https://doi.org/10.31436/iiumej.v25i2.3207

INTEGRATION OF MFCCS AND CNN FOR MULTI-CLASS STRESS SPEECH CLASSIFICATION ON UNSCRIPTED DATASET

NUR AISHAH ZAINAL, ANI LIZA ASNAWI*, AHMAD ZAMANI JUSOH, SITI NOORJANNAH IBRAHIM, HUDA ADIBAH MOHD. RAMLI

Department of Electrical & Computer Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia

**Corresponding author: aniliza@iium.edu.my*

(Received: 3 March 2024; Accepted: 29 May 2024; Published online: 15 July 2024)

ABSTRACT: Stress is an interaction between individuals and their environment, where perceived threats can lead to serious consequences if prolonged and consistently linked to adverse physical and mental health outcomes. Our study explores methods for stress classification via speech, utilizing an unscripted dataset from an experimental study that was able to show the spontaneous reactions of stressed individuals. Mel-Frequency Cepstral Coefficients (MFCCs) emerge as promising speech features, adept at representing the power spectrum crucial to human auditory perception, especially in stress speech recognition. Leveraging deep learning technology, specifically Convolutional Neural Network (CNN), our research optimally combines speech features and CNN algorithms for stress classification. Despite the scarcity of publications on unscripted datasets and multi-class stress classifications, our study advocates their adoption, aiming to enhance performance metrics and contribute to research expansion. The proposed system shows that MFCCs achieve an accuracy of 95.67% in distinguishing among three stress classes (low-stress, medium-stress, and high-stress), surpassing the prior unscripted dataset study by 81.86%. This highlights the efficacy of the proposed MFCCs-CNN system in stress classification.

ABSTRAK: Tekanan merupakan interaksi antara individu dan persekitaran, di mana ancaman akan membawa kepada akibat serius jika berlarutan, dan secara konsisten dikaitkan dengan kesan kesihatan fizikal dan mental yang buruk. Kajian ini mengkaji kaedah pengelasan tekanan melalui pertuturan, menggunakan set data tanpa skrip yang diperoleh daripada kajian eksperimen, iaitu mampu menunjukkan tindak balas spontan individu tertekan. Pekali Septral Frekuensi-Mel (MFCCs) muncul sebagai ciri pertuturan berpotensi, iaitu mahir dalam menunjukkan secara ringkas spektrum kuasa penting bagi persepsi pendengaran manusia, terutama ketika pengecaman pertuturan bertekanan. Memanfaatkan teknologi pembelajaran mendalam, khususnya Rangkaian Neural Lingkaran (CNN), kajian ini menggabungkan ciri pertuturan dan algoritma CNN secara optimum bagi pengelasan tekanan. Walau terdapat kekurangan penerbitan pada set data tanpa skrip dan klasifikasi tekanan pelbagai kelas, kajian ini meningkatkan penggunaannya, bertujuan bagi meningkatkan metrik prestasi dan menyumbang kepada keluasan penyelidikan. Sistem yang dicadangkan ini menunjukkan bahawa MFCC mencapai ketepatan 95.67% dalam membezakan antara tiga kelas tekanan (tekanan rendah, tekanan sederhana dan tekanan tinggi), mengatasi kajian dataset tanpa skrip terdahulu sebanyak 81.86%. Ini menunjukkan keberkesanan sistem MFCCs-CNN dalam pengelasan tekanan.

KEYWORDS: *Multi-class stress classification; Unscripted dataset; Speech stress detection; MFCCs; CNN*

https://doi.org/10.31436/iiumej.v25i2.3207

1. INTRODUCTION

Stress is recognized as an interaction between individuals and their environment, where they perceive something as potentially threatening to their lives [1]. When prolonged, it can escalate to more serious consequences, consistently linked to adverse physical and mental health outcomes [2]. Therefore, the exploration of methods to detect and classify stress is of paramount importance for human well-being. Such endeavors hold the potential to mitigate the risk of severe physical and mental health issues in the future.

In the domain of detection and classification of stress using speech features, Mel-Frequency Cepstral Coefficients (MFCCs) produced promising performance in a few studies [3-8]. As a subset of cepstral features, MFCCs exhibit the capability to succinctly represent the power spectrum, accentuating frequency components crucial to human auditory perception. This attribute is particularly advantageous for stress speech recognition tasks, facilitating the extraction and emphasis of key characteristics across various levels of stress speech productions.

The application of deep learning (DL) technology involves the utilization of a Convolutional Neural Network (CNN) to aid in the classification task. As evidenced by prior research [6, 9-11], CNN has demonstrated superior performance in classifying speech features for binary classification tasks, such as distinguishing between stressed and non-stressed speech. Consequently, our study leverages the advantages of the CNN architecture to determine the optimal combination of speech features and the CNN algorithm for the task of stress classification.

However, the scarcity of publications of unscripted or freestyle datasets, as well as multi-class stress classifications, underscores the necessity of undertaking our study. The motivation behind introducing multi-class stress levels is rooted in its potential to portray the nuanced spectrum of stress perceived by individuals, surpassing the binary classification of stress and no stress. Consequently, the resulting stress classification model is deemed more reliable, given that the provided data is authentic and reflects the actual stress conditions of individuals.

On the other hand, the imperative to generate an unscripted dataset stems from its ability to capture the natural and authentic states of individuals, contrasting with scripted or staged datasets [3]. To the best of our knowledge, only six studies [3,8-12] have utilized unscripted datasets, and one study has employed multi-class stress classification [11]. As a result, our study endeavors to advocate for the adoption of unscripted datasets and multi-class stress classification, enhance existing speech stress performance metrics (e.g., accuracy, F1-score), and ultimately contribute to the expansion of publications in this research domain.

Our study conducted an experimental stress analysis involving 50 tertiary education students who are native speakers of the Malay language in their daily conversations. The rationale behind focusing on native speakers stems from the findings of previous studies [13, 14], which revealed that many students experience apprehension when required to speak in public, particularly in a language other than their native one. This emphasis on native speakers is further justified by concerns raised in the literature, indicating that Malaysian graduates entering the workforce may encounter challenges related to insufficient presentation skills and limited proficiency in the English language. This challenge often arises from a fear of public speaking. Therefore, all these factors contribute to the elicitation of stress among students in the context of the experimental stress analysis.

In addition, the selection of the procedure in the methodology referred to the previous papers that produced the stress speech classification using an unscripted dataset which started

https://doi.org/10.31436/iiumej.v25i2.3207

with data acquisition, data pre-processing, feature extraction, and lastly, stress classification [3, 8-12]. The detailed procedure was further discussed in Section 3.

The structure of this paper unfolds as follows: the second section delves into an analysis of related works employing unscripted datasets for stress classification. Following that, the third section outlines the methodology adopted to fulfill the research objectives. Subsequently, the fourth section scrutinizes the results and engages in a discussion of the findings derived from the conducted study. Finally, the fifth section encapsulates the conclusion and outlines potential avenues for future research in this domain.

2. RELATED WORKS

In this section, recent papers utilizing unscripted datasets with speech features, machine learning (ML), and DL technology are presented. A review of prior works is undertaken to enhance the current performance of speech stress detection and classification by examining the methodology of each study. Subsequent subsections provide detailed insights into the integration of the dataset, speech features, and ML/DL architectures utilized in recent studies.

2.1. 40th Mel-filterbank with LSTM-SVM (40M-LSTM-SVM) and Short-term Fourier Transform with Convolutional Recurrent Neural Network (STFT-CRNN)

A study conducted by [10] served as an extension of the research carried out by [12], who initially utilized the same unscripted dataset, known as The Multimodal dataset. This dataset encompassed both audio and visual recordings and involved 56 participants aged between 20 and 30 years, all native speakers of Korean. Participants were subjected to stress-inducing activities, including an English interview, watching an irritating video, and engaging in a repetitive activity (crossing out the same letter). Notably, both studies exclusively utilized audio data, neglecting video recordings.

Despite using the same dataset, the two studies employed different speech features and deep learning (DL) architectures. The rationale behind this deviation lies in the suboptimal accuracy of the earlier work by [12], which stood at 66.40%. Consequently, [10] aimed to enhance the model's accuracy. [10] extracted the 40th Mel-filterbank coefficients during feature extraction, employing the Hann window for audio signal analysis. The normalization procedure included zero mean and unit variance extraction. Their DL architecture incorporated the fusion of Long-Short Term Memory (LSTM) and Support Vector Machine (SVM), known as LSTM-SVM.

Meanwhile, [10] utilized Short-term Fourier Transform (STFT) as the speech feature and adopted a fused architecture of CNN and Recurrent Neural Network (RNN), referred to as CRNN, for classification. Additionally, multi-head attention and a gradient reversal layer were integrated into their DL architecture. Despite achieving a notable performance boost with a 10.90% increase, reaching a 77.30% accuracy, the outcome remains unsatisfactory due to numerous misclassifications of features into incorrect classes. This issue renders the DL model unreliable and necessitates further improvement.

2.2. Spectrogram images with CNN (Spec-CNN)

In a related study within the same field, [9] employed the Distress Analysis Interview Corpus (DAIC) dataset, curated for clinical interviews tailored to diagnose psychological distress issues. The study utilized spectrogram images derived from the raw audio speech obtained during these interviews, and Convolutional Neural Network (CNN) architecture was selected for the classification task. However, it was observed that the spectrogram feature fell

https://doi.org/10.31436/iiumej.v25i2.3207

short of being the optimal choice, as it failed to effectively emphasize the crucial characteristics inherent in distressed speech production.

2.3. Spectral, Cepstral and Prosodic with LSTM (SCP-LSTM)

Stress speech data was collected from recorded call-center interactions, encompassing 363 audio recordings of service interactions between customers and customer service representatives from two pension service providers in the Netherlands [3]. Initially intended for discrete emotion categorization, the dataset covered a range of emotions, including anger/disgust, neutral/sadness, surprise/fear, and happiness. Subsequently, the data was repurposed for stress binary classification, building upon the earlier discrete emotion categorization.

Various spectral, cepstral, and prosodic features were employed in this study, comprising zero-crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, mel-cepstral coefficients, chroma vector, and chroma deviations. The authors adopted Long Short-Term Memory (LSTM) with a combination of attention layers to facilitate the classification process. However, it is noteworthy that the study encountered challenges due to imbalanced class distribution, resulting in unsatisfactory outcomes, with an accuracy of only 80.00%.

2.4. MFCCs with CNN (MFCCs-CNN)

Ten participants from the International Islamic University Malaysia (IIUM) engaged in individual virtual interviews conducted through the Discord application [8]. During these interviews, participants were directed to verbally answer the Perceived Stress Scale 10 (PSS-10) questionnaire, comprising ten questions with responses rated on a scale from 0 to 4 (0 for Never, 1 for Almost Never, 2 for Sometimes, 3 for Often, and 4 for Very Often).

The collected data was labeled based on the PSS-10 final score. Notably, the authors opted for binary classification instead of the originally introduced three stress level classifications (low-stress, medium-stress, and high-stress). However, the study reported a suboptimal accuracy of only 61.00%. The utilization of CNN as a classifier was deemed unsuitable due to the limited amount of data points, totaling 150.

Considering the low volume of data points, the authors suggested employing a machine learning (ML) classifier instead of CNN to potentially enhance performance. This adjustment was recommended with the anticipation that an ML classifier could yield improved results under these specific data constraints.

2.5. Pitch, Jitter, Short-term energy variance with 10-fold cross-validation (PJS-10fcv)

Thirty participants volunteered to partake in a stress experiment, spanning ages from 18 to 24 years old [15]. The experimental protocol commenced with participants being tasked to respond to three arithmetic questions interspersed with four passage readings, each requiring two minutes for completion.

The speech features incorporated in this study encompassed all four speech types: prosodic, spectral, cepstral, and Teager Energy Operator (TEO)-based features. Notably, the authors identified pitch, jitter, and short-term energy variance as the most effective speech features in highlighting stress characteristics within the produced speech. Furthermore, the stress classification task was executed utilizing the 10-fold cross-validation (10fcv) method.

https://doi.org/10.31436/iiumej.v25i2.3207

Despite the promising findings, the study limited its focus to binary classification. Regrettably, the classifier encountered an overfitting issue in processing female data, achieving a perfect accuracy of 100.00%, indicative of potential classifier limitation.

2.6. TRIpLet Loss network-vectors with CNN (TRILL-CNN)

The StressDat dataset was employed in a study comprising 30 participants, where various potentially stressful scenarios were presented, spanning low, medium, and high stress levels [11]. Each scenario comprised multiple interconnected sentences, and participants were tasked with reading all scenario sentences across the three stress levels.

The study compared various speech features, including traditional ones such as Low-Level Descriptors (LLD), MFCCs, and Perceptual Linear Prediction (PLP). Additionally, modern non-semantic speech features like TRIpLet Loss network (TRILL) vectors and x-vectors were evaluated. Multiple ML and DL architectures were compared, including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), CNN, and LSTM. Notably, our study adopted a multi-class stress classification approach, distinguishing between low-stress, medium-stress, and high-stress categories.

The study identified the TRILL-vectors with CNN as the most effective combination, achieving an accuracy of 81.86%. In our study, the goal is to enhance performance by exploring the synergy between MFCCs and CNN within the same multi-class classification framework.

3. METHODOLOGY

In this section, the methodology encompasses the various activities conducted throughout the study, including data acquisition, pre-processing, feature extraction, and, ultimately, stress classification. The proposed system architecture of our study is illustrated in Fig. 1.

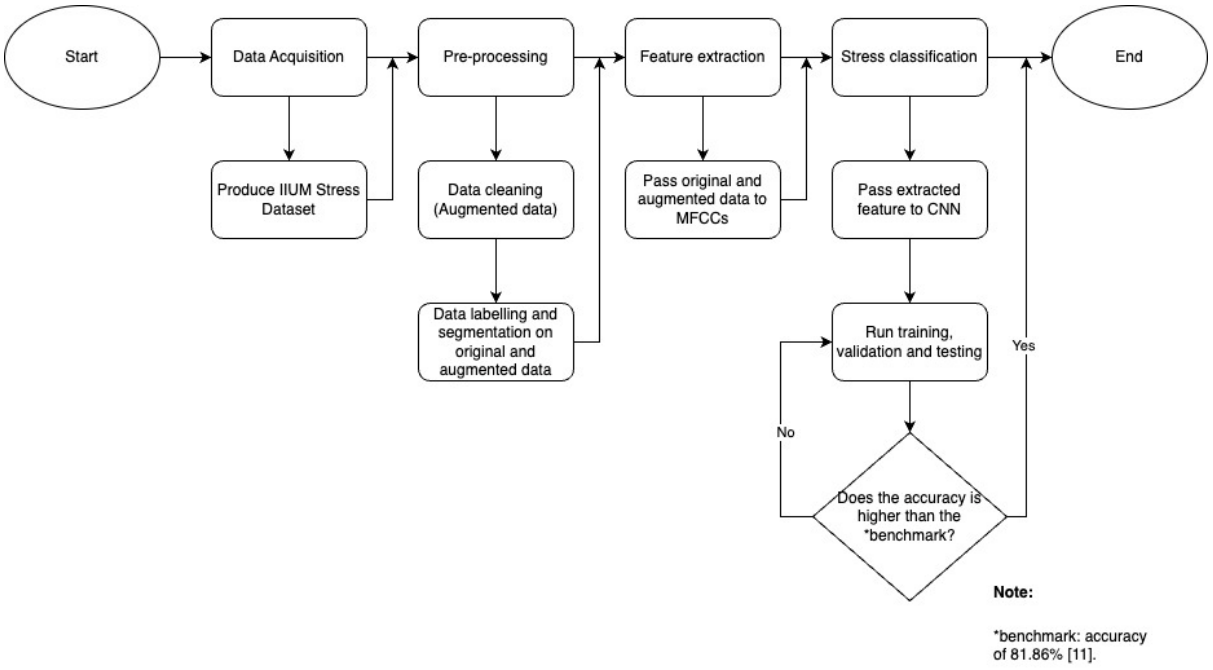

Figure 1. Proposed system architecture of our study

In addition, the study utilized Python programming with several other libraries, such as Pandas, Numpy, OS, and others. Also, the feature extraction and classifier procedures

https://doi.org/10.31436/iiumej.v25i2.3207

utilized the Pytorch framework. The utilization of this library was motivated by the preference to harness the Graphical Processing Unit (GPU) for accelerated learning.

3.1. Data Acquisition

The experiment aimed to elicit stress responses from participants during stress-inducing activities, and the resulting dataset was named the IIUM Stress Dataset. The activities within the experiment were categorized into three parts: pre-stress activities, stress-induced activities, and post-stress activities, as detailed in Table 1.

Table 1. Flow of experiment of IIUM Stress Dataset

Activity	Tasks during activity
Pre-stress	1. Participant self-introduction and experiment setup
	2. Pre-stress measurement: PSS Questionnaire
Stress-induced	1. Arithmetic Questions
	2. Speaking Proficiency Assessment (SPA)
	3. Job Interview Session
Post-stress	1. Peak-stress measurement: PSS Questionnaire
	2. Post-stress recovery: watching breathing a video
	3. Feedback and wrap-up

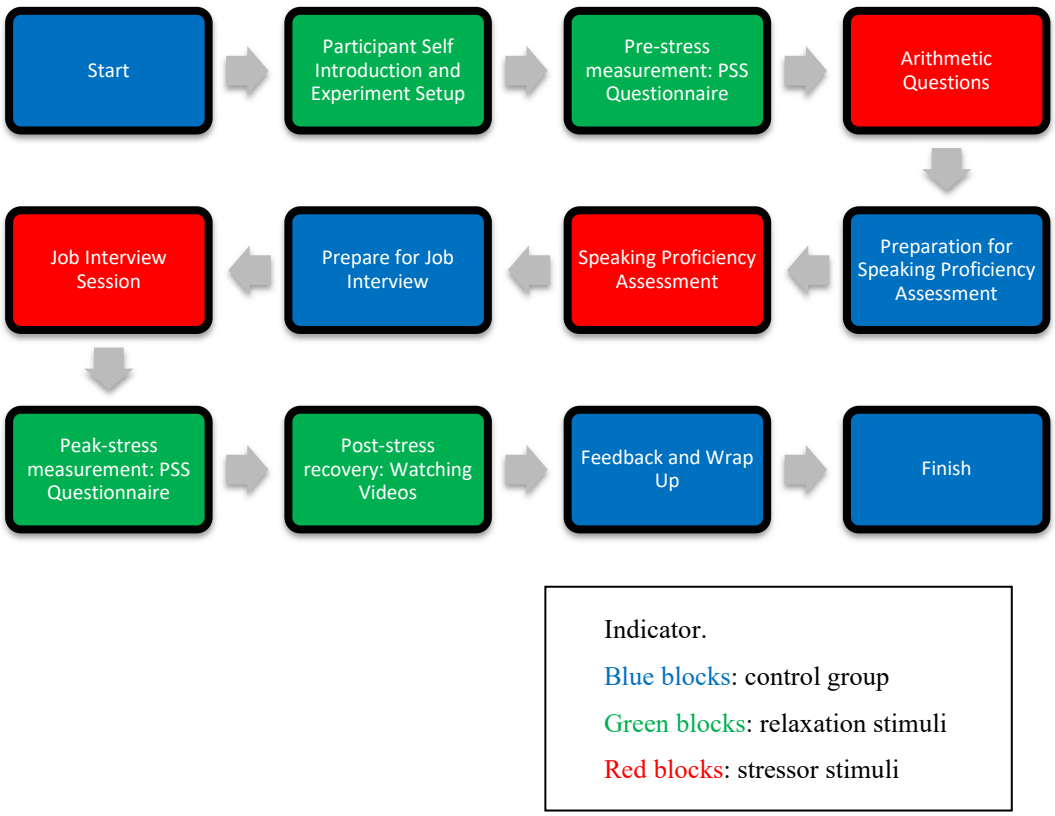

Figure 2. Experiment flow for the IIUM Stress Dataset (pilot and actual studies)

The experimental sequence commenced with pre-stress activities, followed by stress-induced activities, and concluded with post-stress activities. After an in-depth review of the literature, our study selectively incorporated two stress experimental studies that were

https://doi.org/10.31436/iiumej.v25i2.3207

particularly relevant to our case. Firstly, our study employed the same stimuli pattern created by [14], involving relaxation-stressor-relaxation phases. Secondly, the experimental flow was adapted from the Trier Social Stress Test (TSST) conducted by [16], incorporating time-induced stress tasks proven to elicit stress responses in participants. Additionally, the use of the PSS-5 Questionnaire was adapted from the same study by [16].

Fig. 2 illustrates the complete experiment procedure for the IIUM Stress Dataset. The indicator box showcases blue, green, and red blocks representing the control group, relaxation stimuli, and stressor stimuli, respectively. Notably, a pilot study was conducted before the actual study, providing valuable insights into potential flaws in our methodology, which were subsequently refined in the main study. No supplementary experimental figure is deemed necessary, as no modifications were introduced to the experimental flow for both the pilot and actual studies.

3.2. Pre-processing

Several steps were undertaken during the pre-processing stage, encompassing data cleaning, labeling, and segmentation. Our study employed two types of speech data: original (unaltered) and augmented (edited) data. The augmentation process utilized the Adobe Podcast application, which not only underwent the cleaning process but also underwent several changes, including cleaning up speech recordings, removing background noise, enhancing the overall sound quality by adjusting the pitch, and others.

Both sets of data were labeled based on their respective classes. Class identification for each participant's speech production relied on the PSS-5 questionnaire results, serving as the ground truth for the participants' current stress states. For instance, if the pre-stress PSS-5 result indicated low stress, the speech audio recorded during the pre-stress activities was categorized as low stress. Similarly, if the peak-stress PSS-5 result indicated high stress, the speech audio recorded during stress-induced and post-stress activities was categorized as high-stress.

Given the requirement for the same audio clip length as input in DL, the labeled data was further segmented into five-second snippets for use as input to the classifier. These audio snippets were saved according to their respective stress classes: low-stress, medium-stress, and high-stress. The total number of data points and hours of the snippets are detailed in Table 2.

Table 2. The summary of the data points encompasses both the original and augmented stress speech data

Stress Class	Total Data Points	Total Hours
Low-stress	759	1 hr 6 min 10 sec
Medium-stress	766	1 hr 3 min 50 sec
High-stress	794	1 hr 3 min 15 sec
Total	**2319**	**3hr 13 min 15 sec**

3.3. Feature Extraction

The extraction and representation of features play a pivotal role in influencing the performance of algorithms, particularly in the context of deep learning with speech signals. Feature extraction involves the identification and processing of concealed information within a raw data signal. By discerning and eliminating ineffective features while retaining crucial data, this process streamlines data management.

In our investigation, we employed Mel-Frequency Cepstral Coefficients (MFCCs) as the speech feature to elucidate the characteristics of stress speeches. This feature was derived from the Torchaudio library developed by PyTorch. The framework of MFCCs is illustrated in Fig. 3.

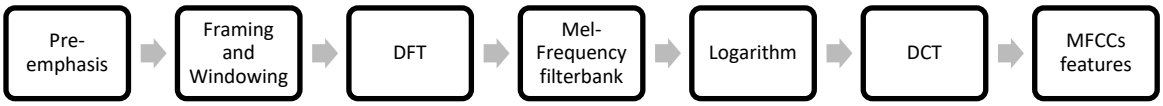

Fig. 3. MFCCs framework [17]

The elaboration of each procedure was as follows [17]:

- **Pre-emphasis**: Pre-emphasis, the first stage of MFCCs is easily achieved by utilizing a high-pass filter with an ordinary filter coefficient (a) of 0.97. The energy distribution across frequencies and the overall energy level were altered by the filtering procedure. Additionally helpful in removing numerical problems during the Fourier Transform operation, this pre-emphasis filter can improve the Signal-to-Noise (SNR) ratio.
  The pre-emphasis filter can be applied to a signal using the first-order filter in Eq. (1) [17].

$$y(t) = x(t) - a(t-1) \quad (1)$$

- **Framing and Windowing**: To make the signal more stationary, the raw signals were divided into frames. Then, speech needs to be assessed briefly enough to have consistent acoustic characteristics. In the speech signal, since there was roughly 20 ms between two glottal closures, 20–30 ms was identified as a Quasi-Stationary Segment (QSS). Conversely, vowel voices were reported to be captured between 40 and 80 ms. Due to this, short-term spectrum measurements were frequently carried out at intervals of 20 ms, with a 10-ms overlap between each frame. Then, tracking the temporal characteristics of the voice signal was possible with frame overlaps of 10 ms.
  A window was employed by each frame to restrict the signal close to the frame's edge. Eq. (2) gave the equation for the Hamming windows, which were chosen for this investigation [17].

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), 0 \le n \le N. \text{ Window length: } L = N + 1. \quad (2)$$

- **Power Spectrum**: A power spectrum, in general, is the signal's frequency components' power distribution. The power spectrum was computed using the DFT. Each frame's power spectrum was determined using Eq. (3). The signal length was N, and the discrete signal was $x(n)$ [17]:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-\frac{2\pi jnk}{N}}, k = 1, 2, 3 \ldots N-1 \quad (3)$$

- **Mel Filter Bank**: The Mel band-pass filter was a collection of filters based on pitch perception. The Mel filter function was designed to extract a non-linear representation of the speech stream, much like the human ear does. On a Mel scale, the standard Mel filter bank consisted of 40 triangular filters. Each triangular filter in the filter bank had an output of 1 at the centre frequency and was linearly lowered towards 0 until it reached the centre frequencies of the two adjoining filters, where the result was 0. Eq. (4) computes the transfer function (TF) of each m-th filter. Where $\sum_{m}^{M-1} H_m(k) = 1$, and the triangle filter's centre frequency is $f(m)$ [17].

https://doi.org/10.31436/iiumej.v25i2.3207

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \le k < f(m) \\ 1, & k = f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (4)$$

Eq. (5) and Eq. (6) were used to compute the conversion of the Mel scale (m) to Hertz (f) and vice versa [17].

$$m = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \quad (5)$$

$$f = 700(10^{\frac{m}{2595}} - 1) \quad (6)$$

- **Discrete Cosine Transform (DCT)**: A limited sequence of data points called a Discrete Cosine Transform (DCT) is used to express a sum of cosine functions that oscillate at different frequencies. To choose the most accelerative coefficients or eliminate the relationship in the log spectral magnitudes from the filter bank, the DCT was applied to the Mel filter bank during the MFCCs process. Eq. (7) was used to compute the DCT. The discrete signal, denoted by x_n in this equation, was of length N [17].

$$X(k) = \sum_{n=0}^{N-1} x_n * \cos\left(\frac{2\pi jnk}{N}\right), k = 1, 2, 3 \ldots N-1 \quad (7)$$

3.4. Stress Classification

Our study utilized Convolutional Neural Network (CNN), a deep learning algorithm, for the multi-class stress classification task. The feature vectors extracted in the previous stage were fed into CNN. The dataset consisted of original and augmented data, and the training set included both. The reason to include the augmented data in the training set was to increase the data points and improve the performance and generalization of the CNN model by having high diversity data on the training set as referred to a finding from a previous study [18]. Meanwhile, the validation and test sets included the original data only. The reason for not using augmented data on the test and validation data was to give an accurate assessment of the model [19].

The dataset was split into three sets: train, validation, and test, with the percentages of each set being 80%, 10%, and 10%, respectively. Both sets of data were combined and used in the train set, while the validation and test sets only comprised the original data. All the data was randomly shuffled before going into CNN. Fig. 4 provides the data point arrangement in each set.

The depicted CNN algorithm, as illustrated in Fig. 5, is composed of six 2-dimensional convolutional layers (Conv2d) interconnected with a Rectified Linear Unit (ReLU) layer and a 2-dimensional Max-pooling (MaxPoold2d) layer placed on each Conv2d layer. Subsequently, after the Conv2d layers, flatten layers were connected to three additional linear layers. A ReLU layer is incorporated at the initiation of the first linear layer, which is then linked to the output layer containing three stress classes. The proposed hyperparameter settings for MFCCs are outlined in Table 3.

https://doi.org/10.31436/iiumej.v25i2.3207

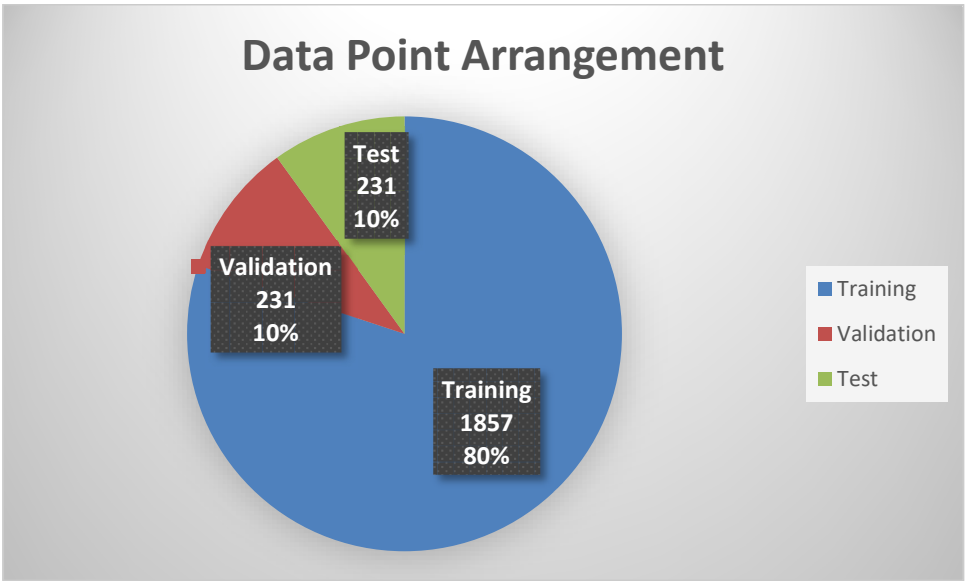

Fig. 4. Arrangement of the data points in each set

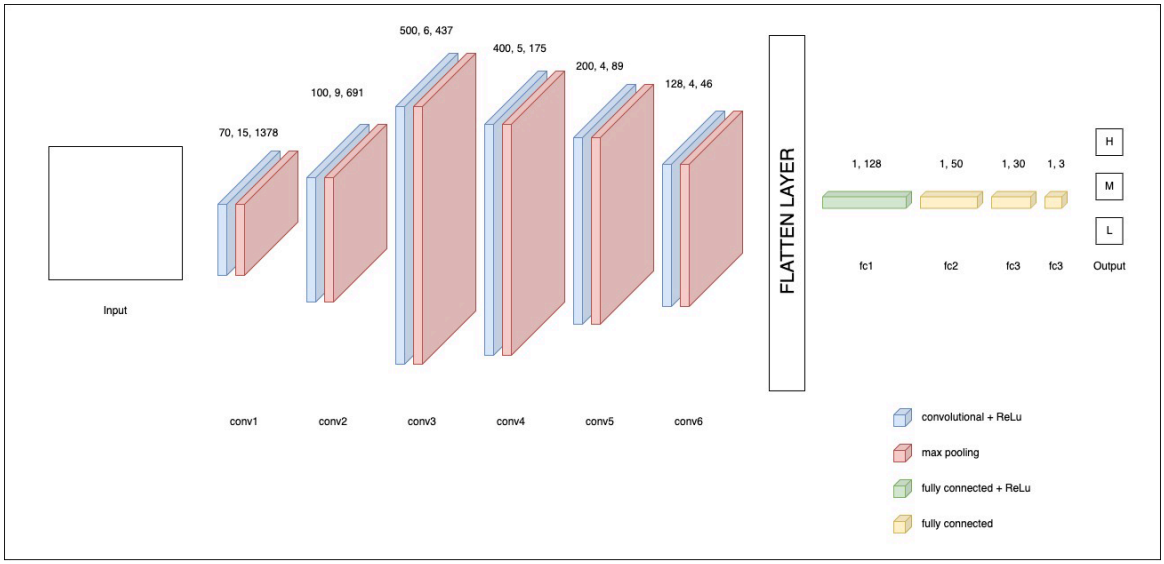

Fig. 5. Proposed CNN algorithm

Table 3: Proposed hyperparameter settings

Epoch	Loss Function	Optimizer	Learning Rate	Momentum
200	Cross-Entropy Loss	Stochastic Gradient Descent	0.0001	0.7

4. RESULT AND DISCUSSION

This section presents a comprehensive examination of the findings, covering the results of the proposed system and a comparative analysis against benchmarks. Firstly, the outcomes from the proposed system include various metrics, such as accuracy, F1-score, individual class accuracies, false positive rate, precision, and recall. Additionally, visual representations of the confusion matrix, accuracy, and loss graphs are provided. Secondly, the comparative analysis against benchmarks is conducted in terms of accuracy only, as the papers available in the literature predominantly reported accuracy metrics in their findings.

https://doi.org/10.31436/iiumej.v25i2.3207

Our study produced better accuracy across all sets: 99.14% for the training set, 92.21% for the validation set, and 95.67% for the test set. As depicted in Fig. 6, the proposed system demonstrated efficient classification of features, exhibiting minimal misclassifications—specifically, fewer than four errors.

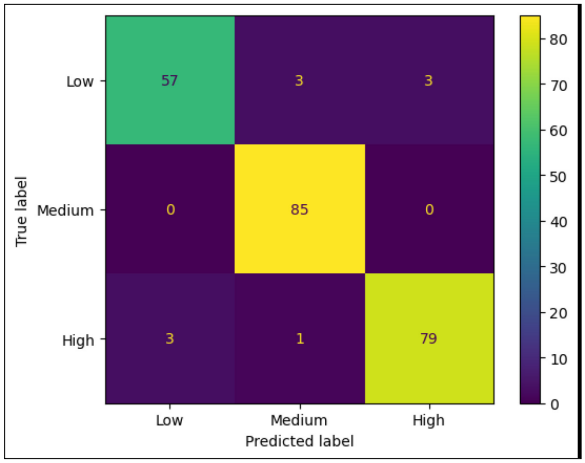

Fig. 6. Confusion matrix of the proposed system

Meanwhile, Table 4 furnishes the performance metrics corresponding to each class for the proposed system, derived from the generated confusion matrix. All scores are normalized to range of 0 to 1. The F1 score for multi-class classification was computed as mean of the F1 scores across individual classes. The proposed system yielded a score approaching 1, indicative of elevated precision and recall, signifying better overall performance.

Table 4: The performance metrics based on each class

Class	Accuracy	False Positive Rate	Precision	Recall	F1-score (for all classes)
Low-stress	0.961	0.018	0.950	0.905	
Medium-stress	0.983	0.027	0.955	1.000	0.960
High-stress	0.970	0.020	0.963	0.952	

Furthermore, as illustrated in Figures 7 and 8, there was a clear positive correlation between the number of epochs and accuracy. The proposed CNN algorithm showed a gradual increase in accuracy over time, unlike the baseline, which showed no improvement. This trend contrasted with the loss curve behavior, where the proposed system experienced a gradual decline in loss as the number of epochs increased, while the baseline system again showed no improvement. This observation supported the claim that the proposed system demonstrated robust generalization capabilities and continuous improvement over time. Notably, the decision to set the final epoch number at 200 was deliberate, as further increases indicated signs of overfitting.

https://doi.org/10.31436/iiumej.v25i2.3207

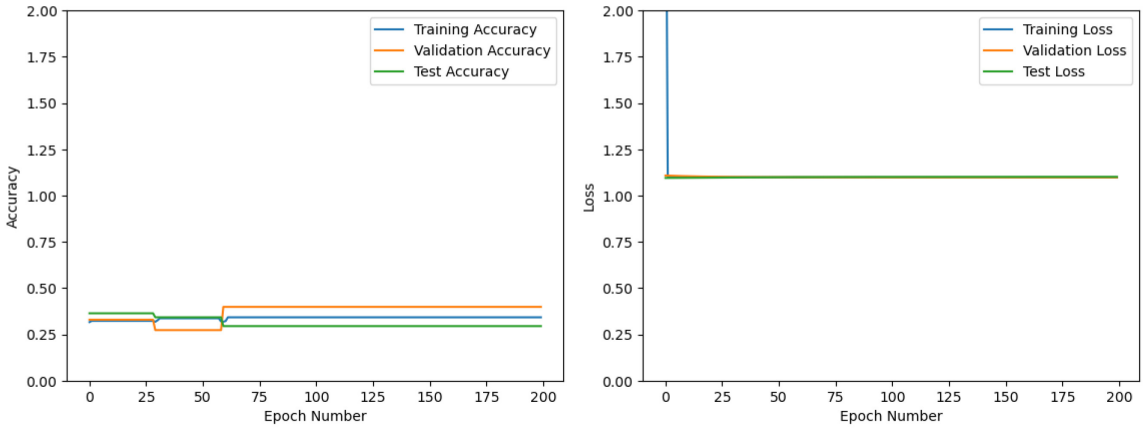

Fig. 7. Accuracy and loss curves for the baseline CNN algorithm, which consists of one layer each of Conv2d, ReLU, MaxPool2d, and Linear

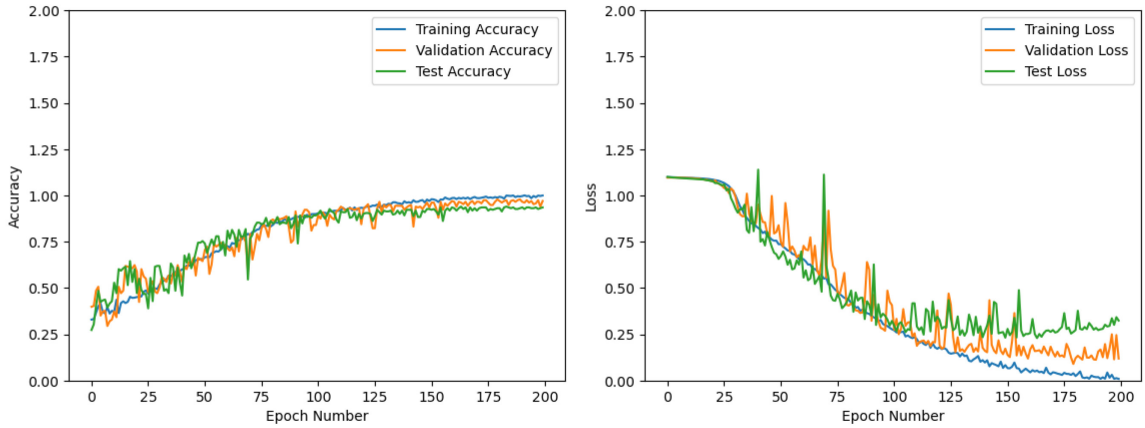

Fig. 8. Accuracy and lost curves for the proposed CNN algorithm

The proposed system in our paper demonstrated the best performance metrics for an unscripted dataset, as evidenced by the results. The study revealed that Mel-Frequency Cepstral Coefficients (MFCCs) exhibit the capability to highlight essential speech characteristics that differentiate between low-stress, medium-stress, and high-stress categories. The implementation of CNN proved instrumental in accurately categorizing the input into the respective stress levels.

Moreover, our work conducted a comparison of performance metrics with several benchmark studies. The papers selected were using one unscripted/freestyle dataset, which was similar to our study. Thus, the comparison was made due to the similar nature of the chosen database. Given that the available metrics from other studies were solely accuracy-related, our study specifically focuses on accuracy, as illustrated in Table 5. Notably, our study achieved an improved accuracy of 95.67%, surpassing previous studies [3, 8-11, 12] with lower accuracy that predominantly engaged in binary classification. While the parameter used in the previous study [8] was similar to our study, variations in the data distribution, including the incorporation of augmented data that increased the number of data points, as well as differences in the selection of CNN parameters' algorithm, collectively contributed to improving the overall performance of our study. In addition, the inclusion of multiple stress classes in our study enhances the accurate categorization of stress speech into respective stress levels. The

https://doi.org/10.31436/iiumej.v25i2.3207

proposed system had better performance compared to previous studies, and this could serve as the framework for multi-class speech stress classification.

Table 5. Comparison between our study and previous research utilizing an unscripted dataset

No.	Speech Features-Classifier	Distribution of Data	Accuracy (%)	Class
1	(MFCCs-CNN) [8]	Training set: 112 datapoints Test set: 38 datapoints	61.00	Two
2	(40M-LSTM-SVM) [12]	Training set: 9.52 hours Test set: 2.36 hours	66.40	Two
3	(Spec-CNN) [9]	Training set: 2627 datapoints Test set: 642 datapoints	77.00	Two
4	(STFT-CRNN) [10]	Training set: 9.52 hours Test set: 2.36 hours	77.30	Two
5	(SCP-LSTM) [3]	Training set: 3738 datapoints Test set: 934 datapoints	80.00	Two
6	(TRILL-CNN) [11]	Training set: 9208 datapoints Test set: 1023 datapoints	81.86	Three
7	**Our study (MFCCs-CNN)**	**Training set: 1857 datapoints Validation set: 231 datapoints Test set: 231 datapoints**	**95.67**	**Three**

As highlighted in Section 2, MFCCs were widely employed in stress speech recognition due to their outstanding performance. The positive outcomes suggested that MFCCs possessed the capability to accentuate stress-related characteristics in the produced speech, leveraging their features that emulated the response of the human auditory system to sound. Subsequently, the extracted features were categorized into three stress classes with the assistance of CNN. Based on the literature [6, 9-11], CNN performed better in classification tasks due to its ability to effectively capture spatial hierarchies and patterns in images. Their ability to acquire features in a hierarchical and translation-invariant way was made possible by these architectural decisions, which were essential for comprehending complicated visual data.

5. CONCLUSION

The research aimed to classify speech stress into three classes—low-stress, medium-stress, and high-stress—using a deep learning approach to improve upon a prior three-class speech classification task. MFCCs were chosen as the speech feature and integrated with CNN for stress classification. The IIUM Stress dataset, derived from an experimental investigation involving Malay students with tertiary education backgrounds, provided unscripted data representing stressed speech characteristics. The dataset parameters ensured the generation of authentic stress-related speech data, subsequently transformed into MFCC features for stress classification using CNN.

The findings demonstrate that the utilization of MFCCs yielded a significantly higher level of accuracy, achieving 95.67% when distinguishing the three stress classes. Importantly, our analysis has been proven to outperform prior studies that used an unscripted dataset in terms of accuracy, highlighting the effectiveness of the proposed system in generating the best stress classification model. In addition, the utilization of an augmented data strategy, which increased the total number of data points, combined with the integration of MFCCs and CNN, proved effective in improving the overall performance.

The significance of conducting our study lies in its impactful contributions. Firstly, our study produced new evidence for the research field, demonstrating the effectiveness of augmented data strategies and the integration of MFCCs with CNN. Secondly, our study produced an unscripted dataset that comprised authentic recordings of stressed speakers' original speeches. Thirdly, our study proved able to categorize the dataset into three distinct stress classes (low-stress, medium-stress, and high-stress), effectively delineating speakers' perceived stress levels. Lastly, the research yielded improved results, achieving a better accuracy rate, surpassing the benchmarks set by the previous studies in the related field.

In future research, a thorough examination and comparison of different speech features can be conducted to discern the most effective feature that results in superior performance metrics for the identification and classification of individual stress levels. This exploration could involve an in-depth analysis of the strengths and weaknesses of various features, shedding light on their respective contributions to accurate stress level categorization. Such a comprehensive investigation would contribute valuable insights to the field, guiding the development of more refined and effective stress classification systems.

ACKNOWLEDGEMENT

This research was made possible by the Fundamental Research Grant Scheme (FRGS), which is administered by the Ministry of Higher Education (MoHE) and was granted funding with the reference code FRGS/1/2021/TK0/UIAM/02/29.

REFERENCES

[1] S. A. Kriakous, K. A. Elliott, C. Lamers, and R. Owen, "The Effectiveness of Mindfulness-Based Stress Reduction on the Psychological Functioning of Healthcare Professionals: a Systematic Review," *Mindfulness (N Y)*, vol. 12, no. 1, pp. 1–28, Jan. 2021, doi: 10.1007/s12671-020-01500-9.

[2] S. Liu, A. Lithopoulos, C.-Q. Zhang, M. A. Garcia-Barrera, and R. E. Rhodes, "Personality and perceived stress during COVID-19 pandemic: Testing the mediating role of perceived threat and efficacy," *Pers Individ Dif*, vol. 168, p. 110351, Jan. 2021, doi: 10.1016/j.paid.2020.110351.

[3] S. Bromuri, A. P. Henkel, D. Iren, and V. Urovi, "Using AI to predict service agent stress from emotion patterns in service interactions," *Journal of Service Management*, vol. 32, no. 4, pp. 581–611, 2020, doi: 10.1108/JOSM-06-2019-0163.

[4] H. Aouani and Y. Ben Ayed, "Speech Emotion Recognition with deep learning," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 251–260. doi: 10.1016/j.procs.2020.08.027.

[5] S. Mihalache, D. Burileanu, and C. Burileanu, "Detecting Psychological Stress from Speech using Deep Neural Networks and Ensemble Classifiers," Institute of Electrical and Electronics Engineers (IEEE), Nov. 2021, pp. 74–79. doi: 10.1109/sped53181.2021.9587430.

[6] P. Chyan, A. Achmad, I. Nurtanio, and I. S. Areni, "A Deep Learning Approach for Stress Detection Through Speech with Audio Feature Analysis," Institute of Electrical and Electronics Engineers (IEEE), Mar. 2023, pp. 1–5. doi: 10.1109/icitisee57756.2022.10057845.

[7] N. A. Zainal, A. L. Asnawi, A. Z. Jusoh, S. N. Ibrahim, H. A. M. Ramli, and N. F. M. Azmin, "MFCCs and TEO-MFCCs for Stress Detection on Women Gender through Deep Learning Analysis," in *2023 9th International Conference on Computer and Communication Engineering (ICCCE)*, IEEE, Aug. 2023, pp. 283–288. doi: 10.1109/ICCCE58854.2023.10246098.

[8] M. S. Hafiy Hilmy *et al.*, "Stress Classification based on Speech Analysis of MFCC Feature via Machine Learning," in *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*, IEEE, Jun. 2021, pp. 339–343. doi: 10.1109/ICCCE50029.2021.9467176.

[9] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 618–628. doi: 10.1016/j.procs.2019.12.228.

[10] H.-K. Shin, H. Han, K. Byun, and H.-G. Kang, "Speaker-invariant Psychological Stress Detection Using Attention-based Network," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 308–313.

[11] J. Kejriwal, S. Benus, and M. Trnka, "Stress detection using non-semantic speech representation," in *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*, IEEE, Apr. 2022, pp. 1–5. doi: 10.1109/RADIOELEKTRONIKA54537.2022.9764916.

[12] H. Han, K. Byun, and H.-G. Kang, "A Deep Learning-based Stress Detection Algorithm with Speech Signal," in *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*, New York, NY, USA: ACM, Oct. 2018, pp. 11–15. doi: 10.1145/3264869.3264875.

[13] M. E. Damayanti and L. Listyani, "AN ANALYSIS OF STUDENTS' SPEAKING ANXIETY IN ACADEMIC SPEAKING CLASS," *ELTR Journal*, vol. 4, no. 2, pp. 152–170, Aug. 2020, doi: 10.37147/eltr.v4i2.70.

[14] X. T. Tee, T. A. T. Joanna, and W. Kamarulzaman, "Self-regulatory Strategies Used by Malaysian University Students in Reducing Public Speaking Anxiety: A Case Study," *Proceedings of the 2nd International Conference on Social Science, Humanities, Education and Society Development (ICONS 2021)*, vol. 629, no. Icons 2021, pp. 146–152, 2022, doi: 10.2991/assehr.k.220101.023.

[15] N. Li, N. Li, M. Guo, and J. Feng, "Research of Speech Biomarkers for Stress Recognition Using Linear and Nonlinear Features," in *2021 7th International Conference on Computer and Communications, ICCC 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 509–513. doi: 10.1109/ICCC54389.2021.9674330.

[16] Y. S. Can, D. Gokay, D. R. Kılıç, D. Ekiz, N. Chalabianloo, and C. Ersoy, "How laboratory experiments can be exploited for monitoring stress in the wild: A bridge between laboratory and daily life," *Sensors (Switzerland)*, vol. 20, no. 3, Feb. 2020, doi: 10.3390/s20030838.

[17] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," *IEEE Access*, vol. 10. Institute of Electrical and Electronics Engineers Inc., pp. 122136–122158, 2022. doi: 10.1109/ACCESS.2022.3223444.

[18] D. C. Marcu and C. Grava, "The Importance of Data Quality in Training a Deep Convolutional Neural Network," in *2023 17th International Conference on Engineering of Modern Electric Systems (EMES)*, IEEE, Jun. 2023, pp. 1–4. doi: 10.1109/EMES58375.2023.10171785.

[19] Saulo Barreto, "Data Augmentation | Baeldung on Computer Science." Accessed: Jul. 24, 2023. [Online]. Available: https://www.baeldung.com/cs/ml-data-augmentation