Exploring Stress among Nurses with Data-driven Insights: A Statistical & Machine Learning Approach

Tanvir Hasan
M.Sc in Computer Science & Engineering
University of Oulu
thasan23@student.oulu.fi
2309496

Qiqi Xie
M.Sc in Business Analytics
University of Oulu
qiqi.xie@student.oulu.fi
2308561

Shafin Alam
M.Sc in Business Analytics
University of Oulu
shalam23@student.oulu.fi
2308917

S. M. Seam Uz Zaman
M.Sc in Business Analytics
University of Oulu
szaman23@student.oulu.fi
2305245

Prof. Pekka Siirtola
University of Oulu
pekka.siirtola@oulu.fi

***Abstract*— This study explores the complex correlation between stress levels among nurses and their physiological reactions, with the goal of deepening our comprehension of this crucial matter and enhancing real-time stress prediction in healthcare environments. Our study examines the relationship between electrodermal activity (EDA), heart rate (HR), skin temperature (TEMP), and self-reported stress levels in 15 female nurses with a total of 1 core datapoints. The dataset consists of biometric signals collected through Empatica E4 wearable devices. Our methodology utilizes correlation analysis and machine learning techniques to investigate important scientific inquiries, such as identifying physiological markers of stress and constructing reliable predictive models. Our study results indicate mild to moderate positive associations between stress levels and skin temperature. In contrast, electrodermal activity shows a very weak positive link, while heart rate displays a weak negative correlation. To improve the reliability of the model on imbalanced datasets, we investigated different cross-validation strategies, such as Leave-one-group-out and K-fold cross-validation and train test addition, we suggest a new method for creating features, which includes dividing the data into time windows and expanding the features. This leads to notable enhancements in the accuracy of stress prediction models. The accuracy of our refined model has increased from 59% to 68% when applied to imbalanced data. This highlights the usefulness of our approach in overcoming the difficulties of assessing stress levels among nurses.**

Keywords— Stress level, Healthcare worker, Leave-one-group-out cross-validation, Correlation Analysis

I. INTRODUCTION

In recent years, the emergence of wearable technologies has given an unparalleled chance to continually monitor many physiological indicators, bringing in a new age in healthcare research. One significant area of focus in this paradigm change is the detection and management of stress, a pervasive concern, particularly among healthcare professionals functioning in high-stakes contexts. As developments in technology permit the constant tracking of physiological markers, understanding the subtle link between these variables and the feeling of stress becomes crucial. Our study presents a detailed exploration into the physiological symptoms of stress among nurses, leveraging a unique multi-modal sensor dataset acquired during the tough circumstances of the COVID-19 pandemic. Nurses, serving as the front lines in the struggle against the COVID-19 epidemic, encountered extraordinary hardships, facing heightened stress levels due to the intricacies of their roles, exposure to the virus, and the stressful work environment. The worldwide healthcare community understands the importance of managing stress among healthcare professionals, not only for the well-being of individuals but also for the maintenance of optimal healthcare standards. The wearable sensor dataset under evaluation was methodically gathered during the height of the epidemic, focused on 15 female nurses working in a hospital, documenting the intricacies of their physiological responses to stress during this vital period.

Our interest in exploring stress among nurses is inspired by numerous compelling considerations. Firstly, the well-being of healthcare personnel significantly affects patient outcomes and overall healthcare system performance. Stress among nurses not only leads to personal suffering but also limits their capacity to give safe and effective care to patients. Understanding the causes contributing to nurse stress and creating appropriate solutions is therefore vital for guaranteeing both nurse and patient welfare. Secondly, traditional methods of assessing stress, such as self-report questionnaires, have limits in capturing real-time and objective measures of stress experienced by individuals. Wearable sensor technology offers a new avenue for gathering biometric data, providing insights into the physiological reactions linked with stress among nursing workers. By exploiting this technology, our research aspires to move beyond subjective assessments of stress and towards a more comprehensive understanding of its biological manifestations. Our dataset encompassing biometric information such as electrodermal activity (EDA), heart rate (HR), skin temperature, and inter-beat intervals (IBI), gives a unique opportunity to examine the physiological correlates of stress among nurses. By studying these physiological characteristics in relation to reported stress levels, we intend to find patterns suggestive of stress and identify potential biomarkers for stress detection.

Using wearable sensor technologies and cutting-edge analytical methods, we examine the association between

physiological indicators and reported stress levels among nurses in this study. We intend to perform statistical studies to investigate the relationships between reported stress levels among nurses and indicators including skin temperature, heart rate, and electrodermal activity (EDA). Additionally, using biometric information gathered from wearable sensors, we will create machine learning models to forecast stress levels. Our initiative aims to offer insights that can guide the creation of tailored treatments targeted at reducing stress and enhancing nurse well-being by evaluating the data and identifying the major causes of nursing stress. Our goal is to address nursing stress in healthcare settings to improve patient care results.

II. Related Work

The prediction of stress and its association with data from wearable devices has garnered considerable attention in healthcare research, notably in the realm of comprehending the stress levels encountered by nurses. This section provides a comprehensive analysis of the pertinent literature concerning the anticipation of stress levels in nurses, the anticipation of stress levels using wearable devices, and the relationship between stress levels and data collected from wearable devices.

A. Nurse Stress Prediction and Management: Enhancing Healthcare

Research into stress prediction among nurses using wearable devices leverages physiological markers like EDA, HR, and skin temperature to provide insights into stress levels in real-time. Eom et al., [1] conducted a study on nurses to find precise stress recognition from bio signals. They created SIM-CNN, a self-supervised learning (SSL) method for tailored stress identification models. Their findings suggested that SIM-CNN outperformed SVM and CNN models, reaching an AUC of 60.96%, an F1 score of 75.22%, and an accuracy of 79.65%. Similarly, a pilot study by Li et al. [2] demonstrates the utility of HRV analysis with wearable ECG devices in assessing workplace stress among nurses, suggesting HRV as a valuable marker for stress detection. The critical need to monitor and manage stress among healthcare professionals is underscored by its direct impact on both their well-being and the quality of care provided to patients. This concern has been particularly magnified during the COVID-19 pandemic, as healthcare workers on the frontlines have faced unprecedented levels of stress, emotional exhaustion, and depersonalization. The implementation of adaptive defense mechanisms and resilience strategies is vital for mitigating these effects and ensuring the mental health and well-being of healthcare professionals [3]. Moreover, the stress experienced by healthcare workers during the pandemic has had profound implications on their personal and professional environments, calling for a comprehensive understanding of the dynamics influencing psychological distress. A model developed by researchers at the Research Institute of the McGill University Health Centre and McGill University proposes a nuanced approach to addressing these dynamics by emphasizing the importance of support and coping strategies in modifying the effect of stressors [4]. In a global context, such as in Nigeria, the situation mirrors the challenges faced by healthcare professionals worldwide, where high levels of occupational stress have been linked to several adverse outcomes, including the hindrance of healthcare advancement. Addressing this issue requires collaborative efforts from policymakers, healthcare institutions, and stakeholders to prioritize the well-being and productivity of healthcare professionals [5].

B. Stress Prediction from Wearable Devices

A stress evaluation methodology using heart rate variability (HRV) and electrocardiography (ECG) signals was described in Palanisamy et al. [6]. Stress-induced fluctuations were tracked using ECG signals by using the Stroop color word test as a stressor. This resulted in a classification accuracy of 79.17%. With high classification accuracies of 94.58% and 94.22% using different wavelet functions, the study demonstrated the efficacy of discrete wavelet transform-based feature extraction and helped to advance the creation of a reliable physiological signal-based stress detection system. In another study authors analyze data from ten participants in the 2013 Hajj pilgrimage to suggest a unique wearable sensor-based technique to evaluate the influence of stress on pilgrims' sleep patterns. Their method uses physical and physiological indicators to gauge stress levels, and it has been shown to be 73% accurate in differentiating between low, moderate, and high levels of perceived stress [7]. Another study by Liapis et al. [8] uses skin conductance data and subjective valence-arousal evaluations to examine gender differences in stress recognition during human-computer interaction (HCI). Utilizing data from five HCI activities performed by 31 volunteers, the study uses a pre-experiment interview to detect stressful circumstances before dividing the data according to gender. Although they lack additional physiological signals like our study, they achieve high-stress recognition accuracy by using Linear Discriminant Analysis (LDA); the mean accuracy for males and females is 94.8% and 98.9%, respectively. Similarly, by using EDA data by Zubair et al. [9], where they created a wearable smart band for healthcare that uses the Internet of Things to identify mental stress by measuring skin conductance. Even though their study's accuracy rate was a high 91.66%, it lacked comprehensive reporting on other performance parameters like precision, recall, AUC, and ROC. Moreover, the lack of feature selection techniques, less number of participants and data impedes their machine learning model's reproducibility and overall comprehension. Moreover, in response to the increased need for devices that can track people's physical and mental health, Sandulescu et al., [10] suggest a machine learning-based method for stress detection utilizing wearable physiological sensors. Their technique successfully categorizes the states of the subjects into "stressful" and "non-stressful" circumstances. The majority of participants showed encouraging results, with accuracy levels above 82% and precision levels over 80%. The study does, however, find a bias in the classification of non-stressful situations as stressed, which may be related to brief stress spikes during the shift from stressful to neutral jobs. Islam and Washington [11] utilized a self-supervised learning (SSL) approach for individualized stress detection, demonstrating the potential of machine learning models to learn complex patterns in bio-signal data, thereby predicting stress more effectively. Their method underscores the importance of personalized machine learning models to accommodate individual differences in physiological responses to stress. Similar research [4; 12] has shown that machine learning techniques, including deep learning and feature extraction from HRV and EDA signals, can be

effectively employed for stress detection. These studies highlight the capability of ML algorithms to analyze intricate physiological signals and identify stress with high precision, emphasizing the relevance of these technologies in real-time stress monitoring.

Our project seeks to extend these efforts by developing machine learning models specifically tailored to the unique stressors encountered by nurses during the pandemic. By integrating data on pandemic-specific stress factors with biometric signals, our models aim to forecast stress levels with even higher accuracy and specificity. This approach not only builds upon the foundational work but also addresses the unique challenges and stressors presented by the pandemic environment. Through this targeted approach, we aspire to offer more effective tools for stress management among nurses, contributing to better mental health outcomes and enhanced patient care during these challenging times.

III. Objectives

The main goal of our project is to acquire a more profound comprehension of stress among nurses and its correlation with physiological variables. To accomplish this overarching objective, our purpose is to address the following research questions.

- How do physiological variables, such as electrodermal activity, heart rate, and skin temperature, correlate with reported stress levels among nurses?
- To improve stress prediction model reliability in real-time healthcare settings.
- To create stress level predictive models while having an imbalance dataset using machine learning techniques.

The expected results of our research encompass a thorough comprehension of the association between physiological variables and stress levels among nurses, a proficient machine learning model for stress prediction, and a precise identification of the primary factors that contribute to stress among nurses. These findings will not only increase the present knowledge base but also have concrete ramifications for boosting the well-being of nurses and the grade of care they give. By understanding the elements that lead to stress, remedies can be devised to decrease these causes, ultimately reducing stress levels among nurses and improving patient care. By creating a robust machine learning model for stress prediction, we can create a tool that can be used in real time to monitor stress levels, allowing for rapid remedies.

IV. Data

The dataset used in this study is from Hosseini et al.'s [13] work, which was initially gathered from a group of female nurses who worked regular shifts at a hospital. Data was methodically collected over two study periods in April-May and November-December of 2020, which lasted roughly one week. The digitally structured dataset is arranged into folders that correspond to individual participants, so enabling a methodical study. Many biometric signals obtained with Empatica E4 wearable devices are included in each participant's dataset. These signals include vital physiological parameters such as skin temperature (TEMP), heart rate (HR), electrodermal activity (EDA), blood volume pulse (BVP), interbeat interval (IBI), and accelerometer data (ACC). The CSV-formatted data shows each row as a data point, with different variables represented by different columns. Figure 1 represents the frequency distribution of the overall physiological data of our study. Apart from the physiological information obtained by means of wearable technology, the dataset also includes survey responses that are kept in an Excel file. This additional feature gives important qualitative insights into the stress levels that nurses describe experiencing as well as contextual details about the type of stress that they are experiencing. In addition to self-reported stress levels, each participant's survey replies provide input about the stresses they faced during their work shifts. These answers encompass a broad spectrum of pressures, such as worries about COVID-19, dealing with patients and coworkers, having a heavy job, technological problems, and environmental stressors.

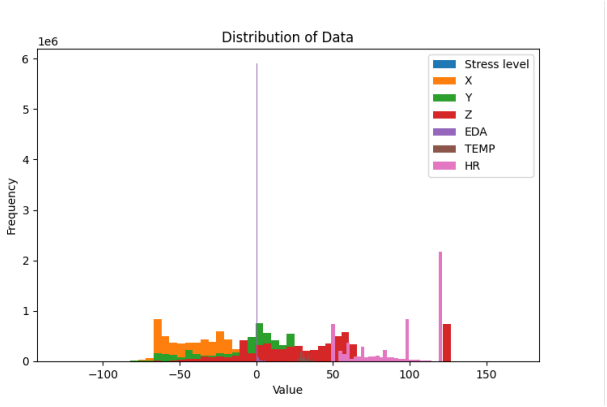

Fig. 1. Frequency distribution of the data

A. Pre-processing

To get the data ready for analysis, there are multiple processes in the data preparation pipeline. The given Figure 2 reflects the pre-processing steps of our study. Initially, we started with the data merging. Data merging was the most challenging part of our study. By extracting all files, including nested zip files, by unzipping them all recursively from a given primary path. Making sure all pertinent data is available for additional processing is the goal of this stage. The code then processes and merges CSV files for each signal type (ACC, EDA, HR, and TEMP) using a shared identifier. This stage makes it easier to analyze the data by combining it into a single file for each type of signal. The method then formats survey data that is kept in an Excel file, adding more columns and producing second timestamps for every row. This makes it possible to analyze the survey data more thoroughly using timestamped entries. After processing, the survey data is stored in a CSV file for later use. Lastly, the code creates a single CSV file with all of the nurse timestamp aggregated data, combining the data from several sources into a single dataset for thorough analysis. The foundation for upcoming analytic and modelling tasks is laid by this data pretreatment pipeline, which guarantees that the data is correctly prepared and consolidated. When the data is merged into a one file we move to data cleaning steps. Finding and addressing missing values in the data frame—which determines the number of missing values for every column—is the first step in the data cleaning process. Following that, the Z-score approach with threshold level 3 is used to identify outliers in numerical columns. Furthermore, measures are included to guarantee uniformity in the 'id' column, such as eliminating leading or trailing white spaces and transforming the IDs into a standard case style. The data frame is then cleared of rows that lack IDs. Because they assist remove

irregularities and guarantee the integrity and dependability of the dataset, these cleaning procedures are essential for getting the data ready for analysis and modelling.

This procedure revealed that the dataset included outliers and missing values, especially in numerical columns, which, if left unchecked, could distort the findings of the analysis. Furthermore, mistakes in data interpretation and modelling could result from discrepancies in the 'id' column. During data merging, issues included timestamp and ID alignment, which required recurrent changes to achieve accurate merging. In order to resolve these problems, outliers were eliminated to stop them from influencing the results of the study, and missing data were either imputed or discarded from rows containing missing IDs. Biased analysis results and erroneous modelling results are expected effects of data problems. Unreliable insights could result from statistical measures and relationships between variables being distorted by missing values and outliers. Inconsistencies in the 'id' column may lead to inaccurate data linkages and make it more difficult to precisely track individual data pieces. Thus, to guarantee the validity and dependability of ensuing studies and conclusions, these faults must be addressed through data cleaning and preprocessing.

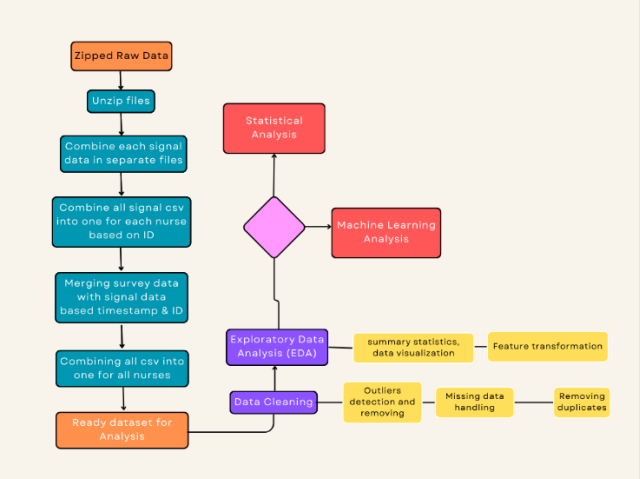

Fig. 2. Data preprocessing steps

B. Dataset

Figure 3 presents the stress level distribution group by the participants' ID. We observe that most of the nurses went through high stress (stress level 2) during the covid, followed by stress level 0 and stress level 1. Only nurse ID 5C experienced all levels of stress in different timestamps.

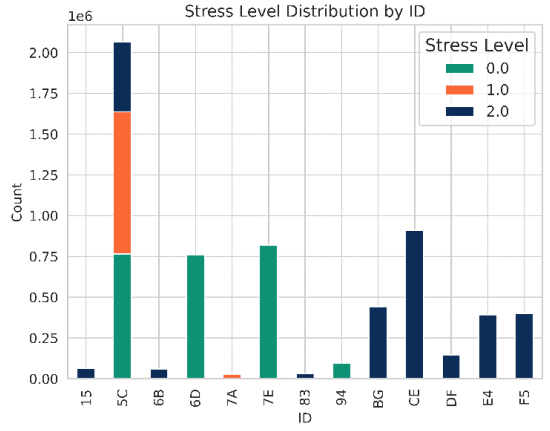

Fig. 3. Stress level distribution of each participant

Moreover, we are also interested in exploring the sensor readings based on the stress level as presented in Figure 4. Thus, we calculated the average readings of each sensor. The figure highlights potential physiological and behavioral variations corresponding to varying stress levels by suggesting that EDA, HR, TEMP, and ACC values fluctuate systematically across stress levels. The EDA readings exhibit an increasing trend with stress level; stress level 2 was the highest, followed by stress levels 0 and 1. On the other hand, the HR readings show the reverse trend, with stress level 0 showing the highest values, followed by levels 1 and 2 although the difference is little. Surprisingly, skin temperature readings show a positive link with stress levels as well, rising with higher stress levels. Nonetheless, there is a clear pattern in the accelerometer (ACC) values for all stress levels. The ACC value is highest at stress level 0, lowest at stress level 1, and least negative at stress level 2 where the sensors data difference is high. This implies that the dataset is highly imbalanced.

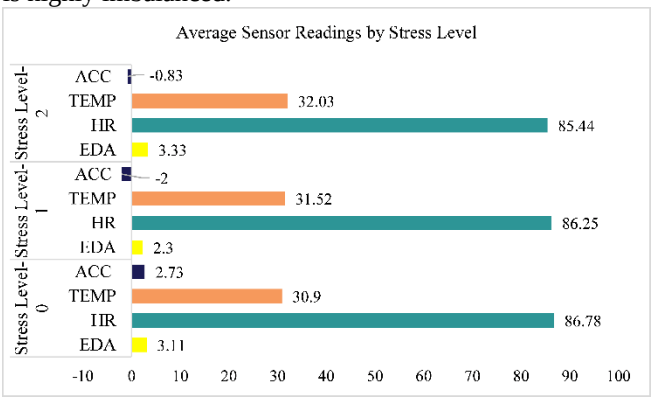

Fig. 4. Average sensor readings by stress levels

V. Methods

A. Statistical model: Method to Solve of our RQ1

After the data distribution of our dataset, we explored correlation analysis. We perform a correlation analysis between each independent variable (i.e., EDA) and stress level as demonstrated in Figure 5. Firstly, we check if the EDA data is normally distributed using the skewness measure. If the absolute skewness is less than 1, indicating normal distribution, the Pearson correlation coefficient and p-value are calculated using the *pearsonr* function. Otherwise, if the data is not normally distributed, the Spearman correlation coefficient and p-value are calculated using the *spearmanr* function. Next, a scatter plot is created using Seaborn's *lmplot* function, which shows the relationship between independent variables and stress level. The plot includes a regression line to visualize the trend in the data. Additionally, the plot is annotated with the correlation coefficient and p-value, providing information about the strength and significance of the correlation.

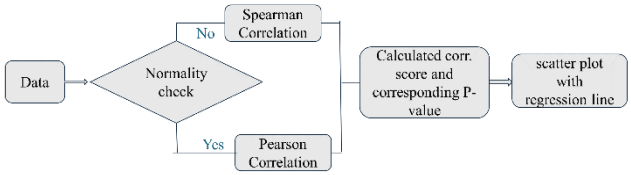

Fig. 5. Correlation analysis of our study.

B. *Machine Learning Method to Solve of our RQ2, RQ3*

In this study, we present a machine-learning model that incorporates Leave-one-group-out cross-validation to investigate the effect of time window segmentation on model performance as shown in Figure 6. The raw dataset consists of nine features taken from participant data. To investigate the impact of feature variability on model efficacy, the dataset is segmented using a 5-second time window. Following that, for each time window, we compute the average, maximum, lowest, and median values across all features, yielding an enlarged feature set of 25 attributes, including spatial coordinates (X, Y, Z), electrodermal activity (EDA), heart rate (HR), and skin temperature (TEMP). Following feature expansion, we organize the data with different features on the X-axis and stress levels on the Y-axis, allowing for easier model training and evaluation. Using Leave-one-group-out cross-validation, we divide the dataset into 15 groups, each representing a unique participant. During each iteration, one participant's data is held back for testing while the model is trained on the remaining 14 participants' data. This procedure enables robust model generalization by assessing performance across a variety of participant profiles. Each test set is evaluated using metrics such as accuracy, precision, recall, and F1 score, which provide information on the model's performance variability across different participant cohorts. After all iterations, the mean evaluation findings are calculated, providing a comprehensive assessment of the model's efficacy. In addition, before training the model, we use data pretreatment approaches to improve its quality and generalizability. Specifically, we use the Synthetic Minority Over-sampling Technique (SMOTE) to solve class imbalance and standard scaler feature scaling to normalize feature values, reducing the likelihood of feature dominance in training. Finally, we used different machine learning algorithms like random forests and decision trees.

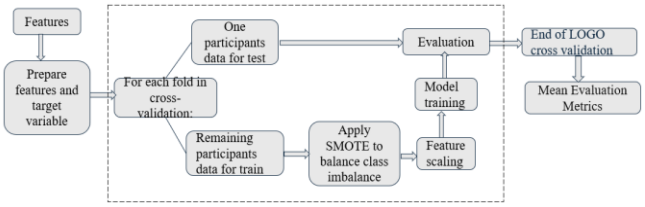

Fig. 6. Our machine-learning model architecture.

C. *Exploratory Data Analysis*

Exploratory data analysis is an approach to analyzing and visualizing datasets to extract meaningful insights, discover patterns, identify trends, and check assumptions. The primary goal of exploratory data analysis is to understand the main characteristics of the data and provide a foundation for further analysis or modeling.

Before doing machine learning, we first plot the distribution of each feature by id and label using histograms in the below figures 7,8, and 9 to understand the characteristics of the features.

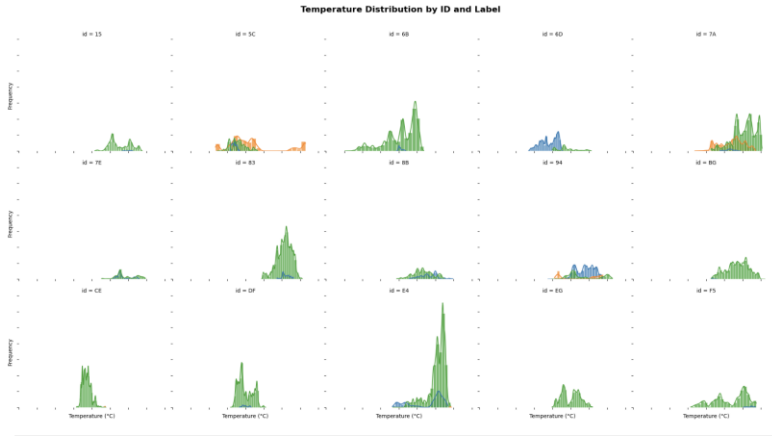

Fig. 7. HR distribution by ID and label

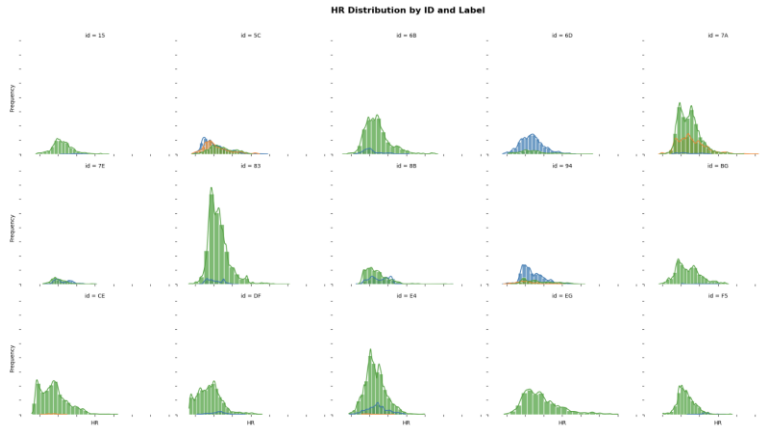

Fig. 8. Temperature distribution by ID and label

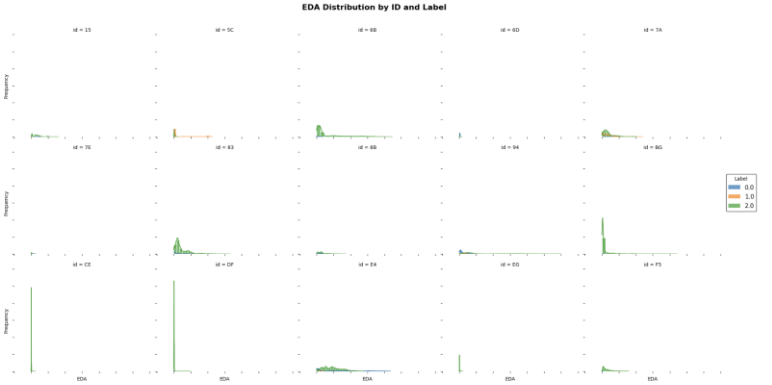

Fig. 9. EDA distribution by ID and label

Classification and cross-validation

A decision tree is a supervised machine learning algorithm that is used for both classification and regression tasks. It is a tree like model where an internal node represents a feature or attribute, the branches represent the decision rules, and the leaves represent the outcome or predicted value. Decision tree is popular due to its simplicity, interpretability, and effectiveness in a wide range of applicants.

The goal of decision tree when it uses in classification is to assign an input to one of several predefined classes or categories. In this case, we select XGBoost as the model, label of nurses' stress is our target, id, X, Y, Z, EDA, TEMP, HR are the features.

To reduce the bias and enhance the performance of model, we utilize the leave one out cross-validation. In this case, we have 15 nurses' data, for each iteration, use the left-out data for validation to evaluate the model's performance.

Evaluating the performance

Accuracy is a metric measures the correctly predicted instances to the total number of instances in the dataset. It provides the overall measure of how well the model performance across all classes.

$$accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predicitons}} \quad (1)$$

Recall is a metric measures the correctly predicted positive instances to the total number of instances in the dataset. It quantifies the model's ability to correctly identify positive instances.

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

Precision is a metric measures the correctly predicted positive instances to the total number of instances predicted as positive by the model. It quantifies the model's ability to avoid false positives.

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

F1 is a metric used to evaluate the balance between precision and recall in a classification model. The F1 score combines precision and recall into a single metric, giving equal weight to both measures.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Macro averaged indicates that all classes equally contribute to the final averaged metrics, it is calculated as an average of precision of all classes.

$$\text{Macro averaged precision} = \frac{1}{Number\ of\ classes} \times \sum Precision \quad (5)$$

Weight average precision is also calculated as an average of precisions of all classes, it indicates that each class's to the average is weighted by it's size.

$$Weighted\ average\ percision = \frac{1}{Total\ number\ of\ samples} \times \sum Precision * Number\ of\ sample \quad (6)$$

Fine-Tuning Model

The label in the dataset represents the stress level of nurses. The count of each label indicates that label = 2 makes a large proportion of the data, label = 1 and label = 0 make a small proportion of the data. With so few numbers of label = 1 and label = 0, the model will spend most of its time on label = 2 examples and not learn enough from the rest ones.

In order to compare whether the variety of features will affect the performance of the model, we processed the data in time window. The original data group by id and label, then we select 5 seconds as a time window to split the data chunk. After calculating the average, max, min, median values, save them in the new columns. After processing, the data size is reduced from 11509051 to 71463, the number of features increases from 7 to 25 (id, min_X, min_Y, min_Z, max_X, max _Y, max _Z, median_X, median_Y, median_Z, avg_X, avg_Y, avg_Z, min_EDA, max_ EDA, median_ EDA, avg_ EDA, min_HR, max_HR, median_HR, avg_HR, min_TEMP, max_ TEMP, median_TEMP, avg_TEMP).

Based on the previous classification report, indicates that if the minority classes (class=0, class = 1) are not well predicted by the model, it can contribute to a low accuracy even if the majority classes are well-predicted. The low precision, recall, and F1-score for class 0 and class 1 suggest that the model struggles to identify instances of these classes. To address the imbalanced dataset, we merged the minority classes into a single class, adjusted the labels accordingly and oversampled to improve the balance between labels. Additionally, we fine-tuned the hyperparameters of the XGBoost model, specifically tuning parameters such as the learning rate, the number of estimators, and the maximum depth of the trees. The final hyperparameters obtained through tuning are as follows: learning rate = 0.01, number of estimators = 20, and maximum depth = 7.

VI. RESULTS

A. Factors associated with stress level

It is essential to comprehend variables like skin temperature (TEMP), heart rate (HR), accelerometer (ACC), and electrodermal activity (EDA) about nurses' stress levels during the COVID-19 pandemic for many reasons. First, it helps identify certain stressors and their physiological effects, allowing focused treatments to promote nurses' well-being and stave off burnout. To adequately support frontline healthcare workers who are dealing with stressors connected to pandemics, it also guides the allocation of resources and execution of policies. Finally, studies into this link help develop technology, evidence-based therapies, and resilience-building strategies that are specific to the requirements of nurses in times of crisis.

The results of our study show some interesting conclusions shown in Table I: for EDA and stress level, the Spearman correlation coefficient (rs= 0.03) and (p-value < 0.01), indicating a very weak positive correlation; for Skin Temp and stress level, the Pearson correlation coefficient is (r=0.19) and (p-value < 0.01), suggesting a weak to moderate positive correlation. These results highlight the significance of taking into account a variety of physiological parameters to assess stress levels and provide insight into the physiological reactions of nurses to stress during the COVID-19 pandemic. Skin Temperature has a somewhat higher link with stress than EDA does, highlighting the complex relationship between physiological reactions and stress in nurses under difficult conditions such as the current pandemic.

Moreover, we also looked at the connection between stress levels and readings from the ACC and HR as presented in Table I. The investigation produced some intriguing results: a very weak negative correlation between stress level and HR was indicated by a Pearson correlation coefficient of r= -0.04 and p-value < 0.01, and a slightly stronger but still weak negative correlation between stress level and ACC was suggested by a Pearson correlation coefficient of r= -0.06 and p-value < 0.01, according to the analysis. These findings imply that stress levels among nurses and both HR and ACC measurements may have a little negative correlation. It's important to remember that the connections are somewhat weak, suggesting that in this particular situation, HR and physical activity as determined by ACC may not be very reliable indicators of stress levels.

TABLE I. **CORRELATION OF PHYSIOLOGICAL VARIABLES AND STRESS LEVEL**

Independent variable	Corr. Coef	P-value
EDA	rs = 0.03	0.00e+00
TEMP	r = 0.19	0.00e+00
ACC	r = -0.06	0.00e+00
HR	r = -0.04	0.00e+00

B. *Improving the Robustness and Performance of Machine Learning Models on Unbalanced Data*

The conventional 80-20 train-test split in machine learning, although widely practiced, is recognized for its propensity to induce overfitting, wherein models become overly specialized to the training data. This method might lead to overfitting since it offers the model limited unseen data during training, forcing it to memorize the training set rather than generalize effectively to new examples. This lack of exposure to different test data might result in an unduly positive evaluation of the model's performance, leading to overfitting when applied to real-world circumstances. We applied this method to our imbalanced dataset. However, we got nearly 98% of the accuracy, precision, recall and f1 score. As there is an issue of overfitting the model, we dive into the popular K-fold cross-validation technique. We applied a Neural network algorithm with 5-fold cross-validation where we can get 81 percent of accuracy, precision, recall and f1 score. Although this method provides less performance compared with train test split, this method is more robust and less overfitted. However, when model hyperparameters are overturned to the validation sets inside each fold of K-fold cross-validation, overfitting might occur, and the model may not be able to generalize to new observations. To overcome this overfitting issue and build a robust machine-learning model, we introduce the Leave-One-Group-Out cross-validation (LOGO-CV) technique. In the following section, we discussed our proposed method to build a robust model for the imbalanced dataset.

Our proposed model:

To find out the impact of feature diversity on model performance, we separated the data into time windows of 5 seconds each, grouping the original data by ID and label. Within each window, we generated average, maximum, minimum, and median values, which were then saved in new columns, increasing the number of features from 7 to 25. After running the XGBoost both in original data and processed data, we got 15 classification reports. In the report, we estimate the performance of the model through precision, recall, F1-score and accuracy. Table II, compares the weighted average value in the original raw data and processed data. The model's performance has shown improvement across all evaluation aspects, resulting in an overall enhancement.

TABLE II. **MODEL PERFORMANCE COMPARISON**

	Precision (raw/processed)		Recall (raw/processed)		F1-score (raw/processed)		Accuracy (raw/processed)	
15	0.91	0.91	0.94	0.93	0.93	0.92	0.94	0.95
5C	0.41	0.48	0.53	0.66	0.44	0.56	0.53	0.66
6B	0.56	0.89	0.11	0.25	0.10	0.34	0.11	0.25
6D	0.90	0.90	0.91	0.91	0.89	0.84	0.91	0.82
7A	0.52	0.51	0.64	0.61	0.57	0.55	0.11	0.61
7E	0.54	0.48	0.05	0.07	0.09	0.13	0.05	0.07
83	0.78	0.78	0.85	0.85	0.81	0.81	0.85	0.84
8B	0.11	0.67	0.32	0.51	0.16	0.52	0.32	0.51
94	0.31	0.55	0.28	0.28	0.17	0.18	0.28	0.28
BG	0.91	0.84	0.07	0.23	0.08	0.36	0.07	0.23
CE	0.95	0.95	0.82	0.82	0.88	0.88	0.82	0.82
DF	0.84	0.85	0.70	0.85	0.77	0.85	0.70	0.85
E4	0.72	0.85	0.67	0.85	0.69	0.81	0.67	0.85
EG	1.00	1.00	0.86	0.94	0.92	0.97	0.86	0.94
F5	0.93	0.86	0.71	0.78	0.78	0.82	0. 71	0.78

After fine-tuning the model, there has been a significant performance improvement. Table III presents a comparison between the initial model and the fine-tuned model. In the context of evaluating a fine-tuned model on an imbalanced dataset, using weighted averaging to represent its performance is a prudent choice. Unlike macro-averaging, which treats each class equally, weighted averaging adjusts for class imbalances by assigning more weight to classes with a larger number of instances. This ensures that the evaluation metric reflects the overall performance of the model while considering the uneven distribution of classes in the dataset. Therefore, by opting for weighted averaging, we can obtain a more representative assessment of the model's effectiveness in handling imbalanced data. From Table 2 we can observe that our fine-tuned model has an overall accuracy of 68% whereas it was around 59% before fine-tuning.

TABLE III. **MODEL PERFORMANCE COMPARISON (PROCESSED)**

	Precision (original/fine-tuned)		Recall (original/fine-tuned)		F1-score (original/fine-tuned)		Accuracy (original/fine-tuned)	
15	0.91	0.91	0.93	0.95	0.92	0.93	0.93	0.95
5C	0.13	0.50	0.28	0.66	0.16	0.56	0.28	0.66
6B	0.89	0.86	0.25	0.43	0.34	0.49	0.25	0.43
6D	0.90	0.91	0.82	0.80	0.84	0.83	0.82	0.80
7A	0.51	0.63	0.61	0.69	0.55	0.63	0.61	0.69
7E	0.48	0.55	0.07	0.48	0.13	0.39	0.07	0.48
83	0.79	0.84	0.84	0.81	0.81	0.82	0.84	0.81
8B	0.67	0.53	0.51	0.44	0.52	0.44	0.51	0.44
94	0.55	0.53	0.28	0.25	0.18	0.20	0.28	0.25
BG	0.84	0.94	0.23	0.45	0.36	0.59	0.23	0.45
CE	0.95	0.95	0.82	0.76	0.88	0.86	0.82	0.79
DF	0.85	0.86	0.85	0.88	0.81	0.87	0.85	0.88
E4	0.85	0.80	0.85	0.82	0.81	0.79	0.85	0.82
EG	1.00	1.00	0.86	0.92	0.92	0.96	0.86	0.92
F5	0.86	0.95	0.78	0.87	0.82	0.90	0. 78	0.87

VII. DISCUSSION

Our study expands upon the results of prior research by introducing innovative machine learning methods specifically designed to tackle the difficulties of predicting stress in real-time healthcare environments [1,2]. By utilizing cross-validation approaches like Leave-one-group-out and K-fold cross-validation, we improve the resilience and applicability of our predictive models, thus surpassing the constraints of traditional training-test divisions. In addition, our method of feature engineering, which includes dividing data into time windows and expanding the features, shows substantial enhancements in stress prediction models, leading to more precise and dependable stress assessment tools.

The model demonstrates a strong ability to identify high stress levels accurately. However, during the fine-tuning process, it was observed that in some cases, the accuracy of the fine-tuned model decreased compared to the original model. This phenomenon can be attributed to the oversampling technique, which emphasizes the small class, potentially leading to a decrease in overall accuracy. In handling imbalanced datasets, selecting an appropriate evaluation metric is crucial. While accuracy may suffice for well-balanced datasets, it may not adequately reflect performance for imbalanced datasets. Other metrics, such as precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC-ROC), provide a more comprehensive assessment of model performance in such scenarios. Therefore, careful consideration of the evaluation metric is essential to ensure accurate and meaningful assessment of model performance, particularly in the context of imbalanced datasets.

However, our model's performance, with an accuracy of 68%, falls short of expectations. This highlights the possibility for improvement and underscores the need for further refinement of our predictive algorithms. Despite the positive outcomes, there is still space for boosting the accuracy and effectiveness of our models through continuous research and development. Despite these limitations, our study provides useful insights into stress prediction among nurses and lays the framework for future research aiming at building more accurate and reliable stress assessment techniques in healthcare settings.

VIII. Reflection on Group Work

Throughout our data mining project, several aspects stood out as particularly engaging and educational. Among the highlights was the implementation of a machine-learning model to address our research question. This not only provided tangible outcomes but also served as a rewarding demonstration of how theoretical knowledge can be applied in practical, impactful ways.

The journey to these outcomes, however, was not without its challenges. The most time-consuming and demanding part of the project was undoubtedly the data preprocessing phase, particularly the data merging process. Our dataset was extensive, comprising various signals and timestamps, which necessitated a meticulous approach to ensure accurate merging and alignment. This phase tested our problem-solving skills and patience, pushing us to innovate and refine our methodologies continuously.

From this project, our learning extended far beyond the confines of our curriculum. The hands-on experience with real-world data was invaluable, teaching us not only about the intricacies of data handling and analysis but also about the importance of resilience and persistence in the face of technical challenges. Our supervisor played a crucial role in this learning process, introducing us to new techniques and providing guidance that was critical in overcoming the hurdles we encountered.

Deciding on the research question was a collaborative effort, grounded in thorough discussions within our group and a review of existing literature. This approach ensured that our work was both relevant and grounded in scientific inquiry, contributing meaningfully to the field of stress detection in healthcare professionals using wearable technology.

Task division among group members was strategically based on individual strengths and interests, yet it was designed in a way that encouraged involvement and contribution from everyone. This not only optimized our workflow but also enhanced the cohesiveness of our team, as each member brought unique perspectives and expertise to the table.

Fortunately, the project unfolded largely as planned. The design and implementation phases proceeded without significant deviations from our initial proposal, reflecting our team's ability to effectively plan and execute a complex research project.

As for the contributions of our team members, each person brought their dedication and skills to ensure the success of the project. Whether it was data analysis, coding, writing, or reviewing, everyone's efforts were pivotal. This equal and active participation fostered a sense of shared ownership and pride in our results, which not only aligned with our goals but also underscored the collaborative spirit of our research.

Reflecting on this project, it is clear that the challenges we faced were as valuable as the successes we achieved. Each step provided us with deeper insights into the field of data science and a greater appreciation for the power of teamwork in tackling complex problems.

References

[1] S. Eom, S. Eom, and P. Washington, "SIM-CNN: Self-supervised individualized multimodal learning for stress prediction on nurses using biosignals," in Machine Learning for Multimodal Healthcare Data, Cham: Springer Nature Switzerland, 2024, pp. 155–171.

[2] X. Li, W. Zhu, X. Sui, A. Zhang, L. Chi, and L. Lv, "Assessing workplace stress among nurses using heart rate variability analysis with wearable ECG device–A pilot study," Front. Public Health, vol. 9, 2022.

[3] M. Di Giuseppe et al., "Stress, burnout, and resilience among healthcare workers during the COVID-19 emergency: The role of defense mechanisms," Int. J. Environ. Res. Public Health, vol. 18, no. 10, p. 5258, 2021.

[4] S. Das, "Stress in healthcare workers: a model to help us rethink challenges," Health e-News, 10-May-2022. [Online]. Available: https://healthenews.mcgill.ca/stress-in-healthcare-workers-a-model-to-help-us-rethink-challenges/. [Accessed: 1-Apr-2024].

[5] E. P. Nwobodo, B. Strukcinskiene, A. Razbadauskas, R. Grigoliene, and C. Agostinis-Sobrinho, "Stress management in healthcare organizations: The Nigerian context," Healthcare (Basel), vol. 11, no. 21, p. 2815, 2023.

[6] P. Karthikeyan, M. Murugappan, and S. Yaacob, "Analysis of stroop color word test-based human stress detection using electrocardiography and heart rate variability signals," Arab. J. Sci. Eng., vol. 39, no. 3, pp. 1835–1847, 2014.

[7] A. Muaremi, A. Bexheti, F. Gravenhorst, B. Arnrich, and G. Troster, "Monitoring the impact of stress on the sleep patterns of pilgrims using wearable sensors," in IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2014.

[8] A. Liapis, C. Katsanos, D. Sotiropoulos, M. Xenos, and N. Karousos, "Stress recognition in human-computer interaction using physiological and self-reported data: A study of gender differences," in Proceedings of the 19th Panhellenic Conference on Informatics, 2015.

[9] M. Zubair, C. Yoon, H. Kim, J. Kim, and J. Kim, "Smart wearable band for stress detection," in 2015 5th International Conference on IT Convergence and Security (ICITCS), 2015.

[10] V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, and O. M. Mozos, "Stress detection using wearable physiological sensors," in Artificial Computation in Biology and Medicine, Cham: Springer International Publishing, 2015, pp. 526–532.

[11] T. Islam and P. Washington, "Individualized stress mobile sensing using self-supervised pre-training," Appl. Sci. (Basel), vol. 13, no. 21, p. 12035, 2023.

[12] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim, “Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings,” in 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007.

[13] S. Hosseini et al., “A multi-modal sensor dataset for continuous stress detection of nurses in a hospital.” Dryad, 2021.