https://doi.org/10.31436/iiumej.v26i1.3411

UTILIZING MFCCS AND TEO-MFCCS TO CLASSIFY STRESS IN FEMALES USING SSNNA

NUR AISHAH ZAINAL[1], ANI LIZA ASNAWI[1*], SITI NOORJANNAH IBRAHIM[1], NOR FADHILLAH MOHAMED AZMIN[1], NORHARYATI HARUM[2], NORA MAT ZIN[3]

[1] Department of Electrical and Computer Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, 53100, Kuala Lumpur, Malaysia
[2] Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, Malaysia
[3] Department of Psychiatry, Kulliyyah of Medicine, International Islamic University Malaysia, Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia

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

(Received: 29 August 2024; Accepted: 4 December 2024; Published online: 10 January 2025)

ABSTRACT: All individuals are susceptible to experiencing stress in their everyday lives. Nevertheless, stress has a greater influence on females due to both biological and environmental factors. This study utilized female speeches to detect and classify stress and no stress in women. Using speech, composed of non-invasive and non-intrusive approaches, helps to identify stress better in females. A comparative analysis was conducted between Mel-frequency Cepstral Coefficients (MFCCs) and Teager Energy Operator- MFCCs (TEO-MFCCs) to determine the best speech feature for classifying emotions associated with stress and no-stress conditions for female voices. With the assistance of the Stress Speech Neural Network Architecture (SSNNA), an improved accuracy of 93.9% was achieved. This research showed that MFCCs enhanced higher-frequency components in stressed speech, distinguishing between stress and no-stress classes. This study shows that SSNNA achieved high accuracy with 14 female voices, confirming its ability to function independently of speaker identity.

ABSTRAK: Semua individu terdedah kepada stres dalam kehidupan seharian mereka. Walau bagaimanapun, stres memberi pengaruh yang lebih besar terhadap wanita akibat faktor biologi dan persekitaran. Kajian ini menggunakan ucapan untuk mengesan dan mengklasifikasikan stres dan tiada stres dalam kalangan wanita. Penggunaan ucapan, yang merupakan pendekatan tidak invasif dan tidak mengganggu, membantu mengenal pasti tekanan dengan lebih baik dalam kalangan wanita. Analisis perbandingan telah dijalankan antara Mel-frequency Cepstral Coefficients (MFCCs) dan Teager Energy Operator-MFCCs (TEO-MFCCs). Tujuannya adalah untuk menentukan ciri ucapan terbaik bagi mengklasifikasikan emosi yang berkaitan dengan keadaan stres dan tiada stres bagi suara wanita. Dengan bantuan Stress Speech Neural Network Architecture (SSNNA), metrik prestasi yang lebih tinggi dengan ketepatan 93.9% telah dicapai. Penyelidikan ini menunjukkan bahawa MFCCs meningkatkan komponen frekuensi tinggi dalam ucapan yang stres, secara efektif membezakan antara kelas stres dan tiada stres. Kajian ini menunjukkan bahawa SSNNA mencapai ketepatan tinggi dengan 14 suara wanita, mengesahkan ia berfungsi secara bebas daripada identiti penutur.

KEYWORDS: *stress detection via speech, stress classification for females, MFCCs, CNN*

https://doi.org/10.31436/iiumej.v26i1.3411

1. INTRODUCTION

Both females and males can experience stress in their daily lives. Stress can arise from various aspects of human life. However, when exposed to the same stressors, females tend to perceive higher stress levels than males. According to a 2023 survey report conducted by a human resources business in the United States, 43% of female workers rated their mental health as poor, compared to 15% of male workers [1]. Also, a study conducted by Costa et al. [2] on unemployment issues in 2021 found that 22.7% of females reported higher stress levels than 11% of males.

According to a report by the Sinar Harian newspaper in 2019, females experienced depression (a result of long-term stress) more frequently than males due to the hormonal transition [2]. Moreover, in 2019, Dr Elinda Tunan, a psychiatrist at a government hospital, concurred that there was a strong correlation between depression in females and hormonal fluctuations, particularly during pregnancy and childbirth [2]. These alterations facilitated a greater tendency for emotional expression. According to her, individuals were likelier to exhibit heightened moodiness, increased anger, and experience several other undesirable emotions [2].

Therefore, it is crucial to detect stress in females early to address its underlying causes and promote a healthy physical and mental lifestyle. Early detection of stress can help prevent physical and emotional disorders, which may lead to serious mental health problems if left unaddressed.

Speech for stress prediction presents an advantage to females by reducing unnecessary stress arising from equipment setup, particularly in healthcare settings where biological markers are used to measure stress levels [3]. Two examples of invasive and intrusive stress detection approaches include the utilization of an electroencephalogram (EEG) to measure brain electrical activity and collect blood samples [3]. These techniques have several drawbacks, including time-consuming procedures, the need for specialized skills, significant financial expenses, and the limitation of being exclusively conducted within healthcare facilities [3]. Therefore, to avoid unnecessary stress caused by these methods, our study uses speech to detect stress from female voices. This approach reduces the potential discomfort linked to invasive methods, as it solely requires participants to provide vocal samples without involving any invasive physical activities.

This study incorporated speech features such as Mel-frequency Cepstral Coefficients (MFCCs) and the fusion of MFCCs with the Teager Energy Operator (TEO)—the comparison aimed to identify the most suitable features that demonstrated significant performance metrics. Our study was motivated by a previous study that compared MFCCs and TEO-MFCCs in distinguishing emotions, which found that MFCCs performed better for female voices [3]. In our study, a binary classification task (e.g., stress and no stress) was performed with the assistance of deep learning technologies developed, known as Stress Speech Neural Network Architecture (SSNNA). This study utilized SSNNA to classify a combined dataset sourced from the Toronto Emotional Speech Set (TESS) [4] and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [5]. Our study aimed to contribute by developing a stress detection model capable of classifying scripted datasets, specifically focusing on identifying stress and non-stress in female voices.

The subsequent sections of this work were structured in the following manner: Section 2 provided an assessment of prior research studies that had focused on detecting speech stress, specifically targeting females, and reviewed the speech features and the classifiers used. The

methodology employed in this study, elaborated in Section 3, encompassed the processes of data acquisition, data preprocessing, feature extraction, data organization into stress and no-stress classes, as well as the classification architecture (SSNNA). The findings and interpretation presented in Section 4 were examined and discussed. The paper's final section presented the conclusion and potential avenues for future research.

2. RELATED WORKS

The trend in this research work could be divided into several key areas, including stress classification using speech features on females, MFCCs and TEO-MFCCs speech features, and the current study utilizing neural network architecture in stress speech classification. The following sub-sections elaborate on each area.

2.1. Stress Classification Using Speech Features on Female

Stress detection is crucial for maintaining good mental health. Prolonged, untreated stress can lead to significant breakdowns in well-being and severe mental health issues. Detecting and classifying stress through speech offers an alternative to traditional methods. While traditional methods can effectively detect and classify stress, they have several drawbacks, such as being invasive and intrusive. For example, inserting a medical instrument into the body to perform a cortisol test is not a favorable technique for some patients, as it is considered invasive. Moreover, it likely increased the patient's stress due to unfamiliar procedures.

Speech-based stress detection and classification have brought significant changes in the medical field. It helped detect and classify stress more favorably, as it was considered a non-invasive and contactless approach. The differences in the speech signals produced during stress and non-stress became an important tool for differentiating between these two conditions. For instance, a stressed person exhibited several symptoms while speaking, including a shaky voice and slurred speech, whereas a stress-free person did not [6]. The similarity of speech produced during stress and the differences between these two conditions helped create a better stress classification tool for healthcare providers. This innovation undoubtedly contributed to society.

Speech signals and vocal characteristics produced from one's utterances offered insights into one's emotional state, facilitating stress identification. In this research context, using speech to classify stress in females provided advantages, as it avoided the discomfort associated with invasive and intrusive clinical procedures. Females typically expressed themselves verbally, which enhanced the comfort and engagement of this method [7]. This approach reduced the unnecessary stress that traditional measurements might have induced. Moreover, it was cost-effective and could be conducted at home, promoting broader participation and accessibility.

2.2. MFCCs and TEO-MFCCs Speech Features

MFCCs were widely used in the image and audio signal processing fields due to their ability to represent the power spectrum of audio signals compactly [8]. Based on findings from previous researchers, MFCCs effectively highlighted the energies of higher-frequency components in speech signals produced by female speakers [3], [8]. Their studies accurately classified several emotions associated with stress, especially female voices. However, their study focused on classifying individual emotions (emotion classification). It did not combine stress and no-stress emotions into the same classes, despite the title referencing speech stress detection. This approach differed from the objective of speech stress classification, which aimed to identify stress and no-stress classes.

Another speech feature often compared with MFCCs was the fusion of TEO-MFCCs. Under stress, muscle tension in the speaker's vocal tract influences the airflow that generates sound. Consequently, non-linear speech characteristics were crucial for accurately detecting stress in speech, which was the essence of the TEO feature [8]. Previous studies found that combining these two features further improved accuracy for both male and female speakers [3], [8]. The studies agreed that TEO enhanced the energies of the emotional contents of speech—nevertheless, the classification focused on individual emotions rather than combining stress and no-stress emotions.

2.3. Current Study Utilizing Neural Network Architecture in Stress Speech Classification

Based on previous works, the usage of neural network architecture in research increased due to the rapid development of the artificial intelligence field [9], [10]. A previous study utilizing an unscripted dataset of 363 recorded call center service interactions [9] employed the Long Short-Term Memory (LSTM) model with an attention layer for the classification task. Their study achieved an accuracy rate of 80% in generating binary classifications for stress and no-stress classes. The authors concluded that the unsatisfactory accuracy was due to the type of dataset used. Further work needed to be done to determine whether the classifier was robust and adaptable to any type of dataset related to the study field.

Another prior study conducted a subsequent investigation utilizing the SUSAS dataset [10]. The study employed deep neural networks (DNNs) operating within an ensemble one-vs-one classification framework. The binary classification achieved the highest level of accuracy, with a single DNN classifier having an accuracy rate of 78%. According to this result, combining many neural networks did not inherently yield favorable performance outcomes. This observation demonstrated that as an algorithm's complexity increased, its accuracy level decreased.

Several key distinctions existed between our work and prior studies. First, our study analysis was limited to the stress prediction of the female class exclusively, with no inclusion of male remarks. To our knowledge, there was a lack of literature on developing classification models for identifying stress in females. Furthermore, our study adhered to the research cycle by replicating a previous study that compared two speech features: MFCCs and the fusion of TEO-MFCCs. This decision was based on the observation that these features could yield good performance metrics, especially for female voices. Finally, our study merged two speech datasets that contained identical emotion classes but with different female speakers to examine the robustness of our model. The subsequent methodology section provides a more detailed explanation of this matter.

3. METHODOLOGY

This section provided an overview of every step undertaken to accomplish the main goal. The following subsections give a detailed description of each step. The TESS and RAVDESS datasets were used for this study, and these two datasets were combined. Subsequently, the acquired data underwent preprocessing procedures. Speech features were extracted from the preprocessed data, and the extracted features were organized with stress and no-stress labels. Finally, the workflow of the deep learning algorithm was established. The schematic representation of our research methodology is depicted in Figure 1.

https://doi.org/10.31436/iiumej.v26i1.3411

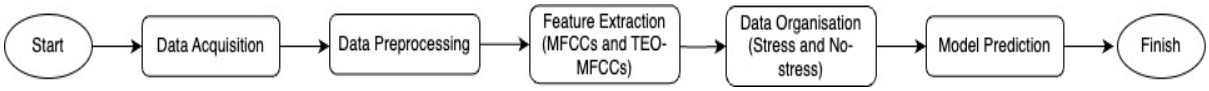

Figure 1. Study flow.

3.1. Data Acquisition

This section details each dataset used in the present investigation. Sections 3.1.1 and 3.1.2 detail the TEST and RAVDESS datasets, respectively.

3.1.1. TESS Dataset

The dataset consisted of female subjects. A total of 200 targeted English phrases were uttered by two actresses, aged 26 and 64, respectively. Recordings were captured for each of the seven emotions: anger, disgust, fear, happiness, neutrality, pleasant surprise, and sadness. In total, a collection of 2800 data points with a duration of 1-2 seconds in each snippet was created.

In the context of deep learning, achieving optimal outcomes necessitated using a balanced dataset across all classes. Consequently, our study decided to use a subset of six emotions instead of incorporating all seven to ensure balanced classes, omitting the disgust emotion. The emotional responses employed to measure stress encompassed anger, fear, and sadness, yielding a cumulative dataset of 1,200 data points. Furthermore, three other emotions were employed to represent the no-stress class: happiness, neutrality, and pleasant surprise. The dataset comprised 600 snippets representing the vocal expressions of young females, along with another 600 snippets representing the vocal expressions of elderly females.

3.1.2. RAVDESS Dataset

An additional dataset was introduced into the system to assess the reliability of our deep learning algorithm in distinguishing between stress and no-stress classes [11]. It was crucial to ensure that both datasets included identical types of emotion classes to achieve dependable and accurate results [11]. In the second dataset, a total of 24 actors and actresses were included, evenly distributed between 12 females and 12 males. These individuals were recorded vocalizing two English phrases. The speech encompassed a range of emotional expressions, including calmness, happiness, sadness, anger, fear, surprise, and disgust.

To achieve the desired outcome, twelve female speeches were selected, encompassing a variety of emotional states. The emotional states associated with stress included sadness, anger, fear, and disgust, totaling 384 instances. Conversely, emotions characterized by no-stress included neutral, calmness, happiness, and surprise, amounting to 336 instances. In total, there were 720 snippets, each lasting 3 seconds.

3.1.3. Summary of the Combined Datasets

Based on the elaboration in Sections 3.1.1 and 3.1.2, Table 1 presents a summary of the combined dataset used in this study. The total number of data points used in this study was 3120 snippets consisting of stress and no-stress speeches spoken by female speakers.

Table 1. Summary of the combined datasets

Classes	TESS dataset	RAVDESS dataset	Total
Stress	1200	384	1584
No-stress	1200	336	1536
		Total	**3120**

3.2. Data Preprocessing

The data was imported into the IDE with the Torchaudio library, with a sample rate 48k and a total of 140k samples. Subsequently, the signals underwent a normalization process using the mean and standard deviation. Finally, the feature vectors were adjusted to have the same length to ensure that the deep learning process could be applied uniformly. This was achieved by padding the shorter signals with zeros. From our observations, RAVDESS had a longer duration of 2 seconds than the TESS dataset. Thus, all the speech signals from the TESS dataset were zero-padded to ensure the same length. The data was not subjected to filtration, as it was collected in a calm environment without external interference.

3.3. Feature Extraction

It was crucial to comprehend that speech involved the filtration of sounds generated by humans through the configuration of the vocal tract, encompassing the tongue, teeth, and other anatomical components. According to previous research [12], when the shape was accurately defined, it accurately represented the phoneme. The selection of the following speech features for this study was based on the promising findings demonstrated in earlier studies [3], [8], [13].

As previously indicated, this research focused on comparing the spectral and TEO speech properties. The first set of speech features employed MFCCs, consisting of 13 coefficients. The second set of speech features combined TEO and MFCCs, consisting of 13 coefficients. Sections 3.3.1 and 3.3.2 provide further details regarding the speech elements employed.

3.3.1. MFCCs

The compact representation known as MFCCs was obtained by expressing a waveform as the summation of a nearly limitless number of sinusoids. Fluctuations in the MFCC coefficients were detected across different spectrum bands. The predominant portion of the spectrum energy was primarily localized within the low-frequency regions, characterized by positive cepstral coefficients. Conversely, a significant portion of the spectrum energy was concentrated at higher frequencies, corresponding to negative values of the cepstral coefficient. A concise explanation of each sequential process involved in the computation of MFCCs was provided, as stated by previous research [8].

i. **Pre-emphasis:** The utilization of pre-emphasis facilitated the amplification of the higher frequencies within the signal.
ii. **Framing:** Using the Fast Fourier Transform (FFT) resulted in errors due to the non-stationary nature of audio processing. To address this, it was assumed that the audio could be considered a transient stationary process. Consequently, the signal was partitioned into tiny frames. To establish connections between frames, each frame overlapped with one another.
iii. **Windowing:** The conversion from the time domain to the frequency domain was achieved by utilizing the FFT technology. To preserve the quality of the audio resumes,

a window function was employed for each frame. Consequently, high-frequency distortions, known as spectral leakage, were mitigated.

iv. **Discrete Fourier Transform for acquiring the Power Spectrum**: The power spectrum, which is a representation of the frequency spectrum, was obtained by calculating the Short-Time Fourier Transform (STFT) using N iterations (N= 512) of the FFT on each frame.

v. **Filter banks**: The ultimate step in computing filter banks involved extracting frequency bands from the power spectrum by utilizing triangle filters on a Mel scale.

vi. **Discrete Cosine Transform**: The interdependence of the filter bank coefficients computed in the preceding step posed challenges in certain deep learning approaches. To achieve decorrelation to the filter bank coefficients and provide a compressed representation of the filter banks, the Discrete Cosine Transform (DCT) was employed.

3.3.2. TEO-MFCCs

The subsequent technique employed in this study involved the integration of TEO with MFCCs. The signal derived from the pre-processing stage was incorporated into the TEO feature. The process of speech production included both linear and nonlinear components. Previous research has revealed that emotional speech, particularly under stress, exhibited significant alterations in its nonlinear components compared to regular speech [14]. An energy measurement was designed to accurately represent the instantaneous energy of nonlinear components as a direct outcome of the speech inquiry founded by Teager, with the initial form of the energy operator being defined by Kaiser [15]. To obtain the TEO-MFCC features, the signals were first retrieved from the pre-processing stage, then TEO features were applied to the signal, and finally, MFCC features were applied.

3.4. Data Organization

The Pandas library was used to differentiate the processed signals generated into stress and no-stress classes, as shown in Figure 2.

	feature	class
0	[[[tensor(-121.2283), tensor(-121.2283), tenso...	0
1	[[[tensor(-137.1273), tensor(-137.1273), tenso...	1
2	[[[tensor(-125.5066), tensor(-125.5066), tenso...	0
3	[[[tensor(-117.8783), tensor(-117.8783), tenso...	1
4	[[[tensor(-115.4972), tensor(-115.4972), tenso...	0
...	...	...
3115	[[[tensor(-18.1040), tensor(0.7011), tensor(9....	1
3116	[[[tensor(-25.5897), tensor(-15.8098), tensor(...	1
3117	[[[tensor(-30.6487), tensor(-24.4433), tensor(...	1
3118	[[[tensor(-52.5320), tensor(-49.4062), tensor(...	1
3119	[[[tensor(-47.3476), tensor(-35.6632), tensor(...	1

3120 rows × 2 columns

	feature	class
0	[[[tensor(-139.3613), tensor(-139.3613), tenso...	0
1	[[[tensor(-157.3328), tensor(-157.3328), tenso...	1
2	[[[tensor(-145.3053), tensor(-145.3053), tenso...	0
3	[[[tensor(-132.0546), tensor(-132.0546), tenso...	1
4	[[[tensor(-143.7691), tensor(-143.7691), tenso...	0
...	...	...
3115	[[[tensor(-169.3116), tensor(-169.3116), tenso...	1
3116	[[[tensor(-188.7097), tensor(-188.7097), tenso...	1
3117	[[[tensor(-178.7599), tensor(-178.7599), tenso...	1
3118	[[[tensor(-184.0467), tensor(-184.0467), tenso...	1
3119	[[[tensor(-178.6420), tensor(-178.6420), tenso...	1

3120 rows × 2 columns

Figure 2. Features of MFCCs (left side) and TEO-MFCCs (right side) that have been put into classes. 0 means "no stress," and 1 means "stress".

https://doi.org/10.31436/iiumej.v26i1.3411

3.5. Model Prediction – Stress Speech Neural Network Architecture (SSNNA)

The PyTorch library performed a binary classification problem: stress and no-stress. This study employed the SSNNA, a CNN architecture developed by this study. Table 2 provides the best possible hyperparameter settings for the SSNNA architecture.

Table 2. Hyperparameter settings for SSNNA

Settings	Total
Batch size	12
Epoch no.	450
Loss Function	Cross-Entropy Loss
Optimizer	Stochastic Gradient Descent
Learning Rate	0.00001
Momentum	0.7

The SSNNA comprised successive convolutional layers implemented using the PyTorch toolkit, as shown in Figure 3. It consisted of six 2-dimensional convolutional layers (Conv2d) connected with Rectified Linear Unit (ReLU) layers, each followed by a 2-dimensional Max-pooling (MaxPoold2d) layer. After these Conv2d layers, flattened layers were appended, leading to three additional linear layers. The first linear layer begins with a ReLU activation and connects to the output layer, which is categorized into two classes: stress and no stress.

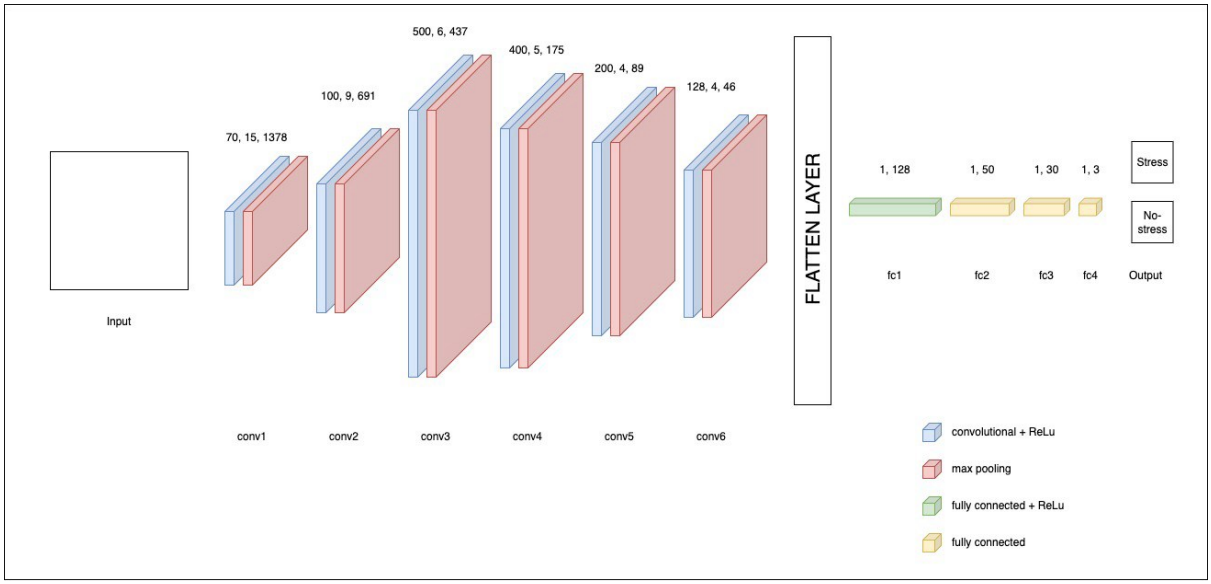

Figure 3. SSNNA architecture.

To ensure the reliability of the results, the data was partitioned into three sets: training, validation, and testing. The training set comprised 80% shuffled data, the validation set comprised 10% non-shuffled data, and the test set comprised 10% non-shuffled data. The model's training was halted once the testing accuracy stabilized, ensuring the prevention of overfitting across all three datasets.

4. RESULT AND DISCUSSION

This section presents the results and outcomes of this study. The dataset used for testing purposes encompassed 312 data points. The subsequent discussion relied on the findings

https://doi.org/10.31436/iiumej.v26i1.3411

derived from the testing set. The study employed a set of performance metrics: F1 score, accuracy, the area under the receiver operating characteristic curve (ROC-AUC curve), and confusion matrix. The confusion matrix comprised four components: True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). The confusion matrix results were based on the labeling shown in Table 3. Sections 4.1 and 4.2 provide the results for MFCCs and TEO-MFCCs.

Table 3. Relationship between the confusion matrix and classes

Classes	No-stress	Stress
No-stress	TN	FP
Stress	FN	TP

4.1. Performance Metrics Based on the MFCC's Speech Feature

This subsection presents the performance metrics for MFCCs. A confusion matrix is constructed based on the categorization of stress and no-stress to visually represent and summarize the overall performance of the model's predictive outputs, as shown in. Figure 4. The presented confusion matrix shows that the model achieves good accuracy for both stress and no-stress female speeches, with fewer than 10 misclassified data points in each class. The achieved outcomes for the F1 score and accuracy are deemed satisfactory, with values of 0.94 and 93.9%, respectively. Additionally, based on the classes, the stress and no-stress classes achieved 94.2% and 93.7% scores, respectively. The accuracy surpassed the average outcomes for the scripted dataset and approached 100%.

The accuracy scores for the training, validation, and testing sets were 94.9%, 91.7%, and 94.2%, respectively. The score difference between these sets was rather small, indicating that the algorithm was not affected by underfitting or overfitting problems. Furthermore, the ROC curve achieved an AUC value of 0.94, indicating a good outcome [18]. The model demonstrated better accuracy, indicates that the use of MFCCs as a speech feature is an effective approach for identifying stress emotions in female speakers.

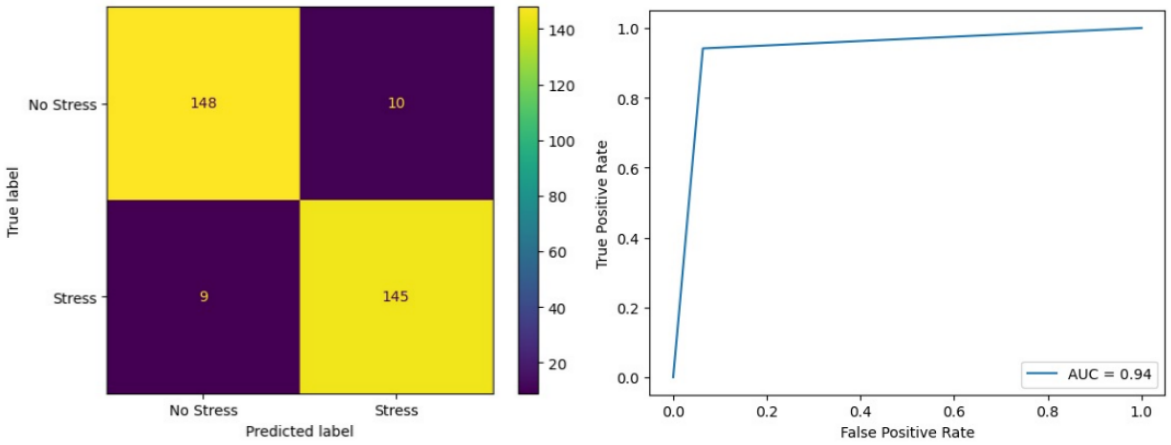

Figure 4. Confusion matrix (left side) and AUC curve (right side) of MFCCs.

4.2. Performance Metrics Based on the TEO-MFCC's Speech Feature

Based on the confusion matrix depicted in Figure 5, the model has achieved good accuracy for both stressful and no-stress classes, with less than 19 instances of misclassification in each class. The obtained results for the F1 score and accuracy were satisfactory, with respective

https://doi.org/10.31436/iiumej.v26i1.3411

values of 0.90 and 90.7%. Furthermore, the classes categorized as stress and no stress demonstrated accuracy rates of 93.6% and 87.7%, respectively. The no-stress class achieved good accuracy, while the stress class achieved almost a fair level of accuracy.

Furthermore, the accuracy scores for the training, validation, and testing sets were 93.7%, 89.4%, and 90.7%, respectively. The observed disparity in scores between the sets was also rather small, suggesting that the algorithm remained unaffected by issues related to underfitting or overfitting. In addition, the ROC curve depicted in Figure 5 acquired an AUC value of 0.91. The model has shown satisfactory performance, suggesting that utilizing TEO-MFCC features is another effective technique for recognizing stressed emotions in female speakers. Nevertheless, it could not surpass the superior performance of the MFCCs' speech feature alone. Based on our observation, this was mainly due to the disruption of the signal when it was padded with zeros, which likely led to inaccurate TEO value results. Those values might not have produced useful information distinguishing between stress and no-stress speeches.

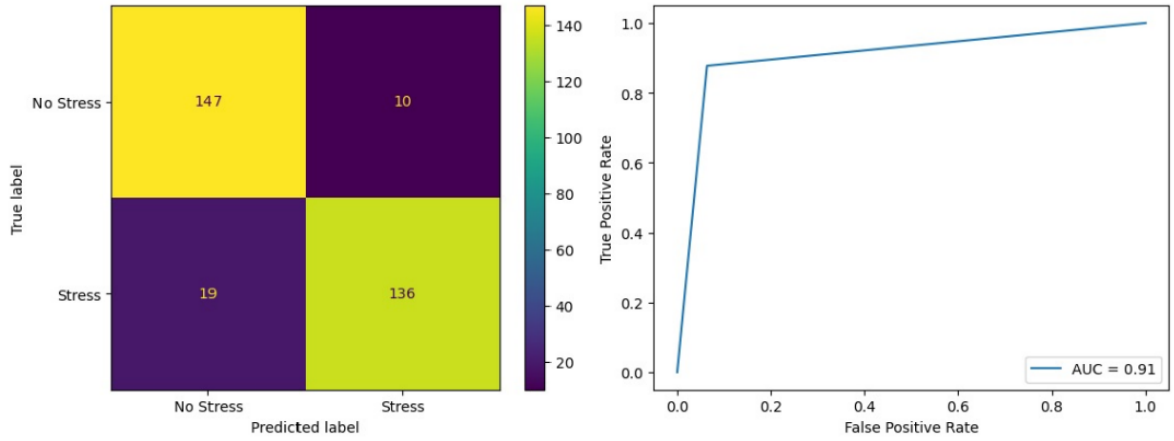

Figure 5. Confusion matrix (left side) and AUC curve (right side) of TEO-MFCCs.

Moreover, based on the analysis of the performance metrics presented in Table 4, it was observed that the speech characteristics derived from MFCCs exhibited favorable performance in stress and no-stress classifications in the female's voice. This was attributed to the characteristics of MFCCs, which were designed to enhance the amplitudes of higher-frequency components in speech signals. The frequencies produced by female vocalizations tend to be naturally higher in intensity compared to males. Besides, this assertion aligned with the results of a prior study [3], wherein the researchers observed that MFCCs yielded highly satisfactory outcomes in female subjects. On another note, the study discovered that the TEO-MFCCs also produced remarkable outcomes in discerning stress in female speakers based on their speech patterns.

Table 4. Relationship between the confusion matrix and classes

Features	MFCCs	TEO-MFCCs
Training Accuracy (%)	93.9	90.7
Accuracy based on classes (%)	Stress: 94.2; No-stress: 93.7	Stress: 87.7; No-stress: 93.6
F1-score (0 – 1)	0.94	0.90
AUC (0 – 1)	0.94	0.91

Additional components that contributed to the overall better outcomes were increased training data and a considerable balance of data between the classes (achieved by omitting

disgust emotion from the TESS dataset). This enhanced the model's learning capabilities and produced unbiased stress classification outcomes. In addition, based on our research, no study had been conducted that integrates several datasets while utilizing only females' speeches. Therefore, in this paper, we opted to exclude a comparative analysis with other studies due to their lack of direct comparability.

5. CONCLUSION AND FUTURE WORKS

The impact of stress on females was a prevalent concern that warranted attention since females were more stressed than males. This was attributed to differences in stress perception, hormonal transitions, and other factors. Efforts were made to mitigate the long-term persistence of this issue, as it could potentially lead to adverse consequences for affected individuals. Speech stress detection and classification provided several advantages, especially for females, as they were non-invasive and non-intrusive approaches. This method also comforted females as they were naturally verbally expressing their thoughts and feelings, adding another positive aspect to this approach. MFCCs demonstrated their effectiveness as speech features by providing valuable information in both stress and no-stress female speech datasets. MFCCs improved the energies of the higher-frequency components of the female speech signals, which could distinguish between these two classes. MFCCs achieved 93.9% accuracy, accompanied by an F1-score of 0.94 and an ROC-AUC of 0.94. Meanwhile, TEO-MFCCs also produced satisfactory results, but they could not surpass the results of MFCCs due to the disruptions of the speech signals during the feature extraction stage. Hence, care needs to be taken if the researchers intend to use TEO-MFCCs as the speech feature where manipulation of speech signals is involved.

The combination of SSNNA and MFCCs could distinguish between the two classes using 14 different female speeches expressing eight emotions. This demonstrated that the classifier was speaker-independent, indicating that the system was more scalable and user-friendly. Additionally, this made it easier for real-world applications since it did not require speaker-specific data collection for adaptation. It improved data efficiency by enabling diverse training datasets, which enhanced the algorithm's generalization and robustness and reduced bias. In conclusion, the combination of MFCCs and SSNNA helped accurately classify stress and no-stress classes, especially in female speeches. For future works, increasing the number of speakers in the dataset was suggested to enhance generalizability, reduce bias, and improve model performance. Additionally, expanding the classification from two to three categories—low, medium, and high stress—would allow a more effective assessment of an individual's stress levels.

ACKNOWLEDGEMENT

The study was financially supported by the Fundamental Research Grant Scheme (FRGS) administered by the Ministry of Higher Education (MoHE), with the reference code FRGS/1/2021/TK0/UIAM/02/29.

REFERENCES

[1] K. Monesson, "Why Are Women More Stressed Out Than Men? | UKG." Accessed: Sep. 11, 2023. [Online]. Available: https://www.ukg.com/blog/life-work-trends/why-are-women-more-stressed-out-men

[2] A. Ghazali, "Wanita tidak tahan tekanan - Sinar Harian," Sinar Harian Newspaper. Accessed: Sep. 04, 2023. [Online]. Available: https://www.sinarharian.com.my/article/10527/sinar-aktif/murung

[3] M. S. Nordin *et al.*, "Stress Detection based on TEO and MFCC speech features using Convolutional Neural Networks (CNN)," in *2022 IEEE International Conference on Computing (ICOCO)*, IEEE, Nov. 2022, pp. 84–89. doi: 10.1109/ICOCO56118.2022.10031771.

[4] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020, *Borealis*. doi: doi/10.5683/SP2/E8H2MF.

[5] 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," 2018, doi: 10.5281/zenodo.1188976.

[6] "Great Speech." Accessed: Jul. 11, 2024. [Online]. Available: https://www.greatspeech.com/can-emotional-stress-cause-speech-problems/

[7] B. Sağlam Topal and A. E. Yavuz Sever, "I love you but I can't say: adaptation of the Measure of Verbally Expressed Emotion (MoVEE) to Turkish and investigation of psychometric properties," *Current Psychology*, vol. 43, no. 24, pp. 20881–20890, Jun. 2024, doi: 10.1007/s12144-024-05861-5.

[8] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2017, pp. 1–5. doi: 10.1109/ICCCNT.2017.8204149.

[9] 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.

[10] 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.

[11] A. De Arriba, M. Oriol, and X. Franch, "Merging Datasets for Emotion Analysis," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, IEEE, Nov. 2021, pp. 227–231. doi: 10.1109/ASEW52652.2021.00051.

[12] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. USA: Prentice Hall PTR, 2001.

[13] 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.

[14] H. Gao, S. Chen, P. An, and G. Su, "Emotion recognition of mandarin speech for different speech corpora based on nonlinear features," in *2012 IEEE 11th International Conference on Signal Processing*, IEEE, Oct. 2012, pp. 567–570. doi: 10.1109/ICoSP.2012.6491552.

[15] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.