Employing Machine Learning Algorithms to Detect Stress with a Specific Emphasis on Commuting Methods

Mhd Saeed Sharif
Intelligent Research Group, ACE School, UEL
University way, E16 2RD, UK
s.sharif@uel.ac.uk

Madhav R. T. Tamang
Intelligent Research Group, ACE School, UEL
University way, E16 2RD, UK
u1430774@uel.ac.uk

Cynthia H.Y. Fu
School of Psychology
UEL
Water Lane, E15 4LZ, UK
c.fu@uel.ac.uk

Wael Elmedany
College of Information Technology, University of Bahrain. Kingdom of Bahrain.

***Abstract*—The regular commute for many individuals could significantly impact their general well-being. The daily commute to work can be linked to chronic stress, which is known to have negative implications on mental health, as well as increased blood pressure, heightened heart rate, and high fatigue. The primary objective of this study is to examine the physiological effects of commuting using machine learning techniques, with a specific emphasis on analysing the impact of different transportation methods. Healthy individuals were recruited to collect various biological signals, such as blood pressure (BP), heart rate, and electroencephalogram (EEG) data. By leveraging multiple machine learning techniques, we examined the effects of different commuting modes, whether short or long. Our findings revealed an increase in objective bio signals following the commute. Furthermore, when comparing stress levels between different commute modes, we observed that driving is more stressful than other modes, like public transport. We obtained highly encouraging outcomes by implementing the support vector machine (SVM) algorithm, which exhibited an impressive accuracy of 93.2%. In comparison, the K-nearest neighbour (KNN) and Naïve Bayes algorithms yielded good accuracy of 87.9%. Similarly, by utilising the PANAS questionnaire, we observed that the positive affect levels were greater before the commute. This suggests that participants demonstrated a higher degree of positivity and enthusiasm towards their work prior to boarding on their commute.**

Keywords—stress assessment, wearable device, commute, machine learning, biosignals

I. INTRODUCTION

Commuting is considered one of the least enjoyable and stressful activities. Approximately 100 million European Community employees commute daily to and from workplaces [1]. On average, workers in England devote about one hour per day to their daily commute [2]. Additionally, one of every seven workers commutes daily for at least two hours. These statistics highlight the significant time many individuals in England allocate to commuting as part of their daily routines. An examination of the existing data reveals that commuting has experienced a rise in recent decades. This increase can be attributed to the growing ownership of cars, which has led to a dispersed distribution of residential areas and job locations. Travelling to work requires considerable time and effort in our daily work schedules and can affect us physically and mentally. It can diminish our productivity at work and potentially lead to higher absenteeism rates [3]. Commuting plays a crucial role in daily work routines, making it a significant factor in the well-being of the working population. However, surprisingly, limited research has been conducted on its influence on employee productivity [4]. Previous research on employee productivity has predominantly concentrated on examining the effects of sociodemographic and attitudinal factors among employees and organisational cultural and environmental aspects. The impact of commuting on productivity has received relatively less attention in these studies [5].

Long commutes have been linked to adverse impacts on one's mental well-being, including conditions like depression, financial concerns, work-related stress, and sleep disturbances. There are assertions that individuals with lengthy commutes are at a higher risk of experiencing burnout, stress, and illness, which can subsequently lead to increased absenteeism and decreased performance in the workplace. These claims suggest a negative relationship between long commutes, well-being, and job effectiveness [2]. The effect of commuting on your health can be influenced by several factors, including the duration of your commute, the mode of commute, and the prevailing weather conditions you encounter. These factors can collectively affect your well-being and potentially contribute to positive or negative health outcomes from your daily commute [6]. A study investigated the correlation between commuting time and perceived stress among young Korean workers [7]. The researchers used a fixed-effects panel data analysis model to collect longitudinal survey data from Korean youth. The empirical analysis conducted in the study revealed a negative association between commuting time and the perceived stress levels of young Korean workers.

Indeed, stress is recognised as a significant risk factor for various diseases and health conditions. Prolonged or chronic stress can develop or exacerbate hypertension (high blood pressure), cardiovascular diseases, heart attacks, strokes, and other related disorders. Additionally, stress can profoundly impact mental health, leading to psychological disorders like anxiety and depression and influencing behavioural patterns and habits. It is essential to address and manage stress effectively to mitigate its potential negative consequences on physical and mental well-being [6]. Experiencing persistent stress can result in severe psychological and physical disorders [8]. Non-invasive physiological measures like heart rate variability (HRV) and electroencephalograms (EEG) have been employed as reliable methods to assess stress objectively. These methods have proven effective in facilitating stress management strategies in various studies [8]. The impact of stress on an individual can be readily observed through noticeable changes in their behaviour. In addition to visible physical manifestations, the medical field utilises various markers to assess stress, including HRV, respiratory activity, Galvanic Skin Response (GSR), BP, and EEG. These markers gauge and monitor stress levels in medical backgrounds [9]. Also, Psychologists widely employ stress questionnaires as an alternative method to determine patients' stress levels [10].

As our world progresses towards intelligent systems, there is an increasing need for collaboration and knowledge-sharing among individuals. In medical treatment, Artificial Intelligence (AI) has found widespread use in facilitating the development of intelligent systems. We aim to harness the collaboration between AI and medical data analysis to create an intelligent system capable of detecting stress. Through this action, individuals or patients can take proactive measures to prevent the initiation of stress. The primary goal of this study is to examine the physiological consequences of commuting by employing a range of qualitative and quantitative measures. The study also aims to investigate the physiological effects of commuting by applying diverse machine-learning techniques. We compared their positive PANAS scores before and after the commute to assess participants' stress levels. If the participant's positive score decreased after the commute, it was interpreted as an indication that some aspects of the commute had induced stress, reducing the positive PANAS score.

This study investigates the impact of commuting on an individual's physical well-being and develops predictive models using machine learning techniques. Additionally, the study aimed to collect and analyse the experiences of commuters, mainly focusing on utilising emerging computing technologies. The research also sought to establish a living lab for conducting multimodal experiments in body sensors, ubiquitous computing, and wireless telehealth. Traditionally, only medical and physiological professionals can determine if someone is stressed, often relying on questionnaires and participants' self-reported reactions. Creating an intelligent model that utilises individuals' bodily data to automatically evaluate stress levels and mitigate potential health risks, such as high blood pressure. The paper is organised as follows: Section-II Literature Review on Stress, Machine learning algorithms and existing methods. Section-III consists of the methodology of this study. Section-IV is about the implementation and details of the proposed framework. Section-V has the results and discussion. Finally, Section-IV concludes the whole work.

II. Literature Review

Stress is a common component of our daily lives that many individuals experience in different situations. However, enduring long periods or intense stress levels can threaten our overall well-being and disrupt our daily routines. In a study, Ghaderi et al. examined stress levels by analysing data from respiration, HR, facial EMG, and GSR from the foot and hand. Their findings indicated that features related to the respiratory process play a significant role in detecting stress [11]. Maria Viqueira et al. conducted a study on predicting mental stress using dedicated stress-sensing hardware. The research involved Galvanic Skin Response (GSR) as the sole physiological sensor for detecting and monitoring stress levels [12]. David Liu et al. conducted an investigation focused on predicting stress levels using Electrocardiogram data [13]. The stress levels can be determined by employing statistical characteristics of heart rate, Galvanic Skin Response, frequency domain features of heart rate and HRV, as well as the power spectral components of the Electrocardiogram [14]. Several recent studies indicate that heart rate is crucial in stress detection. Stress can influence vital control mechanisms related to stroke volumes, including diastolic and systolic functions. Consequently, it leads to an elevation in heart rate and an increase in blood pressure as a response to stress.

In EEG data classification, the extraction of features remains a crucial step. Typical techniques for extracting features include wavelet analysis, statistical analysis, discrete wavelet transforms, and Fourier analysis. The selected parts can be classified using popular classifiers like Support Vector Machine, K-nearest neighbour, random forest, Artificial Neural Network, Bayesian modelling, and k-means clustering[15].

Rahman et al. provided a detailed summary of the influence of mental workload on mental stress, utilising recorded electroencephalogram (EEG) signals and classification methods to identify and measure levels of mental stress. The study involves collecting EEG signals from individuals with specific cognitive workloads, conducting spectrum analysis, and extracting features for classification using K-Nearest Neighbours. The mean power of the beta band is specifically employed as a feature in the category of mental stress. By combining the mean power of the beta band with the K-nearest neighbour classifier, an important accuracy of 91.26% was achieved, surpassing the performance of the conventional Fuzzy K-Nearest Neighbours (FKNN) classifier and other previously suggested methods for mental stress recognition [16]. In addition, Hu et al. propose a classification methodology that integrates correlation-based features with the k-nearest neighbour data mining algorithm, with a specific emphasis on the valence aspect. The valence aspect was divided into three classes within their approach. The results demonstrate a correct classification rate (CCR) of 83% in recognition of attention using EEG data [17].

Likewise, a study examined the impact of OM mantra meditation on the brain. The objective was to observe delta waves in the brains of 23 participants engaged in meditation. The analysis consisted of utilising Support Vector Machine (SVM) algorithms and extracting statistical features from EEG data to classify the existence of delta band brain waves [18].

A model was created for emotion recognition using EEG signals, which employed the empirical mode decomposition method (EMD). This model aimed to categorise various states of human emotion by applying multiple machine learning algorithms [19].

III. Methodology

In this study, we investigate the impact of commuting using different artificial intelligence techniques based on the commute mode. This study provides a comparison of car and train commuters with multiple indicators of stress. We are applying other Artificial intelligence techniques: KNN, SVM, and Naïve Bayes Classifier to investigate this. These methods were employed to examine and analyse the numerical data obtained from the participants. Additionally, we used the Positive and Negative Affect Schedule (PANAS), a subjective self-report questionnaire, to collect supplementary information.

Data were collected from 45 participants, 27 males and 18 females. These participants are from different parts of London and commute to work regularly. To gather biosignals, we employed non-invasive wearable technologies such as MySignals and MindWave EEG headsets. We collected data on participants' blood pressure, heart rate, and EEG signals for five consecutive days. Additionally, we employed the Positive and Negative Affect Schedule (PANAS) as a subjective self-report questionnaire to evaluate the effects experienced by the participants. The questionnaire mentioned has gained widespread acceptance as a dependable assessment tool for emotional states, applicable in community and clinical settings [20]. Heart rate and blood pressure data were recorded using the MySignals device, shown in Figure 1. The EEG signals of the participants were captured using the NeuroSky MindWave headset, as depicted in Figure 2.

Similarly, different parameters such as height, weight, weight, age, alcohol and smoke status, and weather (temperature) were also collected from the participants. The data collection occurred during the participants' commute from home to the workplace. Recording bio-signals in the study posed no risks or hazards to the participants, ensuring their safety throughout the procedure.

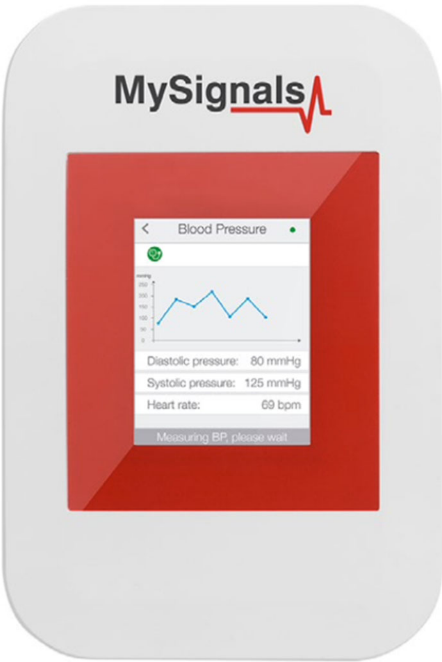

Fig. 1. MySignals Device

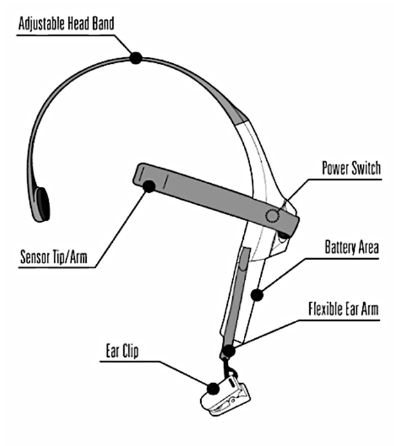

Fig. 2. Electroencephalogram Headset

In this study, two datasets were constructed using the collected participant data. The first dataset consisted of objective parameters: blood pressure, heart rate, and electroencephalogram. The second dataset included these objective parameters (blood pressure, heart rate, and EEG), along with additional variables such as age, gender, height, medication usage, weight, and habits related to smoking and alcohol consumption. Data were pre-processed, selected the relevant features before applying the different machine learning techniques. The complete process is illustrated in Figure 3 below through a graphical representation.

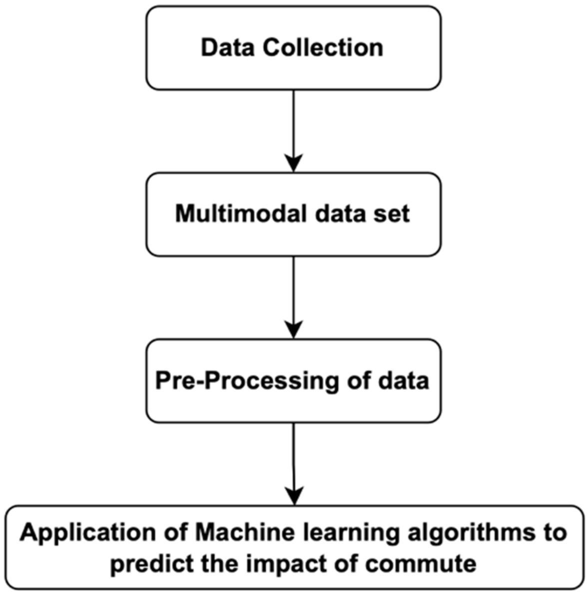

Fig. 3. Stages of this study.

IV. IMPLEMENTATION

This study developed and implemented a data analysis approach based on machine learning. A comprehensive literature review was conducted on the best machine learning algorithm. Support Vector Machine, K-Nearest Neighbour and Naïve Bayes classifier were selected to execute the data analysis. Similarly, confusion matrix Receiver operator characteristic (ROC) curves were used as measures to evaluate the model's performance. ROC curves vividly demonstrate a classification model's remarkable and awe-inspiring performance at every categorisation level, without doubt about its sheer brilliance and effectiveness [21].

Similarly, the performance and efficacy of the selected machine-learning techniques were evaluated and summarised using a confusion matrix [22]. The primary purpose of the confusion matrix was to provide a detailed breakdown, in terms of counts and by class, of the correct and incorrect predictions made by the classification model. By presenting this information, the confusion matrix enables the evaluation of the accuracy and precision of the classification model Likewise, cross-validation was employed to prevent biased decision-making caused by testing the model with the entire dataset [23]. This procedure involves dividing the data into K folds, where K represents the maximum number of possible folds. The test set was employed for each partitioned dataset, while the training set was used for the remaining partitions. This process ensured that each partitioned dataset was tested at least once, and the overall performance was evaluated based on the mean value. Cross-validation demonstrated a decreased level of variability when compared to the conventional approach of splitting the data into separate training and test sets.

V. Results and Discussion

A. *Using only the main parameters: Blood pressure, Heart rate and EEG*

1. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a highly regarded machine learning algorithm appreciated for its competence in addressing regression and classification tasks. After employing this algorithm to classify the data, we achieved an impressive accuracy rate of 90%. Out of a total of 161 predicted values, 146 were correctly identified. Likewise, in the case of the second class, there were only four misclassifications out of 29 instances. The confusion matrix depicted in Figure 4 below illustrates the model's performance.

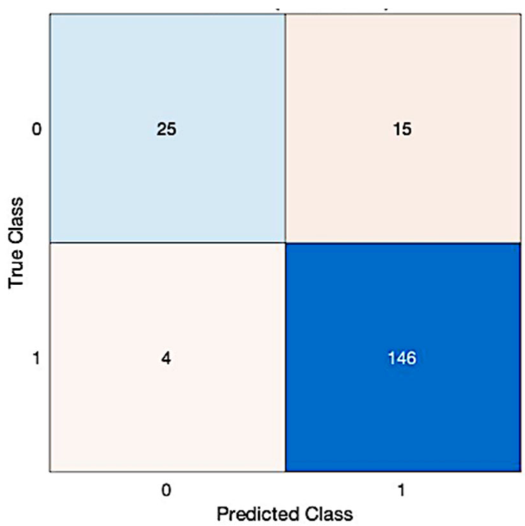

Fig. 4. Confusion matrix of Support Vector Machine for the first dataset.

Similarly, to evaluate the effectiveness of the algorithms, ROC curves were plotted, and the performance of the SVM is depicted in Figure 5 below.

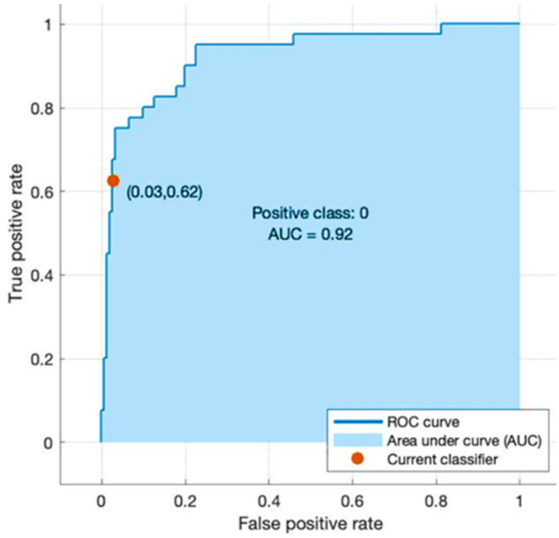

Fig. 5. ROC curve of Support Vector Machine for the first dataset.

2. K-Nearest Neighbour

KNN (K-Nearest Neighbours), a supervised learning technique, is extensively used in regression and classification problems. It operates by categorising a given input by considering its neighbouring data points. As a non-parametric model, KNN calculates the distance between the given data point and each training input. Based on this distance calculation, it then assigns the test input to the class of its K-nearest neighbours. KNN is commonly employed in classification and regression tasks to classify information based on its nearest neighbour. It calculates the distance between the given data point and all training inputs, ultimately assigning the test input to the class of its K-nearest neighbour. Using this algorithm, we obtained an accuracy of 84.7%. As illustrated in Figure 6 below, the classifier only misclassified 27 out of 175 in the first class, while 13 out of 15 values were correctly classified. Additionally, Figure 7 presents the performance of KNN utilising the ROC curve.

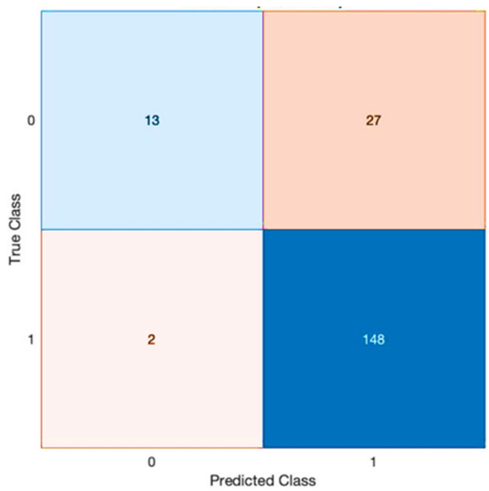

Fig. 6. Confusion Matrix of KNN for the first dataset.

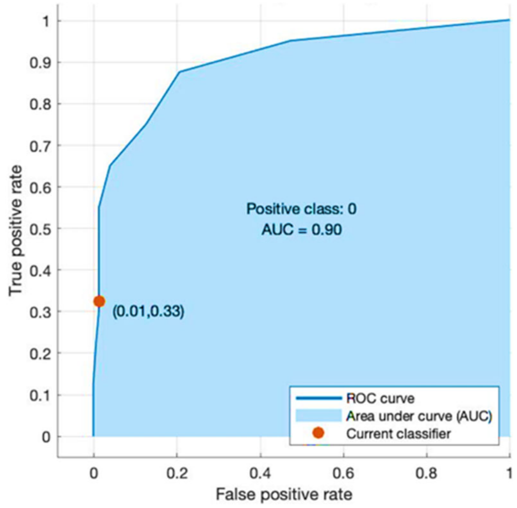

Fig. 7. ROC curve of KNN for the first dataset.

3. Naive Nayes Classifier

The naive Bayes classifier is a supervised learning technique that applies Bayes' theorem to make predictions. It operates under the assumption that the presence of one feature in a specific class is independent of the existence of other elements in different courses. The naive Bayes classifier calculates probabilities for all occurrences without considering their interdependencies. This approach is beneficial when dealing with datasets with multiple features and high dimensionality. In our study, we employed the Gaussian naive Bayes algorithm, which is well-suited for handling continuous data that follows a normal distribution, such as age, height, and weight. The accuracy achieved with this algorithm on the first dataset was 84.2%. Figure 8 illustrates the algorithm's performance using a confusion matrix, while Figure 9 presents the ROC curve, providing a graphical representation of the algorithm's performance.

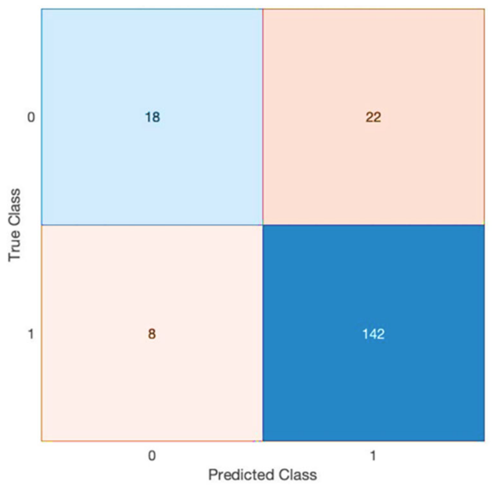

Fig. 8. Confusion matrix of Naïve Bayes Classifier for the first dataset.

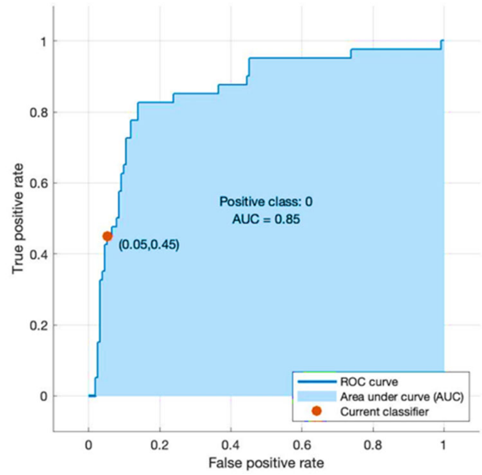

Fig. 9. Performance of Naïve Bayes using ROC curve.

B. Using main parameters (BP, HR, EEG) and other parameters

Similarly, we applied all the selected machine learning algorithms to the second dataset to develop a model that predicts the physiological effects of commute. This comprehensive dataset encompasses essential parameters such as heart rate, BP, and electroencephalogram, along with subjective factors including age, height, weight, gender, medication usage, smoking and alcohol habits, location, weather temperature, and comprehensive medical health information. By considering both objective and subjective aspects, our model aimed to provide a holistic understanding of the potential impact of commuting on physiological well-being. Adding personalised parameters to the first dataset increased the model's performance. We obtained highly encouraging outcomes by implementing the Support Vector Machine (SVM) algorithm, which exhibited an impressive accuracy of 93.2%. In comparison, the K-nearest neighbour (KNN) and Naïve Bayes algorithms yielded good accuracy of 87.9%. These results underscore the efficacy of SVM in capturing patterns and making accurate predictions, positioning it as a promising approach in our study.

Similarly, the classifier performance evaluation, including the accuracy of all classifiers, is effectively represented through a confusion matrix. This matrix enables a visual representation of the comparison between the actual and predicted classes, as shown in the table presented in Table 1 below. This comprehensive evaluation provides valuable insights into the classification accuracy attained by each classifier.

TABLE I. COMPARISON OF THE PERFORMANCE OF DIFFERENT MACHINE LEARNING METHODS.

Technique	Dataset	Accuracy
SVM	First	90.00%
	Second	93.20%
KNN	First	84.70%
	Second	87.90%
Naïve Bayes	First	84.20%
	Second	87.90%

VI. CONCLUSION

In this study, we used various artificial intelligence techniques to develop a classification model for predicting the impact of commuting. Two datasets were used, one with essential parameters like blood pressure, heart rate, and electroencephalogram data and the other incorporating personalised variables. We examined the effects of different commuting modes and found an increase in objective bio signals following the commute. Passive commuters, specifically those who drive, exhibited elevated stress levels in comparison to active commuters. The result obtained from this study suggested that driving is a more stressful commute mode compared to other commuting methods, like public transport.

Similarly, the support vector machine showed the best performance on both datasets, with an accuracy of 90% and 93.20% for the first and second datasets, respectively. The performance of the model improved when personalised parameters were incorporated into the dataset for all machine learning techniques. Likewise, through the PANAS questionnaire, we observed that the level of positive effect decreased after the commute. This indicates that participants exhibited higher positivity and enthusiasm towards their work prior to starting their commute. Conversely, negative affect showed an increase suggesting a decrease in interest or an elevation in stress levels among the participants. The objective findings derived from our machine learning-based approach were in line with the subject alignment obtained from the PANAS assessment, providing additional support and validation for our findings.

In our future work, we plan to expand the scope of the designed model to encompass a broader range, including London and other cities. This will allow for a more comprehensive understanding of the impact of commuting across different urban areas

ACKNOWLEDGEMENT

The authors express their sincere gratitude to the School of Computing and School of Psychology at the University of East London and to all those who provided support and assistance during this research.

REFERENCES

[1] Costa, G., Pickup, L. and Di Martino, V., 1988. Commuting--a further stress factor for working people: evidence from the European Community. II. An empirical study. International archives of occupational and environmental health, 60(5), pp.377-385.

[2] Chatterjee, K., Chng, S., Clark, B., Davis, A., De Vos, J., Ettema, D., Handy, S., Martin, A. and Reardon, L., 2020. Commuting and well-being: a critical overview of the literature with implications for policy and future research. Transport reviews, 40(1), pp.5-34.

[3] Ma, L. and Ye, R., 2019. 'Does daily commuting behavior matter to employee productivity?'. *Journal of Transport Geography*, 76, pp.130-141.

[4] Liao, H. and Chuang, A., 2004. A multilevel investigation of factors influencing employee service performance and customer outcomes. Academy of Management journal, 47(1), pp.41-58.

[5] Evans, G.W., Wener, R.E. and Phillips, D., 2002. The morning rush hour: Predictability and commuter stress. Environment and behavior, 34(4), pp.521-530.

[6] E. T. Attar, V. Balasubramanian, E. Subasi and M. Kaya, "Stress Analysis Based on Simultaneous Heart Rate Variability and EEG Monitoring," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 9, pp. 1-7, 2021, Art no. 2700607, doi: 10.1109/JTEHM.2021.3106803.

[7] Jin, E., Kim, D. and Jin, J., 2022. Commuting time and perceived stress: Evidence from the intra-and inter-city commuting of young workers in Korea. *Journal of Transport Geography*, *104*, p.103436.

[8] Ahn, J.W., Ku, Y. and Kim, H.C., 2019. A novel wearable EEG and ECG recording system for stress assessment. Sensors, 19(9), p.1991.

[9] Pramanta, S.A., Prihatmanto, A.S. and Park, M.G., 2016, October. A study on the stress identification using observed heart beat data. In 2016 6th International Conference on System Engineering and Technology (ICSET) (pp. 149-152). IEEE.

[10] Sano, A. and Picard, R.W., 2013, September. Stress recognition using wearable sensors and mobile phones. In 2013 Humaine association conference on affective computing and intelligent interaction (pp. 671-676). IEEE.

[11] Ghaderi, A., Frounchi, J. and Farnam, A., 2015, November. Machine learning-based signal processing using physiological signals for stress detection. In 2015 22nd Iranian Conference on Biomedical Engineering (ICBME) (pp. 93-98). IEEE.

[12] Villarejo, M.V., Zapirain, B.G. and Zorrilla, A.M., 2012. A stress sensor based on Galvanic Skin Response (GSR) controlled by ZigBee. Sensors, 12(5), pp.6075-6101.

[13] Liu, D. and Ulrich, M., 2014. Listen to your heart: Stress prediction using consumer heart rate sensors. Final. Proj. Standford CS, 229, pp.229-233.

[14] Sriramprakash, S., Prasanna, V.D. and Murthy, O.R., 2017. Stress detection in working people. Procedia computer science, 115, pp.359-366.

[15] O. Bazgir, Z. Mohammadi and S. A. H. Habibi, "Emotion Recognition with Machine Learning Using EEG Signals," 2018 25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME), 2018, pp. 1-5, doi: 10.1109/ICBME.2018.8703559.

[16] Rahman, Tatiur & Ghosh, Apu & Shuvo, Md. Maruf Hossain. 2015. Mental Stress Recognition using K-Nearest Neighbor (KNN) Classifier on EEG Signals.

[17] B. Hu, X. Li, S. Sun and M. Ratcliffe, 2018. "Attention Recognition in EEG-Based Affective Learning Research Using CFS+KNN Algorithm," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 1, pp. 38-45, 1 Jan.-Feb. doi: 10.1109/TCBB.2016.2616395.

[18] B. P. Harne, Y. Bobade, R. S. Dhekekar and A. Hiwale, "SVM classification of EEG signal to analyse the effect of OM Mantra meditation on the brain," *2019 IEEE 16th India Council International Conference (INDICON)*, Rajkot, India, 2019, pp. 1-4, doi: 10.1109/INDICON47234.2019.9030339.

[19] M. Degirmenci, M. A. Ozdemir, R. Sadighzadeh and A. Akan, "Emotion Recognition from EEG Signals by Using Empirical Mode Decomposition," *2018 Medical Technologies National Congress (TIPTEKNO)*, Magusa, 2018, pp. 1-4, doi: 10.1109/TIPTEKNO.2018.8597061.

[20] Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the panas scales. Journalof personality and social psychology 54, 1063.

[21] Kumar, R. and Indrayan, A., 2011. Receiver operating characteristic (ROC) curve for medical researchers. Indian pediatrics, 48, pp.277-287.

[22] Patro, V.M. and Patra, M.R., 2014. Augmenting weighted average with confusion matrix to enhance classification accuracy. Transactions on Machine Learning and Artificial Intelligence, 2(4), pp.77-9.

[23] Malhotra, R. and Meena, S., 2021, August. Empirical validation of cross-version and 10-fold cross-validation for defect prediction. In 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 431-438). IEEE.