Machine Learning-Based Identification of Burnt-out Individuals using ECG data

Arya Sinha - Mohammed Kaif - Amolika Bansal
ML Project Report
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY, DELHI
arya20498@iiitd.ac.in kaif20084@iiitd.ac.in amolika20424@iiitd.ac.in

Abstract

Burnout is a state of emotional, mental, and physical exhaustion caused by prolonged exposure to high levels of stress. The early identification of burnout is crucial in preventing long-term physical and mental health problems. A dataset of 202 individuals with three ECG recordings and Mini-Z 1.0 stress questionnaire responses was collected at the Department of Cardiology in G.B. Pant Hospital in Delhi. This study proposes the use of machine learning algorithms to classify individuals as either burnt-out or not based on their ECG data. The results of this study will help in the early identification and prevention of burnout-related health problems.

Keywords ECG, Mental State, Burn-out, Therapy, Classification, Supervised Learning

1 Introduction

Burnout is a widespread problem in today's fast-paced and demanding society, and it can have severe consequences on an individual's physical and mental health. It is often associated with various health problems, including heart diseases. The electrocardiogram (ECG) is a non-invasive diagnostic tool used to assess the heart's electrical activity and consists of 12 leads - six chest leads and six limb leads - recorded from electrodes on the body's surface. Anomalies in ECG samples have been shown to be predictive of both short- and long-term death. Therefore, early and accurate identification of cardiac ECG irregularities can help prevent burnout-related health problems. However, manual ECG interpretation is time-consuming and requires highly skilled professionals. Therefore, automated prediction of burnout using raw ECG data has been rapidly advancing over the past few years. In this study, we propose the use of machine learning algorithms to classify individuals as either burnt-out or not based on their ECG data. The goal is to develop a model that can identify burnout early so that the necessary interventions can be initiated to prevent long-term negative consequences. The study will utilize a dataset of 202 individuals with three ECG recordings and Mini-Z 1.0 stress questionnaire responses collected at the Department of Cardiology in G.B. Pant Hospital in Delhi. The results of this study could have significant implications for the early detection and prevention of burnout-related health problems.

1.1 Literature review

- The traditional processing pipeline for ECG signals has been subject to several improvements. One such advancement was proposed by Mashrur et al. (2019) [3], who utilized the continuous wavelet transform (CWT) to convert single-lead ECG data into RGB images. This was sent as an input to AlexNet for analysis. In a similar vein, Jun et al. (2018) [2] transformed ECG data into a 2D greyscale image and used this image as an input for classification with AlexNet and VGGNet.
- A research study created an automatic system for detecting mental stress using ECG signals and analyzed with ML classifiers, including decision tree, random forest, Naive Bayes. The most effective model had a `94.1%` accuracy rate[1].
- Another study by Zhang et al. (2014) proposed a classification method that utilized kernel SVM and Genetic Algorithm (GA). The approach had three primary modules: lead-fall detection, feature

extraction, and classification. The method achieved a true positive detection rate of 92% with 5.68% false positives and 94% classification accuracy using a training dataset with 1000 records [4].

- Zhou et al. (2021) conducted research in three phases. In the first phase, they extracted HRV features from ECG signals. In the second phase, they used Gaussian mixture model (GMM) based on HRV features to determine mental state. To reduce the stress classification coefficient (SCC), they applied a clustering approach to process HRV features. They used a Conventional Neural Network to extract features and combined manual and automatic features before passing them to SVM. The average recognition rate was around 95% [5].

1.2 Objective

- **Feature extraction:** Extracting relevant features from the preprocessed ECG data that can help differentiate between burnt-out and non-burnt-out individuals. Some possible features could be heart rate variability, QT interval, P wave duration, and T wave amplitude.
- **Building a classification model:** Developing a machine learning model that can classify individuals into burnt-out and non-burnt-out categories based on their extracted ECG features. This could involve experimenting with different classification algorithms, hyperparameters, and model architectures to achieve the best possible accuracy.
- **Model evaluation and validation:** Evaluating the performance of the developed model using suitable metrics such as accuracy, precision, recall, and F1-score, and validating it on an independent dataset to ensure its generalizability.

1.3 Scope

The identification of burnt-out individuals, using Machine Learning techniques on ECG data, has the potential to open up a new channel for the early detection and intervention of burnout, which is a serious problem in today's society. Machine learning algorithms can be trained to recognise patterns in ECG signals and detect changes related to burnout. ECG data can reveal information on the autonomic nervous system's activity, which is in charge of regulating the body's stress response. Machine learning algorithms can detect changes in heart rate variability, which is a sign of stress, by examining ECG data.
The application of machine learning-based identification of burnt-out individuals using ECG data has a broad scope. It can be employed in a variety of industries, including healthcare, education, and corporate settings. Technology in healthcare can be used to detect burnout in healthcare workers, lowering the risk of medical errors and increasing patient outcomes. Technology can be utilized in the education industry to identify burnt-out pupils and provide timely treatments to improve their mental and emotional wellbeing. In the business environment, technology can be utilized to recognise employee burnout, lowering absenteeism and increasing productivity.

1.4 Impact

The results of this study could have significant implications for the early detection and prevention of burnout-related health problems.
The diagnosis of burnt-out individuals using machine learning techniques on ECG data has the potential to change the way we address burnout in our society. This device can enable early identification and intervention for people who are at risk of burnout, enhancing their mental and emotional well-being and averting severe burnout. Detecting and managing burnout early on can result in increased productivity and lower healthcare expenses in the long run. This technology can improve people's quality of life by preventing health problems and boosting their general well-being.
In conclusion, this can have a tremendous positive impact on our society, increasing individuals' mental and emotional health and enhancing productivity and overall well-being.

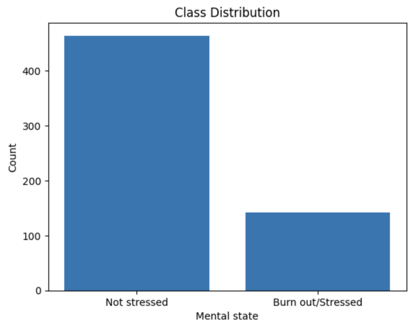
2 Materials

2.1 Dataset

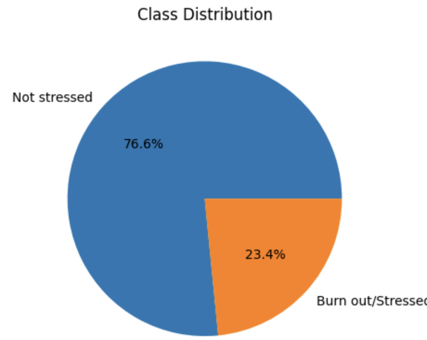
The dataset has been collected from a study consisting of 202 participants. Subjects visited the center three times. The Mini-Z 1.0 stress questionnaire was performed on the subject to determine their mental state. At the same time, their electrocardiogram (ECG) was recorded. The data contains a total of 606 12-lead ECG recordings. In the analysis, the satisfied group is represented by class label 0. In contrast, the burnout class is represented by class label 1.

2.2 Exploratory Data Analysis

The following figure illustrates the heavily imbalanced nature of the dataset.



(a) Bar Graph



(b) Pie Chart

Figure 1: Class Distribution

I Lateral	aVR	V1 Septal	V4 Anterior
II Inferior	aVL Lateral	V2 Septal	V5 Lateral
III Inferior	aVF Inferior	V3 Anterior	V6 Lateral

Figure 2: 12 ECG Leads

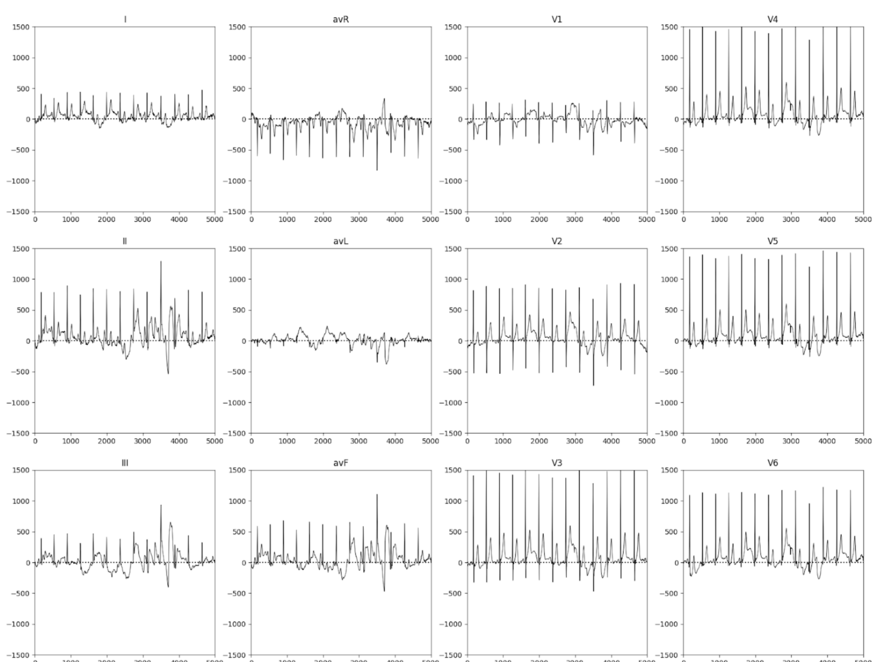


Figure 3: Visualization of ECG Signal

3 Methodology

Figure 4: Work Flow

3.1 Pre-processing

Preprocess the dataset to prepare it for analysis. This involves cleaning the data, removing outliers, and normalizing the ECG readings.
The ecg_clean function from NeuroKit2 has been used to preprocess the ECG data. It performs several preprocessing steps to denoise and remove artifacts from the ECG signal. It removes baseline wander, high-frequency noise, powerline interference, and QRS complexes.

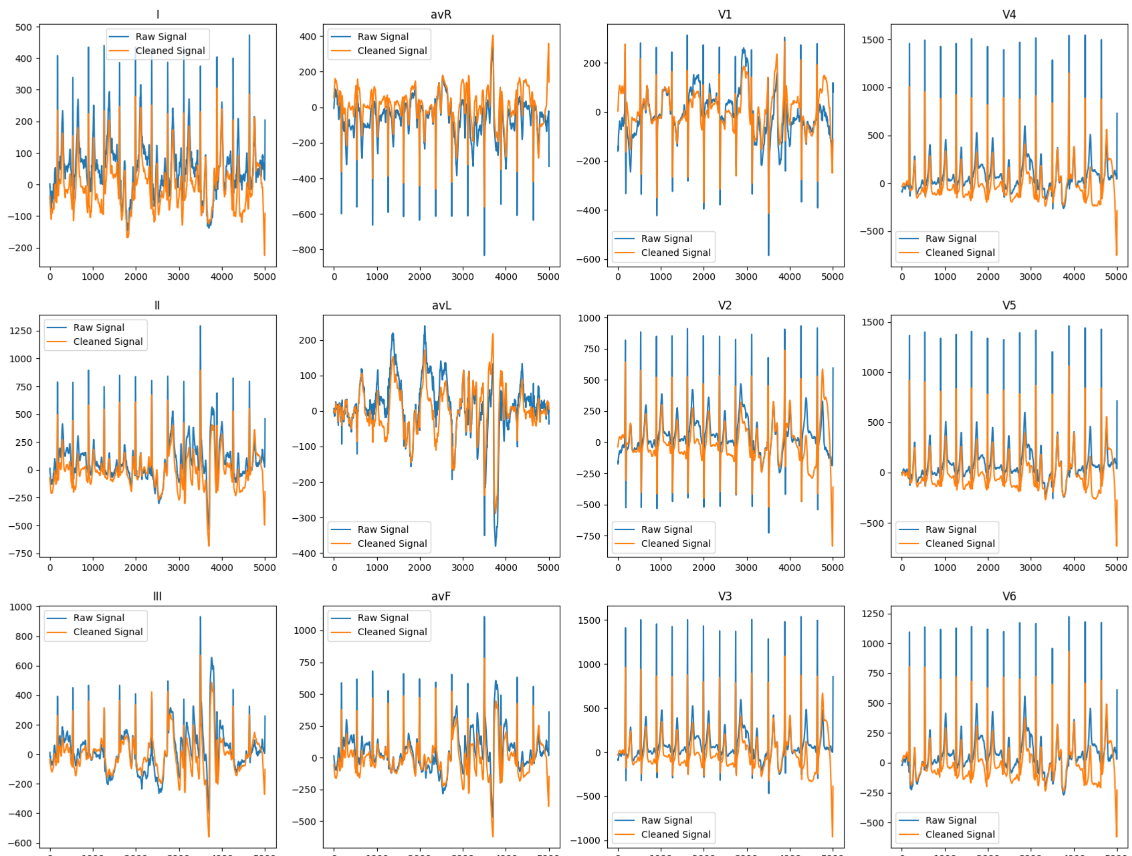


Figure 5: Data Preprocessing

3.2 HRV features extraction

HRV is the variability in the time intervals between consecutive heartbeats, and it is an important indicator of the health of the cardiovascular system. We first clean the ECG data using the Pan-Tompkins algorithm, which is a commonly used method for QRS detection in ECG signals. It then calculates the RR intervals, which are the time intervals between successive R peaks in the ECG signal. Then we apply several HRV analysis techniques using the HRV analysis package (hrvanalysis). The techniques used include time-domain features, frequency-domain features, Poincaré plot features, CSI/CVI features, and geometrical features. The resulting features are stored in a pandas DataFrame and returned. These HRV features are extracted from each lead data of each participant and stored in a dataset. We obtain a total 210 HRV features for each ECG recording ranging from 'II_mean_nni':'V6_triangular_index'

3.3 Data augmentation using Synthetic Minority Oversampling Technique (SMOTE)

Data augmentation has been performed using SMOTE (Synthetic Minority Over-sampling Technique) to handle the imbalance in our dataset. Since our dataset is highly imbalanced, the machine learning model tends to be biased towards the majority class. SMOTE helps to address this problem by generating synthetic samples of the minority class.

3.4 Classification Models

- K-Nearest Neighbours (KNN)
- Decision Tree
- Random Forest
- rbf kernel SVM
- Naive Bayes

3.5 Clustering/Label Re-assignment

Although the Mini-Z 1.0 stress questionnaire is a popular tool used to assess stress levels in individuals, the labels obtained through this questionnaire may not be reliable or trustworthy. This could be due to various reasons such as the subjective nature of stress perception, individual differences in stress response, or the presence of confounding factors that affect stress levels. This issue of fuzzy labels is common in many real-world datasets[1]. This can lead to inaccurate assessments of stress levels and can compromise the decision-making based on the labeled data. Thus, in order to achieve higher consistency in the data, various clustering algorithms such as K-means, Spectral Clustering, Gaussian Mixture Model have been used to refine the labels assigned to the data.

K-Means

# of Clusters	# of PCA features	Silhoutte	calinski_harabasz	davies_bouldin
2	1	0.5692	1178.5388	0.5825
	2	0.4325	612.6506	0.8700
3	1	**0.5724**	**902.1909**	**0.5951**
	2	0.4083	377.5647	0.9565

Gaussian mixture clustering

# of Clusters	# of PCA features	Silhoutte	calinski_harabasz	davies_bouldin
2	1	0.5689	1177.8759	0.5825
	2	0.3926	473.7932	0.9617
3	1	0.5726	938.5771	0.5896
	2	0.3293	247.0012	1.0721

Spectral clustering

# of Clusters	# of PCA features	Silhoutte	calinski_harabasz	davies_bouldin
2	1	0.5692	1178.5388	0.5825
	2	0.4293	21.1821	0.5562
3	1	0.5037	590.5525	0.5354
	2	0.2101	54.1010	1.1356

Figure 6: Evaluating different clustering models

Among other clustering algorithms, K-means clustering performs best with 3 clusters when PCA reduces the number of dimensions to 1. Two clusters out of these are assigned 0 and 1 depending on their frequency correlation with the original labels; the third cluster has been removed as an outlier. Such inferences can be drawn since the labels have an accuracy of 60% with their original labels.

We used K-Means clustering on PCA dimension reduced data to form a new dataset. We clustered the fuzzy data into 3 clusters and chose the target labels similar to the original labels.

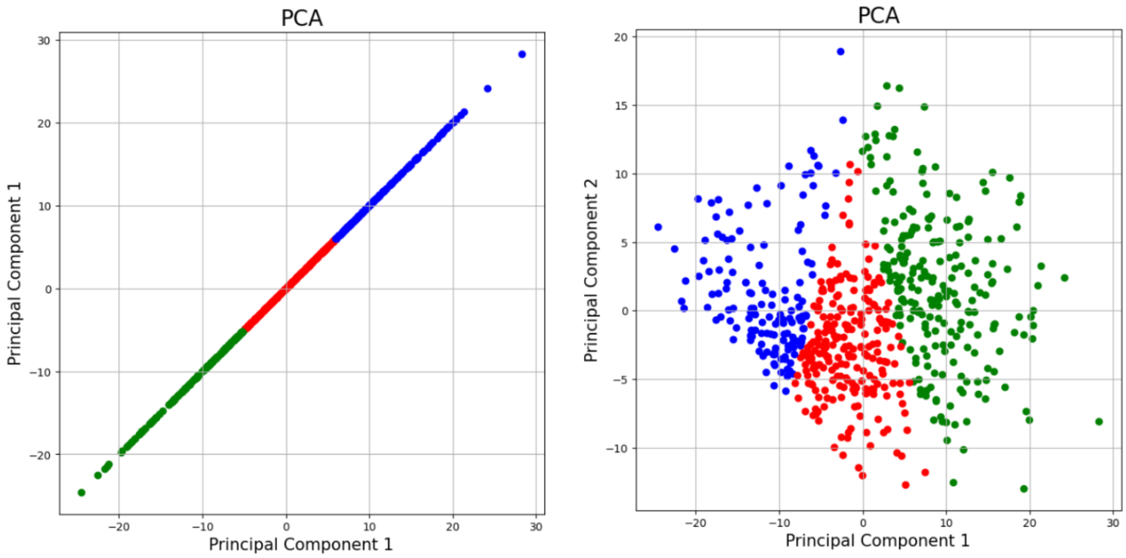
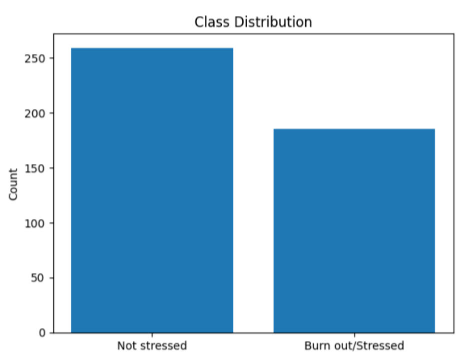
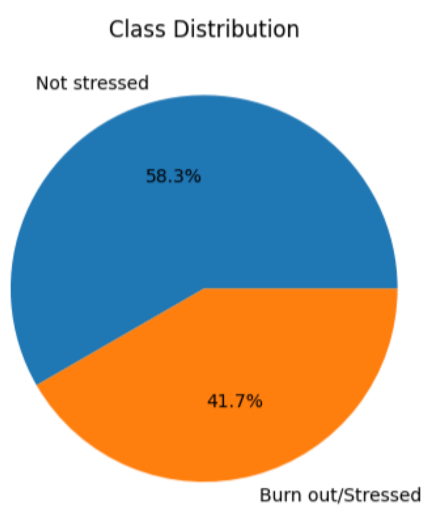


Figure 7: Clustering using K-Means



(a) Bar Graph



(b) Pie Chart

Figure 8: Class Distribution

3.6 Novelty

Our aim is to classify burnt-out individuals based on ECG data, which has not been extensively explored in previous research. While several studies have investigated the relationship between stress and cardiovascular disease, few have used ECG data to identify burnout specifically. Also, the dataset that we will be working on is a real world dataset and thus no previous work has been done using this. The data contains two groups, group 1 is performing Satyam meditation while group 2 is performing control mediation. We can inspect the differences in stress reduction caused by the two techniques.

3.7 Real World Application

Real World Applications include using wearable ECG devices for remote health monitoring, automating ECG interpretation, integrating the model into electronic health record systems, using the predictions for insurance underwriting and providing targeted insurance policies or wellness programmes, and contributing to the expanding body of research on burnout and its related issues.

3.8 Evaluation Metric

We will be computing the **Confusion Matrix** - a table that shows the number of true positives, true negatives, false positives, and false negatives.

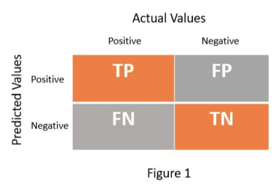


Figure 9: Confusion Matrix

From the confusion matrix - Accuracy, Recall and Precision is evaluated using the following equations:

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

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

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

F1 score: the harmonic mean of precision and recall. It provides a balanced measure of both precision and recall.

ROC AUC: the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate. It measures the ability of the model to distinguish between positive and negative samples.

4 Results And Discussion

In our project, we employed various Machine Learning models such as K-nearest neighbors, Decision Tree, Random forest, Support Vector Machine and Naive bayes classifier to detect the presence of mental stress in an individual using their ECG data. For the purpose of feature extraction, Heart Rate Variability features were extracted. The issue of fuzzy labels was handled using various clustering algorithms. Further, hyperparamter tuning has been performed using GridSearchCV. Following are the results obtained.

4.1 Discussion

After evaluating the above mentioned models, it was seen that SVM with Radial Basis Function kernel performs well and obtains good results when evaluated through stratified cross fold validation as well.

ML method	Accuracy	AUC	F1 score	Precision	Recall
K-nearest neighbors(k=2)	0.705	0.549	0.289	0.333	0.255
Decision Tree	0.645	0.517	0.268	0.26	0.276
Random forest	0.76	0.533	0.518	0.428	0.063
RBF kernel SVM	0.72	0.982	0.522	0.304	0.148
Naive bayes	0.6	0.554	0.554	0.285	0.468

Figure 10: Model selection (stratified test train split, 33% test set, 90% PCA) using Original labels

ML method	Accuracy	AUC	F1 score	Precision	Recall
K-nearest neighbors(k=21)	0.97	0.97	0.971	0.944	1.0
Decision Tree	0.98	0.982	0.98	1.0	0.96
Random forest	**0.982**	**0.982**	**0.982**	**1.0**	**0.964**
RBF kernel SVM	**0.982**	**0.982**	**0.982**	**0.976**	**0.988**
Naive bayes	0.877	0.956	0.877	0.847	0.917

Figure 11: Using Cluster labels

Folds	Accuracy	AUC	F1 score	Precision	Recall
Fold 1	0.986	0.986	0.986	1.0	0.986
Fold 2	0.971	0.971	0.971	0.971	0.971
Fold 3	0.986	0.985	0.986	0.972	0.985
Fold 4	0.986	0.986	0.986	1.0	0.986
Fold 5	0.986	0.985	0.985	1.0	0.985
Average	**0.983**	**0.983**	**0.989**	**0.977**	**0.983**

Figure 12: Stratified Cross fold validation for RBF kernel SVM

5 Conclusion

Working with real-world data was a big learning experience. Handling imbalance data, solving the issue of fuzzy labels, etc while trying to model the data has strengthened our ML knowledge. At first, we faced a lot of problems trying to achieve a good accuracy score since the model used to learn the majority class only. We looked into numerous data augmentation techniques to solve this problem and finally went with SMOTE. Additionally, the presence of fuzzy labels also introduced inconsistencies in our data. We solved this using semi-supervised learning methodologies. We also experimented with using tsfresh to extract time series features, converting 1D ecg signal to image data and classifying using CNN as well as classifying the spectrograms of the ECG data.

6 Distribution of work

- **Arya Sinha** - Data pre-processing, SVM, CNN, Clustering
- **Mohammed Kaif** - Data pre-processing, KNN, SVM, Naive Bayes, Presentation
- **Amolika Bansal** - EDA, CNN, Report Formation

References

[1] Md Belal Bin Heyat, Faijan Akhtar, Syed Jafar Abbas, Mohammed Al-Sarem, Abdulrahman Alqarafi, Antony Stalin, Rashid Abbasi, Abdullah Y. Muaad, Dakun Lai, and Kaishun Wu. Wearable flexible electronics based cardiac electrode for researcher mental stress detection system using machine learning models on single lead electrocardiogram signal. *Biosensors*, 12:427, 06 2022.

[2] Tae Joon Jun, Hoang Minh Nguyen, Daeyoun Kang, Dohyeun Kim, Daeyoung Kim, and Young-Hak Kim. Ecg arrhythmia classification using a 2-d convolutional neural network. *arXiv:1804.06812 [cs]*, 04 2018.

[3] Fazla Rabbi Mashrur, Amit Dutta Roy, and Dabasish Kumar Saha. Automatic identification of arrhythmia from ecg using alexnet convolutional neural network. *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, 12 2019.

[4] Ya-tao Zhang, Cheng-yu Liu, Shou-shui Wei, Chang-zhi Wei, and Fei-fei Liu. Ecg quality assessment based on a kernel support vector machine and genetic algorithm with a feature matrix. *Journal of Zhejiang University SCIENCE C*, 15:564–573, 07 2014.

[5] Ruishi Zhou, Chenshuo Wang, Pengfei Zhang, Xianxiang Chen, Lidong Du, Peng Wang, Zhan Zhao, Mingyan Du, and Zhen Fang. Ecg-based biometric under different psychological stress states. *Computer Methods and Programs in Biomedicine*, 202:106005, 04 2021.