JOIV : Int. J. Inform. Visualization, 7(4) - December 2023 2474-2480

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Hybrid Deep Learning Approach for Stress Detection Model Through Speech Signal

Phie Chyan [a,c], Andani Achmad [a,*], Ingrid Nurtanio [b], Intan Sari Areni [a]

[a] *Department of Electrical Engineering, Hasanuddin University, Bontomarannu, Gowa, 92171, Indonesia*
[b] *Department of Informatics, Hasanuddin University, Bontomarannu, Gowa, 92171, Indonesia*
[c] *Department of Informatics, Atma Jaya Makassar University, Tamalate, Makassar, 90224, Indonesia*
Corresponding author: [*]andani@unhas.ac.id

***Abstract*— Stress is a psychological condition that requires proper treatment due to its potential long-term effects on health and cognitive faculties. This is particularly pertinent when considering pre- and early-school-age children, where stress can yield a range of adverse effects. Furthermore, detection in children requires a particular approach different from adults because of their physical and cognitive limitations. Traditional approaches, such as psychological assessments or the measurement of biosignal parameters prove ineffective in this context. Speech is also one of the approaches used to detect stress without causing discomfort to the subject and does not require prerequisites for a certain level of cognitive ability. Therefore, this study introduced a hybrid deep learning approach using supervised and unsupervised learning in a stress detection model. The model predicted the stress state of the subject and provided positional data point analysis in the form of a cluster map to obtain information on the degree using CNN and GSOM algorithms. The results showed an average accuracy and F1 score of 94.7% and 95%, using the children's voice dataset. To compare with the state-of-the-art, model were tested with the open-source DAIC Woz dataset and obtained average accuracy and F1 scores of 89% and 88%. The cluster map generated by GSOM further underscored the discerning capability in identifying stress and quantifying the degree experienced by the subjects, based on their speech patterns.**

***Keywords*— Stress detection; speech processing; deep learning; CNN; GSOM.**

Manuscript received 2 Aug. 2023; revised 10 Sep. 2023; accepted 30 Oct. 2023. Date of publication 31 Dec. 2023.

I. INTRODUCTION

Stress is a psychological problem that arises in response to the challenges experienced in daily life, affecting individuals of all ages [1], [2]. Stressors or stress triggers can stem from internal factors, such as feelings of inferiority and helplessness, or external factors, including bullying and academic pressure at school [3]. Different cross-disciplinary studies provide evidence of the complex relationship between mental development, social environment, and long-term health conditions [4]. The early years of a child's life often called the Golden Age period, represent a critical phase of development where their biological system rapidly assimilates diverse positive and negative experiences [5]. Intense and prolonged exposure to stress contributes closely to long-term health problems, including heart problems, diabetes, and premature death [6]. Different expressions of negative feelings, such as anger, sadness, nervousness, and fear characterize stress.

Furthermore, this stressful condition also negatively influences the human nervous system. Several studies have shown that high-intensity chronic stress can lead to decreased brain mass, cognitive degradation, and memory problems. A growing child's physical and mental development may be adversely affected due to the occurrence of this circumstance [6], [7].

According to WHO data, the prevalence of mental illnesses in children is estimated to be 13% of the population aged 10 to 19 worldwide, which is roughly equal to disorders in adults at 20 % [8]. Therefore, mental health issues, such as stress, can persist into adulthood when the underlying issues are unaddressed [9]. Stress, considered non-threatening in its mild stages, is a prevalent mental condition encountered in everyday life. However, when not managed effectively, the condition can lead to severe mental health issues [7]. Detecting stress in children, particularly those in the pre-early school age group, presents a considerable challenge. This is primarily due to their limited communication skills and a lack

of awareness regarding the various potential mental health issues encountered [10], [11]. The impact of stress on pre-early school-age children can manifest in various ways, influenced by factors such as their personality, environmental context, and specific stressors [12]. Due to the restricted communication and cognitive abilities, children typically respond to stressful situations by exhibiting alterations in behavior and emotions. Common behavioral responses in children experiencing stress include heightened irritability, aggression, or withdrawal. Concurrently, typical emotional responses are nervousness, sadness, anger, and mood fluctuations.

Detection of Stress in children is a multifaceted challenge, necessitating the exploration of technology-driven solutions capable of autonomously identifying stressors in this vulnerable demographic. Based on current technological developments, various approaches can be used to detect stress, including direct measurements of the human body's biosignal parameters using various sensors. Several related studies include detecting stress through heart rate and skin resistance (Galvanic skin response), monitoring stress through variable heart rate (HRV), and skin conductivity (Electrodermal Activity) using smart watch devices [13], [14]. Other studies have produced stress detection approaches using heart rhythm data through an electrocardiogram (ECG) combined with blood pressure measurements [15] and studies using stress detection through a combination of heart rhythm data and respiration rate (Respiration Rate) [16]. Despite the efficacy of these studies in using biosignal measurements to achieve a high level of accuracy in stress detection, their applicability is limited due to the potential discomfort and the inadvertent introduction of additional stressors associated with affixing sensors to a child's body [17].

An alternative child-friendly approach to stress detection includes the analysis of an individual's speech. Based on medical literature reviews, there is a correlation between stress experienced and human vocal reproduction where the conditions affect various body functions and tension of various muscles for supporting vocal reproduction. Therefore, the output of human vocal sounds can be used as a good marker in detecting stressful conditions [18]–[20]. Studies related to stress detection through speech have been carried out in recent years [17], [21]–[28]. These studies have detected stress through speech with fairly good accuracy, between 75-85%. Furthermore, these studies primarily concentrate on binary classifications of stress status, distinguishing between individuals with and without stress [18]. The model that can provide high accuracy in identifying stressful states experienced by the subject with the level of severity is needed to achieve effective management. This is because the treatment of a person's stressful condition depends on the level of stress experienced. At mild levels of stress, individuals do not necessitate medical intervention but are often sufficient to address and manage the underlying stressors that provoke their distress. Conversely, in cases of severe and protracted stress, medical treatment becomes important. The treatment includes the expertise of specialized medical professionals, including child psychologists, providing the requisite therapeutic assistance [29].

This study introduces a model using a deep learning-based hybrid approach, which integrates supervised and unsupervised approaches. The primary objective is to enhance stress detection accuracy by analyzing speech signals and applying clustering approaches to identify associative relationships between voice characteristics and stress levels. The proposed model constitutes the central contribution of this study. Additionally, a dataset is established for stress detection in children, supporting the contributions in this domain.

According to the proposed detection model, the identification of stress in children can be conducted earlier, enabling the implementation of necessary preventive or therapeutic measures, contingent on the severity level. This proposed model stands apart from previous studies in two key aspects. Firstly, it is distinguished through explicitly constructing a supporting dataset for the stress detection model in children. Secondly, the model uses a novel hybrid approach, combining supervised and unsupervised learning elements to facilitate stress detection through speech analysis. This study is structured into four distinct sections, contributing to a comprehensive understanding of the result, and the introductory section provides the overview. The second section delves into the intricacies of the proposed approach, stating the details for a more comprehensive comprehension. Subsequently, the third section presents the results and engages in a thorough discussion. The fourth and concluding section encapsulates the entirety of the study content.

II. Material and Method

This section describes the dataset, the proposed model, and the performance evaluation approach of the stress detection system.

A. Dataset Preparation

In building the dataset, direct voice samples were obtained from 10 pre-early-school children aged between 5-7 years directly from the school environment, and the parents expressed their consent. The study was accompanied by a child psychologist who designed and supervised the activities based on the Trier Social Stress Test approach. Activities based on the Trier Social Stress Test include working on complex arithmetic questions and public speaking to induce stress [30]. After the activity session, the children were called into an interview session guided by a psychologist. During the session, their voices were recorded, and the psychologist monitored the children directly to observe the symptoms of stress through behavior or gestures. Based on the observations, stress status labeling was carried out on the voice sample recordings in binary, namely Stress or Non-Stressed.

In the case of children experiencing stress, the psychologist documented their levels based on the Kessler standard instrument. This instrument classifies the condition into three categories. The first is mild stress, where the subject exhibits subtle gestures or mild stressful behaviors. The second is moderate stress, characterized by symptoms and noticeable stressful behaviors. The third is severe stress, which indicates intense and pronounced symptoms and behaviors [29], [31]. The voice recordings from each subject were cut into sound samples with a duration of 1 and 2 seconds. After the

validation and labeling process, 106 and 142 voice samples were labeled as Stressed and Non-Stressed.

An open-source dataset known as The Distress Assessment Interview Corpus, abbreviated as DAIC-WOZ, was also used to evaluate the model developed against various state-of-the-art counterparts [32]. The dataset consisted of a voice sample and questionnaire answers from participating subjects labeled with a degree of stress level according to the standard Patient Health Questionnaire (PHQ-8). This questionnaire consisted of eight question items that measured various aspects of depression with a scale of 0 to 3, with response options of "not at all," "several days," "more than half days," and "nearly every day." From the eight questions, a score of 0-24 was obtained, which indicated the degree of stress. This dataset consisted of 59 and 130 samples for Stressed and Non-stressed subjects.

B. The Architecture of the Proposed Model

The proposed model consists of a combination of supervised and unsupervised learning approaches. The supervised approach uses the Convolution neural network (CNN) architecture. In the proposed model, CNN is based on its effectiveness in various tasks related to audio classification, such as speech recognition. This is because of its ability to understand various attribute representations in audio spectrograms, handle various input sizes, and use pre-trained model and transfer learning to provide good classification performance. CNN architecture consists of a Convolution layer, a max-pooling layer, and a fully connected layer [28].

Furthermore, the convolution layer provides the main framework for the entire neural network and a set of kernels to be learned in the training process. The max pooling layer functions to reduce the feature map's dimensions, reducing the data's size and the number of parameters to be studied. The linear layer connects each neuron from the previous to the next layer. Meanwhile, the ReLu activation functions to overcome the vanishing gradient problem, the flattened layer between the convolution and the fully connected layer, the dense layer using SoftMax as activation function, and the drop-out layer, reducing the overfitting problem for the unsupervised approach using GSOM (Growing Self-Organizing Map). GSOM is a type of unsupervised neural network used for dimensionality reduction and visualization of high-dimensional data to support models conducting audio data clustering. The model can categorize voice samples into clusters characterized by shared attributes using the support provided by GSOM, a variant of Self-Organizing Maps (SOMs) to elucidate the topological properties in data.

In addition, providing a visual representation of cluster data is needed for stress analysis. The approach consists of four nodes as the initial configuration and learning using rules based on Euclidean distance. New nodes are formed when quantitation errors accumulate ahead of the growth threshold value. Combining two approaches based on supervised and unsupervised using CNN and GSOM architectures is a hybrid used in developing a stress detection model. Figure 1 summarizes the stress detection model using this hybrid approach.

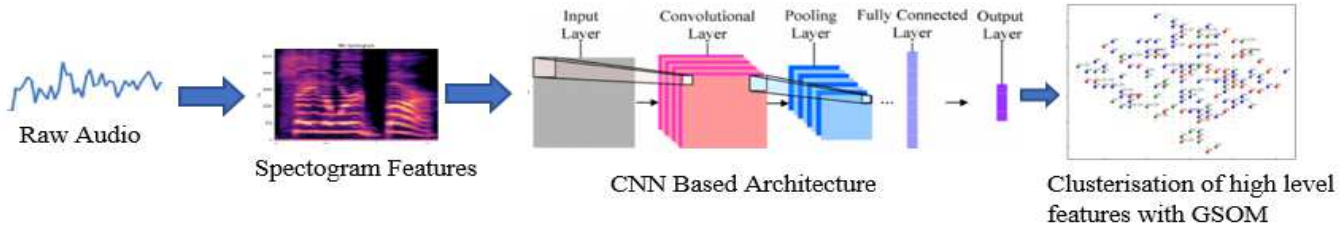

Fig. 1 Architecture of proposed model

The model receives input in the form of voice recordings from the dataset, and the sound samples are converted into a spectrogram, a visual representation of changes in the frequency of signals over time. Before conducting feature extraction, segmentation and data cleaning processes are carried out. This process is one of the critical stages since the sounds acquired are from subjects in a classroom environment susceptible to various noises. Segmentation is conducted to separate the sound signal from noise, including the detected silences on the recording. This study uses the Librosa library to perform segmentation, sound analysis, and feature extraction.

The data balancing procedure is undertaken to rectify the imbalance within the dataset. In the case of DAIC-WOZ, the number of voice samples categorized as non-stressed exceeds stressed samples by more than a twofold margin. Similarly, the child voice dataset exhibits an imbalance, where non-stressed voice is nearly 50% greater than stressed voice. To address this issue, the data imbalance is mitigated by increasing the number of stressed labeled samples through the use of data augmentation approaches. Subsequently, feature extraction is performed on the audio spectrogram to convert the audio signal into a format comprehensible by model. After training, model can classify speech into binary labels, namely `stressed and non-stressed. The output from CNN in the form of high-dimensional features is fed to GSOM to conduct stress clustering by generating and visualizing feature maps to determine the characteristics of model in identifying sounds with stress categories.

C. Data Augmentation and Feature Extraction

The issue of imbalanced data warrants attention due to its potential to introduce bias into the training process, impacting the accuracy of the proposed model's output. In the context of this study, the children's voice dataset and the Daic-Woz dataset encountered challenges associated with data imbalance. The children's voice dataset contains 106 and 142 sound files labeled stressed and non-stressed. Similarly, in the open-source dataset DAIC-WOZ, there are 59 and 130 sound files labeled stressed and non-stressed, respectively. A way to overcome this imbalance problem is to use data augmentation approaches [33]. In addition to overcoming the problem, this

augmentation approach is also useful for increasing the diversity of datasets. In this study, data augmentation is carried out by creating a new synthetic audio file, which is a variation of the original audio file. To do this, we add two new samples of synthetic data for each original audio file. Using the Librosa audio library, we injected artificial noise as the first audio variation and performed pitch shifting as the second audio variation. So that the augmentation process is as bias-free as possible, we do several things to ensure this, including maintaining data balance.

In this case, we ensure that the proportion of each audio file representing both categories, namely stress and non-stress, is balanced so that learning is not biased towards one category. In addition, we ensure that each original audio has two synthetic variations produced with the same method so that the proportion of data after the augmentation process remains consistently maintained. Figure 2 shows the audio signal changes through the graphical waveform of the augmentation process. After the augmentation process on the children's voice dataset, 300 sound samples were obtained, consisting of 145 and 142 labeled stressed and non-stressed. For the open-source DAIC-WOZ dataset, after the augmentation process, a total of 300 samples were obtained, consisting of 142 and 158 sound samples labeled stressed and non-stressed.

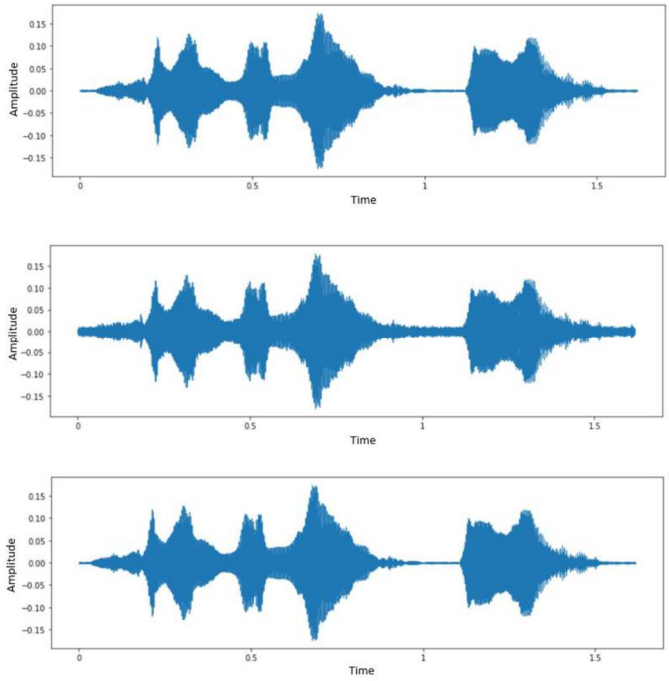

Fig. 2 Waveform of augmented speech sample, original sample (top), with artificial noise (mid), and pitch shifting (bottom)

Audio feature extraction converts the audio signal into a vector form the model understands. Various audio applications, such as audio classification, speech recognition, speech separation, and audio fingerprinting, require proper audio feature extraction to produce good performance. Based on the level of abstraction, audio features can be divided into Low-level, Mid-level, and High-level. High-level audio features, such as keys and rhythms, are abstract features humans can immediately comprehend or understand. The sense of hearing can perceive mid-level audio features, such as pitch, beat descriptor, and MFCC. Low-level audio features are statistical features extracted directly from audio, and these features can only be understood by machines. The features include amplitude envelope, energy, spectral centroid, and zero crossing rate. In addition to the level of abstraction, audio features also depend on the signal domain, which states how the perspective of the signal is represented.

The signal domain is divided into time, frequency, and cepstral [34]. Audio features in the time domain are extracted directly from the waveform represented in time. The amplitude of the sound signal is measured as a function of time, and examples of audio features in the time domain are Zero crossing rate, amplitude envelope, and Root mean square energy. Audio features in the frequency domain are signal characteristics that describe the analysis of the mathematical function of signal to frequency. Signals are converted from the time domain using the Fourier transform, and some examples of audio features in the frequency domain are spectral centroids, band energy ratio, and spectral flux. The audio features in the cepstral domain are obtained by performing an inverse Fourier transform of the logarithm spectrum Fourier. Mel Frequency cepstral coefficients (MFCC) and Mel-spectrogram belong to this cepstral domain.

Furthermore, the Mel-spectrogram represents the cepstral domain for feature extraction in the stress detection model. The Mel-spectrogram has the advantage of imitating human auditory perception suitable for sound analysis. Figure 3 shows an example visualization of sound samples in the dataset represented in the mel-spectrogram.

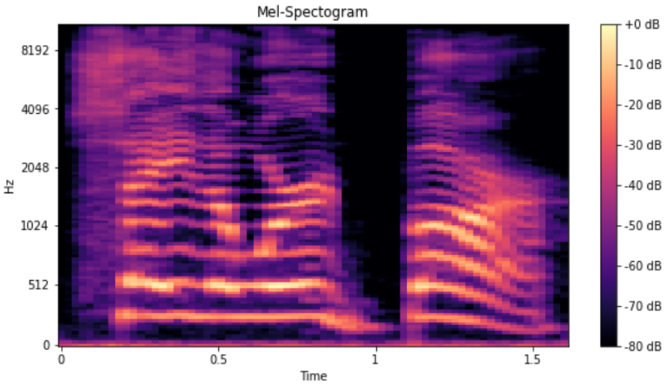

Fig. 3 Visualization of one of the voice samples in mel-spectrogram

D. Stress Detection Hybrid Model

Based on the architecture of the proposed model, the stress detection model uses a hybrid approach through supervised and unsupervised learning. CNN architecture in the proposed model consists of 3 convolution layers mediated by a max pooling layer to maintain the dominant features of the feature map, with a flattened layer followed by two dense layers. In the Mel-spectrogram, a total of 131 features are extracted, and these features are used as inputs into the initial layer, with dimensions (131,1). The output of this layer undergoes a transformation resulting in a (131, 256) shape, achieved through the use of 256 channels with 5x5 filters. Therefore, a dimension-reduction process is implemented using a max-pooling layer with a 5x5 filter and a stride of 2.

The second and third convolutional layers consist of 256 and 128 channels using 5x5 filters. These layers are followed by max-pooling layers, which maintain the same configuration as the initial max-pooling layer. Furthermore, a flattened layer converts the feature map to a linear form.

There is a dense layer with 32 neuron units, and the dropout layer is used with a value of 0.3. The SoftMax layer is configured with two units of neurons in the training phase for 50 epochs and a batch size of 50 using the ADAM optimizer. The training and test distribution is 75:25, with training data of 225 voice samples with 110 and 115 labeled stress and non-stressed. There are 75 voice samples for the test data, with 35 and 40 voice samples labeled Stress and Non-Stressed. In the training process, the ReduceLRonPlateau configuration is used to reduce the learning rate when the metric values begin to slope.

The output from CNN becomes input data to GSOM and is categorized into stress and non-stressed clusters. GSOM is initialized with four parallel nodes forming a quadrilateral with 131 dimensions of input data. The model is dynamically organized until the 131-dimensional mapping to 2-dimensional space is completed. Parameters for GSOM are set to 50 learning iterations, threshold 75, and spread factor from 0.1 to 0.9.

To evaluate model's performance, several metrics are used, namely accuracy, which measures the accuracy of predictions by calculating the ratio between correct predictions compared to those made by model. Furthermore, recall, precision, and F1-Score are also used to understand the effects of model on stress detection. The metric recall offers insight into model's ability to predict stress by examining the ratio of samples correctly predicted as stress to the total number of samples labeled as stress. Precision provides the ratio of correct stress predictions to the entire sample predicted as stress, and F1-Score indicates model performance by combining precision and recall values in a single value, balancing the trade-off between the recall and precision.

III. Result and Discussion

The system platform used is a Windows 10 64-bit computer with Intel Core I7 2.8 GHz CPU specifications, 16GB of RAM, and NVIDIA GTX 1060 4GB GPU, and model training process takes 4 min 25 s. The accuracy of the training model and validation with 50 epochs is 0.97 and 0.94 with a loss of 0.05 and 0.16, respectively. The graph of training accuracy, training loss, validation accuracy, and validation loss is shown in Figure 4. Based on the graph, models have a good learning rate with increasing model accuracy as the number of epochs increases before sloping at 40 epochs. The parameter loss values from the training and validation phases are also very small at 0.05 and 0.16, indicating an effective learning model.

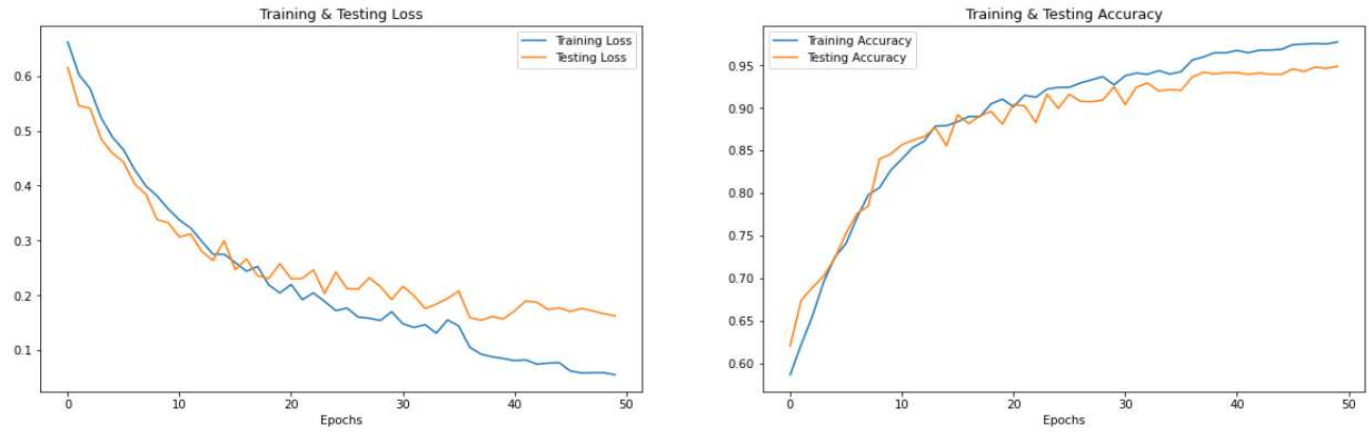

Fig. 4 Graph represents training and testing accuracy and loss vs number of epochs

Model performance results were measured using evaluation metrics for the constructed dataset and from the DAIC-WOZ dataset as presented in Figure 5.

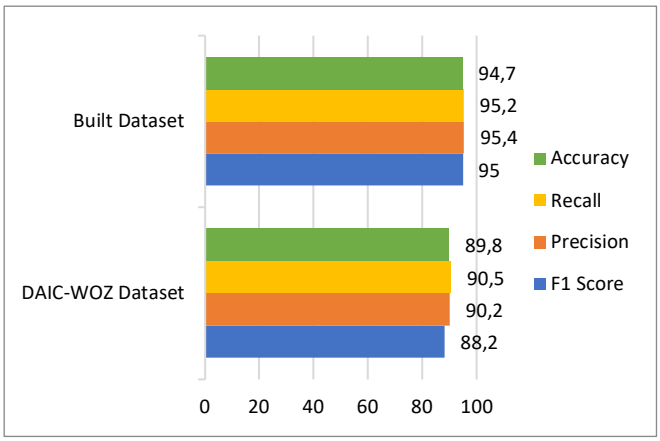

Fig. 5 Model performance with built dataset vs DAIC WOZ dataset

The benchmarking process was carried out to assess the proposed model's performance compared to the state-of-the-art model. Several deep learning model using the DAIC-WOZ dataset were examined, including model discussed in references [26] and [27]. These models use CNN with feature extraction from image spectrograms. Additionally, model [28] was evaluated, which uses a combination of CNN and LSTM while adopting MFCC as the extracted feature. Based on the performance benchmark results in Table 1, the proposed model succeeded in exceeding the performance of the deep learning-based model.

TABLE I
PERFORMANCE BENCHMARK OF THE STATE-OF-THE-ART MODEL USING THE DAIC-WOZ DATASET

Model	Accuracy	F1-Score (N)	F1-Score (S)
RNN [26]	76 %	85%	45%
CNN [27]	-	70%	52%
CNN+LSTM [28]	76%	82%	64%
Our Model	89%	99%	78%

For the DAIC-WOZ dataset of 46 data, 21, 11, and 5 samples have a PHQ-8 score between 0 to 4, 5 to 9, and 10-14 showing non-stressed psychological, mild stress, and moderate symptoms, while the remaining 9 have a PHQ score above 15, indicating severe stress. From 21 sample subjects who showed non-stressed psychological conditions based on a PHQ-8 score <= 4, model correctly classified 19 samples as non-stressed conditions. Meanwhile, 14 samples with a PHQ-8 score >= 10 indicated a real state of stress, and model correctly classified 11 of the samples. These results show model has high sensitivity and specificity in detecting stress.

In the augmentation process for training the classification model using CNN algorithm, synthesis data was added to increase the voice dataset labeled with stress to overcome the problem of imbalanced data. In cluster analysis using GSOM, feature vectors from the original dataset were used without augmentation to avoid bias. The groups of these feature vectors were divided into stressed and non-stressed clusters. Each node in the cluster was labeled according to the data label in the dataset. Nodes from data marked as stressed and non-stressed were given red and blue labels before plotting onto the cluster map. High-density areas with nodes labeled stressed and non-stressed were then identified as stressed and non-stressed clusters.

Fig. 6 shows the analysis of the distribution of nodes on the cluster map. The area marked with a rectangular marker indicates the collection of voice nodes of the subject under stress conditions, and the surrounding area shows the density of subject nodes under stress conditions. Meanwhile, the subject nodes under non-stress conditions are concentrated from clusters with stress nodes. From the comparisons made according to the child's stress level, dense areas with a distribution of stress nodes originated from child subjects with moderate and severe levels. Meanwhile, stress subject nodes scattered in the minority into dense areas with non-stressed are mostly subjects with a mild level.

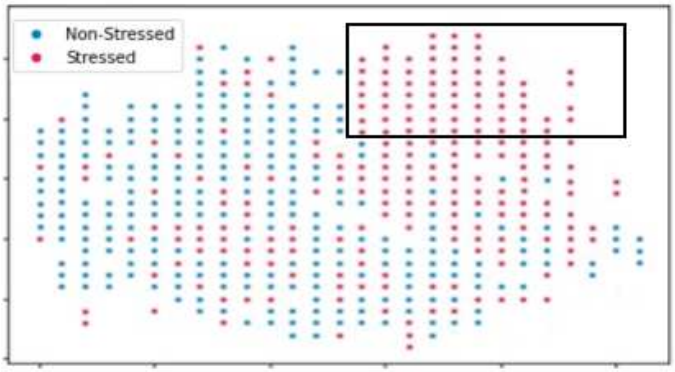

Fig. 6 Distribution node of built dataset

The cluster analysis of the DAIC-WOZ dataset also shows similar results (Fig.7). From the comparisons made according to the degree of stress, dense areas with a distribution of nodes came from subjects with a high PHQ-8 score level above 10, indicating a medium to a high degree. Furthermore, areas densely packed with non-stressed nodes were almost entirely from subjects with low PHQ-8 score levels below 10. The proposed model had a good differentiation ability in identifying stress and the degree of stress level of the subject through speech.

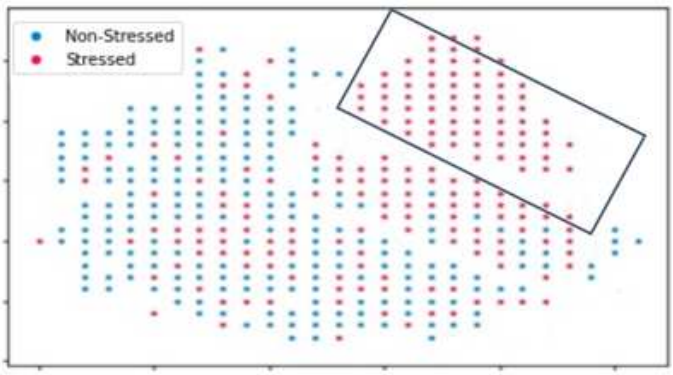

Fig. 7 Distribution node of DAIC-WOZ dataset cluster map

IV. Conclusion

In conclusion, the average accuracy and F1-Score results were 94.7% and 95%, respectively, using the child voice dataset built to support this study. For benchmark testing compared to the state-of-the-art model using the DAIC-WOZ dataset, the model also obtained results that exceeded these results with accuracy and an average F1-score of 89.8 and 88.2. Therefore, the model had generalization abilities over various voice samples. According to the cluster analysis, the model had good differentiation capabilities in identifying the subject's stress level with the unsupervised learning approach using GSOM. This was conducted by grouping nodes representing voice samples into appropriate clusters based on similarity.

Future studies could analyze the viability of detecting and monitoring stress in real-time from the child's activity environment. This will provide more information on the source of stress and stressors in children. However, a challenge to be overcome was optimizing model processing time in more complex audio pre-processing and audio separation approaches. Future model development can also integrate natural language processing methods into the model to detect stress through lexical speech analysis. With this integration, the model can detect stress more accurately by combining the signal and contextual properties of the subject's speech.

Acknowledgment

The authors are grateful to the Directorate of Research, Technology, and Community Service, Director General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for funding this study through the doctoral dissertation research grant (PDD) scheme.

References

[1] E. S. Epel et al., “More than a feeling: A unified view of stress measurement for population science,” Frontiers in Neuroendocrinology, vol. 49, pp. 146–169, Apr. 2018, doi:10.1016/j.yfrne.2018.03.001.

[2] M. Bucci, S. S. Marques, D. Oh, and N. B. Harris, “Toxic Stress in Children and Adolescents,” Advances in Pediatrics, vol. 63, no. 1, pp. 403–428, Aug. 2016, doi: 10.1016/j.yapd.2016.04.002.

[3] M. Kaczmarek and S. Trambacz-Oleszak, “School-Related Stressors and the Intensity of Perceived Stress Experienced by Adolescents in Poland,” International Journal of Environmental Research and Public Health, vol. 18, no. 22, p. 11791, Nov. 2021, doi:10.3390/ijerph182211791.

[4] N. Garmezy, A. S. Masten, and A. Tellegen, “The Study of Stress and Competence in Children: A Building Block for Developmental Psychopathology,” Child Development, vol. 55, no. 1, p. 97, Feb. 1984, doi: 10.2307/1129837.

[5] M. Rohmadi, M. Sudaryanto, C. Ulya, H. Akbariski, and U. Putri, “Case Study: Exploring Golden Age Students’ Ability and Identifying Learning Activities in Kindergarten,” Proceedings of the Proceedings of the First Brawijaya International Conference on Social and Political Sciences, BSPACE, 26-28 November, 2019, Malang, East Java, Indonesia, 2020, doi: 10.4108/eai.26-11-2019.2295218.

[6] H. Yaribeygi, Y. Panahi, H. Sahraei, T. P. Johnston, and A. Sahebkar, "The impact of stress on body function: A review.," *EXCLI J.*, vol. 16, pp. 1057–1072, 2017.

[7] P. Morgado and J. J. Cerqueira, Eds., The Impact of Stress on Cognition and Motivation. Frontiers Media SA, 2019. doi:10.3389/978-2-88945-774-8.

[8] M. Solmi et al., “Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies,” Molecular Psychiatry, vol. 27, no. 1, pp. 281–295, Jun. 2021, doi:10.1038/s41380-021-01161-7.

[9] M. Mohler-Kuo, S. Dzemaili, S. Foster, L. Werlen, and S. Walitza, “Stress and Mental Health among Children/Adolescents, Their Parents, and Young Adults during the First COVID-19 Lockdown in Switzerland,” International Journal of Environmental Research and Public Health, vol. 18, no. 9, p. 4668, Apr. 2021, doi:10.3390/ijerph18094668.

[10] Y. Choi, Y.-M. Jeon, L. Wang, and K. Kim, “A Biological Signal-Based Stress Monitoring Framework for Children Using Wearable Devices,” Sensors, vol. 17, no. 9, p. 1936, Aug. 2017, doi:10.3390/s17091936.

[11] T.-Y. Kim, L. Měsíček, and S.-H. Kim, “Modeling of Child Stress-State Identification Based on Biometric Information in Mobile Environment,” Mobile Information Systems, vol. 2021, pp. 1–13, Apr. 2021, doi: 10.1155/2021/5531770.

[12] K. E. Smith and S. D. Pollak, “Early life stress and development: potential mechanisms for adverse outcomes,” Journal of Neurodevelopmental Disorders, vol. 12, no. 1, Dec. 2020, doi:10.1186/s11689-020-09337-y.

[13] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, “Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory Stress Detection With Smartwatches,” IEEE Access, vol. 8, pp. 38146–38163, 2020, doi: 10.1109/access.2020.2975351.

[14] K. Kyriakou et al., “Detecting Moments of Stress from Measurements of Wearable Physiological Sensors,” Sensors, vol. 19, no. 17, p. 3805, Sep. 2019, doi: 10.3390/s19173805.

[15] S. Gedam and S. Paul, “A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques,” IEEE Access, vol. 9, pp. 84045–84066, 2021, doi: 10.1109/access.2021.3085502.

[16] M. Chauhan, S. V. Vora, and D. Dabhi, “Effective stress detection using physiological parameters,” 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Mar. 2017, doi: 10.1109/iciiecs.2017.8275853.

[17] P. Chyan, A. Andani, I. Nurtanio, and I. Areni, "A Deep Learning Approach for Stress Detection Through Speech with Audio Feature Analysis," in *The 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE-2022)*, IEEE, 2022, pp. 269–273.

[18] G. M. Slavich, S. Taylor, and R. W. Picard, “Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations,” Stress, vol. 22, no. 4, pp. 408–413, Apr. 2019, doi: 10.1080/10253890.2019.1584180.

[19] S. Paulmann, D. Furnes, A. M. Bøkenes, and P. J. Cozzolino, “How Psychological Stress Affects Emotional Prosody,” PLOS ONE, vol. 11, no. 11, p. e0165022, Nov. 2016, doi:10.1371/journal.pone.0165022.

[20] K. Pisanski and P. Sorokowski, “Human Stress Detection: Cortisol Levels in Stressed Speakers Predict Voice-Based Judgments of Stress,” Perception, vol. 50, no. 1, pp. 80–87, Dec. 2020, doi:10.1177/0301006620978378.

[21] K. Tomba, J. Dumoulin, E. Mugellini, O. Abou Khaled, and S. Hawila, “Stress Detection Through Speech Analysis,” Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, 2018, doi: 10.5220/0006855803940398.

[22] H. K. Shin, H. Han, K. Byun, and H. G. Kang, "Speaker-invariant Psychological Stress Detection Using Attention-based Network," *2020 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2020 - Proc.*, no. December, pp. 308–313, 2020.

[23] R. Dillon and A. Ni Teoh, "Real-time Stress Detection Model and Voice Analysis: An Integrated VR-based Game for Training Public Speaking Skills," *IEEE Conf. Games*, pp. 1–4, 2021.

[24] I. Madhavi, S. Chamishka, R. Nawaratne, V. Nanayakkara, D. Alahakoon, and D. De Silva, “A Deep Learning Approach for Work Related Stress Detection from Audio Streams in Cyber Physical Environments,” 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Sep. 2020, doi: 10.1109/etfa46521.2020.9212098.

[25] A. König et al., “Measuring Stress in Health Professionals Over the Phone Using Automatic Speech Analysis During the COVID-19 Pandemic: Observational Pilot Study,” Journal of Medical Internet Research, vol. 23, no. 4, p. e24191, Apr. 2021, doi: 10.2196/24191.

[26] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, “MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech,” Biomedical Signal Processing and Control, vol. 71, p. 103107, Jan. 2022, doi:10.1016/j.bspc.2021.103107.

[27] G. Douzas, F. Bacao, J. Fonseca, and M. Khudinyan, “Imbalanced Learning in Land Cover Classification: Improving Minority Classes’ Prediction Accuracy Using the Geometric SMOTE Algorithm,” Remote Sensing, vol. 11, no. 24, p. 3040, Dec. 2019, doi:10.3390/rs11243040.

[28] Vandana, N. Marriwala, and D. Chaudhary, “A hybrid model for depression detection using deep learning,” Measurement: Sensors, vol. 25, p. 100587, Feb. 2023, doi: 10.1016/j.measen.2022.100587.

[29] N. Rafique, L. I. Al-Asoom, R. Latif, A. Al Sunni, and S. Wasi, “Comparing levels of psychological stress and its inducing factors among medical students,” Journal of Taibah University Medical Sciences, vol. 14, no. 6, pp. 488–494, Dec. 2019, doi: 10.1016/j.jtumed.2019.11.002.

[30] N. F. Narvaez Linares, V. Charron, A. J. Ouimet, P. R. Labelle, and H. Plamondon, “A systematic review of the Trier Social Stress Test methodology: Issues in promoting study comparison and replicable research,” Neurobiology of Stress, vol. 13, p. 100235, Nov. 2020, doi:10.1016/j.ynstr.2020.100235.

[31] Q. Ren, Y. Li, and D. Chen, “Measurement invariance of the Kessler Psychological Distress Scale (K10) among children of Chinese rural-to-urban migrant workers,” Brain and Behavior, vol. 11, no. 12, Nov. 2021, doi: 10.1002/brb3.2417.

[32] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews," *Proc. 9th Int. Conf. Lang. Resour. Eval. Lr. 2014*, pp. 3123–3128, 2014.

[33] A. Défossez, G. Synnaeve, and Y. Adi, "Real-time speech enhancement in the waveform domain," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2020.

[34] S. He-Ping, C. Ji-Hua, and L. Xiao, “Blind Source Separation for Non-stationary Signal Based on Time-Frequency Analysis,” 2011 4th International Conference on Intelligent Networks and Intelligent Systems, Nov. 2011, doi: 10.1109/icinis.2011.12.