

An interpretable machine learning approach for identifying occupational stress in healthcare professionals

Milena Seibert Fernandes¹, Roberto Rodrigues-Filho¹, Iwens Sene-Junior², Stefan Sarkadi³,
Alison R. Panisson¹ and Analúcia Schiaffino Morales¹

¹*Department of Computing, Federal University of Santa Catarina, Santa Catarina, Brazil*

²*Institute of Informatics, Federal University of Goiás, Goiânia, Brazil*

³*Department of Informatics, King's College London, London, United Kingdom*
iwens@ufg.br, stefan.sarkadi@kcl.ac.uk, {roberto.filho, alison.panisson, analucia.morales}@ufsc.br

Keywords: Interpretable ML, Stress, AI for Healthcare

Abstract: In the last few years, several scientific studies have shown that occupational stress has a significant impact on workers, particularly those in the healthcare sector. This stress is caused by an imbalance between work conditions, the worker's ability to perform their tasks, and the social support they receive from colleagues and management professionals. Researchers have explored occupational stress as part of a broader study on affective systems in healthcare, investigating the use of biomarkers and machine learning approaches to identify early conditions and avoid Burnout Syndrome. In this paper, a set of machine learning (ML) algorithms was evaluated using statistical data on biomarkers from the *AffectiveRoad* database to determine whether the use of explanations can help identify stress more objectively. This research integrates explainability and machine learning to aid in the identification of various levels of stress, which has not been previously evaluated for the domain of occupational stress. The Random Forest is the best-performing model for this assignment, followed by k-Nearest Neighbors and Neural Network. Later, explainers were applied to the Random Forest, highlighting feature importance, partial dependencies between characteristics, and a summary of the impact of features on outputs based on their values.

1 INTRODUCTION

Over the last few years, we have studied occupational stress as part of a comprehensive study on affective systems in healthcare to improve decision-making. In seeking alternatives for identifying stress, studies have been conducted on physiological parameters that can assist in the non-invasive identification of occupational stress (Morales et al., 2022b), (Morales et al., 2022c). In the field of affective systems, it is emphasized that recognizing emotions is a complicated task that requires techniques capable of helping to identify various aspects. These include the intensity levels of emotions, the types of emotions, the factors that trigger these emotions, and behavioral and social factors that may affect the expression or suppression of emotional issues (Picard, 2000).

Psychological professionals can identify occupational stress early through standardized responses to questionnaires. A unique way to identify and diagnose stress conditions has been used for several years with this technique. Despite the validity of these

methods, there is a concern about subjective bias as they may generate distortions in the results (Greene et al., 2016). With the advancement of sensor technology, several studies have been conducted to identify stress using physiological signals and machine learning techniques (Morales et al., 2022b): heart rate, brain activity, skin response, body temperature, blood activity, respiratory response, and muscle activity, i.e., biomarkers. There are potential parameters to assist in identifying biomarkers; in fact, many wrist-worn wearables have been tested in data collection in recent years, and their results have been promising (Morales et al., 2022a). The biomarker-based approach, however, does not show much trust because stress identification involves psychological, physiological, and emotional aspects. Stress represents a coordinated activation of multiple biological systems following the introduction of a stressful stimulus (stressors) that represents a disruption of homeostasis or a perceived disruption (Potts et al., 2019). Many studies have examined the association between this point and ML algorithms in the scientific litera-

ture (Morales et al., 2022c). According to Inam et al. (2021), the complexity and sophistication of systems involving artificial intelligence (AI) have grown to the point where humans are not always capable of understanding the reasoning behind the decisions made by ML mechanisms. This fact can be attributed to large datasets composed of a massive volume of information used to train and test increasingly complex systems (Linardatos et al., 2021). Presently, the ability to interpret and understand the mechanisms behind AI is essential for the validation of ML systems in healthcare area (Montavon et al., 2018). Additionally, Guidotti et al. (2018) highlight the possibility of ML components making incorrect decisions without providing the opportunity to detect the learning problem. Systems that apply AI for decision-making can be classified into one of the following three categories:

- **Opaque systems** that do not offer any insight into their algorithmic mechanisms, concealing their internal knowledge from the user (Doran et al., 2017; Guidotti et al., 2018);
- **Interpretable systems**, whose algorithmic mechanisms can be analyzed by their users (Doran et al., 2017);
- **Understandable systems** that emit symbols allowing user-guided explanations on how a conclusion is reached (Doran et al., 2017);

Explainable Artificial Intelligence (XAI) refers to methods and techniques that produce understandable and accurate models, highlighting why a model reaches a specific decision. Therefore, solutions obtained by artificially intelligent systems can be comprehended by humans, providing more transparency and interpretability to instill trust in the results produced by AI-based solutions (Inam et al., 2021). It can be uncomfortable to rely on a decision made without any explanation (Doran et al., 2017; Miller, 2019). The fact that humans are not always capable of understanding the results of black-box algorithms increases the necessity for interpretability, transparency, and explainability of outputs generated by artificial intelligence systems. These factors are crucial for humans to comprehend and trust AI-based systems (Inam et al., 2021). The ability of XAI to identify stress biomarkers from a dataset of stress biomarkers gathered from a study group would be a valuable study in this field. In this way, biomarker-based systems can be made more reliable and the dataset can be optimized, thereby allowing for the identification of individual stressors to be enhanced.

In this paper we present an evaluation of a set of machine learning algorithms trained with statisti-

cal data of biomarkers available in the *AffectiveRoad* database (Lopez-Martinez et al., 2019; Vos et al., 2022) for the detection of different levels of stress. In this study, we investigate whether explanations are useful for identifying stress objectively. Biomarkers used from the database are similar to those used to identify occupational stress in healthworkers (Hosseini et al., 2022): heart rate (HR), electrodermal activity (EDA), and skin temperature (TEMP). The machine learning algorithms used were Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Neural Network (NN), Random Forest (RF), and Logistic Regression (LR). After identifying the algorithm with the best performance, we presented explanations that highlighted the characteristics of the dataset that have the most influence on identifying stress. Explainers employed for identification were Partial Dependency Plot, Feature Importance, and Summary Plot. Lastly, model optimization was conducted considering only the most important characteristics identified in the previous stages. Finally, the algorithms and explainers were tested and evaluated again for the optimized dataset. The relevance of this research lies in combining explainability and machine learning to assist in the identification of different levels of stress. Similar works were identified in the scientific literature (Chalabianloo et al., 2022; Tseng et al., 2020); however, no evidence was found of studies that evaluate and explain machine learning methods specifically for the domain of occupational stress. Thus, this present study contributes to advancing knowledge in both the field of computer engineering (artificial intelligence) and healthcare.

2 MATERIAL AND METHODS

From a preliminary study, articles from the scientific literature were investigated to support the selection criteria among publicly available databases for occupational stress detection. Consideration was given to the article by (Hosseini et al., 2022), which utilizes the *AffectiveRoad* database (Lopez-Martinez et al., 2019; Vos et al., 2022), and also provides pre-processed data with 48 columns of statistical data on the collected biomarkers.

For the selection of explainers, a study was conducted in the scientific literature. Research such as that by (Guidotti et al., 2018) and (Montavon et al., 2018) present evidence for the use of explainers depending on the model's input and output. In the case of this application for stress detection, good choices to clearly highlight the reasons behind the model's decisions are explainers such as Summary Plot (SP),

Partial Dependency Plot (PDP), and Feature Importance (FI). In addition, Shapley Additive Explanators (Antwarg et al., 2021) were used to obtain such explainers. Finally, the 10 most important characteristics of the three models with the best performance during the model evaluation stage were considered. Therefore, all five algorithms were retrained three times:

- First, with the 10 most important features for the first best result;
- Second, with the 10 most important features for the second best result;
- Third, with the 10 most important features for the third best result.

Therefore, three stages were developed: a preliminary assessment of machine learning algorithms, an explanation of the Random Forest algorithm, which showed better performance during the evaluation, and finally, training the algorithms with the most important characteristics identified in the study. The dataset has 49 columns.

3 RESULTS

A performance comparison of black-box machine learning algorithms was conducted. A Python program was developed to obtain statistics related to the models' performance, using classes from the sklearn library to evaluate the selected algorithms in this study. Furthermore, parameters indicating the performance of each algorithm were obtained. These parameters were Accuracy, AUC (area under ROC curve), Recall, and F1 Score.

Figure 1 displays a diagram indicating the inputs and potential outputs of the used algorithms.

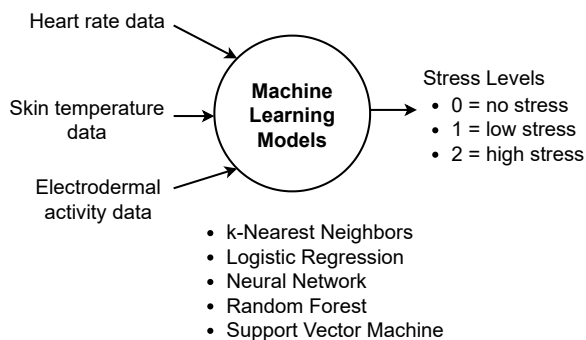


Figure 1: Diagram of inputs and outputs of the algorithms

3.1 Evaluation of the Selected Machine Learning Models

Among the biomarker data, totaling 48 input characteristics, it includes the average of the last 11 measurement cycles, maximum value, minimum value, and standard deviation of the three considered biomarkers. Additionally, other data such as amplitude, duration, kurtosis, standard deviation, and distortion of electrodermal activity, as well as the number of peaks and heartbeats per second of heart rate, are included. On the other hand, the models' output is summarized by the stress level. There were three levels of stress to classify and label: 0 (no stress), 1 (low stress), and 2 (a lot of stress).

3.1.1 Training of models

The parameters used for training the algorithms were also developed using the resources of the sklearn library. In the case of the Random Forest, some parameters were modified to follow the same settings chosen by (Hosseini et al., 2022). In other cases, the parameters were varied and chosen based on the algorithms' performance with parameter settings. The following are the training parameters applied to each model:

- k-Nearest Neighbors was trained with k being 5, using Euclidean metric, and uniform weight.
- Logistic Regression was trained with L2 regularization and C equal to 1.
- Neural Network was trained with 100 neurons in hidden layers, ReLu activation, and Adam optimization.
- Random Forest was trained with 100 trees and a minimum number of branches set to 5.
- Support Vector Machine was trained with a cost of 1, regression loss of 0.1, linear kernel, and numerical tolerance of 0.001.

It's worth noting that Support Vector Machine and Logistic Regression algorithms are more commonly used as binary classifiers, in applications with only two possible outputs. For this reason, approaches were employed to enable the models trained with these algorithms to consider three classes. For the Support Vector Machine algorithm, a One Vs. One approach was used, where 3 classifier models were trained, each using data from two distinct classes. Additionally, the Logistic Regression model used the One Vs. Rest approach, so that three classifiers were trained, with each considering one of the three possible outputs as positive and the other two as negative.

3.1.2 Results of Algorithm Evaluation

The algorithms were then tested, where 70% of the data was selected for training in all cases, and the remaining 30% was used for testing. The division of data for training and testing was randomly performed using the *train_test_split* function from the sklearn library. The evaluation was conducted by comparing the expected outputs with the outputs obtained by each model.

Table 1 shows the results obtained after evaluating the algorithms with the test data.

	Model	AUC	Accuracy	F1	Recall
(1)	RF	0.994	0.955	0.943	0.933
(2)	kNN	0.964	0.878	0.854	0.854
(3)	NN	0.900	0.771	0.702	0.691
(4)	LR	0.747	0.610	0.453	0.496
(5)	SVM	0.735	0.617	0.451	0.500

Table 1: Results of model evaluation.

For all the parameters used, a value closer to 1 indicates better performance. For this reason, the models were ranked in order of AUC, as observed in the assigned ranking. This parameter was chosen because it represents the relationship between true positives and false positives classified by the models. Note that the algorithms Logistic Regression and Support Vector Machine would change their ranking if the decisive parameter were accuracy or recall.

Based on the AUC value, the Random Forest algorithm presents excellent performance in classifying stress levels using biomarkers. Also showing good performance are the k-Nearest Neighbors and Neural Network algorithm. On the other hand, the Support Vector Machine and Logistic Regression algorithms exhibit performance closer to a random classifier, making them less suitable for real-world applications.

3.2 Explanation of the Random Forest

As the Random Forest showed the best result, explainers were applied to the trained model to elucidate how the Random Forest works. The explainers were obtained using the shap library.

3.2.1 The shap library

The shap library employs measures known as *shap values*, introduced by (Lundberg and Lee, 2017). These measures constitute a unified representation of feature importance, based on Shapley Values, which

utilize game theory equations to derive values. While Shapley Values quantify the contribution each player brings to a game, *shap values* quantify the contribution each feature brings to the model's prediction, making shap the most advanced explainability library to date (Mazzanti, 2020).

3.2.2 Partial Dependence Plot

Figure 2 depicts a partial dependence plot of the Random Forest for the feature *TEMPR_Max*, considering the feature *HRR_Min* when the considered output is 2 (high stress).

The horizontal axis of the graph represents the proportion of values of the feature *TEMPR_Max*, the maximum value of skin temperature. The closer to 0.0, the closer it is to the smallest measured value for this feature, and the closer to 1.0, the closer it is to the largest measured value for this feature.

Similarly, the colors of the markers indicate the proportion of the value of *HRR_Min*, the minimum value of heart rate. The more blue, the closer it is to the minimum value measured for this feature, and the closer to red, the closer it is to the maximum value measured for this feature. Thus, the vertical axis indicates the impact of the relationship between the features *TEMPR_Max* and *HRR_Min* on the model.

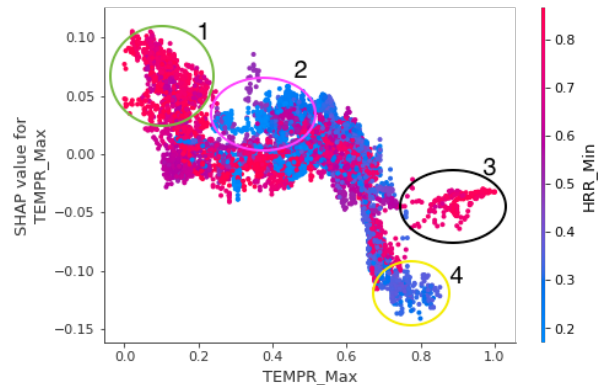


Figure 2: Partial Dependence Plot of the model.

To increase understanding, four regions of the plot deserving attention have been highlighted. These regions are:

- Region 1 indicates that when the minimum heart rate has a high value and the maximum skin temperature has a low value, the skin temperature feature positively contributes to detecting a stress level of 2 (high stress).
- Region 2 suggests that as the minimum heart rate decreases and the maximum skin temperature increases, the impact of maximum skin temperature decreases, approaching 0.

- Region 3 signifies that high values of minimum heart rate and maximum skin temperature have a small negative impact, contributing to diagnosing little or no stress.
- Region 4 suggests that high values of maximum skin temperature, combined with low values of minimum heart rate, contribute to an output indicative of little or no stress.

It is noteworthy that the partial dependence plot is a type of local explanation, as it considers only one target output.

3.2.3 Summary Plot

The summary plot is named as such because it aims to “summarize” the impact of features associated with their values. Red points indicate that the feature value is higher than most other values assigned to that feature. Similarly, blue points correspond to values lower than most other values assigned to a particular feature. Each row corresponds to a feature, indicated on the left, and each point corresponds to a data instance.

The horizontal axis indicates the impact, positive or negative, that the feature has on the considered output value. In the middle of the plot, there’s a vertical line indicating an impact of 0, where features do not influence the model’s output. Points on the left have a negative impact, while points towards the right have a positive impact on the output considered. Figure 3 presents the summary plot of the Random Forest, considering output 2 (high stress).

The plot highlights that the top six features are crucial for the model’s output, as they possess more than double the importance of the others. Additionally, it’s noticeable that the feature *TEMP_Mean*, i.e., average skin temperature, has the most significant impact on the result. Following closely are *TEMP_Max*, which is the maximum skin temperature, and *HRR_Mean*, representing the mean heart rate.

It can be noted that all the top six most important features exhibit the same behavior: with low values, the features contribute to the high stress output, while with high values, there is a tendency towards diagnoses of little or no stress.

It is also notable that concerning less important features, the points are closer to the no-impact line, and the relationship between the feature values and their impact on the model’s output is unclear. Moreover, the summary plot is a local explanation, considering only one of the possible outputs. Next, the results of training the five models using only the top ten most important features from the three models with the best performance are presented.

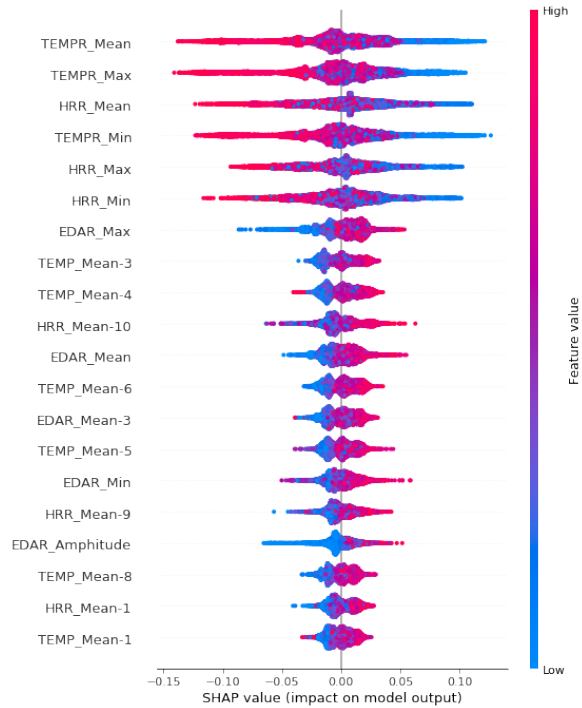


Figure 3: Summary Plot of the *Random Forest*.

3.3 Training Algorithms with Most Important Features

As the final stage of this work’s development, the training of algorithms is conducted using the top 10 most important features for the three highest-rated outcomes. Training with fewer features can help improve model training time and evaluate which features are truly necessary for occupational stress detection, enabling the simplification of data preprocessing. This investigation aims to verify if it is possible to achieve similar results with fewer features in each of the studied algorithms.

For training, it was necessary to first find the 10 most important features from the 3 best-performing algorithms tested in the study: Random Forest, k-Nearest Neighbors, and Neural Network. Figure 3 displays the feature importance assessment for the Random Forest. The top 10 most relevant features for k-Nearest Neighbors and Neural network were also considered for training all five models. With this information, it was possible to perform the training of the five algorithms again, using only the top ten features from each Feature Importance analysis.

All five models were retrained, and the same performance parameters were obtained. Table 2 shows the AUC obtained from all the models trained with sets of features.

The first column of Table 2 displays the AUC

	Model	All	10 RF	10 kNN	10 NN
(1)	RF	0.994	0.992	0.986	0.991
(2)	kNN	0.964	0.990	0.960	0.988
(3)	NN	0.900	0.870	0.841	0.875
(4)	LR	0.747	0.736	0.729	0.739
(5)	SVM	0.735	0.735	0.720	0.732

Table 2: AUC results from training with all features, and the top 10 impactful features for Random Forest, k-Nearest Neighbors and Neural Network.

found for the models trained with all the features in the dataset.

The second column, 10 RF, displays the AUC found for the models trained only with the 10 most important features for Random Forest. When comparing the AUC value of models when trained with the top 10 most important features of RF model with the AUC of models trained with all features, it is observed that there was no drastic reduction in the performance of any of the models. On the other hand, it is worth highlighting the significant improvement in the kNN model’s performance, which considerably approaches the RF model, ranked first. Additionally, the classification of the algorithms remains the same.

The third column, 10 kNN, displays the AUC found for the models trained only with the 10 most important features for k-Nearest Neighbors. When comparing the results from training with the top 10 most important features of kNN model with the results from training with all features, all models showed a decrease in performance, but this decrease was scarcely noticeable in practice.

Finally, the last column, 10 NN, displays the AUC found for the models trained only with the 10 most important features for Neural Network. These results show a slight deterioration in the performance of all results compared to the performance of models trained with all features, except for the kNN model. Similar to what was observed in the results from training with the most impactful features for Random Forest, the *k-Nearest Neighbors* model showed a significant improvement in performance, also approaching the *Random Forest* in the top position. These results also do not alter the classification of the algorithms.

4 DISCUSSION

In this section, we present the main outcomes of this study and compare and discuss them in relation to findings from other works identified in the literature.

The primary outcome of the machine learning algorithm evaluation stage was the superior performance observed in the Random Forest model. In re-

lated studies (Dave et al., 2020), the XGBoost, a tree ensemble model similar to Random Forest introduced by (Chen and Guestrin, 2016), was included in evaluations, unlike in this research.

Moreover, the research by (Tseng et al., 2020) provided an assessment of machine learning algorithms, concluding that the combination of XGBoost with Random Forest into a single algorithm outperformed all other evaluated models, which included SVM, LR, and simple decision trees. Additionally, in (Tseng et al., 2020), XGBoost and Random Forest ranked second and third, respectively, indicating the strong performance of tree ensemble algorithms. This research did not combine algorithms and did not test XGBoost.

Additionally, in (Chalabianloo et al., 2022), four classifiers were identified as the most promising for training. These classifiers were SVM, *Random Forest*, *Extremely Randomized Tree*, and *Light Gradient Boosting Machine*, two of which are similar to those used in this research. The latter two algorithms are also tree-based ensemble methods, added to the study after the authors observed promising results from the Random Forest model. In this case, while Random Forest was not chosen, it outperformed the others, demonstrating its significance.

Furthermore, the study by (Morales et al., 2022c) provided a survey of machine learning model types, considering the analyzed features for stress detection. Despite SVM being a predominant model found in many of the studies surveyed, it did not perform well in this research with the utilized dataset and parameters.

Conversely, in the work by (Bahani et al., 2021), a comparison among machine learning models was conducted, and the Random Forest exhibited one of the worst performances among the evaluated algorithms, which included SVM, kNN, NN, and Naive Bayes. These results contrast with the outcomes of the present study.

4.1 Comparing the Biomarker Impacts

This phase of the development resulted in an explanation of the Random Forest, elucidating the significance of model characteristics, along with graphs associating their values and contributions to possible outputs. Additionally, it can be noted that the importance of characteristics varies depending on the algorithm to be explained.

Among all the analyzed works, the only one addressing stress explainability, allowing for comparisons with the outcomes achieved in this study, is the work by (Chalabianloo et al., 2022). On the other

hand, the work by (Hosseini et al., 2022) does not provide explanations for the obtained outputs. Works such as (Madanu et al., 2022), (Pawar et al., 2020), and (Dave et al., 2020) focus on different areas of medicine, offering no possibility of comparing explanations and biomarker impacts for occupational stress detection.

The work by (Chalabianloo et al., 2022) utilized data collected in a laboratory environment using seven different wearable devices. These data were also categorized among low, medium, and high stress through context analysis during their collection. Therefore, the quantity and types of data are different from those used in this study. Feature selection was performed using recursive feature elimination with cross-validation, resulting in 12 features related to heart rate and entropy. The most important features for the trained algorithms were the interval between heartbeats and approximate entropy. Subsequently, the authors added electrodermal activity measures, resulting in improved model performance. However, there is no presentation of the importance of features associated with electrodermal activity, and models trained with this data are not explained.

The work by (Chalabianloo et al., 2022) also indicates that the importance of features varies according to the explained model, which is supported by the feature importance analysis conducted in this study. Overall, it reinforces the significance of features related to heart rate and electrodermal activity for stress detection, emphasizing that the use of more biomarkers beyond heart rate contributes to making the model more reliable.

4.2 Feature Reduction in Model Training

Training with fewer features resulted in a slight reduction in the performance of the tested algorithms, except for the k-Nearest Neighbors. When trained with the 10 most important features from the Random Forest and Neural Network algorithms, kNN showed a significant improvement in performance, showing efficiency in cases with fewer features, as predicted in the work by (Singh et al., 2016). Despite the performance improvement, kNN did not outperform the Random Forest in any case, which remained in the first position in all evaluations conducted. So far, no research has been found in the survey that utilizes feature importance to reduce the dataset size and optimize algorithm training.

Regarding training time, there was no noticeable difference. However, it is noted that when training algorithms with real-world data, much larger in size,

the difference may be noticeable. Training time tends to be longer for LR and shorter for kNN, RF, and SVM, or indifferent for NN according to characteristics mentioned in the work by (Singh et al., 2016).

5 CONCLUSION

In this paper, we evaluate machine learning models trained with statistical biomarker data extracted from the AffectiveRoad database to detect different stress levels. Our evaluation indicated Random Forest as the best-performing model for this task, followed by k-Nearest Neighbors and Neural Network. Later, explainers were applied to the Random Forest, highlighting partial dependencies between characteristics, and a summary of the impact of features on outputs based on their values.

The feature importance assessment suggests that data related to skin temperature and heart rate hold greater significance for the model. Moreover, the Random Forest summary graph indicates that high values of skin temperature and heart rate suggest low or no stress. Conversely, high electrodermal activity indicates stress, although it holds less importance for the model. Furthermore, the partial dependency graph illustrates that even with high skin temperature, the heart rate value can increase or decrease its impact on the model. Noteworthy regions of the partial dependency graph are highlighted to demonstrate how one characteristic's value influences the impact of another characteristic.

Finally, this study presents some limitations we intend to address in future work. Firstly, machine learning algorithm training was conducted using only statistical features derived from three biomarkers, while there are nine other biomarkers used in stress detection in other studies. Secondly, some algorithms commonly used for diagnostics in the healthcare field were not part of the evaluation, and no combination of algorithms was used for final assessment. Additionally, no emotional factor characteristics were used for model training, making it difficult to distinguish between eustress and distress. Furthermore, there was no translation of explanations to make them easily understandable for healthcare professionals, and there was no integration with a recommendation system to address detected stress. Local explanations, valid for only one data instance, can be applied and translated into a sentence in natural language to facilitate quicker and easier understanding of these explanations, directly assisting healthcare workers to better understand stress identification from biomarker readings.

REFERENCES

- Antwarg, L., Miller, R. M., Shapira, B., and Rokach, L. (2021). Explaining anomalies detected by autoencoders using Shapley Additive Explanations[Formula presented]. *Expert Systems with Applications*, 186.
- Bahani, K., Moujabbir, M., and Ramdani, M. (2021). An accurate fuzzy rule-based classification systems for heart disease diagnosis. *Scientific African*, 14.
- Chalabianloo, N., Can, Y. S., Umair, M., Sas, C., and Ersoy, C. (2022). Application level performance evaluation of wearable devices for stress classification with explainable AI. *Pervasive and Mobile Computing*, 87.
- Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016, pages 785–794. Association for Computing Machinery.
- Dave, D., Naik, H., Singhal, S., and Patel, P. (2020). Explainable AI meets Healthcare: A Study on Heart Disease Dataset. Technical report.
- Doran, D., Schulz, S., and Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives.
- Greene, S., Thapliyal, H., and Caban-Holt, A. (2016). A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. *IEEE Consumer Electronics Magazine*, 5(4):44–56.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Gianotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5).
- Hosseini, S., Gottumukkala, R., Katragadda, S., Bhupatiraju, R. T., Ashkar, Z., Borst, C. W., and Cochran, K. (2022). A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Scientific Data*, 9(1).
- Inam, R., Terra, A., Mujumdar, A., Fersman, E., and Feljan, A. V. (2021). Explainable AI – how humans can trust AI.
- Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods.
- Lopez-Martinez, D., El-Haouij, N., and Picard, R. (2019). Detection of Real-world Driving-induced Affective State Using Physiological Signals and Multi-view Multi-task Machine Learning.
- Lundberg, S. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions.
- Madanu, R., Abbod, M. F., Hsiao, F.-J., Chen, W.-T., and Shieh, J.-S. (2022). Explainable AI (XAI) Applied in Machine Learning for Pain Modeling: A Review. *Technologies*, 10(3):74.
- Mazzanti, S. (2020). SHAP Values Explained Exactly How You Wished Someone Explained to You.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences.
- Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks.
- Morales, A., Barbosa, M., Morás, L., Cazella, S. C., Sgobbi, L. F., Sene, I., and Marques, G. (2022a). Occupational stress monitoring using biomarkers and smartwatches: A systematic review. *Sensors*, 22(17).
- Morales, A. S., de Oliveira Ourique, F., Morás, L. D., Barbosa, M. L. K., and Cazella, S. C. (2022b). *A Biomarker-Based Model to Assist the Identification of Stress in Health Workers Involved in Coping with COVID-19*, pages 485–500. Springer International Publishing, Cham.
- Morales, A. S., de Oliveira Ourique, F., Morás, L. D., and Cazella, S. C. (2022c). *Exploring Interpretable Machine Learning Methods and Biomarkers to Classifying Occupational Stress of the Health Workers*, pages 105–124. Springer International Publishing, Cham.
- Pawar, U., O’shea, D., Rea, S., and O’reilly, R. (2020). Explainable AI in Healthcare. Technical report.
- Picard, R. W. (2000). *Affective computing*. MIT press.
- Potts, S. R., McCuddy, W. T., Jayan, D., and Porcelli, A. J. (2019). To trust, or not to trust? individual differences in physiological reactivity predict trust under acute stress. *Psychoneuroendocrinology*, 100:75–84.
- Singh, A., Thakur, N., and Sharma, A. (2016). *A Review of Supervised Machine Learning Algorithms*.
- Tseng, P. Y., Chen, Y. T., Wang, C. H., Chiu, K. M., Peng, Y. S., Hsu, S. P., Chen, K. L., Yang, C. Y., and Lee, O. K. S. (2020). Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Critical Care*, 24(1).
- Vos, G., Trinh, K., Sarnyai, Z., and Azghadi, M. R. (2022). Machine Learning for Stress Monitoring from Wearable Devices: A Systematic Literature Review.