Semi-supervised Deep Learning for Stress Prediction: A Review and Novel Solutions

Mazin Alshamrani
The Custodian of the Two Holy Mosques Institute for Hajj and Umrah Research (HURI)
Umm Al Qura University, Makkah, Saudi Arabia

***Abstract*—This research introduces a novel self-supervised deep learning model for stress detection using an intelligent solution that detects the stress state using the physiological parameters. The first part of this research represents a concise review of different intelligent techniques for processing physiological data and the emotional states of humans. Also, for all covered methods, special attention is made to semi-supervised learning algorithms. In the second part of the paper, a novel semi-supervised deep learning model for predicting the stress state is proposed. It is the first attempt of using contrastive learning for the stress prediction tasks. The model is based on utilizing generative and contrastive features specially tailored for treating time-series data. A widely popular multimodal WESAD (Wearable Stress and Affect Detection) data set is exploited for experimental purposes. It consists of physiological and motion data recorded from the wrist and chest-worn devices. To provide an intelligent solution that will be widely applicable, only the wrist data recorded from smartwatches is exploited during the model's training. The proposed model in this research is tested on a single subject's data and predicts the stress and non-stress events. Keeping in mind that the initial data was unbalanced with only 11% of the stress data, data augmentation techniques are applied within the model to provide additional reliable training information. The model shows significant potential in clustering stress conditions, and it presents accuracy in the range with other state-of-the-art solutions. The most significant benefits of using this model are its prediction capabilities when dealing with unlabeled data and performances when undersized data cannot be processed optimally by traditional intelligent methods.**

Keywords*—*Deep learning; semi-supervised learning; contrastive learning; physiological data; stress prediction

I. Introduction and Literature Review

The coronavirus pandemic has directly affected the health of millions of people and living conditions all over the planet. To reduce the impact of the infection, worldwide governments temporarily closed schools, malls, restaurants, sports and art facilities, and all other public and private institutions in which a large number of people are generally present. Like a domino effect, such situations have affected the loss of jobs of millions of people, almost the closure of some professions, the closure of small and large companies, and finally, significant disruptions in the world economic system. In addition to the negative effects that this situation is causing, so far, not much attention is paid to the health problem that will affect a much larger number of people than the virus itself did: stress. To contribute to the diagnosis of stress and the detection of stress in humans, this research addresses the development of an intelligent method for the early detection of stress and the timely prevention of more serious health problems.

Until now, stress and emotions, in general, have been examined in numerous researches. In [1] an emotion recognition system was presented that is applicable when limited data resources are available. The system identified emotions with a 65% success rate and with the confidence of 57%. Further, in [2] the system that discovers classroom emotions and moods of students was introduced. Wristband sensors from Empatica E4 and smart-phones were used to detect all emotions using physical activities, event tags data, and various physiological parameters. The results were exploited for finding associations and correlations between students' data and extracting meaningful insights. In another research [3], a deep learning approach for the classification of emotions was presented. This approach was based on processing data acquired from three sensor modalities: locations into the global model, environmental and on-body. It was also proven that deep learning algorithms can be very effective in classifying human emotions, especially when many sensors are utilized. The average accuracy of the proposed model was 73%. Next, the importance of emotions in user-modeling and multimodal computer interaction was presented in [4]. Additionally, in this paper, the results of different supervised learning algorithms for categorizing physiological signals from the autonomous nervous system were shown. The multimodal system for recognizing users' emotions and generating responses to recognized emotions was presented in [5]. The experiment is thoroughly explained, and the mapping principles of physiological signals to certain emotions were introduced. The utilization of positive and negative emotional electroencephalogram (EEG) signals was described in [6] to research emotions. The support vector machine algorithm was exploited for data analysis, and an accuracy of 58.3% was achieved. Besides using EEG signals, an important role in emotion recognition is the electrocardiogram (ECG) signal. In [7] wearable ECG device was used to follow four kinds of emotional states recorded while involved participants watched prepared movie clips. Features from different analysis domains (time, frequency, static analysis) were sensed and recorded, and the most relevant features for evaluating human's emotions were highlighted. An interesting approach to establishing emotional connections between SMART TVs and the audience was presented in [8]. EEG signals of involving participants were recorded and analyzed, and three emotional states were registered: relaxation, neutral, and horror. These signals were classified by using the support vector machine with an accuracy of 92% for all subjects.

Keeping in mind that the paper's novel research is based on using smartwatch data to detect emotions, it is beneficial to highlight the paper of [9]. The paper examined if smartwatches or wrist bands can be useful in collecting valuable information to recognize emotions. The primary limitations, potential problems, and crucial research steps in this domain were successfully found. Further, in [10], an automatic emotion recognition system was proposed. The system is based on using a wearable wristband, human emotions were evoked artificially, and multimodal physiological signals were collected by exploiting three different sensors. Finally, support vector machines were used once again for classifying emotions, achieving an accuracy of 76%. The usage of wearable sensors was also verified in [11], where they were used within Ambient Intelligence systems to provide affect-based adaptations. Wearable sensors were also utilized in [12], where a real-time mobile biosystem "iAware" was proposed. The system was efficient in depicting five basic emotional states and emotional feedback information was provided to users. A smartwatch application that integrates heart rate, motion, and light data to sense mental health was used in [13], and the PRISM-Passive platform was proposed. Both supervised and unsupervised learning algorithms were applied, and it is proven that smartwatch data could be useful in evaluating and predicting mental health. Finally, the usage of the pervasive wearable devices within the "emotional IoT" concept for recognizing emotions was presented in [14]. This paper presented the end-to-end real-time solution based on smartwatch and smartphone devices that showed great applicability potential in consumers' everyday lives.

In the following papers, the research cover techniques to estimate physiological signals, their processing, and signal quality improvements. In [15] psychophysiological signal quality estimators were proposed that were utilized to affect recognition systems. Further, in [16] findings in the domain of estimations of affective states of users' optimal experiences were presented. Estimated signals through end-to-end intelligent architecture possessed 67.5% accuracy in recognizing different affective states, including stress. The difference in emotion recognition accuracy between laboratory and wearable sensors was examined in [17]. The results showed a similar level of accuracy between the two approaches, which implies that wearable sensors' usage is reliable enough for serious considerations and accurate collection of physiological information of a user outside of a laboratory. Another research [18] covered a framework for signal processing pertaining to clarifying patterns of humans' physiological changes. An urban environment was taken as a scientific background, and the framework included signal unification, filtering, quantification, and the usage of techniques for data labeling. Finally, one interesting research of transformation of emotional signals is presented in [19], where the biosensing prototype for transforming emotions into music was proposed. Four emotional states were covered within the research (neutral, anger, sadness, and happiness). The appropriate EEG signals were recorded, and Audiolize Emotion was used to transform collected data into audio files. In the cases when it is difficult to assign labels to training input data consistently, Multiple Instance Learning from [20] allowed the training of classifiers from not precisely defined labeled data. A potential guideline for increasing the accuracy of labels was presented in [21]. However, in the cases when the labels cannot be completely provided, a self-supervised approach from [22] can be applied. The approach was designed in a way to learn valuable representations from unlabeled sensor inputs as blood volume pulse, electroencephalography, accelerometer, etc. The proposed methodology showed performances in the range with fully-supervised networks and improved generalization capabilities in semi-supervised settings.

As one more interesting topic for proposing the novel research in this paper, deep learning techniques for processing physiological and emotional parameters should also be presented. In [23] deep learning techniques for real-time stress and affect detection was examined. New models based on Multiple Instance Learning were proposed and applied, showing the performances 10% better in terms of accuracy. Further, the hyperparameter optimization framework of long short-term memory networks in the context of emotion classification was presented in [24]. It was shown that the framework provided an improved recognition rate accuracy of more than 10% compared to other state-of-the-art optimization methods. One more classification model built with deep neural networks was presented in [25]. A fully convolutional network was proposed and achieved performances were in the range or better than other state-of-the-art time series classification algorithms. A convolutional neural network architecture was also used in [26] to classify the biosignals, achieving the precision across all the classes equal to 97.65%. Finally, in [27, 28] novel techniques for optimizing network architectures to improve their processing power were presented. Specifically, in [27] rectifying neurons as improved models of biological neurons were presented. The structures based on these neurons are suitable for sparse data, they do not require unsupervised pre-trainings, and deep rectifier networks were very efficient in environments where there was a lack of labeled data. Additionally, classification models can be improved by making a normalization process an integral part of a model architecture [28]. This method performed the normalization process for each training batch, provided the same accuracy with 10-15 times fewer training iterations than some traditional network structures. Two more successful approaches of utilizing deep learning techniques for prediction purposes were given in [29, 30].

In this novel research, a novel self-supervised learning (SSL) algorithm is proposed and utilized for processing a widely popular WESAD data set. The goal is to accurately detect stress by using smartwatch sensor data. To describe the research methodology and proposed framework, the mentioned dataset must first be introduced properly. In general, WESAD is a multimodal dataset for wearable stress and affect detection [31]. The set includes information recorded both from chest and wrist measurements of sensors. It was based on three different affective states: neutral, amusement, and stress. Fifteen subjects participated in the experiment, 12 males and three females. The average age of participants was 27.5. The stress condition occurred as a response from public speaking exercises, and no other types of stress environments were included. Linear Discriminant Analysis (LDA) model was used

in [31] as a stress classification model, which achieved 93% accuracy. The next important paper is [32], in which the WESAD data were classified into four classes: neutral, amusement, and stress as in the previous paper, and meditation was an additional class. In comparison to [31], [32] only used the wrist sensor data for classification purposes. A machine learning model was trained for each subject separately (logistic regression, decision tree, and random forest models). The best performances were achieved using random forest models: accuracy between 88% and 99% depending on the examined subject. Whether it was possible to perform stress detection using only a smartwatch sensor data from the WESAD data set was examined in [33]. Three different models were used: LDA, Quadratic Discriminant Analysis (QDA), and Random Forest (RF). The best performances were achieved with LDA in combination with the next sensors' data: heart rate (HR), blood volume pulse (BVP), and skin temperature (ST). The next paper [34] relied on using deep learning techniques for processing the WESAD data. The primary model processed inputs of different sampling rates by utilizing four different sub-models as classifiers that individually process per one different sampling rate. The final model was based on the RF algorithm and generated the final classifications following the fusion mechanism in [35]. Finally, the research in [36] used a self-supervised methodology and deep learning for processing only the ECG signal from four data sets (including WESAD). The methodology was developed using data augmentation techniques, including stacked convolutional network layers and a final "SoftMax" dense classification layer. By examining previously presented WESAD research papers, a major issue that significantly influences the dataset's classification results was identified: a lack of data diversity. An attempt to solve this problem will be made in the research by introducing a novel SSL methodology. In the next section, the research background is presented, and fundamental SSL concepts are introduced.

II. Research Background

For this research application purposes, the future model's output labels are two physiological states: "stress" and "neutral". The labels represent the outputs of a model for corresponding input data points. To determine when these two labels occur, it is required to know whether a subject of examination is stressed or not stressed. This can only be accurately determined in an experimental environment by making long-term observations by expert knowledge from the field. If a supervised learning algorithm is selected for processing the data, it is required to provide output labels for all the input data. For the situations when complex and time-consuming experimental procedures should be performed (as in this case of the stress prediction), much effort should be made and costs covered to collect the complete database of labels. One possible way to optimize this process is to use an SSL approach for the training process and reduce the need for the number of prepared labels. Unlike supervised learning that relies on using pre-prepared input/output pairs of data, SSL techniques create their output labels from available input information. These techniques reduce the dependency on labels by constructing meaningful and invariant representations that capture the original data's high-level information. SSL techniques are based on two essential concepts: the pretext task and making appropriate representations. The key to the pretext task is to use information about the input data to construct pretext labels. On the other hand, a representation is a way to simplify the data while keeping relevant information. For sensor data that is processed within this research, a useful representation preserves emotional and physical state information and discards redundant information like noise. Additionally, the previous SSL applications show that representations with significant invariance are often more robust and provide better quality than other types of representations.

A review of SSL methods for treating images, text files, and graph data was presented in [37]. The comparisons between supervised and unsupervised learning algorithms were made, and three main categories of SSL were explained: Generative, Contrastive, and Adversarial SSL. One efficient approach of contrastive learning in the form of a simple framework called SimCLR was proposed in [38]. The approach was based on utilizing the data augmentation techniques and using two different networks within the architecture: the first one to create representations of different inputs and the second one for comparison of representations and preserving important information. Examined research showed that data augmentation could be a key tool to build accurate SSL models. Augmentation techniques apply input data transformations to create new relevant samples from the existing ones and increase the size of data sets when needed. One example of using data augmentation for creating the models with significant accuracy was presented in [39]. In this paper, a popular contrastive method called Dimensionality Reduction by Learning an Invariant Mapping (DrLIM) was proposed. The method is efficient in creating relevant features out of complex data, and it was commonly applied in previous years for solving a variety of practical problems. Further, the effectiveness of data augmentation techniques can be observed within [36], where six different augmentation techniques were applied, and the accuracy of the SSL algorithm on the WESAD data was more than 95%.

In this research, two related SSL approaches within a novel algorithm will be used: the first one based on temporal classification and the second one on contrastive learning. It can be considered that the temporal classification is a pre-step toward contrastive learning. Generally speaking, temporal classification [40] is an SSL methodology specially tailored for time series data. The methodology's basic principle is that the data features vary slowly compared to the sampling time of recorded measurements. The methodology does not require labels because the classification is performed using time intervals (seconds, minutes) generated via a random data preparation procedure. Another advantage of temporal classification is that it transforms complex unsupervised features into simpler classifying segments.

This scientific approach represents one of the first implementation attempts of contrastive learning to stress classification. Contrastive learning is an SSL technique that provides a model's training without requirements for output labels [41], and it is based on learning a similarity metric between data samples. Using a similar technique, SimCLR from [38] reached state-of-the-art accuracy and even matched

some supervised models' performances. In [42], the representation learning method with contrastive predictive coding that applies to different data modalities was presented. The method's predictive coding component relies on training the model to predict the representation in time instance t+1 by using the history of the specific representation until time t. In other words, the model must understand what activity the subject is currently doing to generate future predictions accurately. The efficient methodology from [42] was successfully used to improve a speech recognition algorithm based on SSL representations in [43]. The methodology was utilized on high-frequency sensor data and represented an excellent example for the novel research in this work from the perspective of the existence of similar research environments in both cases. The research from [43] was based on using an encoder network that produces representations that are further mixed by the context network to create a context vector. The vector was finally used to predict the next representation. Another way of solving speech recognition tasks was presented in [44] when an unsupervised learning algorithm was used. It is another proof that these complex kinds of tasks can be successfully solved without supervised learning algorithms. Significant results within [43] and other related papers represented the motivation for the authors of this new research to implement contrastive learning on the WESAD data. However, besides all the advantages of contrastive learning that have been proven experimentally through introduced papers, its implementation remains a challenge because of the potential difficulties with creating the pairs, evaluating the model performance, and implementing and evaluating a proper loss function [45]. In the following sections of this paper, the research attempts to prevent all these difficulties and proposes a novel intelligent SSL solution.

III. Data Preparation and Research Methodology

The initial step of almost every intelligent approach is data exploration and pre-processing. As previously explained, in this research, the WESAD data set was exploited. It was created and maintained by the University of California Irvine and stored within their open-source machine learning repository. WESAD is the multimodal dataset that consists of motion and physiological data recorded from the chest and wrist-worn devices. For this novel research, only the wrist data coming from smartwatch sensors were used. Complete data were collected by recording vital parameters of 15 involved subjects during the study (labeled with S1 to S15, accordingly).

Besides measured parameters, three affective states were also registered during the experiment: neutral, stress, and amusement. The primary deficiencies of the WESAD dataset that should be mentioned are the lack of examined subjects (only 15 participated in the experiment) and a single type of evaluated stress activity. Collecting new stress labels would be expensive from the perspectives of the required time for new laboratory experiments, and the costs of assigning new participants required the recreation of the WESAD experiment and increase the database. Application of SSL techniques can solve this problem and give optimal performances from already available data and provide maximum possible accuracy in stress prediction. Besides described approaches of classifying emotions by recording physiological parameters, an interesting language-independent acoustic emotion classification was presented in [46].

Initial data preparation work in this novel research was based on putting all the sensors on the same timeline (700Hz) and merging all the subject data. The data set was split into train, validation, and test set by following standard machine learning procedures. Further, the data exploration process was performed on the training data composed of subjects S2 to S15. Within 40 million rows within the training data, only 11% (around 4 hours) was collected from the stress state. Such a small percentage of the stress information made this set imbalanced and presented a real challenge to make an accurate system for detection and prediction of it. One possible approach for treating imbalanced data was given in [47], where an intelligent algorithm based on utilizing a genetic algorithm was proposed. Finally, 3% of data was considered invalid.

Keeping in mind the main research task to detect and predict a subject's stress status, it was essential to determine which sensors were the most correlated with stress. After performing a basic correlation analysis, it was concluded that acceleration and electrodermal activities were the most correlated with stress. The next effort was made in data pre-processing, where the outliers were removed and the signals denoised. For the outlier removal process, each sensor was assigned to a valid range of values. The values outside of the specified ranges were deleted and replaced by the closest valid values. Finally, for dealing with the noise components, a low-pass filter was utilized to remove undesirable frequencies. For each sensor value, a cutoff frequency (the highest frequency that is meaningful for a specific sensor) was specified. A Butterworth low-pass filter of the second order was then used with the corresponding cutoff to process the signal.

After a brief introduction of the WESAD data and finishing basic pre-processing tasks, the next research phase was to utilize a novel SSL methodology to a small subset of the data and progressively increased the size of exploited information and the complexity of processing. Considering that WESAD was a representative of time series data, in [48] how an intelligent approach was utilized for treating and classifying time series data was researched. Following the progressive experimental approach, the research task was to train the model on a single subject of data, which corresponded to 1 hour and 30-minute readings from sensors. The features used for processing purposes were wrist acceleration, blood volume pulse, and wrist temperature measurements.

SSL techniques were already successfully applied to processing the WESAD data in [22, 36]. However, the novelty of the new research that will make the novel approach unique was that it was the first attempt to use contrastive learning for the stress prediction tasks. Further, the approach was wholly based on using commercial smartwatch data and making it available for a broad audience. In this research, two previously established SSL methods were implemented: the temporal classification approach from [40] and the contrastive learning from [38]. As in [40], it was important to mention that the features of interest for this novel research also varied slowly (every few minutes) compared to the sampling rates of including sensors (a few milliseconds). The research data was

split into one-minute segments, and the model was trained to classify these segments by their natural belongings. The model was based on three complementary techniques: contrastive learning, the utilization of slow-moving features, and data augmentation techniques. These techniques and the overall algorithm of the model are presented in Fig. 1.

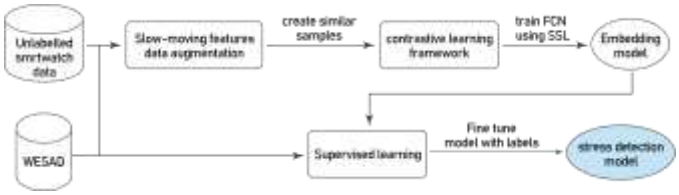

Fig. 1. The Algorithm of the New Model.

IV. Model Design

To set up the temporal classification part of the algorithm, the initial process was divided into three steps: generating one-minute segments, associating a corresponding label to each window, and training the deep learning model to predict the segment from offered windows. A similar methodology was applied in [43], where segments of 12,5 seconds were formed, and a linear classifier was used on top of the representations to predict which stimuli were experienced by a subject. One of the methodology's observed drawbacks was the requirement of creating the segments strictly before their usage within machine learning and deep learning models. It was not possible to add new data later if it was generated during an online learning process. Another deficiency was in terms of data size restrictions and the applied number of segments to prevent losing the speed of a model. The number of segments matched the output of the final layer of the model, implying that the number of segments was directly proportional to the model's size.

To evaluate the initial setup from Fig. 1, a dataset consisting of a single subject's data was processed. A simple neural network with a single hidden layer and 20 neurons were utilized as the supervised learning model (Fig. 2). The network was based on SoftMax activation functions, and its purpose was to classify the labeled data. Finally, the model was trained using stochastic gradient descent with the Adam optimizer, while the cross-entropy was used as the loss function.

Next, Fully Convolutional Network from [25] is implemented as the embedding model from Fig. 1. The network algorithm is adjusted to this specific research case. Three layers within the network are proposed, and the graphical representation of its structure is presented in Fig. 3.

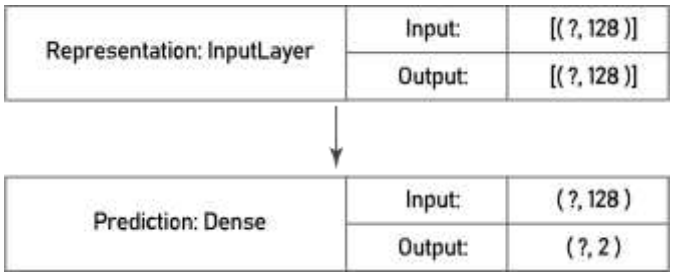

Fig. 2. The Supervised Learning Model Algorithm.

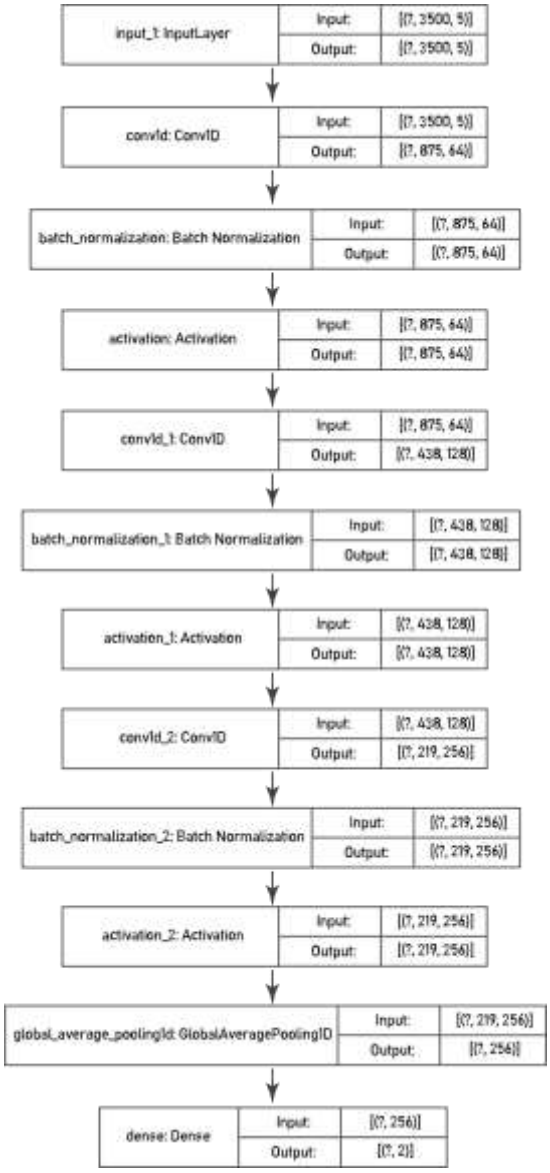

Fig. 3. Fully Convolutional Network (FCN) Architecture.

The contrastive method's key was to create a local classification objective that does not depend on the whole dataset. The task is to classify pairs as similar and different following their temporal closeness. To create similar pairs, the following data preparation algorithm was used. First, the data was split into 5 seconds windows of five 700 Hz sensors. The splitting phase resulted in 17500 elements for each input time point. A pair was created for each window by associating it randomly with another window that occurred less than a minute before. Each such pair was called a positive pair because it was considered temporally close enough to be similar. Further, ten pairs were merged to form a batch of 20 pairs of windows. If two windows belonged to the same positive pair, they were considered similar; otherwise, they were dissimilar. Initiated labels were local because they were created within each batch of data separately.

Once the labels were created, a loss function from [38] was used to estimate how well the model classified the pairs.

TensorFlow 2, as one of the most famous Python libraries, was used for this purpose. The loss function called the Noise Contrastive Estimation loss (NCE loss), and was presented in the following form:

$$L_i = -\log \frac{\exp(z_l(\mathrm{i}) z_r(\mathrm{i})^T)}{\sum_{i \neq j} \exp(z_l(\mathrm{i}) z_r(\mathrm{i})^T)} \quad (1)$$

Where z_l and z_r are the representations obtained from the model of the left and right samples of each pair. To accurately measure the loss, it was essential to evaluate the loss on small subsets of the data individually. When each small sample was selected, the binary cross-entropy was computed using only elements from this subset. Because of the small size of any subset, the cross-entropy computation was fast. To learn from the whole dataset, all subsets were processed one by one. Finally, to get the final loss, the losses from each pair were averaged. The model was trained to utilize this loss function within the gradient descent algorithm on the pairs of previously created batches. Additionally, Adam optimizer was used to update the weights at each pass.

V. Results

The goal of this case study was to recognize stress and neutral states from the input representation data on a single subject's data. The model was trained to learn two-dimensional representations to validate the ability to learn relevant features. The training time of a single model on the machine with Nvidia K80 GPU was 10 minutes approximately. The features that were generated after the training process are presented in Fig. 4. The train set (left) and the test set (right) for Subject 15 is shown with the separation between the different reported activities. For the test set for Subject 15 the stress and meditation results are shown only because the data is split by time into train and test where the first 70% of this subject data goes to train and the rest goes to test. For this subject, the final 30% of data only have these two activities which is a very interesting result considering only a very limited amount of data that been used for this training. Therefore, by using unlabeled data, the model is efficiently capable of clustering different activities as shown.

It is observable from the previous figures that the model efficiently learned to cluster different activities. It learned to separate activities without knowing them in advance, which is the temporal classification paradigm's success. If Fig. 4 is presented from the perspective of only stress and no stress activities, Fig. 5 can be generated as well. It can be concluded from the figure that the model is efficient in separating stress from no stress data. It showed 85% accuracy demonstrating that generated representations can capture most of the relevant information in some simpler cases.

In the final part of this research, the performances of the proposed model were compared to five similar researches by other authors. Table I present information about the advantages and limitations of developed models, as well as exploited types of data during the training processes. Finally, the results and achieved accuracies are shown as the evaluation measures for all the models. Final examination showed that the proposed model provides the accuracy in the range with other state-of-the-art solutions, and an advantage in the term of processing unlabeled data and augmenting existing data. Through this conducted research and the previous one cited within this paper, it can be concluded that the contrastive methods could be very efficient in processing large data sets. It is even possible to parallelize the computations to train such sets, to provide online training, and add additional data within the training process. All these topics will be examined in detail in the future work of the authors.

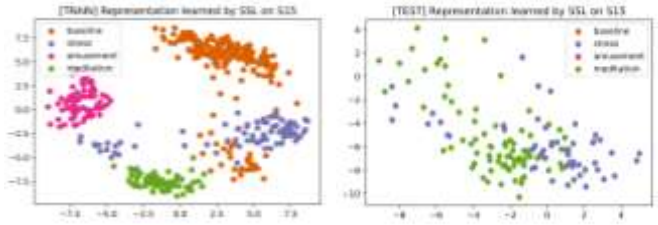

Fig. 4. Data Representations Learned by Contrastive Learning by Examining Subject 15.

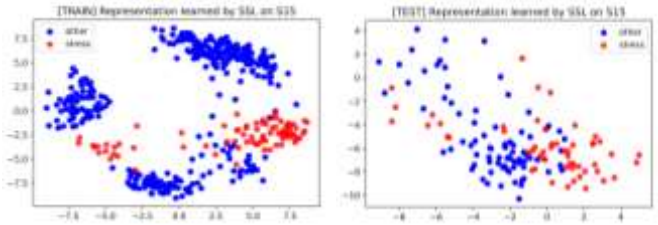

Fig. 5. Representations of Stress and no Stress Features of S15.

TABLE I. Comparison of Six different Approaches for Stress Detection

Reference number	Method	Advantages	Limitations	Exploited data	Results, accuracy
[31]	Feature-based ML approach	Speed, interpretability, low compute power	Required expert knowledge of the features	All the sensors	93% (stress detection)
[32]	ML approach, per subject	Same as [1] + tailored to a single subject	Same as [1] + need to be retrained for each subject	All the sensors	88-99% (per subject)
[33]	ML approach	Same as [1] + applicable to commercial smartwatches	Same as [1]	Smartwatch compatible data: wrist only without EDA	87% (stress detection, balanced accuracy)
[34]	DL approach	High-accuracy, no expert knowledge	High compute, complex model	All sensors	85% (3 classes)
[36]	DL, self-supervised learning	Very high accuracy, no expert knowledge	High compute, complex method	ECG only + additional data sets	97% (4 classes)
New research	DL, self-supervised learning	Can leverage unlabeled data	More complex than supervised learning	WESAD smartwatch data (all wrist data without EDA as in [3])	85% (stress detection)

VI. Conclusion

In this paper, the SSL concept of deep learning was presented and analyzed, and a novel SSL solution was proposed. As a popular case study nowadays, emotional states and stress detection were selected as test cases for this research. In the first section of the paper, a review of popular scientific papers dealing with intelligent techniques for processing emotions was made. It was proven that deep learning and machine learning approaches can threaten emotion data effectively and produce desirable results in the form of a prediction, label detection, classification, or clusterization. This paper's contribution was the utilization of only wrist sensor data (from smartwatches) in the processing phase, without the requirement for any additional data that should be collected by using any intrusive method. State-of-the-art research papers concerning smartwatch sensor data applications were also provided, highlighting the smart approaches for treating such data. Special attention was made to deep learning techniques in the field of emotion recognition and solving similar tasks. It was shown that different supervised and unsupervised learning techniques could be effectively applied for processing physiological data and providing valuable insights. Finally, the WESAD data set, as a base for the case study in this paper, was presented, and the most important research papers were introduced and described. In the second section, a literature review concerning SSL techniques was provided, and the main features were exposed. Special attention was made to the introduction of generative and contrastive SSL algorithms, as they represented the basis for the future model. Additionally, temporal classification as a pre-step toward contrastive learning was also highlighted, as it was an efficient methodology specially tailored for time series data, as was the WESAD data set. In the third section, the nature and features of the data and applied pre-processing techniques were described. All outliers from the data were removed, and important correlations between the features were discovered. A small subset of optimized data (measurement information for a single subject) was finally used to train and test the proposed model in section IV. The model was based on using slow-moving features and data augmentation techniques to increase available data and create similar samples. Then, the contrastive learning framework was used to train the developed network by using the SSL approach. Finally, the embedding model outputs were used within the supervised learning algorithm to provide fine-tuning of the proposed stress detection model.

The novelty of the proposed solution was in utilizing pairs of samples instead of state-of-the-art models that processed a single sample at a time. The SSL algorithm's potential was also shown by the developed ability to cluster human activities without knowing their specifics. The model was able to recognize features very efficiently and abstract concepts, such as meditation, stress, and amusement using only the raw sensor data. Finally, generated representations of the model were evaluated in two ways: the first one using the accuracy metric on pairs of batches and the second one utilizing a shallow supervised model on the top of the representations. In the next steps and future research work in this domain, new functionalities will be added and the updated algorithm utilized on the complete WESAD data set. Novel research should answer if SSL models are capable of processing a large quantity of data and providing accurate stress predictions when multiple subjects are treated at once.

References

[1] Pollreisz, D., and Taheri, N. (2017) A simple algorithm for emotion recognition, using physiological signals of a smart watch. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju Island, Korea, 11-15 July, 2353– 2356, IEEE, New York, USA.

[2] Dao, M., Nguyen, D., Tien, D., and Kasemet, A. (2018) Healthyclassroom-a proof-of-concept study for discovering students' daily moods and classroom emotions to enhance a learning-teaching process using heterogeneous sensors. 7th International Conference on Pattern Recognition Applications and Methods, 16 - 18 Jan, 2018, Funchal, Madeira, Portugal. ISBN 978-989-758-276-9.

[3] Kanjo, E., Younis, E., and Ang, C. (2019) Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. Information Fusion. 2019, 49, 46–56.

[4] Lisetti, C., and Nasoz, F. (2006) Categorizing autonomic nervous system (ANS) emotional signals using bio-sensors for HRI within the MAUI paradigm. 15th IEEE International Symposium on Robot and Human Interactive Communication. 20116, 277–284.

[5] Lisetti, C., and Nasoz, F. (2004) Using non-invasive wearable computers to recognize human emotions from physiological signals. EURASIP journal on applied signal processing. 2004, 1672–1687.

[6] Nie, Y., Wu, Y., Yang, Z., Sun, G., Yang, Y., and Hong, X. (2017) Emotional evaluation based on svm. 2nd International Conference on Automation, Mechanical Control and Computational Engineering.

[7] Guo, H., Huang, Y., Chien, J., and Shieh, J. (2015) Short-term analysis of heart rate variability for emotion recognition via a wearable ecg device. International Conference on Intelligent Informatics and Biomedical Sciences. 262–265.

[8] Jalilifard, A., da Silva, A.G., and Islam, K. (2017) Brain-tv connection: Toward establishing emotional connection with smart tvs. IEEE Region 10 Humanitarian Technology Conference. 2017, 726–729.

[9] Saganowski, S., Dutkowiak, A., Dziadek, A., Dziezyc, M., Komoszyńska, J., Michalska, W., Polak, A., Ujma, M., and Kazienko, P. (2019) Emotion recognition using wearables: A systematic literature review - Work in progress. IEEE International Conference on Pervasive Computing and Communications Workshops. 2019, 1-6.

[10] Zhao, B., Wang, Z., Yu, Z., and Guo, B. (2018) Emotion sense: Emotion recognition based on wearable wristband. IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. 2018, 346–355.

[11] Nalepa, G., Kutt, K., Gizycka, B., Jemioło, P., and Bobek, S. (2019) Analysis and use of the emotional context with wearable devices for games and intelligent assistants. Sensors. 2019, 1-24.

[12] Albraikan, A., Hafidh, B., and El Saddik, A. (2018) iaware: A real-time emotional biofeedback system based on physiological signals. IEEE Access. 2018, 6, 78 780–78 789.

[13] Kamdar, M.R., and Wu, M.J. (2016) Prism: a data-driven platform for monitoring mental health. Biocomputing 2016: The Pacific Symposium.- World Scientific. 2016, 333–344.

[14] Setiawan, F., Khowaja, S.A., Prabono, A.G., Yahya, B.N., and Lee, S. (2018) A framework for real time emotion recognition based on human ansusing pervasive device. IEEE 42nd Annual Computer Software and Applications Conference. 2018, 1, 805–806.

[15] Gupta, R., Khomami Abadi, M., Cárdenes Cabré, J.A., Morreale, F., Falk, T.H., and Sebe, N. (2016) A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. ACM 2016 on international conference on multimedia retrieval. 2016, 317–320.

[16] Maier, M., Marouane, C., and Elsner, D. (2019) Deepflow: Detecting optimal user experience from physiological data using deep neural networks. 18th International Conference on Autonomous Agents and Multi Agent Systems. International Foundation for Autonomous Agents and Multiagent Systems. 2019, 2108–2110.

[17] Ragot, M., Martin, N., Em, S., Pallamin, N., and Diverrez, J.M. (2017) Emotion recognition using physiological signals: laboratory vs. wearable sensors. International Conference on Applied Human Factors and Ergonomics.Springer. 2017, 15–22.

[18] Ojha, V., Griego, D., Kuliga, S., Bielik, M., Buš, P., Schaeben, C., Treyer, L., Standfest, M., Schneider, S., Koenig, R., Donath, D., and Schmitt, G. (2019) Machine Learning Approaches to Understand the Influence of Urban Environments on Human's Physiological Response. Information Sciences. 474, 154–169.

[19] Lu, X., Liu, X., and Bergqvist, E. (2019) It sounds like she is sad: Introducing a biosensing prototype that transforms emotions into real-time music and facilitates social interaction. CHI Conference on Human Factors in Computing Systems. 2019, 1-6.

[20] Babenko, B. (2008) Multiple Instance Learning: Algorithms and Applications. Technical Report, San Diego, USA.

[21] Schmidt, P., Reiss, A., Dürichen, R., and Van Laerhoven, K. (2018) Labelling affective states in the wild: Practical guidelines and lessons learned. ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. 2018, 654–659.

[22] Saeed, A., Salim, F., Ozcelebi, T., and Lukkien, J. (2020) Federated Self-Supervised Learning of Multi-Sensor Representations for Embedded Intelligence. IEEE Internet of Things Journal. Early access, 2020, 1-11.

[23] Ragav, A. (2019) Scalable Deep Learning for Stress and Affect Detection on Resource-Constrained Devices. 18th IEEE International Conference On Machine Learning And Applications. 2019, 1585–1592.

[24] Nakisa, B., Rastgoo, M.N., Rakotonirainy, A., Maire, F., and Chandran, V. (2018) Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. IEEE Access. 6, 49 325–49 338.

[25] Wang, Z., Yan, W., and Oates, T. (2017) Time series classification from scratch with deep neural networks: A strong baseline. International Joint Conference on Neural Networks. 2017, 1578-1585.

[26] Chakraborty, S. (2019) A Multichannel Convolutional Neural Network Architecture for the Detection of the State of Mind Using Physiological Signals from Wearable Devices. Journal of Healthcare Engineering. 2019, 1–17.

[27] Glorot, X., Bordes, A., and Bengio, Y. (2010) Deep Sparse Rectifier Neural Networks. Journal of Machine Learning Research. 15, 315-323.

[28] Ioffe, S., and Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 32nd International Conference on Machine Learning. 37, 448-456.

[29] Taplak, H., Erkaya, S., Yildirim, Ş. (2014) The Use of Neural Network Predictors for Analyzing the Elevator Vibrations. Arabian Journal for Science and Engineering. 39, 1157–1170.

[30] Elkatatny, S. (2019) A Self-Adaptive Artificial Neural Network Technique to Predict Total Organic Carbon (TOC) Based on Well Logs. Arabian Journal for Science and Engineering. 44, 6127–6137.

[31] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. (2018) Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. 20th ACM International Conference on Multimodal Interaction. 400-408.

[32] Indikawati, F., and Winiari, S. (2020) Stress Detection from Multimodal Wearable Sensor Data. IOP Conference Series: Materials Science and Engineering. 1-6.

[33] Siirtola, P. (2019) Continuous stress detection using the sensors of commercial smartwatch. ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 ACM International Symposium. 1198-1201.

[34] Lin, J., Pan, S., Lee, C., and Oviatt, S. (2019) An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals. 28th ACM International Conference on Information and Knowledge Management. 2069-2072.

[35] Ruta, D., and Gabrys, B. (2000) An Overview of Classifier Fusion Methods. Computing and Information Systems. 7, 1-10.

[36] Sarkar, P., Etemad, A. (2020) Self-supervised ECG Representation Learning for Emotion Recognition. IEEE Transactions on Affective Computing. Early access, 1-13.

[37] Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., and Tang J. (2006) Self-supervised Learning: Generative or Contrastive. arXiv:2006.08218. 1-23

[38] Chen, T. (2020) A Simple Framework for Contrastive Learning of Visual Representations. 37th International Conference on Machine Learning. 119, 1597-1607.

[39] Hadsell, R., Chopra, S., and Lecun, Y. (2006) Dimensionality Reduction by Learning an Invariant Mapping. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1735-1742.

[40] Hyvarinen, A., and Morioka, H. (2016) Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. 30th International Conference on Neural Information Processing Systems. 3772–3780.

[41] Kreuk, F., Keshet, J., and Adi, Y. (2010) Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation. Electrical Engineering and Systems Science : Audio and Speech Processing. Early access, 1-5.

[42] Oord, A., Li, Y., and Vinyals, O. (2018) Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748.

[43] Gutmann, M., and Hyvärinen, A. (2010) Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Journal of Machine Learning Research - Proceedings Track. 9, 297-304.

[44] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019) wav2vec: Unsupervised Pre-Training for Speech Recognition. Interspeech 2019, 3465-3469.

[45] Alshamrani, M. (2021) IoT and artificial intelligence implementations for remote healthcare monitoring systems: A survey, Journal of King Saud University - Computer and Information Sciences, 2021, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2021.06.005.

[46] Singh, R., Puri, H., and Aggarwal, N. (2020) An Efficient Language-Independent Acoustic Emotion Classification System. Arabian Journal for Science and Engineering. 45, 3111–3121.

[47] Jiang, K., Lu, J., and Xia, K. (2016) A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE. Arabian Journal for Science and Engineering. 41, 3255–3266.

[48] Dempster, A., Petitjean, F., and Webb, G. (2020) ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery. 1-42.