A Deep Learning Approach for Work Related Stress Detection from Audio Streams in Cyber Physical Environments

Ishara Madhavi
Department of Computer Science & Engineering,
University of Moratuwa,
Katubadda, Sri Lanka.
madhavi.16@cse.mrt.ac.lk

Sadil Chamishka
Department of Computer Science & Engineering,
University of Moratuwa,
Katubadda, Sri Lanka.
sadilchamishka.16@cse.mrt.ac.lk

Rashmika Nawaratne
Centre for Data Analytics and Cognition,
La Trobe University,
Melbourne, Austrlia.
b.nawaratne@latrobe.edu.au

Vishaka Nanayakkara
Department of Computer Science & Engineering,
University of Moratuwa,
Katubadda, Sri Lanka.
vishaka@cse.mrt.ac.lk

Damminda Alahakoon
Centre for Data Analytics and Cognition,
La Trobe University,
Melbourne, Austrlia.
d.alahakoon@latrobe.edu.au

Daswin De Silva
Centre for Data Analytics and Cognition,
La Trobe University,
Melbourne, Austrlia.
d.desilva@latrobe.edu.au

***Abstract*—Work-related stress is an uncompromising burden with compounding effects on individuals, communities, organizations, and the economy. Highly automated digital environments, such as smart factories and cyber-physical ecosystems, are significantly impacted by these negative effects due to the inherent constraints on human engagement and social interaction. Verbal communication between human operators is a critical resource in such environments, as it can be effectively utilized to address this challenge. A mix of statistical and artificial intelligence (AI) techniques have been reported in recent literature for the detection of stress-related indicators from audio recordings. However, none of these studies has focused on the challenges of detecting work-related stress in cyber-physical environments, where verbal communication is not only constrained but also impaired by background noise and other disturbances common to industrial settings. In this paper, we address these challenges by proposing a novel deep learning approach, based on the workings of the Convolutional Neural Network (CNN) and Growing Self-Organizing Map (GSOM) algorithms. In addition to this novel AI approach, sampling strategy and speaker diarization techniques are used for noise reduction. Pitch and speech augmentation techniques address the data imbalance issue common in most real-world dataset. The accuracy and effectiveness of the proposed approach is demonstrated using a benchmark dataset (DAIC-WOZ) which reports F1 scores, 82% and 64% for normal and distressed classes respectively, outperforming the state-of-the-art models. We conclude with a discussion on the empirical evaluation of the proposed approach in cyber-physical environments and directions for future work.**

Keywords— Work related stress, Human operators in smart factory, Audio stream in smart factory, Industry 4.0, Operator 4.0, cyber physical environment, distress detection, deep learning, unsupervised machine learning, Growing Self-Organizing Map, Cyber Physical Systems, Artificial Intelligence

I. INTRODUCTION

Cyber Physical environments designate the advancement of the Internet of Things (IoT) for physical devices and the cyber components to function collaboratively. Industry 4.0 smart factories are the most effective settings for cyber physical environments, but it also extends across healthcare systems, building management, energy and transportation [1]. This tightly controlled and real-time monitored automation has led to innovative work environments [2], but on the other hand, extended work hours, limited autonomy and lack of social interactions and human engagement has led to work-related stress among human operators in such environments [3],[4].

Work-related stress has been defined as the adverse reaction individuals have to excessive pressures or other types of demand placed on them at work [5]. It encapsulates a variety of feelings of vulnerability, fear and sadness that can cause depression, anxiety, social isolation and mental health deterioration [6]. Depression is closely associated with distress in clinical references, thereby most assessment tools for distress are based on depression, specifically the Patient Health Questionnaire (PHQ), ranging from 2 to 15 questions [7],[8]. Conventional approaches which involve health care questionnaires or discussions with patients, for the detection and diagnosis of work-related stress are primarily subjective, inconsistently implemented, and expensive for the individual in urgent need of support [9],[10]. These issues are further augmented in a cyber physical environment where there is limited scope for a medical professional to conduct interviews or questionnaires, the absence of a peer support network of co-workers, and the increased human to machine interactions severely restricts how individuals experiencing work-related stress express themselves.

Recent studies have recognized the role of artificial intelligence in detecting work-related stress, as well as the indicators of depression [11]. In these studies, diverse modalities of human expressions such as facial expressions, verbal expressions, conversations and physical behaviors have been postulated as mediums for the application of artificial intelligence algorithms. From these modalities, audio streams of both verbal expressions and conversations have been found to be more effective than the others, mainly due to the rich content of semantics, themes and emotion [12],[13],[14]. Several techniques have been proposed for feature extraction and model selection strategies from audio streams, based on hand-crafted features [15],[16] and deep

learning techniques [17],[18]. In deep learning techniques, either the raw audio signals can be directly supplied to the model or hand-crafted features are extracted and then input to the model [17]. The features predominately used for stress detection include prosodic features, formants, spectral features, as well as the variations of prosodic features of the voice.

However, the challenges of detecting work-related stress and indicators of depression in cyber physical environments has not been explored. Building on the success of the aforementioned studies, we propose a deep learning approach for work related stress detection from audio streams in cyber physical environments. We specifically focus on addressing the constraints of brief speech, limited verbal exchanges, as well as background noise and other disturbances common to industrial settings.

The proposed approach extracts low-level features using Fourier transform with incremental hop size, followed by a convolution neural network consisting of two convolution, two dense and a fully connected layer for high-level feature extraction. Pitch and speech augmentation is used for addressing the sampling bias and the extracted high-level features are learned using the Growing Self-Organizing Map (GSOM) algorithm [19], which generates a topological map of this feature space using a dynamic unsupervised machine learning process. More specifically, the selection of an incremental hop size, filtered range of frequencies, transition of low-level features to high-level features using deep learning and the dynamic self-organization of these features on a topological map address the challenges of audio streams in cyber physical environments. In this paper, we make the following research contributions,

1) Development of a noise removal and feature extraction pipeline for cyber physical environments impacted by the constraints of brief speech, limited verbal exchanges, and background noise

2) A novel deep learning approach based on the workings of the Convolutional Neural Network (CNN) and Growing Self-Organizing Map (GSOM) algorithms for the detection of work-related stress detection from audio streams

3) Evaluation of the accuracy and effectiveness of the proposed approach using a benchmark dataset.

The rest of the paper is organized as follows; Section II presents related work on stress detection from audio streams. Section III presents the noise removal and feature extraction pipeline as well as the proposed novel deep learning approach. Section IV documents the experiments and results conducted on the benchmark dataset (DAIC-WOZ). Section V sees to the conclusion of the paper.

II. RELATED WORK

Related work on stress detection from audio streams has broadly focused on two key themes, salient feature extraction techniques and stress detection techniques.

A. Feature extraction techniques

Spectral and temporal features of audio streams represent short-time spectrum and the temporal evolution of signals respectively. Spectral feature extraction is influenced by the quasi-frequency analysis performed inside the cochlea of the human ear as described by Kurzekar et al. [20]. Spectral features are generated by transforming the time-domain signal into frequency-domain to recognize pitch, rhythm, and melody related information. Mel Frequency Cepstral Coefficient (MFCC) has been shown to be useful in recognizing the presence of distress among individuals in the work of Pan et al. [21]. According to Lu et al. [22], the most widely investigated acoustic feature for stress is the pitch of the human voice (F0). Pitch reflects the fundamental frequency of vocal cord vibration during speech production in which the mean and average statistics increases with high stress levels. Temporal features extraction is relatively straightforward as it is based on the energy and zero crossing rate of the audio signal, which are direct indicators of the level of distress. Scherer [23] suggests that stressed, gloomy or moody human nature is blended with low amounts of energy in voice. A relationship between the harmonic-to-noise ratio (HNR) which can be calculated using either the time or the frequency base, and the depression associated with an individual has also been identified in previous research.

Besides spectral and temporal features, there have been recent attempts to extract hidden representative features of distress. Prosodic features determine the variation of tone, stress or rhythm embedded in the voice. Being voice quality features, jitter and shimmer represent irregular vocal fold vibrations and underlying depression according to Nunes et al. [24]. As a high-level feature, gender has been identified as an important factor for identifying distressed individuals. Gender dependent models (GDM) and gender independent models (GIM) have been proposed and discovered that GDM outperforms GIM in the work of Nolen-Hoeksema [25]. When identifying the distress status of teenagers, this work explains, teenage girls face more challenges than the boys and appear to be more of a risk factor for depression. Even the female having a higher pitch of their speech than male, has to be considered when developing models to identify distress status. It has also been noted that the acoustic feature set is proven to be effective in capturing the affective nature of the individual speakers in comparison to the knowledge gained through vocabulary or grammatical semantics involved with the spoken content [21].

In terms of extraction techniques, there are broadly two types, statistical methods for hand-crafted acoustic feature extraction and directly feeding raw audio inputs to deep neural networks for feature learning. The key issues of finding a methodology to extract audio related hand-crafted features from a speech sequence are addressed in [26], where genetic programming and support vector machines (SVM) are used.

B. Stress detection techniques

In terms of hand-crafted acoustic features, various classification techniques are proposed for distress detection. Comparative analysis of various classification techniques including SVM, Gaussian Mixture Model (GMM) and Linear Regression (LR) over combination of multiple features sets are analyzed in [27]. Improved classification accuracies have been achieved with the proposed ensemble logistic regression model over other basic classifiers.

In the deep learning literature, several architectures are adapted as Deep Convolutional Neural Network (DCNN), the Deep Convolutional Neural Network followed by a Deep Network (DCNN-DNN) and the Long Short Term Memory network (LSTM) [18],[28]. A Recurrent Neural Network (RNN) based solution is proposed by Rejaibi [29], dividing

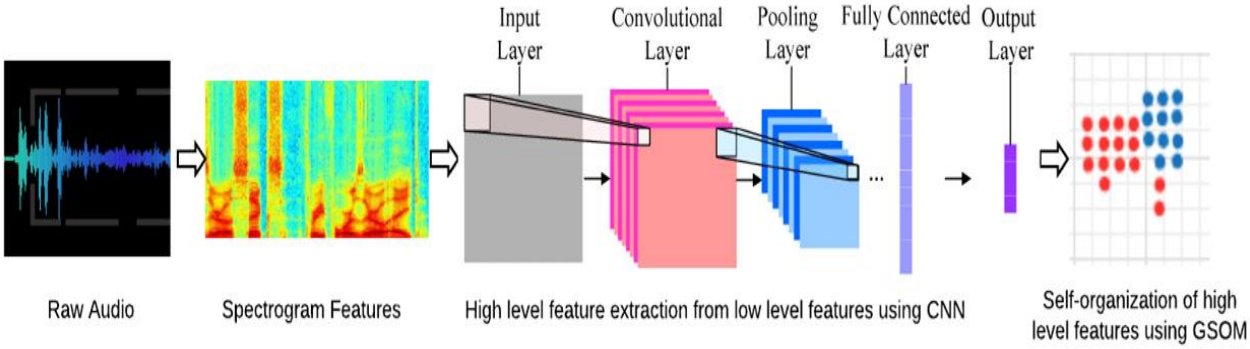

Fig.1. The speech signals are preprocessed and spectrograms are generated. The convolution filters extract high-level features. The dimension-reduced features of CNN are input to the self-organizing algorithm.

audio signal into 2.5 seconds long frames and extracting 60 cepstral coefficients for each 500 milliseconds. The proposed RNN is used to extract the high-level audio features. In addition, data augmentation techniques have been used to handle the data imbalance of minority samples and transfer learning from emotion recognition models is also utilized as an aid for distress detection. DepAudioNet is a CNN based LSTM network [18], using Mel-scale filter bank features. Feature extraction is organized for extracting short term, mid-term and long-term features. A random sampling strategy is introduced to address the data imbalance issue by selecting a limited number of voice frames from majority samples while selecting a higher number of voice frames from minority (distress) samples. A combination of spectrogram-based CNN model and MFCC based CNN model has been proposed as EmoAudioNet in the work of Rejaibi et al. [28].

III. PROPOSED METHODOLOGY

An overview of the proposed method is illustrated in Fig.1. Incoming audio streams are initially preprocessed and prepared for the extraction of low-level audio features of spectrograms. Next, these spectrogram features are fed into a CNN to extract high level features such as temporal patterns in frequency levels and to reduce the dimensionality of the extracted features. Finally, the output from the CNN is presented to the GSOM algorithm for the generation of a topology preserving feature map which is then used for unsupervised classification of distress. This approach is delineated in two phases, 1) noise removal and feature extraction and 2) deep learning and dynamic self-organizing phase.

A. Noise Removal and Feature Extraction

As the cyber physical environments are inherently prone to background noises of machinery, a noise removal step is proposed to be implemented with the support of noisereduce python library [30] which utilizes spectral gating technique. Speech signals are denoised by analyzing fourier statistics of an identified noise signal and applying the identified, noise-specific mask over the noisy speech signals. Speech segments of the participants are separately identified from silence and audio signals of other speakers with the help of the pyAudioAnalysis segmentation module [31]. After performing basic preprocessing, continuous speech streams of each individual are input to the feature extraction process. A spectrogram is a 2-dimensional representation of the spectrum of frequencies that vary across time. Spectrograms can be generated by applying the short time Fourier transform over the preprocessed audio signals with a proper hop size having the ability of extracting low-level features of the signal, based on empirical analysis related to our work. Fourier transform of each time step provides the frequency information in a time-localized manner. The fundamental frequency of females is approximately 350Hz and harmonics vary up to 17kHz. The fundamental frequency of the males is 100Hz - 900Hz and its harmonics rise up to 8kHz [32]. Considering both statistics, the sampling rate of 16kHz and a window size of 64 milliseconds (with a 0.5 window overlap) is configured to capture the frequency information in the 0-8 kHz range of audio signals. This approach leads to attenuate the effect of high-frequency background noises in cyber-physical environments. The decibel intensities of the frequency components are taken as the features. The feature matrix consists of frequency bins for each time step. A windowing function is applied over the matrix to extract 4 seconds of data to analyze the variations of the frequency bins with the assumption that the distress or normal status of a person is constantly maintained throughout the audio signal and a 4 seconds time interval is sufficient to capture the information.

B. Deep Learning and Dynamic Self-organizing phase

As noted earlier, the proposed deep learning approach is based on the workings of the CNN and GSOM algorithms.The CNN architecture consists of 2 convolution layers and 2 successive dense layers followed by a fully connected layer. The convolution filters are configured in a way that they identify the temporal variations of the intensity levels of each frequency component. After each convolution layer, there exists a max-pooling layer to capture the high-level variations of those patterns. From each 4 seconds frame of the audio sample, 512 feature arrays are extracted using the CNN. The intermediate output from the CNN model is fed as input data into the GSOM algorithm to identify the distressed and non-distressed clusters in an unsupervised manner. Each input vector consists of 512 dimensions of high-level features extracted from the preceding convolution layers and mapped on a 2-dimensional grid. The GSOM algorithm starts with four nodes and learns the structure of the input data space using a Euclidean distance-based learning rule. When the accumulated quantization error of a single node exceeds the growth threshold (GT), this specific node is deemed unsuitable to represent the input data, and new nodes are generated from this node, to better represent this quantization error [33]. Following the training, testing and smoothing out phases, the topological map generated by the GSOM algorithm is used to identify clusters of nodes that exhibit similar audio properties indicative of distress. The complete workings of the GSOM algorithm are reported in [34].

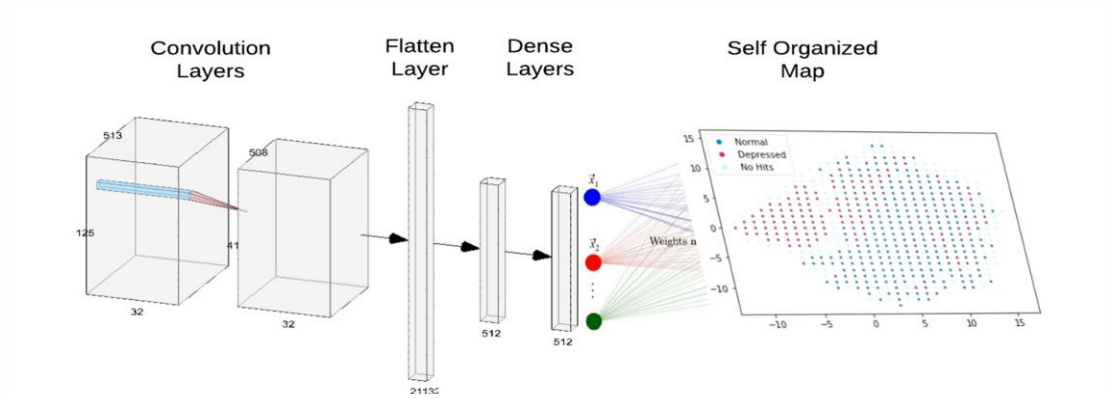

Fig.2. The model architecture of the CNN-GSOM combined implementation. The two convolution filters extract low-level features, the succeeding lay flatten and compact the features to be visualized on self-organizing map.

In the context of distress detection, most of the datasets are imbalanced having a majority of data for normal individuals and a minority for distressed individuals. We evaluated several approaches to address this data imbalance, such as oversampling the minority data, under-sampling the majority data or combination of both undersample and oversample. Based on this evaluation, we selected augmenting the raw audio to be the most effective approach for audio streams in a cyber physical environment. We specifically used pitch and speech augmentation to generate new audio samples which resulted in doubling the minority sample space, thus matching the size of the control space.

IV. EXPERIMENTS AND RESULTS

We used a benchmark dataset to evaluate the accuracy and effectiveness of the proposed deep learning approach. The Distress Assessment Interview Corpus (DAIC-WOZ) dataset [35] was released as part of the 2016 Audio/Visual Emotional Challenge and Workshop (AVEC 2016). As mentioned in Section I, depression is closely associated with distress and work-stress in clinical references, thereby most assessment tools for distress are based on depression, specifically the PHQ. This dataset comprises an audio file of a conversation with the participant and a label based on the participant's responses to the eight-item Patient Health Questionnaire scale (PHQ-8). PHQ-8 is an accepted diagnostic and severity measurement for work-related stress and other types of mental distress as noted in large clinical studies [36]. The DAIC-WOZ dataset is imbalanced as the number of non-distressed participants is 3 times higher than distressed participants. The training-testing split was approximately 75 – 25, for training 116 audio recordings consisting of 37 distressed and 79 non- distressed, for testing 37 audio recordings consisting of 11 distressed samples and 26 non-distressed samples.

Short term Fourier Transform operation performed on signals with a hanning window of size 64 milliseconds and a hop size of 32 milliseconds was used to extract the low- level features of the signal. For each time step, 513 frequency components are calculated to interpret the 32 milliseconds long frames of the signal. The decibel intensities of corresponding frequency bins are taken as the features. A matrix of size (513, N) is generated at the end of the feature extraction process in which N varies depending on the length of the session. In order to prevent personal characteristics being emphasized from long duration speeches, 46 numbers of 4s long segments from each audio clip is taken and labeled. A width of 4 seconds of audio results in a window size of 4/0.032=125 pixels. Therefore, a randomly selected 46 number of (513, 125) chunks from each feature matrix (513, N) is fed to the CNN.

The overall model architecture, consisting of the CNN and the GSOM models is shown in Fig.2. The CNN contains an input layer which accepts feature matrices of shape (513, 125, 1) and outputs (511, 123, 32) feature maps via 32 amounts of 3x3 filters, using the ReLU activation function. Then the feature maps undergo dimensionality reduction with a max-pooling layer of 4x3 filters and a stride of size 1x3. The second convolution layer consists of 32 filters of size 1x3 getting activated by the ReLU function and dimensionality is further reduced by the max-pooling layer of size 1x3 filters having a stride of 1x3. The features are flattened at the fully connected layer and the next 2 dense layers are configured with 512 neurons and 50% dropouts.

For the CNN training and validating phase, a SoftMax layer is configured with a dense layer having two neurons. The initial CNN is trained with nearly 10 epochs, batch size of 32 and using the Adadelta optimizer which dynamically adapts the learning rate based on the gradient. When the CNN is yielding a stable validation, the SoftMax layer is replaced with the proposed GSOM layer.

Initially the GSOM consists of 4 nodes placed in a rectangular shape and input data samples are of 512 dimensions. The map dynamically self-organizes until it can map the 512-dimension data points into 2-dimensional space.

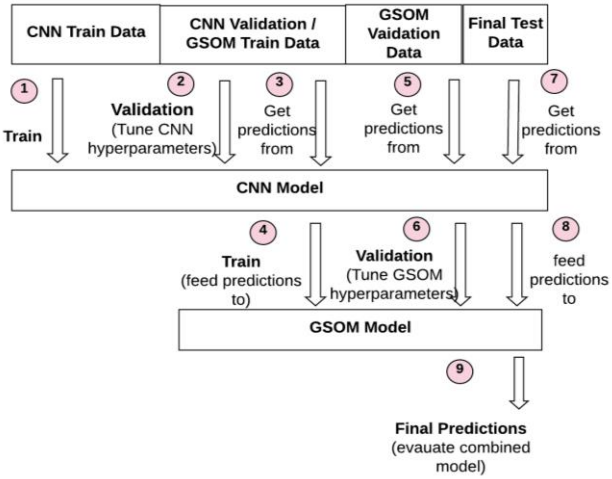

Fig.3. Splitting of the dataset for training, testing and validation in order to avoid overfitting

A set of GSOM node maps are analyzed for performance, by configuring the hyperparameters (100 learning iterations, 50 smoothing iterations, forget threshold of 80, spread factor from 0.1 to 0.9) which govern the growth of the map.

There are 116 sessions for the training and validation altogether. 37 sessions are reserved for testing. Fig.3 indicates how data is separated for each of models' training and validation without causing overfitting. From the 79 control sessions, a sample space of 3634 audio segments are generated by randomly extracting 46 segments from each session. As the number of distress sessions are 37, the number is increased up to 74 by using the pitch and speed augmentation technique that doubles the minority sample space. As a result, the distressed sample space consists of 3404 audio segments, after the random extraction of 46 samples from each session.

In order to train the CNN model, approximately 50% of the data is utilized, in which 1802 segments are control samples while 1702 segments are from the distressed sample space. From the remaining sample spaces, 1600 and 1500 segments are extracted from the control and distressed samples respectively, to validate the CNN model as well as to train the GSOM model. The remaining control and distressed audio segments are kept aside for validating the GSOM model, which are counted as 232 and 202 segments in particular.

A. Results

The test set contains 37 sessions that can be utilized for predictions in which 11 are distressed and 26 are control samples. The test set creates a sample space of 1702, after selecting 46 random samples from each session. After the frame-wise predictions for each session is found, the majority voting procedure is followed to detect the whole session as a distressed session or a normal session. Each test session consists of 46 frames, each frame having an audio segment of 4 seconds. If more than 50% frames are predicted as distressed, then the overall session is identified as distressed. If less than 50% frames are predicted as distressed, then the overall session is identified as non-distressed.

The accuracies of the model predictions are evaluated using precision (P), recall (R) and F1-Score (F). Precision is the fraction of correctly identified instances out of all instances identified as positive. Recall is the fraction of correctly identified instances over all the positive instances in the dataset. F1-Score corresponds to the weighted harmonic mean of precision and recall metrics. The results taken after the majority voting procedure is shown in Table I and Table II. In order to illustrate the impact of speech augmentation, test data are evaluated for the models created without data augmentation (model A) and with data augmentation (model B). Table III indicates that increasing the data samples via data augmentation to train the model has improved the performance of the proposed model (model A outperforms model B).

For the Depression Classification Challenge (DCC) in [37] in which the same DAIC-WOZ corpus is offered, it requires measuring an average F1-Score using both distressed and non- distressed classes. We evaluate the accuracy of distress detection, with respect to 7 benchmark approaches. First, the baseline model following a linear support vector machine with stochastic gradient descent using the baseline features

TABLE I. TEST SET PREDICTIONS USING MAJORITY VOTE

Confusion matrix	***Actual Yes***	***Actual No***
Predicted Yes	20 (TP)	3 (FP)
Predicted No	6 (FN)	8 (TN)

TABLE II. TEST SET STATISTICS USING MAJORITY VOTE

	F1 score	***Precision***	***Recall***	***Accuracy***
Non-distressed	0.82	0.87	0.77	77%
Distressed	0.64	0.57	0.73	73%

TABLE III. TEST SET PREDICTIONS FOR INDIVIDUAL SAMPLES

Confusion matrix	***Before Augmentation (Model A)***		***After Augmentation (Model B)***	
	Actual Yes	***Actual No***	***Actual Yes***	***Actual No***
Predicted Yes	621(TP)	279(FP)	715(TP)	230(FP)
Predicted No	575(FN)	227(TN)	481(FN)	276(TN)

provided in the challenge [37]. Second, a combination of CNN and Long Short Term Memory (LSTM) network is used for raw audio with Mel-Filter bank features in [18]. Third [38] and fourth [39] are CNNs that utilize spectrogram features of audio data. Fifth is one SVM based model and the sixth work in [29] presents a Recurrent Neural Network (RNN) architecture, both learning from MFCC data. The EmoAudioNet model in [28] consists of a Deep Neural Network and utilizes MFCC and spectrogram data of audio. All the benchmark models implement supervised learning techniques. As illustrated in Table IV, our method (with augmentation) outperforms the other benchmark models and methodologies by achieving an average F1-Score of 73%.

Having many data samples with the PHQ-8 score near the benchmark value of 10, can affect the learning of the model. Also, the test results can be misinterpreted by those samples.Therefore, it was focused to investigate the predictive power of the model by specially examining the samples having very high and very low values for PHQ-8 corresponding to highly distressed and control samples respectively. From the 26 normal test samples, there are 10 samples which have the distressed severity levels (PHQ-8 score) of 0,1,2. Out of those 10 highly normal samples, 8 samples are classified as normal by the proposed model. From the 11 distressed test samples, there are 8 samples which have the severity level over 15. Out of those 8 highly distressed samples, 6 samples are classified as distressed by the model. The results indicate the high predictive power of the proposed model.

B. GSOM Cluster Analysis

By analyzing the results, it becomes obvious that increasing the amount of data samples by following the data augmentation technique has led to a better training of convolution filters and neuron weights of the CNN. As the self-organizing map undergoes unsupervised learning, it has separated data with augmented (having the audio speed increase of 1.07) and data without augmented into two mutually exclusive clusters. The cluster consisting of data points with no augmentation, has been further separated into normal and distress related clusters. As only the original

distress data samples are augmented, the cluster of augmented data is not further categorized.

According to the node labeling strategy, the nodes of the map are labeled as a distressed node or a normal node using the training dataset. The regions with higher density of distressed nodes are identified as the distress clusters and others are identified as normal clusters. The map visualizes a separation of non-distressed and distressed clusters accordingly.

Comparative analysis of how test data points are spread over the map near the distress cluster (marked area) is shown in Fig.4. Data points of distress individuals are gathered with higher density around the identified distress cluster and normal data points are spread away from the distress cluster.

TABLE IV. PERFORMANCE COMPARISON FOR DAIC-WOZ DATASET

Reference	***Model***	***Accuracy***	***F1(N)***	***F1(D)***	***F1(A)***
[36]	SVM + SGD	-	58%	41%	49%
[18]	DepAudioNet (CNN+LSTM)	-	70%	52%	61%
[38]	CNN	77%	87%	25%	56%
[39]	CNN	64%	66%	61%	63%
[16]	SVM	-	85%	48%	66%
[29]	RNN	76%	85%	46%	65%
[28]	EmoAudioNet (DNN)	74%	83%	50%	66%
Ours (Model A)	CNN + GSOM	57%	65%	44%	54%
Ours (Model B)	CNN + GSOM with Data Augmentation	**76%**	**82%**	**64%**	**73%**

We can compare the distress levels of different individuals based on their corresponding positioning (mapping) on the map. The mappings inside the marked distress cluster represent individuals with high distress levels (high PHQ-8 scores) and the locations external to the marked area contain normal individuals (low PHQ-8 scores). The red points diffused into the region of normal samples represent the ones with average distress conditions. This demonstrates the ability of our approach to separate out distress data points from normal data points as well as the high comparability and interpretability of the test samples.

C. Impact of Industrial Noise

In order to justify the strength of the model in the presence of noise, we furthermore conduct an experiment by injecting artificial noise that is likely to be common in cyber physical environments. One experiment utilizes the impact of additive Gaussian noise to mimic the effect of random disturbances occurring in cyber physical environments. The second experiment mimics real-world noise by injecting disturbances emitted from a typical ventilation equipment. Based on the results in Table V, our approach has accurately identified highly distressed individuals even after the introduction of background noise similar to that in cyber physical environments.

TABLE V. TEST SET STATISTICS WITH NOISE INJECTION

	Distressed F1 score	***Non-Distressed F1 score***
Gaussian Noise with SNR = 4	0.35	0.71
Gaussian Noise with SNR = 2	0.25	0.79
Noise of an Air Conditioner	0.42	0.71

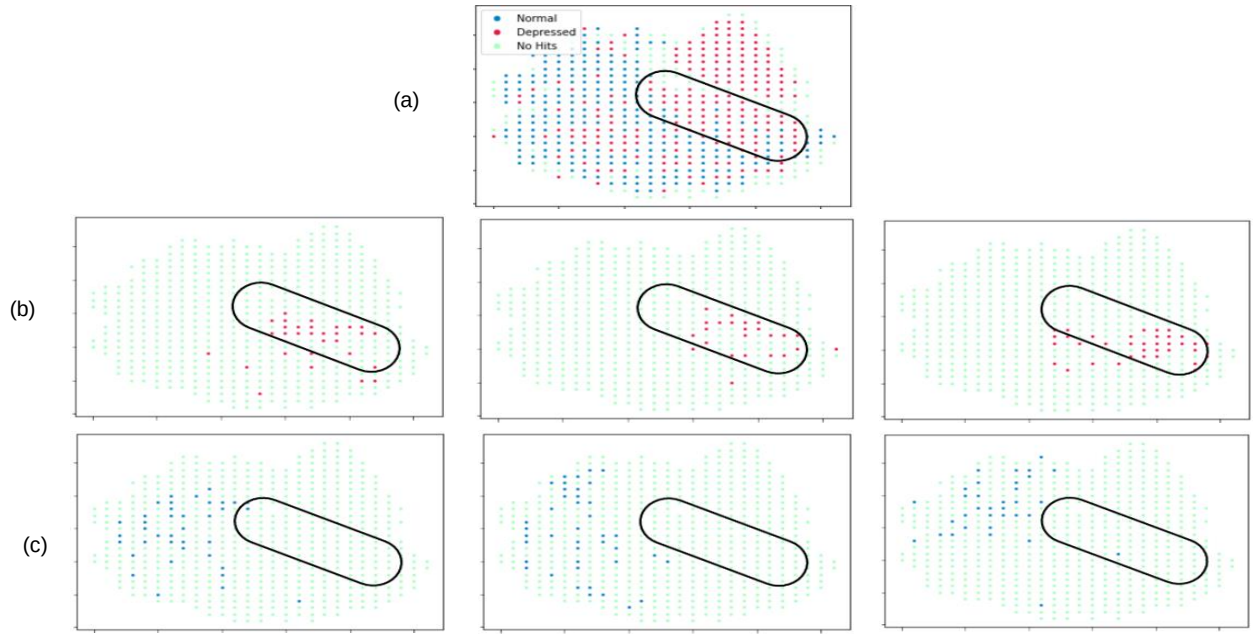

Fig.4. Illustration of the test data distribution on the original map. The first row (a) contains the original map generated from training data. The second (b) and the third (c) rows show the test data mappings of depressed and normal samples on the map respectively

V. CONCLUSION

In this paper we have proposed a novel deep learning approach for work related stress detection from audio streams in cyber physical environments. The proposed approach consists of a noise removal and feature extraction phase followed by a distress detection phase which is based on the workings of the CNN and GSOM algorithms. Low-level features are extracted using Fourier transform with incremental hop size, pitch and speech augmentation techniques are used to address the dataset imbalance, which leads to a feature matrix consists of decibel intensities for each time step. The CNN learns high level features from these low-level spectrogram features to extract temporal patterns in frequency levels and to reduce dimensionality. The high-level features are input to the GSOM algorithm to generate a topology preserving feature map that consists of clusters of nodes used for unsupervised classification of distressed audio streams.

Given the autonomous and intelligent features of the proposed approach, it can be effectively deployed in cyber physical environments consisting of Industry 4.0 technologies and Operator 4.0 behaviours. The restricted nature of social engagements and interactions and the impact of industrial noise does not impede the performance of this approach which can cohesively integrate with existing systems and platforms of CPS and IoT devices. The proposed approach is robust to be implemented in cyber physical environments, due to its ability to distinguish distressed individuals in a high accuracy compared to the state-of-the-art models.

In terms of future work, firstly we will focus on the empirical evaluation of the proposed approach in real-world cyber-physical environments, such as a smart factory, smart building or smart campus. In a smart factory setting, we will deploy this technique as a modularised component of the overall CPS architecture. We have been working on the container technologies required for this deployment. Secondly, we will expand the detection process by identifying emotions of the individuals in addition to the current features. This will be effective for the human monitoring systems placed in cyber-physical environments to precisely detect distress of the workers and support them in maintaining a good mental health in workspaces. In addition to the spectrograms taken into consideration, there is substantial evidence that audio related features including MFCCs, jitter, shimmer, zero-crossing rate are also useful for work-related distress detection [40], just as gender has been found to be an indicative variable to be considered [41]. A further direction of future work is to incorporate these into the detection approach. The model is trained and evaluated on a dataset which consists of audio samples collected with much care for mitigating background noises as much as possible. However, depending on the nature of the cyber-physical environment, noise impairment levels might vary differently from one system to the other. This approach can be further modified to tolerate a large variety of noises associated in industry settings using more sophisticated strategies.

REFERENCES

[1] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, “Cyber-physical systems: The next computing revolution,” Proc. - Des. Autom. Conf., pp. 731–736, 2010.

[2] H. Nakashima, H. Aghajan, and J. C. Augusto, Handbook of Ambient Intelligence and Smart Environments, 1st ed. Springer Publishing Company, Incorporated, 2009.

[3] P. Boxall, M. Huo and J. Winterton, "How do workers benefit from skill utilisation and how can these benefits be enhanced?", Journal of Industrial Relations, vol. 61, no. 5, pp. 704-725, 2019.

[4] Romero, D., Stahre, J., Wuest, T., Noran, O., Bernus, P., Fast-Berglund, Å., & Gorecky, D. (2016, October). Towards an operator 4.0 typology: a human-centric perspective on the fourth industrial revolution technologies. In Proceedings of the International Conference on Computers and Industrial Engineering (CIE46), Tianjin, China (pp. 29-31).

[5] "What is work-related stress?", Health and Safety Executive Northern Ireland,2020.

[6] J. Peŕezgonzaĺez, T. Cox, A. Griffiths and E. Rial-Gonzales, Research on work-related stress. United States: [Lulu, Inc.], 2003.

[7] K. Kroenke, R. L. Spitzer, J. B. W. Williams, and B. Löwe, “The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review,” Gen. Hosp. Psychiatry, vol. 32, no. 4, pp. 345–359, 2010.

[8] M. Idris and M. Dollard, "Psychosocial safety climate, work conditions, and emotions in the workplace: A Malaysian population-based work stress study.", International Journal of Stress Management, vol. 18, no. 4, pp. 324-347, 2011.

[9] A. Mitchell, A. Vaze and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis", The Lancet, vol. 374, no. 9690, pp. 609-619, 2009.

[10] A. Adikari, D. De Silva, D. Alahakoon and X. Yu, "A Cognitive Model for Emotion Awareness in Industrial Chatbots," 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), Helsinki, Finland, 2019, pp. 183-186.

[11] Tran, B., McIntyre, R., Latkin, C., Phan, H., Vu, G., Nguyen, H., Gwee, K., Ho, C. and Ho, R., 2019. The Current Research Landscape on the Artificial Intelligence Application in the Management of Depressive Disorders: A Bibliometric Analysis. International Journal of Environmental Research and Public Health, 16(12), p.2150.

[12] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, L.-P. Morency,Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews, IEEE Transactions on Affective Computing (1) (2015) 59–73.

[13] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber,P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, T. F. Quatieri, Detecting depression using vocal, facial and semantic communication cues, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 11–18.

[14] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, D. Jiang, Hybrid depression classification and estimation from audio video and text information, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, 2017, pp. 45–51.

[15] S. Song, L. Shen, and M. Valstar, “Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features,” Proc. - 13th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2018, pp. 158–165, 2018.

[16] M. Nasir, A. Jati, P. G. Shivakumar, S. N. Chakravarthula, and P. Georgiou, “Multimodal and multiresolution depression detection from speech and facial landmark features,” AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, co-located with ACM Multimed. 2016, pp. 43–50, 2016.

[17] L. Yang, E. Pei, D. Jiang, M. C. Oveneke, X. Xia, and H. Sahli, “Multimodal measurement of depression using deep learning models,” AVEC 2017 - Proc. 7th Annu. Work. Audio/Visual Emot. Challenge, co-located with MM 2017, pp. 53–54, 2017.

[18] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, “DepAudioNet: An efficient deep model for audio based depression classification,” *AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, co-located with ACM Multimed. 2016*, pp. 35–42, 2016.

[19] D. Alahakoon, S. K. Halgamuge and B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery," in IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 601-614, May 2000.

[20] P. K. Kurzekar, R. R. Deshmukh, V. B. Waghmare, and P. P. Shrishrimal, “A Comparative Study of Feature Extraction Techniques for Speech Recognition System,” Int. J. Innov. Res. Sci. Eng. Technol., vol. 03, no. 12, pp. 18006–18016, 2014.

[21] W. Pan et al., “Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders,” PLoS One, vol. 14, no. 6, pp. 1–14, 2019.

[22] H. Lu et al., “StressSense: Detecting stress in unconstrained acoustic environments using smartphones,” UbiComp’12 - Proc. 2012 ACM Conf. Ubiquitous Comput., pp. 351–360, 2012.

[23] K.R. Scherer, “Vocal assessment of affective disorders,” in Depression and Expressive Behavior, J.D. Maser, Ed., pp. 57–82. Lawrence Erlbaum Associates, 1987.

[24] A Nunes, L Coimbra, and A Teixeira, “Voice quality of european portuguese emotional speech corresponding author,” Computational Processing of the Portuguese Language Lecture Notes in Computer Science, vol. 6001/2010, pp. 142–151,2010.

[25] S. Nolen-Hoeksema, "An Interactive Model for the Emergence of Gender Differences in Depression in Adolescence", Journal of Research on Adolescence, vol. 4, no. 4, pp. 519-534, 1994.

[26] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, in: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2016, pp. 5200– 5204.

[27] Jiang, H., Hu, B., Liu, Z., Wang, G., Zhang, L., Li, X. and Kang, H.,2018. Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features. Computational and Mathematical Methods in Medicine, 2018, pp.1-9

[28] E. Rejaibi et al., “Clinical Depression and Affect Recognition with EmoAudioNet,” pp. 1–13, 2019.

[29] 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", arXiv.org, 2020.

[30] "timsainb/noisereduce", GitHub, 2020.

[31] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis", PLOS ONE, vol. 10, no. 12, p. e0144610, 2015.

[32] "Human Voice Frequency Range - Sound Engineering Academy", Sound Engineering Academy, 2020.

[33] R. Nawaratne, D. Alahakoon, D. De Silva, H. Kumara and X. Yu, "Hierarchical Two-Stream Growing Self-Organizing Maps with Transience for Human Activity Recognition," in IEEE Transactions on Industrial Informatics.

[34] Nawaratne, R., Adikari, A., Alahakoon, D., De Silva, D., & Chilamkurti, N. (2020). Recurrent Self-Structuring Machine Learning for Video Processing using Multi-Stream Hierarchical Growing Self-Organizing Maps. MULTIMEDIA TOOLS AND APPLICATIONS.

[35] 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.

[36] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, “The PHQ-8 as a measure of current depression in the general population,” J. Affect. Disord., vol. 114, no. 1–3, pp. 163–173, 2009.

[37] M. Valstar et al., “AVEC 2016 - Depression, mood, and emotion recognition workshop and challenge,” AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, co-located with ACM Multimedia. 2016, pp. 3–10, 2016.

[38] 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 Sens., vol. 11, no. 24, 2019.

[39] K. Kiefer, DepressionDetect : A Machine Learning Approach for Audio based Depression Classification. Accessed on: June 16, 2020.

[40] Malviya, A., Meharkure, R., Narsinghani, R., Sheth, V. and Meshram, P., 2019. Depression Detection Through Speech Analysis : A Survey. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp.712-716.

[41] S. Alghowinem et al., “A comparative study of different classifiers for detecting depression from spontaneous speech,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp. 8022–8026, 2013.