples. The model generates pseudo-labels from the raw input audio data and learns to predict missing or transformed parts of the data. In this context, SSL enables models trained on general speech to better understand specific speech patterns associated with stress and emotions by adapting their representations to this domain, thereby reducing the need for large labeled data.

For a deeper understanding of SSL algorithms, it should be noted that a comprehensive summary of various algorithms, their applications, and emerging trends can be found, for example, in [62]. Furthermore, the application of SLL in biomedical signal processing is discussed in [63].

In this research, we focused on the application of the BYOL-S method to generate speech representations for subsequent classification of the stress level using the SVM algorithm and fine-tuning the pre-trained Wav2Vec 2.0 model using existing datasets of speech under emotional stress.

1) BYOL-S

Bootstrap Your Own Latent for Speech (BYOL-S) [15] is a SSL method for audio representation learning, inspired by the BYOL framework from computer vision. It trains an audio model without labeled data by using two networks – a predictor and a target network – that learn to generate similar embeddings for augmented versions of the same audio input. Unlike contrastive methods, BYOL-S does not require negative samples, making it more data-efficient. The learned representations capture meaningful speech features, which can be used for tasks like speaker identification or emotion detection. BYOL-S is particularly useful in scenarios where labeled speech data is scarce or expensive to obtain.

The input to the BYOL-S algorithm comprised raw audio recordings, which were initially processed into 64-band log-scaled mel spectrograms. Spectrograms were derived from speech recordings constrained to the frequency range of 60 to 7800 Hz, with a sampling rate of 16 kHz, a window length of 25 ms, and a hop size of 10 ms, consistent with the configurations of the VGGish and YAMNet models. The input data were subsequently augmented through the application of the *MixUp*, *pitch-shifting*, and *time-stretching* techniques.

As was said before, the BYOL-S architecture consists of two networks: an online network and a target network. Each network incorporates an encoder and a projection head, with differing weight settings. The encoder extracts speech representations from the augmented input, while the projection head maps these representations into a low-dimensional latent space. Both networks are comprised of three convolutional blocks (each containing convolution, batch normalization, ReLU activation, and max pooling), followed by two fully connected layers that project the output into a final 2, 048-dimensional embedding.

The objective of the online branch is to predict the representation that the target network generates for the same signal, with the loss measured via, for example, cosine similarity. This process involves training the online network such that its output (post-prediction) closely aligns with the output generated by the target network.

The BYOL-S approach employs three distinct types of pre-trained models, each differing in the architecture of the encoder. These include a standard CNN model, as well as two variants: a convolutional LSTM network (ResNetish) and a convolutional visual transformer (CvT). Fine-tuning of these pre-trained models is performed by freezing all convolutional blocks and continuing training within the BYOL architecture, while fine-tuning the classification layer to accommodate new labels. The final output, represented as a 2, 048-dimensional embedding, is subsequently passed to the classifier.

[6]Rectified Linear Unit (ReLU) is a popular activation functions used in neural networks, especially in deep learning models.

For this research, we have used the following BYOL-S models, available in the GitHub repository[7]:

- *BYOL-S/AudioNTT model* - uses a lightweight CNN architecture designed for speech representation learning;
- *BYOL-S/ResNetish-34 model* - based on a Convolutional LSTM network ResNetish-34, originally adapted from image processing to handle audio spectrograms;
- *Hybrid BYOL-S/CvT model* - combines Convolutional vision Transformers (CvT) with BYOL-S to leverage self-attention mechanisms for capturing long-range dependencies in audio [16].

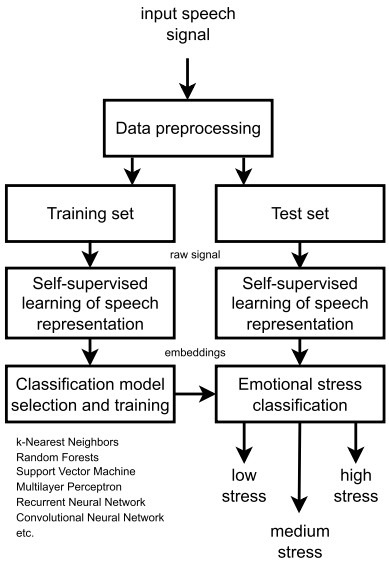

FIGURE 4. The block diagram of self-supervised learning at the feature extraction level.

2) Wav2Vec 2.0

Wav2Vec 2.0 [64] is a SSL model originally designed for automatic speech recognition (ASR) by learning representations directly from raw audio waveforms. It consists of a convolutional feature encoder that extracts speech features, followed by a transformer-based context network that learns temporal dependencies. The model is trained by masking portions of the audio input and predicting missing segments, enabling it to learn rich speech representations without labeled transcriptions. Fine-tuning on small labeled datasets allows Wav2Vec 2.0 to achieve state-of-the-art ASR performance with minimal supervision. Its ability to learn from unlabeled speech data makes it highly effective for domain adaptation, e.g. for speech-based stress detection.

The advantage of the Wav2Vec 2.0 approach is that it extracts embeddings directly from the raw speech, eliminating the need to generate log-scaled mel spectrograms. In this case, we chose the pretrained model as a baseline, retaining its speech representation extractor. We partially unfroze the Wav2Vec 2.0 encoder and added a classification head consisting of a fully connected (dense) layer with three outputs and a Softmax activation function. The input speech data were preprocessed by converting them to the correct sampling frequency (16 kHz) and normalizing the amplitudes. The models were then further trained using a *cross-entropy loss* function, suitable for classification tasks. The weights were optimized using the *Adam optimizer*, the *number of epochs* was set to 5, and the *learning rate* to 1e-5.

We compared the effectiveness of three different Wav2Vec 2.0 models available in the HuggingFace repository:

- *Wav2Vec 2.0 English base model*[8] - base model pretrained on a subset of 24.1k unlabeled English data from the VoxPopuli dataset;
- *fine-tuned Wav2Vec 2.0 English base model*[9] - base model pretrained on the subset of 10k unlabeled English data from the VoxPopuli dataset, fine-tuned with the transcribed data;
- *multilingual Wav2Vec 2.0 base model*[10] - large-scale multilingual base model pretrained on 436k hours of unlabeled speech in 128 languages.

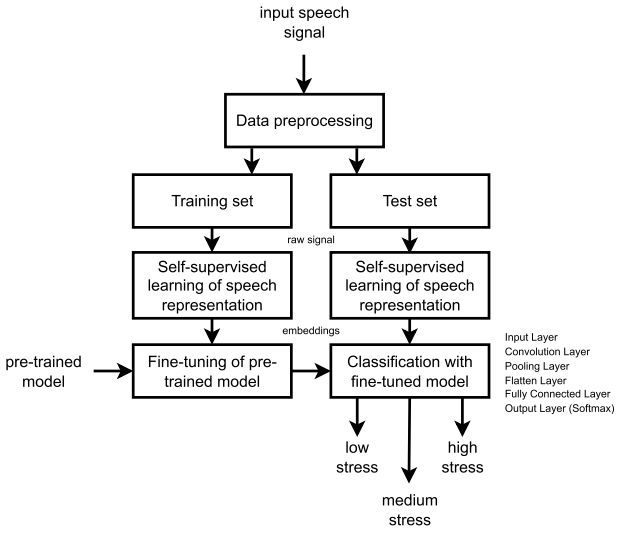

FIGURE 5. The block diagram of self-supervised learning with pre-trained models.

D. CLASSIFICATION

As part of this research, we tested a number of classifiers, ranging from standard Naïve Bayes, or k-Nearest Neighbors, through Decision Tress and Ensemble Methods, up to Artificial Neural Networks. Among all the algorithms we tested, we achieved the best results with classifiers based on k-Nearest Neighbors, Bagged Trees, and Support Vector Machines. In the following, we will briefly introduce each of them, along with their optimal parameter settings [65]:

- *Weighted k-Nearest Neighbors* (*kNN*) classifies a data point based on the majority class of its k nearest neighbors, but assigns higher weights to closer neighbors.

[7]https://github.com/GasserElbanna/serab-byols

[8]https://huggingface.co/facebook/wav2vec2-base-en-voxpopuli-v2

[9]https://huggingface.co/facebook/wav2vec2-base-10k-voxpopuli-ft-en

[10]https://huggingface.co/facebook/wav2vec2-xls-r-300m

The weight is typically the inverse of the distance, meaning closer neighbors have a stronger influence on the classification. This approach improves accuracy by giving more importance to relevant points while reducing the impact of distant, potentially misleading neighbors. In our setup, the number of neighbors was set to 10.

- *Fine kNN* uses a small k value, typically 1, making it highly sensitive to local variations in data. It classifies a test point by assigning the most frequent class among its closest neighbors, often using Euclidean distance. Due to its fine granularity, it captures detailed decision boundaries but may be prone to overfitting. Fine kNN is effective in datasets with well-separated clusters but requires careful tuning for stability.
- *Subspace kNN* improves the traditional kNN classifier by using feature subspaces, where multiple kNN classifiers are trained on randomly selected subsets of features. Each classifier makes a prediction, and the final decision is obtained by aggregating their outputs, often using majority voting. This method enhances robustness and reduces the impact of irrelevant or redundant features, improving classification accuracy. In our setup, the ensemble consists of 30 learners, each trained on a randomly selected subspace of 3 features, which helps to balance model diversity and computational efficiency.
- *Bagged Trees* (or *Bootstrap Aggregated Trees*) are an ensemble learning method that builds multiple decision trees on different bootstrap samples of the training data. The final classification is obtained by majority voting across all trees. This approach reduces overfitting and variance, leading to a more stable and accurate model. In this case, the number of learners was also set to 30.
- *Support Vector Machine* (SVM) classifies data by finding the optimal hyperplane that maximizes the margin between different classes in a high-dimensional space. It uses support vectors, which are the critical data points near the decision boundary, to determine this hyperplane. By using kernel functions, SVM can handle non-linearly separable data by transforming it into a higher-dimensional space. A common choice for such problems is a SVM classifier with a Radial Basis Function (RBF) kernel and a penalty parameter $C = 1.0$. This configuration is suitable for non-linear classification tasks, as it allows the model to capture complex decision boundaries while maintaining a balance between margin maximization and classification accuracy.

V. EXPERIMENTS AND RESULTS

The first experiment was oriented toward evaluating three different approaches for classifying speech stress levels: feature-based methods using standard acoustic features and simple classifiers, transfer learning with fine-tuned CNN models (YAMNet and VGGish), and self-supervised learning using BYOL-S and Wav2Vec 2.0 models. Each method was tested on three stress-related speech datasets (CRISIS, StressDat, SUSAS) to compare their classification performance.

The second experiment was focused on evaluating the performance of selected algorithms on a merged and balanced dataset that combined the three stress-related speech corpora. The goal was to determine how well these algorithms could classify stress levels under consistent conditions.

The third experiment aimed to assess the impact of incorporating emotional speech datasets mapped to stress levels using two different approaches on stress classification performance. It tested whether emotional data could improve model accuracy and whether different mapping strategies influenced the effectiveness of feature-based, transfer learning, and self-supervised learning methods.

A. EXPERIMENT SETUP

The training and testing of all classification models were performed on a server with the following configuration: Intel Xeon W-2245, 3.90 GHz 8-core CPU, 64 GB RAM, 2 TB HDD, 2x NVIDIA RTX A4500 GPGPU (20 GB), Linux OS.

Before training, all speech databases were split, with 80% used for training and validation and 20% for testing. Five-fold cross-validation was applied during training. Speech samples for training and testing were chosen randomly, while ensuring speaker independence, meaning that samples from the same speaker were not included in both the training and test sets simultaneously.

We used the F1-score, the harmonic mean of Precision and Recall, to evaluate classification results, according to the formula:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100, \tag{1}$$

where

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{2}$$

and

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{3}$$

Precision measures the ratio of correctly predicted positive observations (True Positives) to the total number of predicted positive observations (True Positives with False Positives), while Recall measures the ratio of correctly predicted positive observations to all actual positive cases (True Positives with False Negatives).

Unlike Accuracy (Acc), the F1-score does not take True Negatives into account and evaluates the model's performance solely in recognizing positive cases.

B. EXPERIMENTS WITH INDIVIDUAL STRESS DATASETS

1) FEATURE-BASED APPROACHES

In the first approach, we extracted standard frequency features from speech. We selected ten extraction algorithms that are most frequently found in similar studies, primarily focused on the detection of emotions in speech. In addition to

TABLE 5. The experimental results based on averaged F1-score [%].

Feature extraction or embeddings	Classifier	Stress datasets		
		CRISIS	StressDat	SUSAS
Feature-based approaches				
IMFCC	SVM	75.4	59.5	55.8
DWT	Bagged Trees	77.6	70.0	64.1
STFT	Bagged Trees	80.5	68.2	64.6
BFCC	Weighted kNN	83.9	77.3	60.9
MFCC	Bagged Trees	83.3	74.7	64.6
GFCC	Weighted kNN	86.6	77.7	60.5
LFCC	Weighted kNN	84.7	81.2	60.5
MHFCC	Subspace kNN	84.8	77.3	65.3
HFCC	Subspace kNN	86.5	77.4	66.6
GTCC	Subspace kNN	**91.0**	**81.6**	**73.1**
DWT+HFCC	Bagged Trees	88.4	78.5	71.9
DWT+MFCC	Bagged Trees	88.5	79.2	70.9
DWT+GTCC	Subspace kNN	**91.1**	**82.4**	72.7
MFCC+GTCC	Subspace kNN	89.0	80.4	72.8
DWT+MFCC+GTCC	Bagged Trees	89.7	81.8	**73.6**
Transfer learning approaches				
VGGish CNN	Softmax	73.9	73.0	57.3
YAMNet CNN	Softmax	69.6	68.4	56.3
Self-supervised learning approaches				
BYOL-S/AudioNTT	SVM	92.0	83.0	**78.0**
BYOL-S/Resnetish34	SVM	90.0	80.0	72.0
Hybrid BYOL-S/CvT	SVM	**93.0**	**84.0**	77.0
Wav2Vec2-VoxPopuli-v2	Softmax	93.0	82.0	72.0
Wav2Vec2-VoxPopuli-ft	Softmax	**94.0**	87.0	74.0
Wav2Vec2-XLS-R	Softmax	85.0	**88.0**	**77.0**

DWT and STFT, algorithms that extract cepstral features use filter banks that differ in how frequency levels are distributed, whether on a linear scale, mel scale, Bark scale, or others. The STFT method serves as the basis for all these algorithms.

In audio and speech processing, the mel filter bank is the most widely used. It employs triangular filters that are unevenly distributed along the frequency axis, simulating sound perception by the human ear. Other filter banks, such as Bark, gammatone, ERB, etc., were designed to achieve a more accurate distribution, also taking into account the presence of noise in the speech. For example, the distribution of frequency bands in the gammatone filter bank was inspired by the response of the cochlear membrane in the HAS.

A non-linear function is then applied to the frequency components, either in the form of a logarithm or a cube root function. The goal is to emphasize the essential components of the spectrum (peaks) while suppressing the noise components. By subsequently applying a discrete orthogonal transformation, such as DCT, we obtain the cepstral coefficients, which can be used as input for a classification algorithm. The aforementioned speech features were extracted using the *librosa*[11] and *spafe*[12] libraries in the Python language.

Since these feature extraction methods produce only a small number of coefficients, it was not necessary to use a very complex classification algorithm. As shown in the results Table 5, the best classification results for stress detection across three levels were achieved for all stress databases using the GTCC algorithm and the Subspace kNN classifier. As previously mentioned, the GTCC algorithm is more robust to noise than GFCC or MFCC.

We also evaluated combinations of several approaches for extracting features from speech (see Table 8). For example, by combining the DWT and MFCC algorithms with GTCC, we achieved improved classification results based on the F1-score for individual datasets. For the CRISIS and StressDat datasets, the combination of DWT and GTCC appeared to be the most effective. In the case of the SUSAS database, which also includes recordings from real-world situations, the combination of GTCC, MFCC, and DWT coefficients seemed to be the most suitable feature set. Moreover, the algorithm based on Bagged Trees proved to be the most effective classifier for this feature set.

2) TRANSFER LEARNING APPROACHES

The second approach focused on fine-tuning the already pre-trained convolutional neural network models VGGish and YAMNet using recordings of speech under stress.

From the classification results summarized in Table 5, it can be observed that the F1-score values are lower compared to the previous approach across all stress datasets. This may be due to the fact that the models were pre-trained on general acoustic data rather than specifically on speech data. The AudioSet database does not even include speech emotions among its categories, let alone stress levels.

[11] https://librosa.org/doc/latest/index.html
[12] https://spafe.readthedocs.io/en/latest/

Nevertheless, the results of the fine-tuned models are quite satisfactory.

3) SELF-SUPERVISED LEARNING APPROACHES

In the case of self-supervised learning, we applied two algorithms. The first focused on extracting embeddings from the speech signal (BYOL-S), which were then fed into a standard classifier. The second used transfer learning with pre-trained models (Wav2Vec 2.0), where self-supervised learning was also applied to the raw speech recordings.

Several classifiers were evaluated using BYOL-S embeddings, with the best performance achieved by the SVM. The classification results on the CRISIS, StressDat, and SUSAS databases are presented in Table 5. Notably, the hybrid BYOL-S model, based on a CvT, delivered the best performance in classifying stress into three levels. For the SUSAS database, which includes real-world recordings of speech under stress, the baseline BYOL-S/AudioNTT model achieved an F1-score that was 1% higher.

Regarding the Wav2Vec 2.0 approach, a total of three pre-trained base models were fine-tuned. A model trained on a subset of the VoxPopuli dataset and fine-tuned for speech recognition (see Table 5) demonstrated stable results across all stress datasets. We also obtained promising results with the multilingual model.

C. EXPERIMENT WITH MERGED AND BALANCED STRESS DATASET

In the next experiment, we combined all three stress datasets – CRISIS, StressDat, and SUSAS – into a single merged and balanced dataset. This dataset contained 10,108 speech utterances per category. After merging the stress datasets, we reanalyzed the basic speech characteristics. As shown in Table 7, the trend remains consistent: mean F_0 and voice intensity increase with higher stress levels. Additionally, increases in formant F_1 and shimmer values can be observed. In contrast, HNR values decrease as stress levels rise.

Only those speech feature extraction or vector representation algorithms, as well as self-supervised learning algorithms that achieved the best results on the individual unbalanced datasets in previous experiments, were included in this experiment. The results are shown in Table 8. The self-supervised learning algorithms again delivered the best performance, with F1-scores ranging from 82% to 85%. The standard feature-based approach, relying on the extraction of GTCC coefficients combined with the Subspace kNN classifier, also achieved a very similar result.

D. THE EFFECT OF INCLUDING EMOTIONAL DATASETS-MAPPING EMOTIONAL STATES TO STRESS LEVELS

There are several options for mapping emotional speech recordings to stress levels. If we already have a trained model for stress classification, we can use it to redistribute emotional recordings into individual stress levels. If such a model is not available, we can estimate the level of stress based on the calculation of basic voice characteristics or assign an appropriate stress level to individual emotional states. In previous research [66], we found that calculating basic voice characteristics may not yield the desired results. We can only rely on the calculation of fundamental frequency F_0, voice intensity, and speech tempo, as other speech characteristics produce debatable results. For this reason, we decided to find a suitable mapping of emotional states to stress levels. Several studies focus solely on negative emotions, labeling them as stressful. However, since our research classifies stress into three levels and also considers stress as a potentially positive and motivating factor, we needed to establish a suitable mapping that accounts for these aspects.

1) NATURAL DISTRIBUTION USING FINE-TUNED WAV2VEC 2.0 MODEL

Each emotional recording was assigned a stress level using the pre-trained "*Wav2Vec2-VoxPopuli-base-en-v2*" model, which produced the best overall stress classification results after being trained on the merged and balanced CRISIS, StressDat, and SUSAS stress datasets (see Table 8).

After remapping the RAVDESS, SAVEE, SUS, and TESS emotional datasets to corresponding stress levels, we balanced them in terms of the number of recordings per stress level, merged them with the previously created stress dataset, and thus created a mixed dataset of emotionally stressed recordings. The resulting dataset contained 10,812 speech utterances for each category (low, medium, and high).

Similarly, in this case, we analyzed the behavior of the basic speech characteristics (see Table 7). After adding emotional recordings that were reclassified into individual stress levels, the trend remains consistent: mean F_0, formant F_1, average voice intensity, and shimmer values increase with rising stress levels, while HNR decreases.

The results of stress classification are presented in Table 8, in the column titled "1st approach". The outcomes show either the same performance as when the classifier was trained solely on a balanced set of stress recordings, or a slight deterioration or improvement. Only in the case of models using the Wav2Vec 2.0 architecture was there a slight improvement, suggesting that this self-supervised learning approach has some potential. However, for the effect to be more substantial, significantly more data would be required than the 704 emotional recordings (representing 7% additional data) that were included in the training after remapping to stress levels. Additionally, the accuracy of the mapping from emotional states to stress levels is limited by the performance of the classification model "*Wav2Vec2-VoxPopuli-base-en-v2*" after fine-tuning on stress-related speech recordings, which in our case reached only 85%.

2) MAPPING THROUGH EMOTIONAL CLASSES

In previous work, we tested the distribution of emotions using Mikels' emotional wheel [67], which considers pleasant and unpleasant valence as well as activated and deactivated

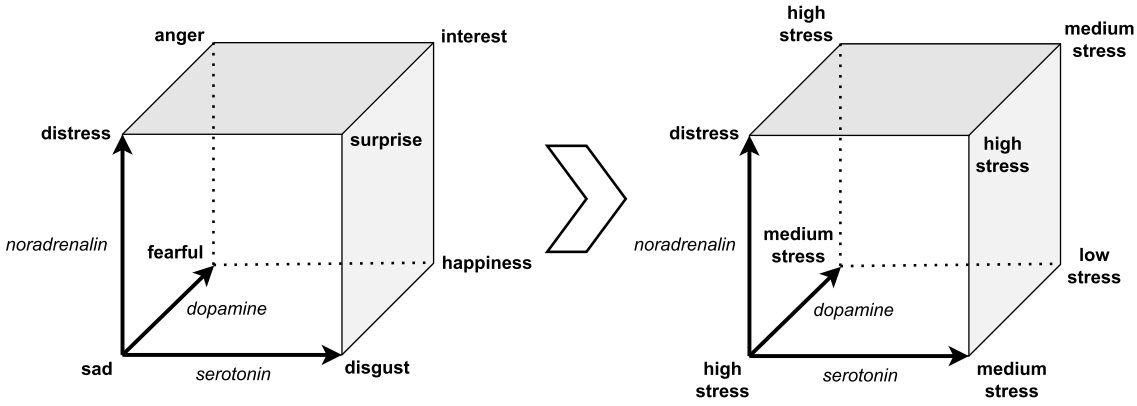

FIGURE 6. Mapping emotional classes to stress levels using Lövheim's cube.

arousal. However, this approach did not yield the expected results. When emotional data were added to stress data, a slight deterioration in stress classification performance was observed. Therefore, we explored an alternative approach – mapping emotions based on neurotransmitter activity using Lövheim's emotion cube [68].

Lövheim's cube describes the relationship between emotions and the levels of three key neurotransmitters in the brain: serotonin, dopamine, and noradrenaline (see Figure 6). According to Hugo Lövheim, the emotion of distress is located at the front-left upper vertex of the cube. From this perspective, distress is characterized by low levels of serotonin and dopamine and a high level of noradrenaline.

We defined the following categorization (see Figure 6): emotions positioned on the cube's body diagonal farthest from the distress vertex – specifically happiness – were classified as low stress. Emotions located on the wall diagonal adjacent to the distress vertex – such as disgust, fearful, and interest – were classified as medium stress. Emotions nearest to the distress vertex along an edge – namely, anger, surprise, and sadness – were classified as high stress. A neutral emotion was assigned to the low stress category.

When examining neurotransmitter levels, the emotions classified as high stress (anger, surprise, and sadness) differ from distress in only one neurotransmitter. Emotions in the medium stress category (disgust, fearful, and interest) differ in two neurotransmitters, while happiness differs from distress in all three (see Table 6).

The resulting dataset, after remapping emotional categories to stress levels, contained 11,682 speech utterances in each stress level following combination with the balanced stress dataset. This means that the balanced dataset of speech recordings under stress was expanded by 1,574 new emotional recordings, representing a 15.5% increase.

We once again examined whether the new mapping disrupted the trend in the behavior of the basic speech characteristics. As shown in Table 7, the trend remains consistent with the previous two cases. Except for HNR, the values of mean F_0, formant F_1, shimmer, and voice intensity increase with increasing stress levels.

TABLE 6. Mapping emotional states to stress levels based on Lövheim's cube of emotions.

Base emotion	Level of serotonin	Level of dopamine	Level of noradrenaline	Stress level
distress	low	low	high	
sad	low	low	low	high
anger	low	high	high	
surprise	high	low	high	
fearful	low	high	low	medium
disgust	high	low	low	
interest	high	high	high	
happy	high	high	low	low

When we examine the results of classifying stress into three levels using the emotional stress dataset created in this way (see Table 8, column titled "2nd approach"), we observe a significant improvement in F1-score values with feature-based approaches, with the results even slightly outperforming self-supervised learning methods. In contrast, the performance of the self-supervised learning methods shows a slight deterioration. Nevertheless, this is an interesting finding that can be further explored.

VI. DISCUSSION

Comparing the presented research results with other approaches is very challenging. In the case of the SUSAS database, which was one of the databases used in our experiments, the authors of [7], [10], and [32] utilized it in a different research context. Since the SUSAS database itself is not balanced across stress levels, the authors used only a portion of the database in their experiments. Often, they considered only the simulated part of the database, where they also achieved accuracy values above 90%. In our research, we used both subcorpora, simulated and actual parts, which negatively influenced the results. Although it is not explicitly mentioned here, on the actual subcorpora we achieved, for example, an average F1-score of up to 82% using the BYOL-S algorithm and the SVM classifier.

If we focus on three-class stress classification, only Prasetio et al. [10] and Duvvuri et al. [11] address this. Even

TABLE 7. Extracted speech characteristics across merged and mixed emotional stress datasets (averaged values).

Speech characteristic	Stress level	Merged and balanced stress dataset	Mixed emotional stress dataset	
			1st approach	2nd approach
mean F_0 [Hz]	low	141.67	143.53	160.86
	medium	174.52	176.24	176.30
	high	187.60	190.11	189.30
pitch range [Hz]	low	88.01	89.97	122.50
	medium	120.25	122.83	125.52
	high	129.09	131.71	132.46
formant F_1 [Hz]	low	578.91	588.59	626.88
	medium	637.37	643.72	645.32
	high	647.68	654.17	654.11
formant F_2 [Hz]	low	1557.66	1572.50	1682.48
	medium	1634.61	1642.72	1643.89
	high	1631.46	1639.77	1643.42
formant F_3 [Hz]	low	2457.56	2478.48	2645.69
	medium	2565.45	2578.19	2582.19
	high	2550.70	2564.52	2574.03
average vocal intensity [dB]	low	71.39	70.37	70.32
	medium	72.07	71.41	70.95
	high	75.04	74.53	73.68
HNR [dB]	low	10.95	11.03	11.86
	medium	10.40	10.44	10.59
	high	10.06	10.21	10.32
jitter [%]	low	2.31	2.30	2.31
	medium	2.21	2.20	2.18
	high	2.25	2.23	2.22
shimmer [%]	low	9.01	9.03	8.97
	medium	9.27	9.31	9.19
	high	9.55	9.59	9.47

TABLE 8. The experimental results based on averaged F1-score.

Feature extraction or embeddings	Classifier	Stress datasets			Merged and balanced stress dataset	Mixed emotional stress dataset	
		CRISIS	StressDat	SUSAS		1st approach	2nd approach
GTCC	Subspace kNN	91.0	81.6	73.1	80.1	77.6	**85.5**
DWT+GTCC	Fine kNN	91.1	82.4	72.7	79.3	76.9	85.4
DWT+MFCC+GTCC	Bagged Trees	89.7	81.8	73.6	79.8	78.0	85.3
Hybrid BYOL-S/CvT	SVM	93.0	84.0	**77.0**	83.0	83.0	82.0
Wav2Vec2-VoxPopuli-v2	Softmax	93.0	82.0	72.0	**85.0**	**85.0**	84.0
Wav2Vec2-VoxPopuli-ft	Softmax	**94.0**	87.0	74.0	82.0	83.0	82.0
Wav2Vec2-XLS-R	Softmax	85.0	**88.0**	**77.0**	67.0	78.0	70.0

when employing advanced convolutional neural networks for classification, combined with MFCC-based feature extraction as in Duvvuri et al. [11], the resulting F1-scores are significantly lower than those achieved with our simpler classifier.

Throughout the entire research, we worked with stress datasets that had already been divided into three stress levels. However, with the exception of the StressDat database, we were not well-informed about the manual labeling process used by annotators. We observed that the classifiers had significant difficulty distinguishing between medium and high stress levels. For this reason, it would be appropriate to revisit the recordings included in the medium stress category and assess the extent to which they correlate with negative and positive emotions.

Also, the quality and heterogeneity of the data are crucial for the robustness of the models. Using multiple datasets increases the generalizability of the findings and enables the identification of approaches that perform consistently across different data collection conditions.

When examining feature-based approaches, this research demonstrated that the gammatone filter bank models the HAS more accurately than other filter banks, such as mel or Bark, particularly in the low-frequency range. Unlike the GFCC algorithm, which uses a logarithmic function, GTCC employs a cubic root function after applying the gammatone filter bank. This approach more effectively simulates the non-linear response of the HAS to variations in sound intensity and captures subtle emotional changes in speech, such as shifts in pitch or speech rate caused by stress [69]. Similar findings were highlighted by Lim [59], who noted that the cubic root function better preserves spectral peaks and valleys in the signal compared to the logarithmic function.

Feature-based approaches perform well under controlled conditions, are less computationally demanding, and offer better interpretability. In contrast, self-supervised learning

with the Wav2Vec 2.0 architecture generally yields excellent results even with small amounts of data, but the training process can take several hours to tens of hours, depending on the model's hyperparameter settings. A promising compromise is the use of self-supervised learning approaches for feature extraction from speech, such as BYOL-S. When combined with a suitable classifier, BYOL-S can be applied in simpler, fast-response systems and is more interpretable than Wav2Vec 2.0. Notably, feature extraction using the BYOL-S algorithm has been tested on several natural speech and language processing tasks with excellent results [15].

A significant advantage of self-supervised learning architectures like Wav2Vec 2.0 or BYOL-S is their capability of learning "*general-purpose representations*" and their ability to offer "*efficient learning without labels*". This suggests that self-supervised models can learn more robust and generalizable speech representations from large amounts of unlabeled data, which contributes to their better performance across different detection tasks compared to methods relying on handcrafted features.

Of course, there are other versions of Wav2Vec 2.0 models, whether large models or those pre-trained on massive data, such as MMS (Massively Multilingual Speech) [70]. However, we encountered limitations in terms of computational requirements during model training.

VII. CONCLUSION

In this paper, we comprehensively evaluated both conventional feature-based methods and modern deep learning techniques for classifying emotional stress levels in speech.

Our analysis confirms that certain acoustic features – particularly fundamental frequency (F_0) and vocal intensity – strongly correlate with stress levels.

Among the feature-based approaches, GTCC features combined with Subspace kNN classifiers demonstrated strong performance, especially on datasets collected under controlled conditions. Additionally, feature extraction using gammatone filter banks provided a closer approximation of the HAS, thereby enhancing stress detection sensitivity. The combination of different feature extraction techniques, such as DWT with GTCC, further improved classification performance. These findings underscore the importance of careful feature selection, classifier design, and dataset diversity in achieving high accuracy and generalizability.

In the realm of deep learning, self-supervised learning models – particularly Wav2Vec 2.0 – consistently outperformed other approaches across multiple datasets, achieving up to a 94% F1-score on CRISIS, over 87% on StressDat, and 77% on SUSAS. BYOL-S also proved effective when coupled with suitable classifiers like SVM, delivering robust results even under acoustically diverse conditions. Transfer learning models such as VGGish and YAMNet were also useful, though slightly less effective.

Evaluation on the merged and balanced stress dataset showed that Wav2Vec 2.0 maintained high performance, achieving up to an 85% F1-score, supporting its potential for deployment in general-purpose stress detection systems. Integrating both real and simulated stress data helped reduce dataset bias and improved the generalization capabilities of the models. Furthermore, augmenting stress datasets with emotional speech data enhanced robustness, particularly in real-world environments. However, distinguishing between moderate and high stress levels remained challenging, likely due to ambiguities in dataset labeling.

Overall, this study advances the field of emotional stress detection in several important ways:

- It achieves state-of-the-art F1-scores using a fine-tuned Wav2Vec 2.0 model, outperforming prior studies that reported significantly lower scores (e.g., 65% in similar multi-class settings).
- It introduces novel feature combinations, such as DWT+GTCC with a Subspace kNN classifier.
- It supports integrating self-supervised learning methods with robust classifiers to improve the generalization and effectiveness of speech-based stress detection across diverse acoustic conditions.

VIII. FUTURE WORK

While the results of this research are promising, there are still several open questions that require further investigation.

There is a limited number of stressed speech databases, and even fewer that reflect stress in real-world conditions. In future studies, we aim to broaden the training dataset by incorporating as many databases containing genuine emotions as possible, as these are much more abundant, and to map them to specific stress levels. This will require finding a more efficient method for associating emotions with stress levels by taking multiple aspects of speech into account.

Another way to increase the volume of data is through the use of data augmentation techniques. When augmenting data for stress detection from human voice, it is important to preserve the prosodic, spectral, and temporal features that convey stress (e.g., pitch, speaking rate, and energy). Some standard augmentation techniques, such as adding noise, reverberation, pitch-shifting, and speed or volume perturbation, can still be applied, but only with caution and to a limited extent. In this research, data augmentation methods were used only within the BYOL-S self-supervised learning algorithm, which includes built-in techniques such as MixUp, pitch-shifting, and time-stretching. In future research, we therefore aim to explore more suitable data augmentation techniques.

In common practice, models often perform well on the data they were trained on (in-domain performance), but tend to fail under real-world conditions that differ from the training data (out-of-domain performance). This is often because the model adapts to dataset-specific features rather than learning to recognize the underlying phenomenon itself (e.g., stress). Therefore, in future research, we aim to focus on cross-corpus evaluation, which assesses the model's ability to generalize beyond the training data. This involves training the model on one corpus and testing it on a different, independent

corpus featuring different speakers, recording conditions, and acoustic environments. We also plan to explore several alternative approaches, whose outputs could be fused to further improve classification performance.

If we look at the feature-based approaches, we hypothesize that by further tuning the parameters of the Subspace *k*NN algorithm or by employing other classification approaches, we could achieve F1-scores comparable to those of the self-supervised learning algorithms. Additionally, we can experiment with other feature combinations, possibly supplemented with basic speech characteristics such as fundamental frequency, voice intensity, or speech rate. Alternatively, we could explore existing sets of speech features, such as GeMAPS [71] or eGeMAPS [72], which are primarily used in the detection of speech disorders and combine basic speech characteristics with energy coefficients and low-level descriptors. This could be the subject of future research.

In this research, we also utilized pre-trained models, such as VGGish, to generate embeddings, which were then used to train the classifiers. In future work, we aim to explore approaches that focus on extracting deep spectral features, including auDeep [73], BEATs [74], and others.

Self-supervised learning enables new AI capabilities across various domains. Beyond Wav2Vec 2.0, several other architectures, like HuBERT [75], remain untested.

These improvements should enable the approach to generalize speech-based stress detection under real conditions.

ACKNOWLEDGMENT

The authors acknowledge the use of AI-based tools, such as ChatGPT, for assistance in editing, grammar enhancement, and spelling checks during the preparation of this manuscript.

REFERENCES

[1] C. Kirchhübel, D. M. Howard, and A. W. Stedmon, "Acoustic correlates of speech when under stress: Research, methods and future directions," *Int. J. Speech, Lang. Law*, vol. 18, no. 1, pp. 75–98, Sep. 2011. [Online]. Available: https://journal.equinoxpub.com/IJSLL/article/view/5979

[2] S. Z. Bong, M. Murugappan, and S. Yaacob, "Methods and approaches on inferring human emotional stress changes through physiological signals: A review," *Int. J. Med. Eng. Informat.*, vol. 5, no. 2, pp. 152–162, 2013. [Online]. Available: https://www.inderscienceonline.com/doi/abs/10.1504/IJMEI.2013.053332

[3] W. A. D. Perera, M. J. Perera, and M. K. Tharaniyawarma, "AI based speech analysis framework," *Int. Res. J. Innov. Eng. Technol.*, vol. 8, no. 1, pp. 94–104, 2024. [Online]. Available: https://www.proquest.com/scholarly-journals/ai-based-speech-analysis-framework/docview/2933913748/se-2

[4] M. Sigmund, "Changes in frequency spectrum of vowels due to psychological stress," in *Proc. 20th Int. Conf. Radioelektronika*, Brno, Brno, Czech Republic, Apr. 2010, pp. 1–4.

[5] M. Sigmund, "Introducing the database ExamStress for speech under stress," in *Proc. NORSIG*, Jun. 2006, pp. 290–293.

[6] B. Jena and S. S. Singh, "Psychological stress speech analysis: A review," *Int. J. Eng. Res. Technol.*, vol. 4, no. 28, pp. 1–4, Apr. 2018.

[7] L. He, M. Lech, S. Memon, and N. Allen, "Recognition of stress in speech using wavelet analysis and teager energy operator," in *Proc. Interspeech*, Sep. 2008, pp. 605–608.

[8] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. Eurospeech*, Sep. 1997, pp. 1743–1746.

[9] H. Han, K. Byun, and H.-G. Kang, "A deep learning-based stress detection algorithm with speech signal," in *Proc. Workshop Audio-Visual Scene Understand. Immersive Multimedia*, Oct. 2018, pp. 11–15, doi: 10.1145/3264869.3264875.

[10] B. H. Prasetio, D. Syauqy, and E. R. Widasari, "Hilbert-huang mel frequency cepstral coefficient for speech stress recognition system," in *Proc. 9th Int. Conf. Inf. Technol., Comput., Electr. Eng. (ICITACEE)*, Semarang, Indonesia, Aug. 2022, pp. 111–114.

[11] K. Duvvuri, H. Kanisettypalli, T. N. Masabattula, S. Vekkot, D. Gupta, and M. Zakariah, "Unravelling stress levels in continuous speech through optimal feature selection and deep learning," *Proc. Comput. Sci.*, vol. 235, pp. 1722–1731, Jan. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924008391

[12] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC*, May 2014, pp. 3123–3128. [Online]. Available: https://aclanthology.org/L14-1421/

[13] S. Chen, X. Xing, G. Liang, and X. Xu, "I feel stressed out: A Mandarin speech stress dataset with new paradigm," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Chiang Mai, Thailand, Nov. 2022, pp. 583–589.

[14] Z. Wu, N. Scheidwasser-Clow, K. E. Hajal, and M. Cerňak, "Speaker embeddings as individuality proxy for voice stress detection," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 1838–1842.

[15] G. Elbanna, N. Scheidwasser-Clow, M. Kegler, P. Beckmann, K. E. Hajal, and M. Cerňak, "BYOL-S: Learning self-supervised speech representations by bootstrapping," in *Proc. HEAR*, Jan. 2022, pp. 25–47.

[16] G. Elbanna, A. Biryukov, N. Scheidwasser-Clow, L. Orlandic, P. Mainar, M. Kegler, P. Beckmann, and M. Cerňak, "Hybrid handcrafted and learnable audio representation for analysis of speech under cognitive and physical load," in *Proc. Interspeech*, Sep. 2022, pp. 386–390.

[17] D. Ghose, O. Gitelson, and B. Scassellati, "Integrating multimodal affective signals for stress detection from audio-visual data," in *Proc. ICMI*, Oct. 2024, pp. 22–32.

[18] E. Rituerto-González, A. Gallardo-Antolín, and C. Peláez-Moreno, "Speaker recognition under stress conditions," in *Proc. IberSPEECH*, Barcelona, Spain, Nov. 2018, pp. 15–19.

[19] A. Aguiar, M. Kaiseler, H. Meinedo, P. R. Almeida, M. Cunha, and J. Silva, "VOCE corpus: Ecologically collected speech annotated with physiological and psychological stress assessments," in *Proc. LREC*, Jun. 2014, pp. 1568–1574. [Online]. Available: https://aclanthology.org/L14-1514/

[20] A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, S. C. Sturmbauer, E.-M. Meßner, B. M. Kudielka, N. Rohleder, H. Baumeister, and B. W. Schuller, "An evaluation of speech-based recognition of emotional and physiological markers of stress," *Frontiers Comput. Sci.*, vol. 3, Dec. 2021, Art. no. 750284.

[21] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Meßner, E. Cambria, G. Zhao, and B. W. Schuller, "The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," in *Proc. 2nd Multimodal Sentiment Anal. Challenge*, Virtual Event, China, Oct. 2021, pp. 5–14, doi: 10.1145/3475957.3484450.

[22] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, Mar. 2017, pp. 776–780.

[23] J. Vamsinath, B. Varshini, T. Sandeep, V. Meghana, and B. Latha, "A survey on stress detection through speech analysis using machine learning," *Int. J. Sci. Res. Sci. Technol.*, vol. 9, no. 4, pp. 326–333, Jul. 2022, doi: 10.32628/ijsrst229436.

[24] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391, doi: 10.1371/journal.pone.0196391.

[25] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiß, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.

[26] S. Vaikole, S. Mulajkar, A. More, P. Jayaswal, and S. Dhas, "Stress detection through speech analysis using machine learning," *Int. J. Creative Res. Thoughts*, vol. 8, no. 5, pp. 2239–2244, 2020.

[27] M. S. Nordin, A. L. Asnawi, N. A. Zainal, R. F. Olanrewaju, A. Z. Jusoh, S. N. Ibrahim, and N. F. M. Azmin, "Stress detection based on TEO and MFCC speech features using convolutional neural networks (CNN)," in *Proc. IEEE Int. Conf. Comput. (ICOCO)*, Nov. 2022, pp. 84–89.

[28] S. Nagarajan, S. S. S. Nettimi, L. S. Kumar, M. K. Nath, and A. Kanhe, "Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales," *Digit. Signal Process.*, vol. 104, Sep. 2020, Art. no. 102763. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200420301081

[29] S. Haq and P. J. B. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proc. AVSP*, Sep. 2009, pp. 53–58.

[30] H. Rahali, Z. Hajaiej, and N. Ellouze, "Feature extraction method human factor cepstral coefficients in automatic speech recognition," in *Proc. 9th Int. Symp. Commun. Syst., Netw. Digit. Sign (CSNDSP)*, Jul. 2014, pp. 266–270.

[31] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP)*, Paris, France, Oct. 2000, pp. 181–188.

[32] V. V. Yerigeri and L. K. Ragha, "Speech stress recognition using semi-eager learning," *Cognit. Syst. Res.*, vol. 65, pp. 79–97, Oct. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389041720300735

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210.

[34] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

[35] M. Hosain, M. Yeasmin Arafat, G. Zahirul Islam, J. Uddin, M. Mobarak Hossain, and F. Alam, "Emotional expression detection in spoken language employing machine learning algorithms," 2023, *arXiv:2304.11040*.

[36] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," Borealis, V1, Univ. Toronto Mississauga, Mississauga, Canada, doi: 10.5683/SP2/E8H2MF.

[37] M. Rusko, S. Darjaa, M. Trnka, R. Sabo, and M. Ritomský, "Expressive speech synthesis for critical situations," *Comput. Informat.*, vol. 33, no. 6, pp. 1312–1332, Jan. 2014. [Online]. Available: https://www.cai.sk/ojs/index.php/cai/article/view/2816

[38] R. Sabo, Š. Beňuš, M. Trnka, M. Ritomský, M. Rusko, M. Schaper, and J. Szabo, "StressDat-database of speech under stress in Slovak," *J. Linguistics/Jazykovedný Časopis*, vol. 72, no. 2, pp. 579–589, Dec. 2021, doi: 10.2478/jazcas-2021-0053.

[39] L. L. Reddy and S. Kuchibhotla, "Survey on stress emotion recognition in speech," in *Proc. ICCCIS*, Oct. 2019, pp. 1–4.

[40] M. Rahurkar, J. H. L. Hansen, J. L. Meyerhoff, G. A. Saviolakis, and M. L. Koenig, "Frequency band analysis for stress detection using a teager energy operator based feature," in *Proc. ICSLP*, Sep. 2002, pp. 2021–2024.

[41] R. C. Fernandez and R. W. Picard, "Modeling drivers' speech under stress," *Speech Commun.*, vol. 40, nos. 1–2, pp. 145–159, Jan. 2003.

[42] P. Angkititrakul, J. H. L. Hansen, S. Choi, T. Creek, J. Hayes, J. Kim, D. Kwak, L. T. Noecker, and A. Phan, "UTDrive: The smart vehicle project," in *In-Vehicle Corpus and Signal Processing for Driver Behavior*. Boston, MA, USA: Springer, 2009, pp. 55–67, doi: 10.1007/978-0-387-79582-9_5.

[43] J. Luig and A. Sontacchi, "A speech database for stress monitoring in the cockpit," *Proc. Inst. Mech. Eng., G, J. Aerosp. Eng.*, vol. 228, no. 2, pp. 284–296, Feb. 2014, doi: 10.1177/0954410012467944.

[44] S. Shukla, S. Dandapat, and S. R. M. Prasanna, "A subspace projection approach for analysis of speech under stressed condition," *Circuits, Syst., Signal Process.*, vol. 35, no. 12, pp. 4486–4500, Dec. 2016, doi: 10.1007/s00034-016-0284-9.

[45] M. Jaiswal, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, "MuSE: A multimodal dataset of stressed emotion," in *Proc. LREC*, 2020, pp. 1499–1510. [Online]. Available: https://aclanthology.org/2020.lrec-1.187/

[46] H. Chaptoukaev, V. Strizhkova, M. Panariello, B. D'alpaos, A. Reka, V. Manera, S. Thümmler, E. Ismailova, N. Evans, F. Brémond, M. Todisco, M. A. Zuluaga, and L. M. Ferrari, "StressID: A multimodal dataset for stress identification," in *Proc. NeurIPS*, Dec. 2023, pp. 1–12.

[47] J. Pešán, V. Juřík, M. Karafiát, and J. Černocký, "BESST dataset: A multimodal resource for speech-based stress detection and analysis," in *Proc. Interspeech*, Sep. 2024, pp. 1355–1359. [Online]. Available: https://www.fit.vut.cz/research/publication/13324

[48] X. Zuo, T. Li, and P. Fung, "A multilingual natural stress emotion database," in *Proc. LREC*, May 2012, pp. 1174–1178. [Online]. Available: https://aclanthology.org/L12-1338/

[49] D. Ververidis and K. Kotropoulos, "A state of the art review on emotional speech databases," in *Proc. 1st Richmedia Conf.*, Jan. 2003, pp. 1–21.

[50] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, Jun. 2012, doi: 10.1007/s10772-011-9125-1.

[51] G. H. Mohmad Dar and R. Delhibabu, "Speech databases, speech features, and classifiers in speech emotion recognition: A review," *IEEE Access*, vol. 12, pp. 151122–151152, 2024.

[52] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," U.S. Dept. Commerce, Nat. Inst. Standards Technol., Comput. Syst. Lab., Adv. Syst. Division, Gaithersburg, MD, USA, Tech. Rep. NIST Speech Disc 1-1.1, NISTIR 4930, 1993.

[53] L. Macková, A. Čižmár, and J. Juhár, "A study of acoustic features for emotional speaker recognition in I-Vector representation," *Acta Electrotechnica et Inf.*, vol. 15, no. 2, pp. 15–20, Jun. 2015.

[54] G. Demenko and M. Jastrzębska, "Analysis of natural speech under stress," *Acta Phys. Polonica A*, vol. 121, no. 1, pp. 92–95, Jan. 2012, doi: 10.12693/aphyspola.121.a-92.

[55] M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn, "Voice stress analysis: A new framework for voice and effort in human performance," *Frontiers Psychol.*, vol. 9, pp. 1–25, Nov. 2018. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.01994

[56] A. C. N. D. Felippe, M. H. M. M. Grillo, and T. H. Grechi, "Standardization of acoustic measures for normal voice patterns," *Brazilian J. Otorhinolaryngology*, vol. 72, no. 5, pp. 659–664, Sep. 2006.

[57] P. Singh and E. G. Rajan, "Application of different filters in mel frequency cepstral coefficients feature extraction and fuzzy vector quantization approach in speaker recognition," *Int. J. Eng. Res. Technol.*, vol. 2, no. 6, pp. 3171–3182, Jun. 2013.

[58] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2011, pp. 559–564.

[59] J. S. Lim, "Spectral root homomorphic deconvolution system," *IEEE Trans. Acoust.*, vol. A-27, no. 3, pp. 223–233, Jun. 1979.

[60] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, Mar. 2017, pp. 131–135.

[61] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[62] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9052–9071, Dec. 2024.

[63] F. D. Pup and M. Atzori, "Applications of self-supervised learning to biomedical signals: A survey," *IEEE Access*, vol. 11, pp. 144180–144203, 2023.

[64] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, Jan. 2020, pp. 1–12.

[65] Z. Ceylan, "Diagnosis of breast cancer using improved machine learning algorithms based on Bayesian optimization," *Int. J. Intell. Syst. Appl. Eng.*, vol. 8, no. 3, pp. 121–130, Sep. 2020.

[66] J. Staš, D. Hládek, Z. Sokolová, M. Čech, K. Škotková, and P. Poremba, "Analysis and detection of speech under emotional stress," in *Proc. ICETA*, Oct. 2023, pp. 493–498.

[67] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behav. Res. Methods*, vol. 37, no. 4, pp. 626–630, Nov. 2005.

[68] H. Lövheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Med. Hypotheses*, vol. 78, no. 2, pp. 341–348, Dec. 2011.
[69] B. H. Prasetio, L. O. A. Hazmar, D. Syauqy, and E. R. Widasari, "Gammatone-frequency cepstral coefficients based fear emotion level recognition system," *Revista Mexicana de Ingenieria Biomedica*, vol. 45, no. 2, pp. 6–22, May 2024.
[70] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *J. Mach. Learn. Res.*, vol. 25, no. 1, pp. 1–52, 2024.
[71] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.
[72] W. Xue, C. Cucchiarini, R. V. Hout, and H. Strik, "Acoustic correlates of speech intelligibility: The usability of the eGeMAPS feature set for atypical speech," in *Proc. SLaTE*, Coimbra, Portugal, Sep. 2019, pp. 48–52.
[73] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. W. Schuller, "AuDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6340–6344, Jan. 2017.
[74] S. Chen, Y. Wu, C. Wang, S. Liu, D. M. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, Jan. 2022, pp. 1–22.
[75] W.-N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.

JÁN STAŠ was born in Bardejov, Slovakia, in 1984. He received the M.Sc. (Ing.) degree in electronics and telecommunications from the Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia, in 2007, and the Ph.D. degree in telecommunications, in 2011. From 2011 to 2015, he was a Research Assistant with the Laboratory of Speech Communication Technologies. Since 2015, he has been an Assistant Professor with the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice. He is the author or co-author of more than 130 articles, with more than 65 indexed in Scopus and 43 in Web of Science. His research interests include natural language processing, text classification, statistical language modeling, speaker diarization, emotion recognition, and automatic speech recognition. He also explores topics related to hate speech and offensive language detection, voice stress analysis, and the early detection of Alzheimer's disease from speech.

STANISLAV ONDÁŠ was born in Prešov, Slovakia, in 1981. He received the M.Sc. (Ing.) and Ph.D. degrees in telecommunications from the Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia, in 2004 and 2008, respectively. He is currently an Associate Professor with the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice. He is the author or co-author of more than 123 articles, with more than 73 indexed in Scopus and 47 in Web of Science. His research interests include human–machine interaction, dialogue modeling and management, and supportive technologies for children with speech and hearing disorders.

JOZEF JUHÁR was born in Poproč, Slovakia, in 1956. He received the M.Sc. (Ing.) and the Ph.D. degrees in radioelectronics from the Technical University of Košice, in 1980 and 1991, respectively. He is currently a Full Professor with the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice. He is also the Founder of the Laboratory of Speech Communication Technologies, Department of Electronics and Multimedia Communications. He is the author or co-author of more than 430 articles, with more than 186 indexed in Scopus and 131 in the Web of Science. His research interests include speech and audio processing, speech analysis and synthesis, speech acoustics, acoustic event detection, speech enhancement and dereverberation, acoustic modeling, speaker recognition, and the research and development of speech recognition systems, spoken dialogue systems, and human–computer interfaces.