IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 April 2023, accepted 4 May 2023, date of publication 8 May 2023, date of current version 14 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274478

RESEARCH ARTICLE

Multi-Class Stress Detection Through Heart Rate Variability: A Deep Neural Network Based Study

JON ANDREAS MORTENSEN[1], MARTIN EFREMOV MOLLOV[1], AYAN CHATTERJEE[1,2], DEBASISH GHOSE[1,3], (Senior Member, IEEE), AND FRANK Y. LI[1]

[1]Department of Information and Communication Technology, University of Agder, N-4898 Grimstad, Norway
[2]Department of Holistic Systems, Simula Metropolitan Center for Digital Engineering, N-0167 Oslo, Norway
[3]School of Economics, Innovation, and Technology, Kristiania University College, N-5022 Bergen, Norway

Corresponding author: Debasish Ghose (debasish.ghose@kristiania.no)

This work was supported by the Research Council of Norway through the Orchestrating Internet of Things and Machine Learning for Early Risk Detection to Ensure Inpatients Safety (StaySafe) Program under Grant 309257.

ABSTRACT Stress is a natural human reaction to demands or pressure, usually when perceived as harmful or/and toxic. When stress becomes constantly overwhelmed and prolonged, it increases the risk of mental health and physiological uneasiness. Furthermore, chronic stress raises the likelihood of mental health plagues such as anxiety, depression, and sleep disorder. Although measuring stress using physiological parameters such as heart rate variability (HRV) is a common approach, how to achieve ultra-high accuracy based on HRV measurements remains as a challenging task. HRV is not equivalent to heart rate. While heart rate is the average value of heartbeats per minute, HRV represents the variation of the time interval between successive heartbeats. The HRV measurements are related to the variance of RR intervals which stand for the time between successive R peaks. In this study, we investigate the role of HRV features as stress detection bio-markers and develop a machine learning-based model for multi-class stress detection. More specifically, a convolution neural network (CNN) based model is developed to detect multi-class stress, namely, *no stress, interruption stress, and time pressure stress*, based on both time- and frequency-domain features of HRV. Validated through a publicly available dataset, SWELL−KW, the achieved accuracy score of our model has reached 99.9% ($Precision = 1, Recall = 1, F1-score = 1$, and $MCC = 0.99$), thus outperforming the existing methods in the literature. In addition, this study demonstrates the effectiveness of essential HRV features for stress detection using a feature extraction technique, i.e., analysis of variance.

INDEX TERMS Stress detection, heart rate variability, convolution neural network, feature extraction.

I. INTRODUCTION

Physical or mental imbalances caused by noxious stimuli trigger stress to maintain homeostasis. Under chronic stress, the sympathetic nervous system becomes overactive, leading to physical, psychological, and behavioral abnormalities [1]. Stress levels are often measured using subjective methods to extract perceptions of stress. Stress level measurement based on collected heart rate viability (HRV) data can help to remove the presence of stress by observing its effects on the autonomic nervous system (ANS) [2].

Typically, people with anxiety disorders have chronically lower resting HRV compared with healthy people. As revealed in [2] and [3], HRV increases with relaxation and decreases with stress. Indeed, HRV is usually higher when a heart is beating slowly and vice versa. Therefore, heart rate and HRV generally have an inverse relationship [2], [3]. HRV varies over time based on activity levels and the amount of work-related stress.

Furthermore, stress is usually associated with a negative notion of a person and is considered to be a subjective feeling of human beings that might affect emotional and physical well-being. It is described as a psychological and biological reaction to internal or external stressors [4], including a biological or chemical agent and environmental stimulation that

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

induce stress in an organism [5]. On a molecular scale, stress impacts the ANS [6], which uses sympathetic and parasympathetic components to regulate the cardiovascular system. The sympathetic component in a human body [7] works analogously to a car's gas pedal. It activates the fight-or-flight response, giving the body a boost of energy to respond to negative influences. In contrast, the parasympathetic component is the brake for a body. It stimulates the body's *rest and digests* reaction by relaxing the body when a threat has passed. Given the fact that the ANS regulates the mental stress level of a human being, physiological measurements such as electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), HRV, heart rate, blood pressure, breath frequency, and respiration rate can be used to assess mental stress [8].

ECG signals are commonly adopted to extract HRV [9]. HRV is defined as the variation across intervals between consecutive regular RR intervals,[1] and it is measured by determining the length between two successive heartbeat peaks from an ECG reading. Conventionally, HRV has been accepted as a term to describe variations of both instantaneous heart rate and RR intervals [12].

Obtaining HRV from ECG readings requires clinical settings and specialized technical knowledge for data interpretation. Thanks to the recent technological advances on the Internet of medical things (IoMT) [17], it is possible to deploy a commercially available wearable or non-wearable IoMT devices to monitor and record heart rate measurements.

Based on ECG data analysis (or HRV features, various machine learning (ML) and deep learning (DL) algorithms have been developed in recent years for stress prediction [20], [21], [22], [23], [24], [25], [26], [27] (see more details in Sec. II). Among the publicly available datasets for stress detection, SWELL−KW developed in [13] and [14] one of the two most popular ones. However, none of the existing ML and DL studies based on the SWELL−KW dataset for multi-class stress classification have achieved ultra-high accuracy, especially for multi-class stress level classification [15], [16]. Therefore, there exists a research gap on developing novel ML models which are able to achieve ultra-high accurate prediction.

Motivated by various existing applied ML and DL based studies on HRV feature processing for stress level classifications, we have designed and developed a one-dimensional convolutional neural network (1D CNN) model for multi-class stress classification and demonstrate its superiority over the state-of-the-art models based on the SWELL-KW dataset in term of prediction accuracy. More specifically, we have performed studies on stress detection using both traditional machine learning algorithms and/or multi-layer perceptron (MLP) algorithms which are inspired from the fully connected neural network (FCNN) architecture. In our work, we have developed a 1D CNN model which is based on the convolution operation. CNN reduces number of training parameters as MLP takes vector as input and CNN takes tensor as input so that CNN can understand spatial relation.

While the accuracy achieved with full features is nearly 100%, we have also introduced a feature reduction algorithm based on *analysis of variance (ANOVA)* F-test and demonstrate that it is possible to achieve an accuracy score of 96.5% with less than half of the features that are available in the SWELL−KW dataset. Such a feature extraction reduces the computational load during the model training phase.

In a nutshell, the novelty and the main contributions of this study are summarized as follows:

- We have developed a novel 1D CNN model to detect multi-class stress status with outstanding performance, achieving 99.9% accuracy with a *Precision, F1-score*, and *Recall* score of 1.0 respectively and a *Matthews correlation coefficient (MCC)* score of 99.9%. We believe this is the first study that achieves such a high score of accuracy for multi-class stress classification.
- Furthermore, we reveal that not all 34 HRV features are necessary to accurately classify multi-class stress. We have performed feature optimization to select an optimized feature set to train a 1D CNN classifier, achieving a performance score that beats the existing classification models based on the SWELL-KW dataset.
- Our model with selected top-ranked HRV features does not require resource-intensive computation and it achieves also excellent accuracy without sacrificing critical information.

The remainder of the paper is organized as follows. After summarizing related work and pointing out the distinction between our work and the existing work in Sec. II, we introduce briefly the framework for stress status classification, dataset, and data preprocessing in Sec. III. Then the developed CNN model is presented in Sec. IV. Afterwards, Sec. V defines the performance metrics to evaluate the proposed classifier and Sec. VI presents the numerical results. Further discussions are provided in Sec. VII. Finally, the paper is concluded in Sec. VIII.

II. RELATED WORK

The related work considered in this study covers HRV data quality and various state-of-the-art ML/DL algorithms developed for stress detection.

For HRV data quality, a detailed review on data received from ECG and IoMT devices such as Elite HRV, H7, Polar, and Motorola Droid can be found in [18]. 23 studies indicated minor errors when comparing the HRV values obtained from commercially available IoMT devices with ECG instrument-based measurements. In practice, such a small-scale error in HRV measurements is reasonable, as getting HRVs using portable IoMT devices is more practical, cost-effective, and no laboratory/clinical equipment is required [18], [19].

[1] An RR internal represents the time from an R-peak to the next R-peak [10]. It defines the time elapsed between two successive R-waves of the Q-wave, R-wave and S-wave (QRS) signal on the electrocardiogram [11].

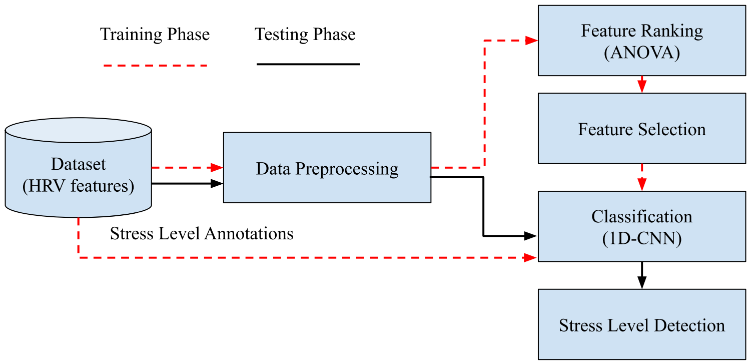


FIGURE 1. Framework of the proposed stress status classification model: From data collection to stress level classification.

On the other hand, there have been a lot of recent research efforts on ECG data analysis to classify stress through ML and DL algorithms [20], [21], [22], [23]. Existing algorithms have focused mainly on binary (stress versus non-stress) and multi-class stress classifications. For instance, the authors in [4] classified HRV data into stressed and normal physiological states. The authors compared different ML approaches for classifying stress, such as naive Bayes, k-nearest neighbour (KNN), support vector machine (SVM), MLP, random forest, and gradient boosting. The best recall score they achieved was 80%. A similar comparison study was performed in [27], where the authors showed that SVM with radial basis function (RBF) provided an accuracy score of 83.33% and 66.66% respectively, using the time-domain and frequency-domain features of HRV. Moreover, dimension reduction techniques have been applied to select best temporal and frequency domain features in HRV [24]. Binary classification, i.e., stressed versus not stressed, was performed using CNN in [25] through which the authors achieved an accuracy score of 98.4%. Another study, StressClick [26], employed a random forest algorithm to classify stressed versus not stressed based on mouse-click events, i.e., the gaze-click pattern collected from the commercial computer webcam and mouse.

In [14], tasks for multi-class stress classification (e.g., no stress, interruption stress, and time pressure stress) were performed using SVM based on the SWELL−KW dataset. The highest accuracy they achieved was 90%. Furthermore, another publicly available dataset, WESAD, was used in [27] for multi-class (amusement versus baseline versus stress) and binary (stress versus non-stress) classifications. In their investigations, ML algorithms achieved accuracy scores up to 81.65% for three-class categorization. The authors also checked the performance of deep learning algorithms, where they achieved an accuracy level of 84.32% for three-class stress classification. Furthermore, it is worth mentioning that novel deep learning techniques, such as genetic deep learning convolutional neural networks (GDCNNs) [38], [39], have appeared as a powerful tool for two-dimensional data classification tasks. To apply GDCNN to 1D data, however, comprehensive modifications or adaptations are required and such a topic is beyond the scope of this paper.

As summarized in Tab. 5 of [15], in a fresh study published online in August 2022, the best results for stress detection based on the SWELL−KW dataset for the single-dataset models developed therein are 88.64% (Accuracy), 93.01% (Precision), 92.68% (Recall), and 82.75% (F1-scores) respectively. Compared with these state-of-the-art models, the model developed in this study has achieved much better performance (see more details in Subsec. VI-F especially Tab. 3 of this paper).

III. FRAMEWORK OVERVIEW AND DATA PREPROCESSING

In this section, we give an overview about the framework for multi-class stress classification. While the overview and model preparation (including data collection, dataset, and data preprocessing) are outlined in this section, the CNN model itself is presented in the next section.

A. FRAMEWORK OVERVIEW

Fig. 1 illustrates the schematic diagram of the proposed stress level classification framework. Briefly, the framework constitutes the following procedures.

- Data collection and datasets. HRV signals are collected and separated into a training dataset and a testing dataset. They will use to define the model's architecture and to assess the proposed model's effectiveness.
- Data preprocessing and feature extraction. Data are preprocessed to fit into the feature ranking algorithm. In this study, ANOVA F-tests [28] and forward sequential feature selection are employed for feature ranking and selection respectively.
- Classification and validation. The designed DL-based multi-class classifier is trained, tested, and validated with significant features and annotations (e.g., *no stress, interruption condition*, and *time pressure*) labeled by medical professionals.
- Testing. In the testing phase, distinctive features are considered from the new test samples, and the class label is resolved using all classification parameters estimated

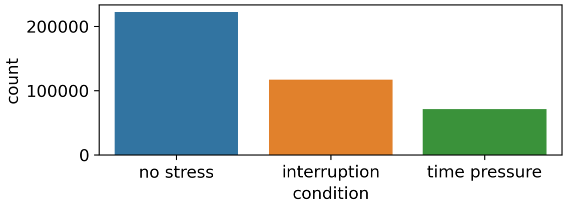


FIGURE 2. Distribution of data in SWELL−KW [13].

in training. Different numbers of features are extracted and tested.

- Performance assessment. The performance of the classifier is measured against discrimination analysis metrics, such as *Accuracy, Precision, Recall, F1-score, and MCC*.

B. DATA COLLECTION AND DATASET

We adopt the SWELL−KW dataset, which was collected in a study reported in [13] and [14]. Various types of data have been recorded, including computer logging, facial expression from camera recordings, body postures from a Kinect 3-dimensional (3D) sensor, heart rate (variability), and skin conductance from body sensors.

In the experiments, 25 volunteers performed typical knowledge tasks (writing reports, making presentations, reading emails, searching for information) during which their psychological and biological status data were recorded. The working conditions of the participants were manipulated with two types of stressors: email interruptions and time pressure. The SWELL−KW dataset comprises HRV computed for stress and user modeling. The subjective experiences of participants with task load, mental effort, mood, and perceived stress were also recorded. Each participant was exposed to three different working environments and the data are then labeled by medical professionals as follows.

- *No stress*: The participants are permitted to work on the activities for as long as they need, up to 45 minutes. However, they are unaware of the maximum duration of the task.
- *Time pressure*: Under time pressure, the time to complete the same job was decreased to 2/3 of its time in the normal condition.
- *Interruption*: The participants were interrupted when they received 8 emails in the middle of a given activity. Some emails were pertinent to their tasks, and the participants were asked to take particular actions, whereas others were totally irrelevant to the ongoing tasks.

The distribution of the collected data with three different stress classes is presented in Fig. 2. The HRV indices were computed by extracting an inter-beat interval (IBI) signal from each participant's peaks of the ECG signals. For each participant, the experiment lasted for approximately 3 hours. From the HRV data, various time-domain and frequency-domain features are extracted, as presented in Tab. 1. Furthermore, we illustrate in Fig. 3 the time-domain features, e.g., time intervals between consecutive heart beats (RR interval) and hear rate of HRV signals. Correspondingly, the frequency-domain features, i.e., the signal power levels with respect to low frequency (LF) and high frequency (HF), are illustrated in Fig. 4. These plots are generated using the first 1000 samples from the SWELL−KW dataset.

TABLE 1. Explanation.

No.	Feature	Meaning
1	MEAN_RR	Mean of RR intervals
2	MEDIAN_RR	Median of RR intervals
3	SDRR	SD of RR intervals
4	RMSSD	Root mean square of successive RR interval differences
5	SDSD	SD of successive RR interval differences
6	SDRR_RMSSD	Ratio of SDRR over RMSSD
7	HR	Heart rate
8	pNN25	Percentage of successive RR intervals that differ more than 25 ms
9	pNN50	Percentage of successive RR intervals that differ more than 50 ms
10	SD1	Measures short-term HRV in ms and correlates with baroreflex sensitivity (BRS)
11	SD2	Measures of long-term HRV in ms and correlates with BRS
12	KURT	Kurtosis of RR intervals
13	SKEW	Skewness of RR intervals
14	MEAN_REL_RR	RR Mean of relative RR intervals
15	MEDIAN_REL_RR	Median of relative RR intervals
16	SDRR_REL_RR	SD of relative RR intervals
17	RMSSD_REL_RR	Square root of the mean of the sum of the squares of the difference between adjacent relative RR intervals
18	SDSD_REL_RR	SD of interval of differences between adjacent relative RR intervals
19	SDRR_RMSSD_REL_RR	Ratio of SDRR_REL over RMSSD_REL
20	KURT_REL_RR	Kurtosis of relative RR intervals
21	SKEW_REL_RR	Skewness of relative RR intervals
22-23	VLF; VLF_PCT	Very low (0.003 Hz - 0.04 Hz) frequency activity of the HRV spectrum
24-26	LF; LF_PCT; LF_NU	Low frequency activity in the 0.04 - 0.15 Hz range
27-29	HF; HF_PCT; HF_NU	High-frequency activity in the 0.15 - 0.40 Hz range
30	TP	Total HRV power spectrum
31	LF_HF	Ratio of low to high frequency
32	HF_LF	Ratio of high to low frequency
33	sampen	Sample entropy of the RR sign
34	higuci	Higuchi Fractal Dimension

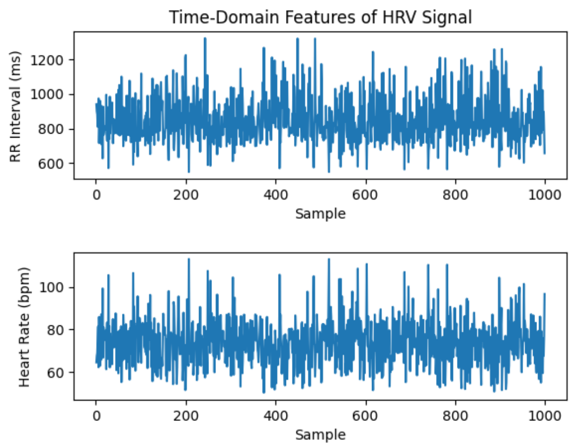


FIGURE 3. Time-domain features of HRV.

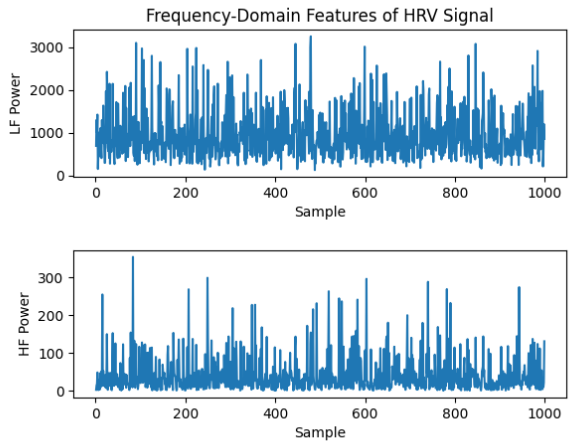


FIGURE 4. Frequency-domain features of HRV.

C. DATA PREPROCESSING

The collected HRV data in the SWELL−KW dataset are time-variant. For classification, we re-construct the HRV data, which was a discrete time series with timestamps, to a series indexed with sequence numbers without timestamps. Moreover, we convert all data into the numerical format. We also remove participants' noisy, incomplete, or missing data. These processing steps result in 25 participant's data with 410322 number of records and 34 number of features for stress level classification.

Moreover, we perform normality tests using methods, such as Shapir–Wilk [29], on each feature of the datasets and the results reveal that the data samples do not look like *Gaussian*. The normality tests are performed following the standard hypothesis testing method with a P-value $\alpha \geq 0.05$ (i.e., sample looks like Gaussian). Further data preprocessing steps are performed as follows.

- Splitting data for training and testing as 80|20 for train|test datasets, respectively;
- Normalization with a standard scalar method to confine the feature values within the range of {0,1}, as some of the selected features were in different magnitudes; and
- Reshaping of each row of the training features into a 1D vector so that it becomes an input to the input layer of the deep learning model.

IV. A CNN MODEL FOR STRESS STATUS CLASSIFICATION

In this section, we present the developed deep learning model for stress status classification. As shown on the right-side hand of Fig. 1, the model consists of feature ranking, feature extraction, and tress level classification.

A. FEATURE RANKING AND EXTRACTION

Firstly, we rank the essential features based on their relevance to the classification task. To do so, the ANOVA [31] F-test is adopted to select the significant features from the SWELL−KW dataset for feature ranking and extraction. ANOVA is a popular tool to perform a parametric statistical hypothesis test that assesses whether the means of two or more data samples (typically three or more) are from the same distribution or not. An F-statistic or F-test is a statistical test method that adopts ANOVA to calculate the ratio between variance values, such as variance from two different samples, or explained and unexplained variance. Furthermore, ANOVA can be used when one variable is numeric, and the other one is categorical, such as when a numerical input data and a classification outcome variable are compared in a classification task.

In this study, we first employ all features for stress classification and then drop the minor significant features based on the importance of features (i.e., feature ranking) before performing the classification task. In the latter case, the training time is shortened while keeping the accuracy of the model.

B. A CNN DL MODEL FOR STRESS CLASSIFICATION

The designed DL model for stress level classification is developed based on the conventional, well-known CNN architectures [32]. CNN is a powerful tool for automatic feature extraction and learning from 1D data sequences. The HRV features of the CNN architecture that are used in our model are illustrated in Tab. 1. For our model design, we retain a reasonable number of neurons in each layer based on the common heuristics (e.g., validation loss, hidden units are a fraction of the input). The CNN kernels slide over the components of the 1D input pattern during convolution.

More specifically, our 1D CNN model consists of an input layer, multiple hidden layers, a max-pooling layer, a flattening layer, and an output layer, as depicted in Fig. 5. The input layer is a 1D convolutional layer, and it consists of 64 filters, a kernel of size 2, and a relative light unit (ReLU) activation function. The ReLU activation function helps to avoid the vanishing gradient so that a faster convergence can be obtained. The 1D max-pooling layer has been introduced to reduce the dimensions of the feature maps. The flattening layer has been adopted to convert the down-sampled data into a 1D vector that acts as an input to the output layer. A softmax activation function has been adopted in the output layer for

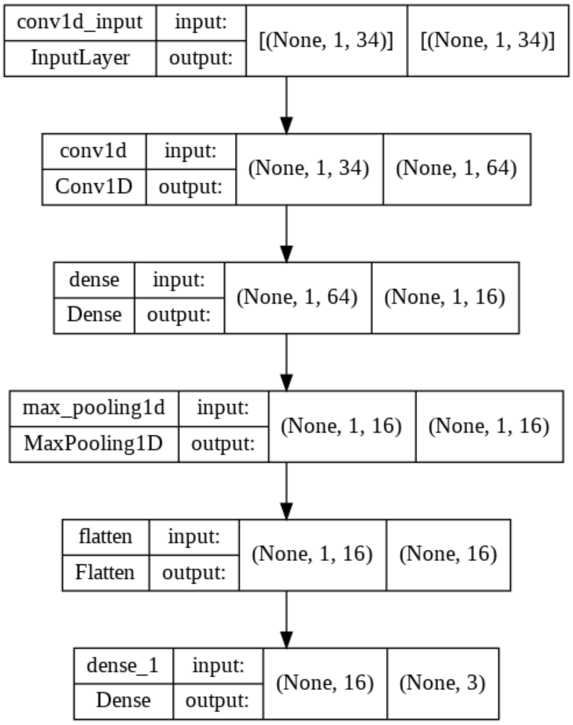


FIGURE 5. The structure of the developed 1D CNN model for stress classification.

multi-class, i.e., no stress, time pressure, and interruption classification based on probability distribution.

For loss calculation, we introduce the *categorical cross-entropy* loss function to compile our 1D CNN model. For model training, we adopt the *adaptive moment estimation (ADAM)* optimizer, as it is computationally efficient and claims less memory. To reduce the learning rate and improve the performance of our model, a validation split step of 0.05 is configured.

As the platform to train and validate the developed model, we rely on Google Colab. Specifically, the model is trained with the default configuration of Google Colab, e.g., Intel(R) Xeon(R) central processing unit (CPU)@2.20 GHz and 12 GB random access memory (RAM). The initial input data shape is (328257, 34). Then the input data is reshaped to (328257, 1, 34) where each row of the input data is formed into a one-dimensional vector. The *Fit()* generator turns training data into many batches, each with a size 64, for training.

V. PERFORMANCE METRICS

The performance of the developed 1D CNN model for multi-class stress classification has been evaluated through discrimination analysis based on the SWELL−KW dataset. The discrimination analysis metrics are *Precision* (eq. (1)), *Recall* (eq. (2)), *Accuracy* (eq. (3)), *F1-score* (eq. (4)), *MCC* (eq. (5)), classification report, and confusion matrix [29], [30]. A confusion matrix is a 2-dimensional table (*actual* versus *predicted*) and both dimensions have four options, namely, *true positives (TP)*, *false positives (FP)*, *true negatives (TN)*, and *false negatives (FN)*.

The cells, or a collection of cells, considered by the ratios for a particular class in multi-class classification are explained as follows [33]. *TP* is an outcome where the model estimates the positive class accurately; *TN* is an outcome in which the model correctly predicts the negative class; *FP* is an outcome where the model estimates the positive class inaccurately; and *FN* is an outcome in which the model forecasts the negative class incorrectly. Accordingly, The performance metrics for a given class are expressed respectively as follows [29].

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1\text{-}score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4}$$

A higher value from the above expressions represents better performance of a model, and this applies to all performance metrics. On the other hand, *bias* is an error due to erroneous assumptions in the learning algorithm, and *variance* is an error from sensitivity to small fluctuations in the training set. While high bias leads to under-fitting, high variance results in overfitting. *Accuracy* and *F1-scores* can be misleading because they do not fully account for the sizes of the four categories of the confusion matrix in the final score calculation. In comparison, the *MCC* is more informative than the *F1-score* and *Accuracy* because it considers the balanced ratios of the four confusion matrix categories (i.e., *TP, TN, FP*, and *FN*). The *F1-score* depends on which class is defined as a positive class. However, *MCC* does not depend on which class is the positive class, and it has an advantage over the *F1-score* as it avoids incorrectly defining the positive class [34]. The *MCC* is expressed as follows [30].

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

VI. CLASSIFICATION RESULTS AND DISCUSSIONS

In this section, we present the experimental results and reveal the importance of ANOVA-based feature selection.

A. FEATURE RANKING AND SELECTION FOR SWELL−KW

In this study, we have considered all 34 features provided by the SWELL−KW dataset. However, some of the features are irrelevant and act as outliers. With this regard, the ANOVA method has been very significant. Initially, it ranks the 34 features based on their F-values. Fig. 6 presents the ranking of the HRV features that are available in the SWELL−KW dataset. Typically, features with higher F-values are more important for final stress level categorization. The most relevant and important subset of

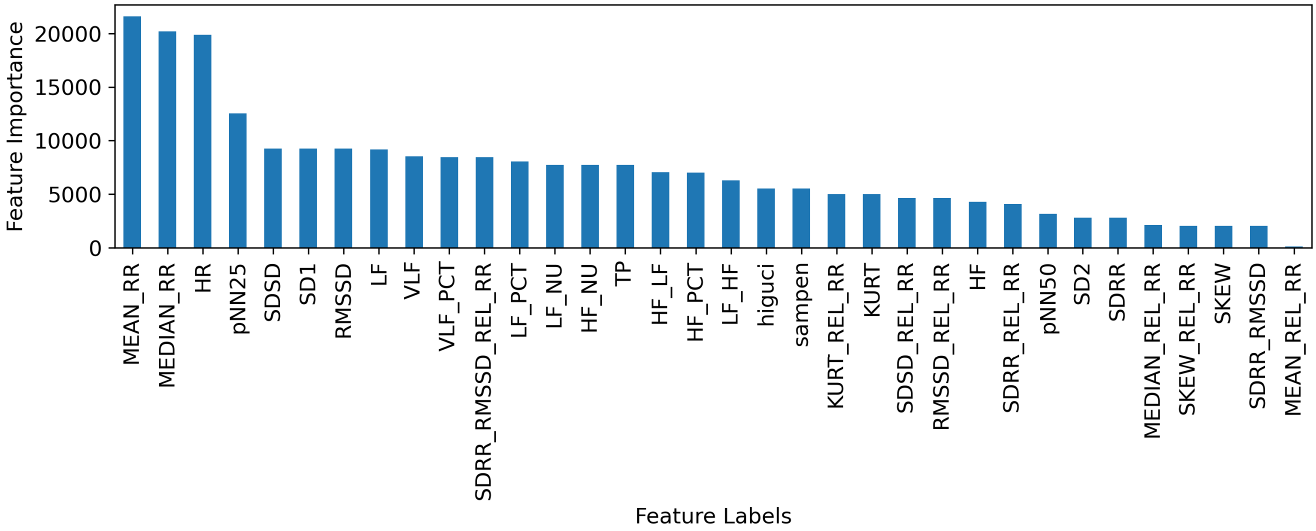


FIGURE 6. Feature ranking of the 34 features using ANOVA.

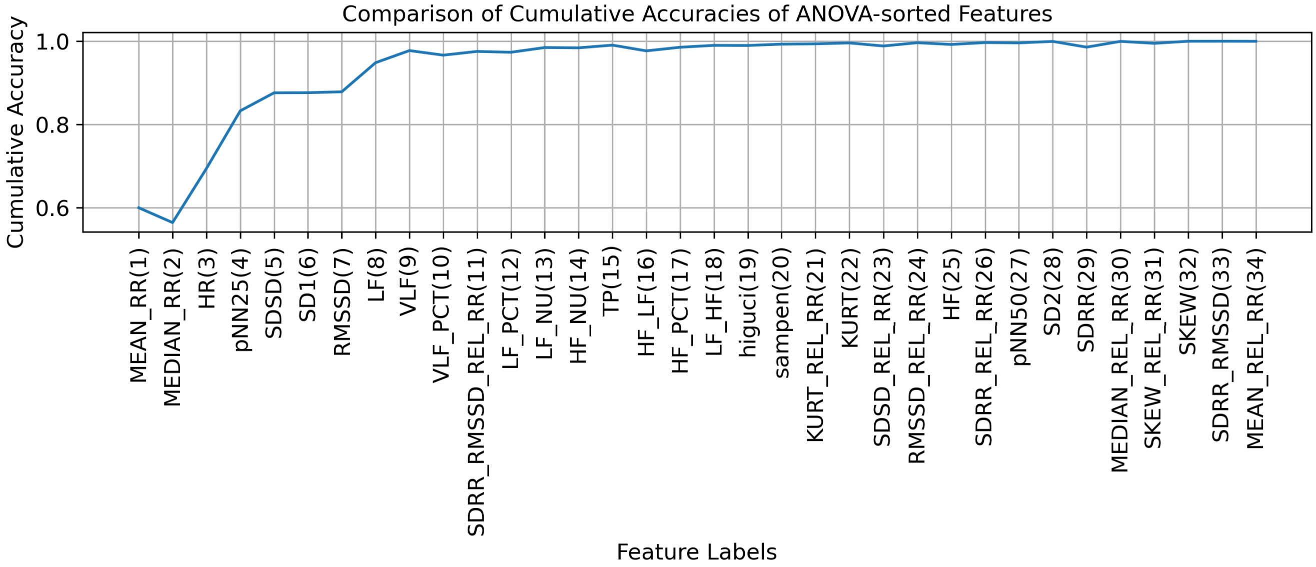


FIGURE 7. Accuracies with ANOVA-sorted features.

the rated features is further identified via a *forward sequential feature selection* method. The forward sequential feature selection forms the optimal subset of features from the 34 features in their ranked order by sequentially selecting the features.

In Fig. 7, we demonstrate the accuracy scores by sequentially selecting the ANOVA-sorted features. It can be observed that accuracy increases with the number of features adopted for model training. More specifically, the developed model achieves above 95% accuracies with less than half of the ANOVA-sorted features, i.e., less than 17 features. In the following two subsections, we first evaluate the performance of our model in terms of *Precision, Recall, F1-score*, and *MCC* when *all available features are applied* to the classifier and then demonstrate the efficacy of the feature reduction algorithm for stress level detection when *the top 15 features are selected.*

TABLE 2. Performance of the proposed 1D CNN model for three level classifications with all features.

Classification Level	Precision	Recall	F1-score
No stress	1.0	1.0	1.0
Time pressure	1.0	1.0	1.0
Interruption	1.0	1.0	1.0

B. PERFORMANCE WHEN ALL FEATURES ARE APPLIED

The developed CNN model has classified the SWELL−KW dataset into the following three stress categories based on emotional states, i.e., *no stress, time pressure*, and *interruption*, and it has obtained an extremely high level of accuracy.

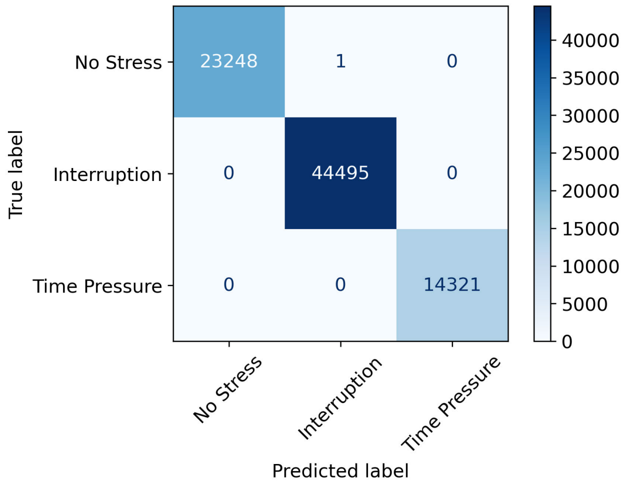


FIGURE 8. Confusion matrix obtained based on stress class classification.

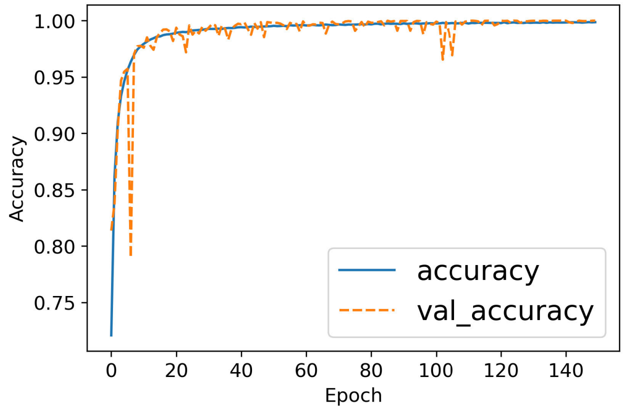


FIGURE 9. Training versus validation accuracy.

More specifically, Tab. 2 demonstrates the performance of the developed 1D CNN model on stress level classifications. Clearly, we have achieved the highest accuracy score of 0.99 with *Precision* $= 1$, *recall* $= 1$, *F1−score* $= 1$, and *MCC* $= 0.99$ respectively. Overall, the accuracy of the developed 1D CNN model reaches an accuracy level of 99.9% for all three classification levels.

Fig. 8 presents the confusion matrix obtained from the developed 1D CNN model based on the SWELL−KW dataset. It is evident from the figure that the proposed classifier correctly predicts the true label with less than 0.01% error for all three classes.

Furthermore, we have verified whether the proposed model is overfitted or not. Fig. 9 illustrates the training versus validation accuracy obtained through our experiments. From this figure, it is clear that the validation accuracy and training accuracy are nearly identical, with the validation loss being slightly higher than the training loss. In other words, the model is not overfitted, and it meets the criteria for a good fit model.

C. PERFORMANCE WITH TOP FIFTEEN FEATURES

We further investigate the performance of the model by employing only the top 15 ANOVA-sorted features, and the obtained results are listed in Tab. 4. Through the values shown in the table, we demonstrate that the average scores for *Precision, Recall, F1-score*, and *MCC* achieved by the proposed model are still excellent, reaching a score of 96.5%, 94.6%, 97.0% and 92.9%, respectively. Overall, we have achieved a score of 96.5% accuracies on average. Furthermore, the performance of the model using a 70/30 train-test split resulted in an accuracy of 0.961, precision of 0.960, recall of 0.956, F1 score of 0.957, and MCC of 0.935.

On the other hand, it is worth reiterating that the performance of our 1D CNN model with all features is extraordinary, outperforming the case with top 15 features. However, such a benefit comes at a cost of a longer training time, specially when the size of a dataset is massive. In general, there is always a trade-off between performance and resource consumption. Therefore, whether to select all features or not depends on the key performance requirements of a system or service. In our experiments, the model training time with 15 features is 1733 seconds, which is 8 seconds less than the model training time with all features.

D. K-FOLD CROSS-VALIDATION

To validate the obtained results with the top 15 features, a k-fold cross-validation procedure has been performed and the results are compared with the ones obtained from the developed 1D CNN model. K-fold cross-validation divides the dataset into k equal-sized folds, training and evaluating the model k times, with each fold serving as the test set once and the remaining k-1 folds serving as the training set. The evaluation scores are then averaged across the k folds to obtain a more robust estimate of the model's performance.

For our validation, the default value, i.e., 5 splits is configured. In each split, the model is trained and evaluated on the test data, and performance metrics in terms of *Precision, Recall, Accuracy, F1 score*, and *MCC* are calculated. The evaluation results based on these five splits show that the model achieves an average score of *Precision* $= 0.944$, *Accuracy* $= 0.945$, *Recall* $= 0.933$, *F1* $= 0.908$, *and MCC* $= 0.908$, obtained based on the same test dataset. As such, it is evident that the developed model is capable of classifying the samples into their respective classes with ultra-high accuracy.

E. HYPERPARAMETER OPTIMIZATION

Initially the model parameters are selected based on experience (as explained in Sec. IV-B). In what follows, we further investigate the impact of hyperparameter optimization on the performance of the developed model, using the *Hyperband Tuning* technique.

Using the top 15 features of the SWELL−KW dataset, hyperband [40] tuning is employed to optimize the hyperparameters of our model. The purpose of the tuning process

TABLE 3. Quantitative comparison of the results with other state-of-the-art models.

Reference	Dataset	Binary/ Multilevel	No. of features	Model	Accuracy	Precision	Recall	F1-score
[24]	SWELL−KW	Binary	17	SVM	92.75%	N.A.	N.A.	N.A.
[25]	SWELL−KW, AMIGOS [5]	3 class	N.A.	CNN	98.30%	96.00%	96.30%	95.80%
[13]	SWELL−KW	3 class	N.A.	SVM	90.00%	N.A.	N.A.	N.A.
[27]	WESAD [36]	Binary/3 class	7	ML/ANN	84.32%/ 95.21%	N.A.	N.A.	78.71%/ 94.24%
[37]	From experiment	Binary/3 class	7	MLP	92.85%/ 64.28%	N.A.	N.A.	N.A.
[15]	SWELL−KW	Binary	34	MLP	88.64%	93.01%	92.68%	82.75%
This study	SWELL−KW	3 class	34	1D-CNN	99.99%	100%	100%	100%
This study	SWELL−KW	3 class	15	1D-CNN	96.50%	94.60%	97.00%	96.00%

TABLE 4. Performance of the proposed 1D CNN model for three level classifications with top 15 ANOVA-sorted features.

Classification Level	Precision	Recall	F1-score
No stress	0.96	0.97	0.97
Time pressure	0.99	0.95	0.97
Interruption	0.89	1.0	0.94

is to maximize the model's validation accuracy. Through the validation procedure illustrated in Appendix A, the best set of hyperparameters is found by the algorithm to be *filters* = *160, kernel size* = *5*, and *dense units* = *48*, resulting in a validation accuracy of 0.99.

On the other hand, it is worth noting that, although hyper-parameter tuning can be effective in improving the performance of ML models, it can be a challenging task to apply it in real-life applications. This is due to its demand for a significant amount of computational resources, especially for large-volume datasets and complex models which may not always be available. Additionally, the optimal set of hyperparameters may be specific to the dataset, model, and the problem at hand, making it difficult to develop a generalizable approach to hyperparameter tuning [41], [42]. Thus, default hyperparameters or a small set of manually tuned hyperparameters may suffice in many cases including this study to achieve satisfactory performance.

F. QUANTITATIVE COMPARISON WITH EXISTING STUDIES

Finally, we make a quantitative comparison of our model versus other related studies appeared in the literature. In Tab. 3, the performance indicators from a few recent studies for automatic classification of stress levels are compared with our 1D CNN model.

Existing studies that are based on publicly accessible datasets such as SWELL−KW, WESAD, and AMIGOS concentrated on binary and multi-class stress detection when assessing the effectiveness of their ML/DL models. It is worth mentioning that we used the SWELL−KW dataset for multi-class stress detection. Regarding performance evaluation, prior studies, e.g., [13] and [24], considered merely the accuracy score as the key performance metric. Although *accuracy* is a popular indicator, it is sufficient only if the false positive and false negative rates are essentially similar, and the dataset is symmetric.

Furthermore, Tab. 3 reveals that, *when all features are considered during model training, none of the existing ML/DL models reported in the literature outperform the one developed in this study* in terms of *Accuracy, Precision, Recall, F1-score*, and *MCC* for categorizing stress levels.

When a subset of features is selected for model training, the model presented in [25] shows higher performance than the proposed model in this study with top 15 ANOVA-sorted features. The reason is that the authors in [25] considered all available features in the datasets, and they did not apply any dimension reduction technique for performance evaluation of their model.

VII. FURTHER DISCUSSIONS

Execution time of full features versus top-15 features: The execution time difference between the all feature-based model and the top-15 feature-based model reported in Subsec. VI-C seems small. There are two reasons for this result. 1) The SWELL−KW dataset which serves as the basis for this study has a moderate amount of data (410322 number of records and 34 features as mentioned in Subsec. III-C) and 2) our training and validation procedures are performed based on Google Colab which has powerful CPUs and graphics processing unit (GPUs) as well as a huge amount of RAMs. When the volume of a dataset becomes huge which is typical for big data processing, or/and the data processing machine is less powerful, e.g., based on a personal computer or a server located at a clinic, the benefit of our model with feature reduction will be more significant, specially for validation. This is because, after the data collection phase, data training can be still performed offline based on powerful CPUs/GPUs.

Model Applicability: The model developed in this study is built based on the SWELL−KW dataset. Nevertheless, we believe that, with proper parameter tuning or enhancement, the model may be applicable to other datasets that target at similar mental health status analysis. Within the framework of an ongoing research project acknowledged below, we are collecting real-life data including HR and RR for mental health inpatients in a Norwegian hospital based on non-wearable Internet of things (IoT) devices. We plan to assess the performance of the developed model based on our own datasets. However, to include the validation results based on these inpatient datasets is beyond the scope of this paper.

VIII. CONCLUDING REMARKS

In this study, we have developed novel a 1D CNN model for stress level classification using HRV signals and validated the proposed model based on a publicly available dataset, SWELL−KW. In our model, we also applied an ANOVA feature selection technique for dimension reduction. Through extensive training and validation, we demonstrate that our model outperforms the state-of-the-art models in terms of major performance metrics, i.e., *Accuracy, Precision, Recall, F1-score*, and *MCC* when all features are employed. Furthermore, our approach with ANOVA feature reduction also achieves excellent performance. For future work, we plan to further investigate the feasibility of optimizing the model to fit it into edge devices so that real-time stress detection can become a reality.

REFERENCES

[1] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Invest.*, vol. 15, no. 3, pp. 235–245, Mar. 2018.

[2] D. Muhajir, F. Mahananto, and N. A. Sani, "Stress level measurements using heart rate variability analysis on Android based application," *Proc. Comput. Sci.*, vol. 197, pp. 189–197, Jan. 2022.

[3] J. Held, A. Vîslă, C. Wolfer, N. Messerli-Bürgy, and C. Flückiger, "Heart rate variability change during a stressful cognitive task in individuals with anxiety and control participants," *BMC Psychol.*, vol. 9, no. 1, p. 44, Mar. 2021.

[4] K. M. Dalmeida and G. L. Masala, "HRV features as viable physiological markers for stress detection using wearable devices," *Sensors*, vol. 21, no. 8, p. 2873, Apr. 2021.

[5] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr./Jun. 2021.

[6] E. Won and Y.-K. Kim, "Stress, the autonomic nervous system, and the immune-kynurenine pathway in the etiology of depression," *Current Neuropharmacol.*, vol. 14, no. 7, pp. 665–673, Aug. 2016.

[7] B. Olshansky, H. N. Sabbah, P. J. Hauptman, and W. S. Colucci, "Parasympathetic nervous system and heart failure: Pathophysiology and potential implications for therapy," *Circulation*, vol. 118, no. 8, pp. 863–871, Aug. 2008.

[8] S. Goel, P. Tomar, and G. Kaur, "ECG feature extraction for stress recognition in automobile drivers," *Electron. J. Biol.*, vol. 12, no. 2, pp. 156–165, Mar. 2016.

[9] V. N. Hegde, R. Deekshit, and P. S. Satyanarayana, "A review on ECG signal processing and HRV analysis," *J. Med. Imag. Health Informat.*, vol. 3, no. 2, pp. 270–279, Jun. 2013.

[10] M. Vollmer, "A robust, simple and reliable measure of heart rate variability using relative RR intervals," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2015, pp. 609–612.

[11] M. H. Kryger, T. Roth, and W. C. Dement, *Principles and Practice of Sleep Medicine*, 5th ed. Amsterdam, The Netherlands: Elsevier, 2011.

[12] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, Mar. 1996.

[13] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The SWELL knowledge work dataset for stress and user modeling research," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 291–298.

[14] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 227–239, Apr. 2018.

[15] M. Albaladejo-González, J. A. Ruipérez-Valiente, and F. G. Mármol, "Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate," *J. Ambient Intell. Human. Comput.*, pp. 1–11, Aug. 2022, doi: 10.1007/s12652-022-04365-z.

[16] R. Walambe, P. Nayak, A. Bhardwaj, and K. Kotecha, "Employing multimodal machine learning for stress detection," *J. Healthcare Eng.*, vol. 2021, Oct. 2021, Art. no. 9356452.

[17] A. Ibaida, A. Abuadbba, and N. Chilamkurti, "Privacy-preserving compression model for efficient IoMT ECG sharing," *Comput. Commun.*, vol. 166, pp. 1–8, Jan. 2021.

[18] W. C. Dobbs, M. V. Fedewa, H. V. MacDonald, C. J. Holmes, Z. S. Cicone, D. J. Plews, and M. R. Esco, "The accuracy of acquiring heart rate variability from portable devices: A systematic review and meta-analysis," *Sports Med.*, vol. 49, no. 3, pp. 417–435, Mar. 2019.

[19] C.-M. Chen, S. Anastasova, K. Zhang, B. G. Rosa, B. P. L. Lo, H. E. Assender, and G.-Z. Yang, "Towards wearable and flexible sensors and circuits integration for stress monitoring," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2208–2215, Aug. 2020.

[20] R. A. Rahman, K. Omar, S. A. M. Noah, M. S. N. M. Danuri, and M. A. Al-Garadi, "Application of machine learning methods in mental health detection: A systematic review," *IEEE Access*, vol. 8, pp. 183952–183964, 2020.

[21] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati, "Application of machine learning methods in mental health detection: A systematic review," in *Proc. Int. Conf. Adv. Comput. Eng. Appl.*, 2015, pp. 714–721.

[22] S. Celin and K. Vasanth, "ECG signal classification using various machine learning techniques," *J. Med. Syst.*, vol. 42, no. 12, p. 241, Oct. 2018.

[23] A. Padha and A. Sahoo, "A parametrized quantum LSTM model for continuous stress monitoring," in *Proc. 9th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2022, pp. 261–266.

[24] S. Sriramprakash, V. D. Prasanna, and O. V. R. Murthy, "Stress detection in working people," *Proc. Comput. Sci.*, vol. 115, pp. 359–366, Dec. 2017.

[25] P. Sarkar and A. Etemad, "Self-supervised learning for ECG-based emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3217–3221.

[26] M. X. Huang, J. Li, G. Ngai, and H. V. Leong, "Stressclick: Sensing stress from gaze-click patterns," in *Proc. 24th ACM Int. Conf. Multimedia (MM)*, Oct. 2016, pp. 1395–1404.

[27] P. Bobade and M. Vani, "Stress detection with machine learning and deep learning using multimodal physiological data," in *Proc. 2nd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2020, pp. 51–57.

[28] B. J. Feir-Walsh and L. E. Toothaker, "An empirical comparison of the ANOVA *F*-test, normal scores test and Kruskal–Wallis test under violation of assumptions," *Educ. Psychol. Meas.*, vol. 34, no. 4, pp. 789–799, Dec. 1974.

[29] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, "Identification of risk factors associated with obesity and overweight—A machine learning overview," *Sensors*, vol. 20, no. 9, art., p. 2734, May 2020.

[30] A. Chatterjee, N. Pahari, A. Prinz, and M. Riegler, "Machine learning and ontology in eCoaching for personalized activity level monitoring and recommendation generation," *Sci. Rep.*, vol. 12, no. 1, pp. 1–26, Nov. 2022.

[31] L. Stahle and S. Wold, "Analysis of variance (ANOVA)," *Chemometrics Intell. Lab. Syst.*, vol. 6, no. 4, pp. 259–272, Nov. 1989.

[32] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107398.

[33] F. Mattioli, C. Porcaro, and G. Baldassarre, "A 1D CNN for high accuracy classification and transfer learning in motor imagery EEG-based brain-computer interface," *J. Neural Eng.*, vol. 18, no. 6, Jan. 2022, Art. no. 066053.

[34] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020.

[35] K. Nkurikiyeyezu, K. Shoji, A. Yokokubo, and G. Lopez, "Thermal comfort and stress recognition in office environment," in *Proc. 12th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2019, pp. 256–263.

[36] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 400–408.

[37] A. Arsalan, M. Majid, A. R. Butt, and S. M. Anwar, "Classification of perceived mental stress using a commercially available EEG headband," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2257–2264, Nov. 2019.

[38] R. G. Babukarthik, V. A. K. Adiga, G. Sambasivam, D. Chandramohan, and J. Amudhavel, "Prediction of COVID-19 using genetic deep learning convolutional neural network (GDCNN)," *IEEE Access*, vol. 8, pp. 177647–177666, 2020.

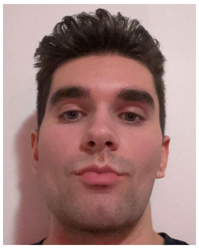
[39] R. G. Babukarthik, D. Chandramohan, D. Tripathi, M. Kumar, and G. Sambasivam, "COVID-19 identification in chest X-ray images using intelligent multi-level classification scenario," *Comput. Electr. Eng.*, vol. 104, Dec. 2022, Art. no. 108405.

[40] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6765–6816, 2017.

[41] M. Feurer, L. Kotthoff, and J. Vanschoren, "Hyperparameter optimization," in *Automated Machine Learning: Methods, Systems, Challenges*. Springer, 2019.

[42] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

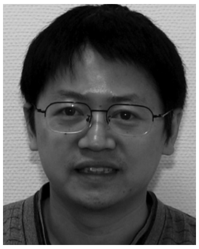
JON ANDREAS MORTENSEN is currently pursuing the bachelor's degree in computer science with the Department of Information and Communication Technology, University of Agder, Norway. His research interests include data analytics and machine learning.



MARTIN EFREMOV MOLLOV is currently pursuing the bachelor's degree in computer science with the Department of Information and Communication Technology, University of Agder, Norway. His research interests include data analytics and machine learning.

AYAN CHATTERJEE received the B.Eng. degree in computer science and engineering (CSE) from the West Bengal University of Technology, India, in 2009, the master's degree in information technology from Jadavpur University, India, in 2016, and the Ph.D. degree from the University of Agder, Norway, in 2022. His Ph.D. thesis was on ICT-eHealth. He worked as an Associate Consultant with Tata Consultancy Services, Ltd., India, from 2009 to 2019, and was deputed to Denmark and the Netherlands for 3.4 years as a Java Solution Designer and a Data Analyst. He is currently a Senior Researcher in AI and semantics with the Simula Research Laboratory (SimulaMet), Oslo, Norway, and an Adjunct Associate Professor of object-oriented programming with the University of Agder, Kristiansand, Norway. He has a strong aptitude for object-oriented programming concepts. His research interests include AI, eHealth, recommendation technology, semantics, human-centered design, software engineering, and bioinformatics.

DEBASISH GHOSE (Senior Member, IEEE) received the Ph.D. degree in information and communication technology from the University of Agder, Grimstad, Norway, in 2019. He was a System Developer with Confirmit, Grimstad, Norway, from 2020 to 2021. From 2021 to 2022, he was a Post-Doctoral Researcher with the University of Agder. He is currently an Associate Professor with the School of Economics, Innovation, and Technology, Kristiania University College, Bergen, Norway. His research interests include protocol design, modeling, and performance evaluation of the Internet of Things, edge and fog computing, data analytics, cyber security, and machine learning.



FRANK Y. LI received the Ph.D. degree from the Department of Telematics (now the Department of Information Security and Communication Technology), Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2003. He was a Senior Researcher with the UniK-University Graduate Center (now the Department of Technology Systems), University of Oslo, Norway, before joining the Department of Information and Communication Technology, University of Agder, Norway, in 2007, as an Associate Professor and then a Full Professor. From 2017 to 2018, he was a Visiting Professor with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. During the past few years, he has been an active participant in multiple Norwegian and EU research projects. His research interests include MAC mechanisms and routing protocols in 5G and beyond mobile systems and wireless networks, the Internet of Things, mesh and ad-hoc networks, wireless sensor networks, D2D communications, cooperative communications, cognitive radio networks, green wireless communications, dependability and reliability in wireless networks, QoS, resource management, traffic engineering in wired and wireless IP-based networks, and the analysis, simulation, and performance evaluation of communication protocols and networks. He was listed as a Lead Scientist by the European Commission DG RTD Unit A.03—Evaluation and Monitoring of Program in 2007.