Machine Learning-Based Neuroergonomic Stress Quantification of Factory Workers in Human–Robot Collaboration Environments Using Multimodal Data

Arshia Arif *, Zohreh Zakeri, Ahmet Omurtag, Philip Breedon, Azfar Khalid

Posted Date: 7 April 2026

doi: 10.20944/preprints202604.0485.v1

Keywords: machine learning; mental stress quantification; human robot collaboration (HRC); EEG; fNIRS; occupational health and safety

 doi:10.20944/preprints202604.0485.v1

Article

Machine Learning-Based Neuroergonomic Stress Quantification of Factory Workers in Human–Robot Collaboration Environments Using Multimodal Data

Arshia Arif *, Zohreh Zakeri, Ahmet Omurtag, Philip Breedon and Azfar Khalid

Department of Engineering, School of Science & Technology, Nottingham Trent University, Clifton, Nottingham NG11 8NS, UK

* Correspondence: arshia.arif2021@my.ntu.ac.uk

Abstract

Mental stress is a common issue in demanding occupational setups, such as smart industrial settings, particularly from working with robots, being one of the primary reasons for decreased performance and productivity. Quantifying and evaluating stress are critical for worker safety, performance, and overall well-being. A novel integrated framework is proposed in this research for digitising and assessing cognitive stress that combines neuroimaging (EEG and fNIRS), gaze tracking and machine learning. A factory workers' stress assessment experiment is designed and implemented, which employs physiological, behavioural and subjective measures to assess mental stress from different perspectives. Physiological features extracted from multimodal data are used for training supervised classification and regression models. To further optimise the pipeline, multiple feature selection algorithms are tested, followed by ensemble learning approaches, and the best one is chosen for stress prediction. After implementing the novel stress quantification framework for the factory workers' stress assessment experiment, the ensemble learning approach produced the best results for both regression (RMSE: 10.86) and classification (accuracy: 84.1%) techniques using the STAI score as the target. The behavioural and subjective measures demonstrate the effect of varying process variables (light, noise, task speed, and complexity) during the experiment. Multimodal data, machine learning, and other computational approaches are integrated in this study to objectively quantify cognitive stress, utilising the novel stress quantification framework presented in this research, thereby bridging the gap between research and practical application. This study proposes a multi-domain framework for measuring stress, providing a promising solution for worker well-being in occupational setups.

Keywords: machine learning; mental stress quantification; human robot collaboration (HRC); EEG; fNIRS; occupational health and safety

1. Introduction

Digitisation and automation have reshaped manufacturing systems, leading to the smart factories where Human–Robot Collaboration (HRC) is critical for optimum and desired productivity, flexibility, and quality standards [1]. In these settings, although combining human cognitive strengths with robotic precision offers significant benefits [2], constant interaction with smart machines and adapting to dynamic robot behaviour can also create psychological challenges [3]. These factors may increase cognitive stress among workers in human-robot collaborative environments [3,4].

Technostress, mental strain caused by rapid technological change [5], is particularly evident in human–robot collaborative environments. Although cobots are designed for safety and efficiency [6,7], their close proximity can pose psychological pressure [4]. Continuous monitoring of robot behaviour, fear of unexpected malfunctions, and the need to synchronise with automated systems

 doi:10.20944/preprints202604.0485.v1

can induce anxiety. Trusting complicated automated systems for personal safety and productivity can cause mental overload, fatigue, and heightened stress, affecting workers' cognitive health and job satisfaction [8].

In smart factories, incorporating neuroergonomics and human factors is crucial to create human-centred, stress-free workspaces for employees collaborating with cobots [9,10]. However, the rapid evolution of smart manufacturing has raised concerns about worker health, safety, and prevalent ergonomic problems [11]. Improving employee well-being remains an ongoing challenge for both SMEs and large industrial organisations [12].

Conventionally, mental stress has been evaluated using subjective measures, such as self-report questionnaires and rating systems [13–16] and behavioural measures, including task accuracy, reaction time, speech features, facial expressions, and gaze behaviour [17–20]. While these measures provide useful insights, physiological measures offer a more objective assessment by recording stress-induced bodily changes [21,22]. Common physiological indicators include heart rate variability (HRV) [23,24], cortisol levels [25], electrodermal activity (EDA), body temperature, and ocular metrics such as pupil dilation and blink rate [25–27]. Advances in sensing technology and data analytics have facilitated the integration of physiological signals, such as electrocardiography (ECG) [28], facial expressions, and speech features [29], into multimodal stress-evaluation frameworks, often combined with machine learning for continuous, real-time monitoring [30,31]. Among these indicators, brain signals, including electroencephalography (EEG) [32], functional magnetic resonance imaging (fMRI) [33]and functional near-infrared spectroscopy (fNIRS) [34] are important for understanding the neurological mechanisms underlying mental stress.

In this study, an experiment is designed to quantify factory workers' cognitive stress using EEG and fNIRS neuroimaging, along with gaze data such as pupil diameter. Error rate is used as the behavioural indicator, while the NASA Task Load Index (NASA-TLX) [14] and State-Trait Anxiety Inventory (STAI) [15] scores provide subjective assessments. Additional gaze metrics, including fixation duration, number of visits, and visit duration, reflect behavioural responses.

Existing literature shows significant efforts to detect and analyse mental stress; however, to the best of the authors' knowledge, cognitive stress has not yet been quantified as a continuous, measurable variable. Most prior studies rely on qualitative assessments, self-reports, or observational methods, and recent advances mainly classify stress into binary or multiple categories rather than tracking real-time stress intensity as a quantifiable and measurable factor. Therefore, this research aims to address this gap by developing a novel, comprehensive framework that quantifies cognitive stress and provides meaningful insights for both employers and employees.

This study aims to quantify factory workers' cognitive stress within a human–machine collaborative environment by simulating a production line and using multidimensional data, including behavioural, physiological, and subjective parameters. A novel stress-quantification framework is proposed, representing one of the first studies to measure mental stress as a measurable variable rather than in categorical levels. Another novelty of this study is that the feature selection followed by an ensemble learning approach has been employed towards the end of the stress quantification pipeline to obtain a final model with improved accuracy (for the classification approach) and lower RMSE (for the regression approach). This approach has also reduced the bias for both classification and regression approaches. The interpretations from this research can be applied to any workspace where humans and cobots coexist.

2. Methodology

This study uses an eight-step pipeline (Figure 1) to digitise and quantify cognitive stress. The process begins with collecting subjective, behavioural, and physiological data during stress-inducing tasks that mimic real workplace situations. The data preprocessing is then done to remove noise and artefacts, followed by the extraction of statistical, frequency-domain, and time-domain features. A key novelty of the framework is the mapping of subjective stress scores to physiological features to

 doi:10.20944/preprints202604.0485.v1

generate labelled data for supervised learning. Feature selection is applied to reduce dimensionality and enhance model efficiency, after which multiple machine-learning models are trained and tested to identify the best predictors of stress. Finally, ensemble learning is employed to improve robustness, accuracy, and reduce overfitting across models.

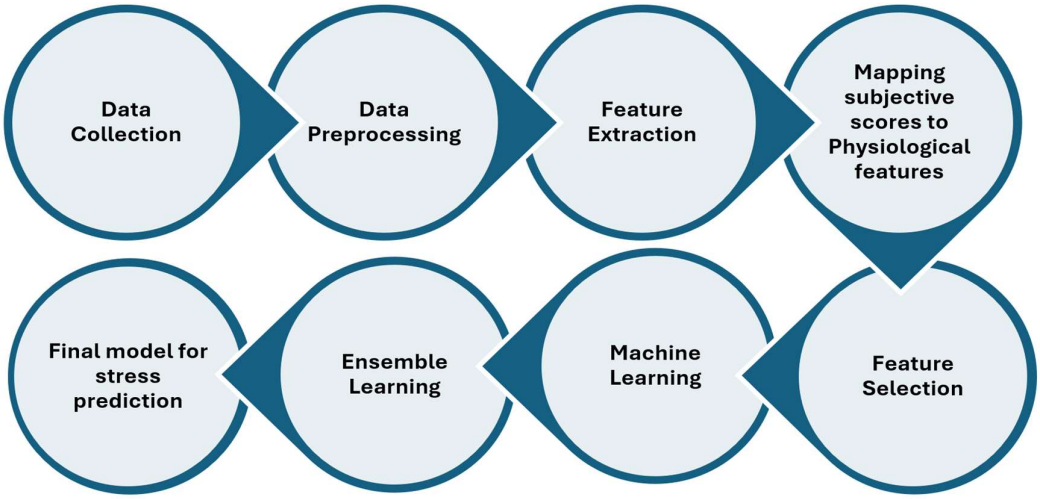


Figure 1. Overview of the proposed stress quantification framework, illustrating the main processing stages.

The ensemble model is ultimately tested and chosen as the framework's output, providing a predictive model that quantifies mental stress using multidimensional data. This model supports practical applications in real-time occupational stress monitoring. Overall, the pipeline not only enables stress quantification but also contributes to mental health technology by incorporating machine intelligence into mental state analysis.

2.1. Flow of Experiment

In this experiment, crisp packets are delivered to the participant on a conveyor belt, each displaying a cognitive question. The participant picks up the packets, answers the questions, and drops them into a box. Once the required number of packets is collected, the box is placed in front of the cobot, which completes the packing process before the participant places it on the pallet. As illustrated in Figure 2, this workflow simulates a realistic production scenario. Cognitive stress is evaluated by varying task complexity and speed, noise level, and light intensity.

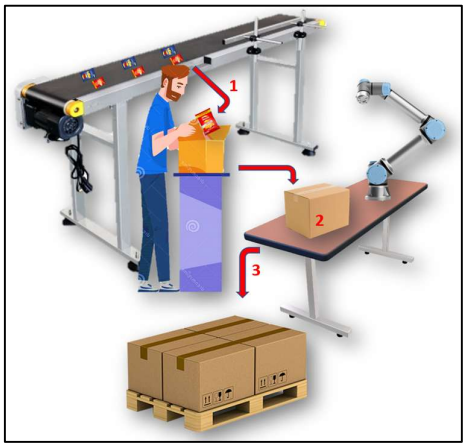


Figure 2. The flow of the factory workers' stress assessment experiment.

 doi:10.20944/preprints202604.0485.v1

A secondary task is introduced to increase task complexity. The participant receives a visual stimulus in the form of a damaged packet and must separate these from the intact ones. Damaged packets must not be placed in the box; doing so is counted as an error.

2.2. Experimental Paradigm

The experiment is designed to record participants' brain activity continuously for 60 minutes. Before data collection, participants receive instructions, provide informed consent, and have all devices calibrated and fitted. The session begins with a two-minute rest period, followed by the first task episode of approximately 3.5 minutes. After each episode, participants complete the NASA-TLX and STAI questionnaires. As shown in Figure 3 and detailed in Table 1, each episode is conducted under predefined low or high parameter settings. In the ninth episode, the cobot is replaced by a human operator to compare mental stress during human–robot and human–human interaction.

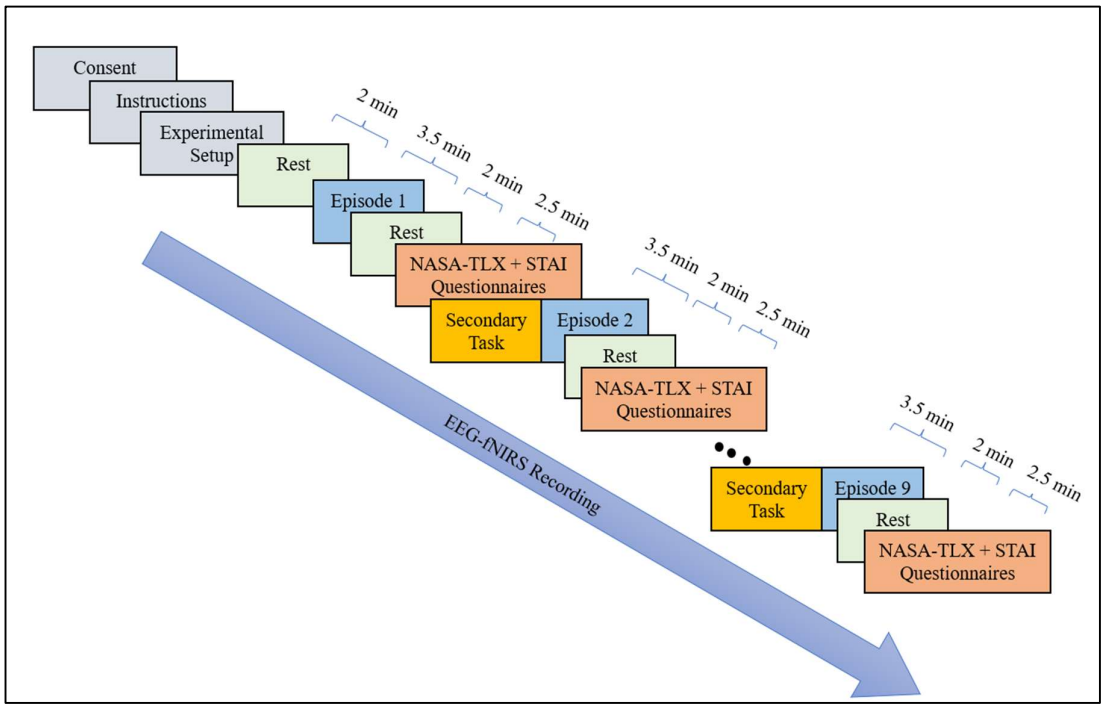


Figure 3. Experimental paradigm for the factory workers' stress assessment experiment, demonstrating 9 task episodes (3.5 minutes each) with rest episodes (2 minutes each) and questionnaire-filling periods (2.5 minutes each) in between.

 doi:10.20944/preprints202604.0485.v1

Table 1. Experimental episodes for the factory workers' stress assessment experiment with 'H' and 'L' indicating high and low levels, respectively, for each process variable. Light levels of 50–60 lux are considered low, and levels above 250 lux are considered high [35]; noise below 55 dB is classified as low, while 80–85 dB is considered high [36].

Episode No.	Task speed and complexity	Noise level	Light intensity
1	L	L	L
2	L	L	H
3	L	H	L
4	L	H	H
5	H	L	L
6	H	L	H
7	H	H	L
8	H	H	H
9 (Sub-episode: Human-human interaction)	H	H	H

2.3. Data Acquisition

EEG and fNIRS data are gathered to measure neural and hemodynamic activity, respectively. EEG signals are recorded using the TMSi Saga wireless system at a 4000 Hz sampling rate [37], with 19 electrodes positioned according to the international 10–20 system. fNIRS data is acquired using the Artinis Brite device with 21 channels at a 75 Hz sampling frequency [38], with transmitter–receiver pairs spaced 20–30 mm apart. Gaze tracking is conducted throughout the experiment using Tobii Pro Glasses 3, recorded at a sampling frequency of 50 Hz and 25 fps with a 1920×1080 resolution [39]. Subjective measures are recorded after each episode using the NASA-TLX and STAI questionnaires.

2.4. Data Pre-Processing

Both EEG and fNIRS data are pre-processed to obtain clean signals (Figures 4 and 5). EEG recordings are first cleaned using an ICA-based method to remove artefacts and non-brain activity, then down-sampled to 250 Hz and filtered with a 0.5–40 Hz band-pass filter to eliminate drift and high-frequency noise. The EEG frequency bands considered include delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–28 Hz), and gamma (28–50 Hz) [40].

 doi:10.20944/preprints202604.0485.v1

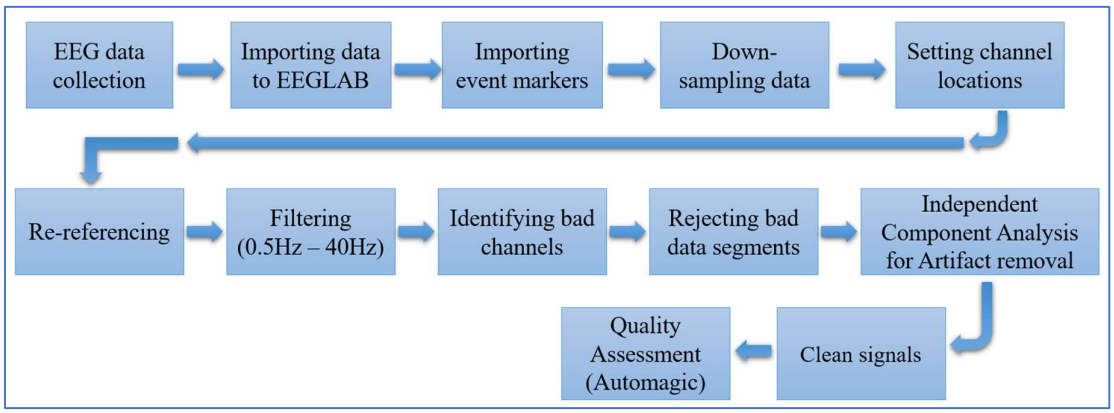


Figure 4. EEG data preprocessing.

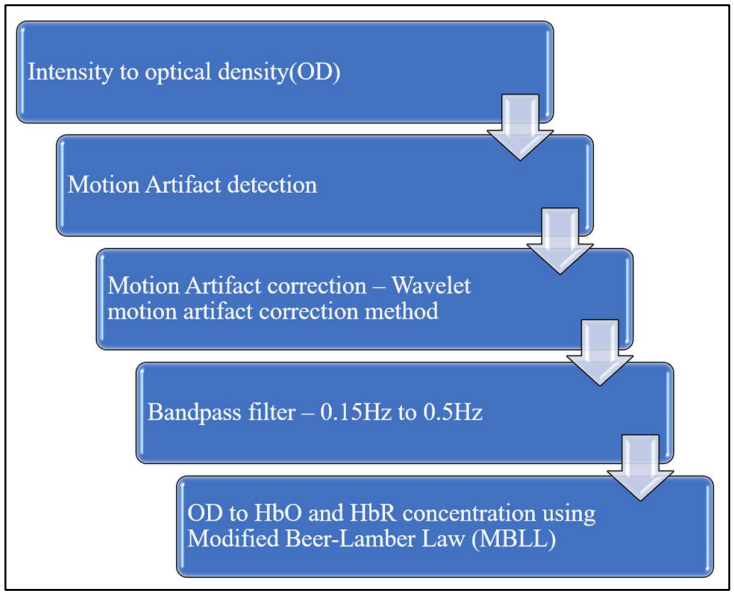


Figure 5. fNIRS data preprocessing.

To obtain artefact-free fNIRS signals, the intensity is first transformed to optical density (OD). Motion artefacts are then rectified using a wavelet-based motion artefact correction approach, after which each channel is filtered between 0.15 and 0.5 Hz. Finally, the Modified Beer-Lambert law is applied to convert OD to HbO and HbR concentrations [41].

2.5. Feature Extraction

Features are extracted from the pre-processed EEG and fNIRS signals. For EEG, spectral features are computed by normalising each clean signal and applying five Kaiser-window-based FIR filters to generate the standard EEG sub-bands (Figure 6) [42]. The average band powers for five frequency bands are then calculated and used as spectral features.

 doi:10.20944/preprints202604.0485.v1

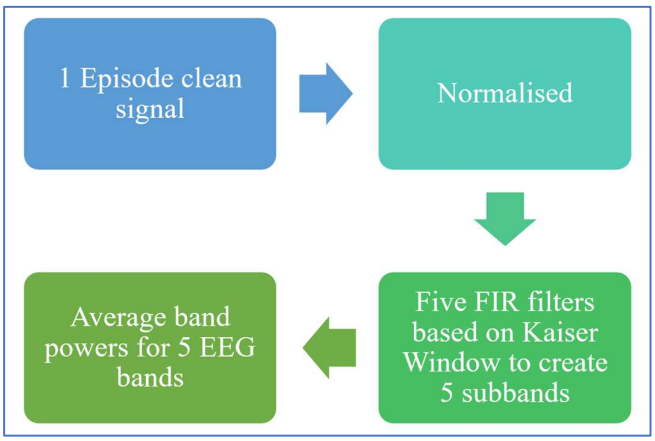


Figure 6. EEG spectral features extraction.

Several statistical features, including mean, standard deviation, variance, kurtosis, skewness, root mean square, shape factor, impulse factor, first and second differences and their normalised forms, are extracted from the EEG data [43,44]. For fNIRS signals, both linear features (mean, skewness, kurtosis, standard deviation, variance) and nonlinear features (entropy, fractal dimension (box-count), and detrended fluctuation analysis (alpha)) are computed [45].

2.6. Data Selection

For EEG data, all recordings are initially imported into the Automagic toolbox [46], where signal quality is assessed using four measures: the ratio of bad channels, overall high-amplitude data (OHA), timepoints of high variance (THV), and channels of high variance (CHV). Using amplitude and standard-deviation thresholds of 30 μV, found suitable for this dataset, the toolbox classifies each recording as *Good*, *OK*, or *Bad*. This evaluation enables consistent comparison across recordings and supports logical inclusion or exclusion of data in subsequent analysis.

The pertinent EEG data is then selected by visually examining the clean data. The following criteria are applied:

- Only data labelled “Good” or “OK” by the Automagic toolbox are retained, as illustrated in Figure 7.
- Only those participants’ data are eliminated where all episodes are marked as “Bad”.
- Noisy channels identified through visual inspection are removed.
- Only the bad channels’ feature data is excluded instead of discarding an entire episode.

For fNIRS data, the following criteria are adopted to select the relevant data:

- Bad channels affecting signal quality are identified through visual inspection.
- Only the features from bad channels are removed, rather than discarding entire episodes.
- Episodes with brief signal-loss or connection dropouts are excluded.

 doi:10.20944/preprints202604.0485.v1

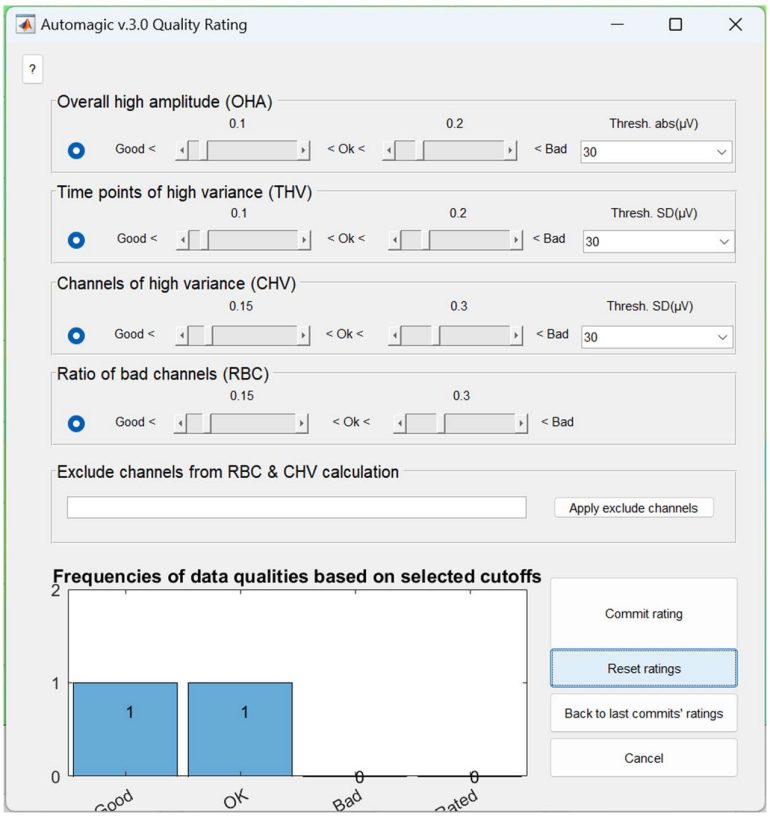


Figure 7. Automagic toolbox for data quality assessment of EEG signals [46].

A few measures are taken to further improve the outcomes:

- Rejecting the first 5 seconds from all episodes, as the participant is not performing any cognitive task during this period, and only the conveyor belt is in motion to deliver the crisp.
- An undefined gap, generated in the feature matrix due to rejecting bad channels, is filled in using the averaging approach.
- Min–max normalisation [47] is applied to scale each feature in the matrix between 0 and 1, using the formula:

$$Z_{norm} = \frac{Z - Z_{min}}{Z_{max} - Z_{min}} \quad (1)$$

Machine learning approaches, especially those relying on distance or gradient calculations, can be more biased toward features with wider numerical ranges; however, normalisation ensures that each feature contributes equally [48].

- The ensemble model approach, covered in more detail in section 2.10, is used to determine the optimal feature combination.

2.7. Data Labelling

Mental stress is quantified by mapping the feature matrix to two target scores: NASA-TLX (0–100) [49] and STAI (1–4) [50]. These scores serve as target variables for supervised machine-learning models in both regression and classification approaches. For regression, the feature matrix is labelled with the NASA-TLX score and the STAI score is rescaled to 0–100 to ensure comparability. For classification, NASA-TLX scores are grouped into three stress levels (Table 2). STAI scores are first mapped from 1–4 to 20–80 and then divided into three anxiety levels, as shown in Table 3 [45].

 doi:10.20944/preprints202604.0485.v1

Table 2. Levels of stress based on NASA-TLX score.

NASA_TLX Score	Level of Stress
0-50	Low
50-80	Moderate
80-100	High

Table 3. Levels of anxiety based on STAI score.

STAI Score	Level of Anxiety
20-37	Low
38-44	Moderate
45-80	High

2.8. Machine Learning for Mental Stress Prediction

Machine learning algorithms are used to predict cognitive stress and anxiety based on physiological and subjective indicators. Both regression and classification approaches are applied, using 70% of the data for training and 30% for testing, with 5-fold cross-validation to reduce overfitting. As shown in Figure 8, the workflow includes data collection, feature extraction, model selection, training, testing, prediction, and optimal model deployment. Seven regression models, tree, linear regression, including SVM, Gaussian process regression, ensemble, neural networks, and kernel models, are evaluated. For classification, eight techniques are tested, including linear regression, SVM, tree, ensemble, Naïve Bayes, neural networks, kernel, and K-nearest neighbours (KNN). Additionally, other variants of these primary regression and classification methods are examined.

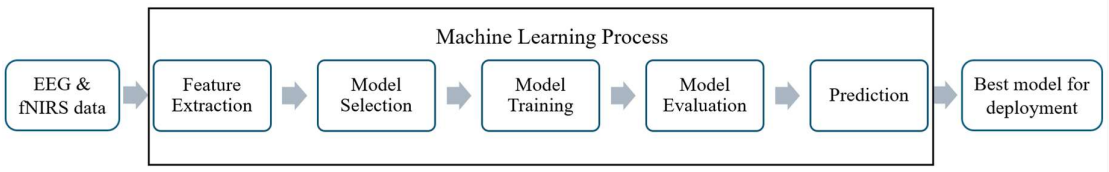


Figure 8. Machine learning approach for predicting conventional measures employing physiological parameters.

A range of EEG (spectral and statistical) and fNIRS (linear and nonlinear) feature combinations is used to predict both NASA-TLX and STAI targets. All regression and classification models are tested across multiple predictor sets, using accuracy as the metric for classification and RMSE for regression. For each set, the model achieving the lowest RMSE or highest accuracy is selected, and the best feature combination is selected for each target.

2.9. Feature Selection

Feature selection is crucial for developing effective machine learning models, as it identifies the most informative features while removing redundant ones. This process enhances accuracy, minimises overfitting, speeds up computation, and improves interpretability, especially important for dealing with high-dimensional datasets. In this study, ReliefF [51] and the F-test [52] are used for

 doi:10.20944/preprints202604.0485.v1

regression, while Kruskal–Wallis [53], ANOVA [54], Chi² [55], and ReliefF [56], are applied for classification.

2.10. Ensemble Learning Approach

Ensemble learning is a machine learning approach in which multiple base or weak learners are trained and their outputs combined to improve performance [57], as illustrated in Figure 9. By aggregating diverse models, ensemble methods typically outperform individual models in accuracy, robustness, and resistance to overfitting. Although each model has its own advantages and disadvantages, combining them produces more stable predictions and reduces bias and variance [58]. In this study, four ensemble variants, averaging, weighted averaging, majority voting, and weighted majority voting, are employed to identify the best-performing configuration.

These ensemble techniques, including averaging, weighted averaging, majority voting, and weighted majority voting, improve model performance by utilising the strengths of several base learners. The detailed results of the mental stress quantification framework applied to the stress assessment experiment for the factory workers are discussed in Section 3.

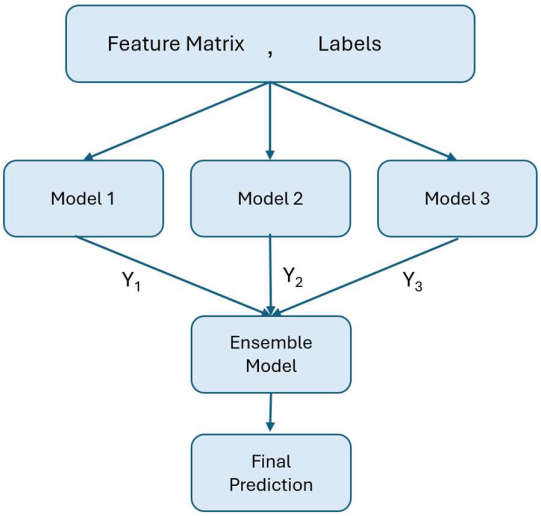


Figure 9. Ensemble learning approach.

2.11. Stress Score Prediction and Visualisation

Another dimension of this study's scope is the development of a digital solution in the form of an Android application, *Stress Sense*, to monitor workers' stress levels, enabling preventive measures and seamless communication between workers and supervisors. For this purpose, the mental stress score is predicted using the best-performing machine learning model. The predicted stress score is translated into a visual stress scale, which is integrated within the Android app to support mental stress interpretation.

3. Results and Discussion

The suggested stress-quantification framework is demonstrated through a factory workers' stress assessment experiment, where participants perform mentally challenging tasks alongside a pick-and-place activity in a smart factory setting, and multidimensional data have been gathered. The goal is to identify critical stress indicators and predict mental stress in a digitised format using machine learning to help employers and employees better understand employees' mental condition.

 doi:10.20944/preprints202604.0485.v1

3.1. Regression

The regression approach was initially applied to two datasets, one using the complete feature matrix (EEG and fNIRS) and another with data rejection based on data selection criteria, to compare the results and observe the effects of data selection. Using NASA-TLX and STAI scores as target variables, this comparison has enabled evaluation of how selective data inclusion influences the accuracy and reliability of regression results.

Initially, regression is applied separately to EEG spectral features, EEG statistical features, and fNIRS linear and nonlinear features, followed by combined EEG and fNIRS predictors for both targets. All seven regression techniques are tested for each feature set, and the best-performing model is reported. Using the entire dataset without data rejection, the ensemble model achieves the lowest validation RMSE (18.48) and a test RMSE of 13.59 for the STAI score as the target, as shown in Figure 10.

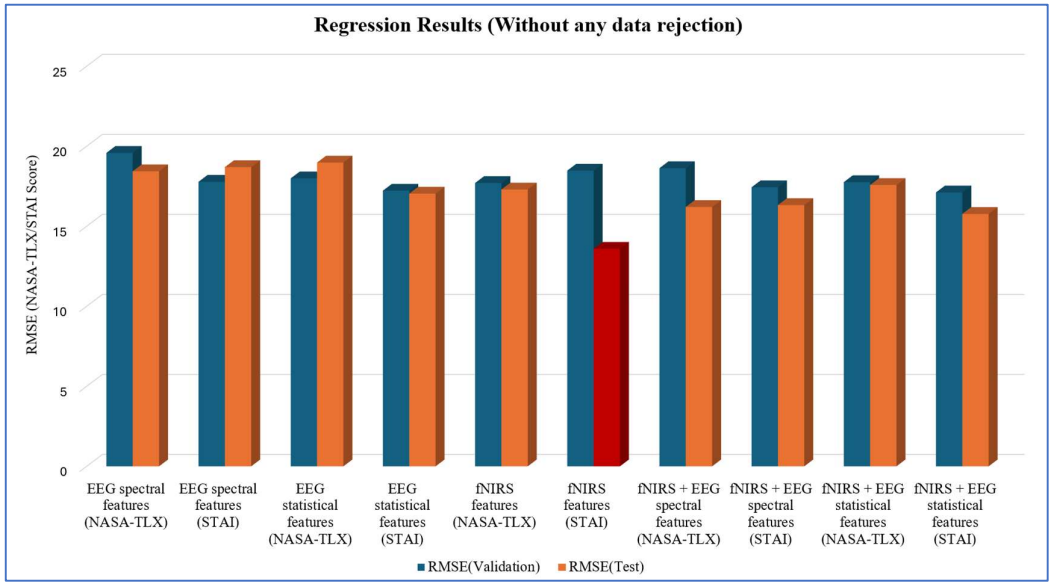


Figure 10. Regression results for the dataset without any data rejection. NASA-TLX and STAI scores are used as labels. The red bar illustrates the best-performing predictor set.

Figure 10 indicates that fNIRS features alone achieve the best performance for the full, unrejected dataset, with the ensemble model performing optimally and highlighting the significance of linear and nonlinear fNIRS features. Combining fNIRS with EEG statistical features shows the best performance for the STAI score as the target, emphasising the importance of EEG statistical predictors. In contrast, EEG spectral features consistently produce higher RMSE values for NASA-TLX, while fNIRS features provide low RMSE for the STAI score as the target.

After applying data-selection criteria, multiple models are tested using EEG-only, fNIRS-only, and combined feature sets for both targets, NASA-TLX and STAI scores. The best performance is achieved by combining EEG statistical and fNIRS features, with the ensemble model (Figure 11) reaching a validation RMSE of 16.22 and a test RMSE of 16.18 for the STAI score as the target. Overall, EEG spectral features continue to show higher RMSE values than statistical features across both targets.

In the unrejected dataset, most feature sets produce a minimum test RMSE of around 18, with the best case reaching approximately 14. However, this does not indicate that the no-rejection technique is superior. Noise and artefacts can deceptively boost performance and reduce real-world generalisability. Data selection ensures cleaner inputs, lowers bias from poor-quality segments, and yields more consistent performance across feature sets [59]. Although the test RMSE is slightly higher after data selection, the reduced gap between validation and test suggests improved robustness and

 doi:10.20944/preprints202604.0485.v1

reduced overfitting [60]. Overall, data selection is essential for obtaining reliable and generalisable regression results.

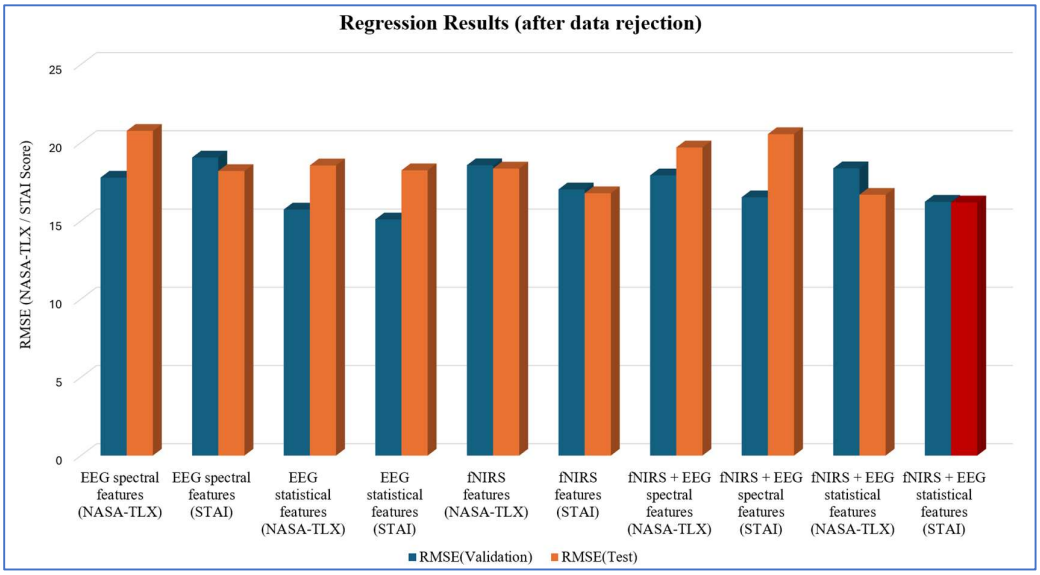


Figure 11. Regression Results for the dataset after data rejection. NASA-TLX and STAI Scores are used as labels. The red bar illustrates the best-performing predictor set.

A few strategies have been investigated to observe their effect on the outcomes, as mentioned in Section 2.6. After rejecting the initial five seconds of each episode in the fNIRS and EEG data, the feature matrix's undefined values are filled using the averaging approach. The best results were obtained by combining EEG spectral features and fNIRS features, with the STAI score as the target, and adopting Gaussian Process Regression as the regression algorithm (validation RMSE: 16.074), resulting in a minimum test RMSE of 11.747.

Normalisation is then applied to improve model performance by reducing scale-related bias in the selected feature set (combined fNIRS and EEG spectral features). Using the STAI score as the target, seven regression models were tested on the normalised data, with SVM achieving the best results (validation RMSE: 15.884; test RMSE: 11.563). The lower test RMSE indicates better generalisation, showing that normalisation can help maintain a consistent scale for all features while reducing bias from overly large-valued features. Overall, normalisation enhances feature comparability, stabilises training, and helps the model more effectively capture stress patterns in multimodal physiological data.

3.2. Classification

Similarly, the classification approach has been employed for two datasets, the dataset without any data rejected and the dataset selected based on data selection criteria stated in Section 2.6, to compare the outcomes. The results for the full dataset have been obtained first, followed by the selected dataset after data rejection. Like regression, the two targets for this strategy are NASA-TLX and STAI scores.

This investigation followed a similar approach to regression analysis. For the first dataset, EEG features are employed first, followed by fNIRS features, and then combined feature sets. As shown in Figure 12, eight classifiers are tested for each predictor set, and the best-performing model is demonstrated. The highest test accuracy of 67.30% (validation accuracy 60.97%) is achieved using combined fNIRS and EEG spectral features with the STAI score as the target, and neural networks as the best performing model.

 doi:10.20944/preprints202604.0485.v1

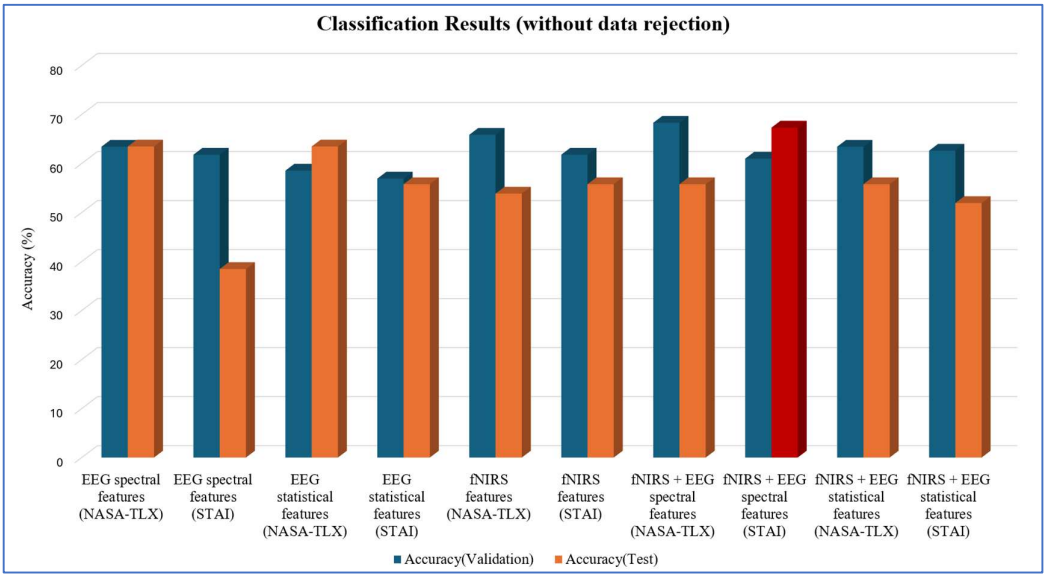


Figure 12. Classification results for the dataset without any data rejection. NASA-TLX and STAI scores are used as labels. The red bar illustrates the best-performing predictor set.

Figure 12 depicts the optimal performance achieved by combining EEG and fNIRS features for several classifiers, using the first dataset with the STAI score as the target, demonstrating the contribution of EEG features. In most cases, accuracy is higher using NASA-TLX as the target.

After applying data-selection criteria, the classification approach is repeated across all EEG, fNIRS, and combined feature sets using eight classifiers. As shown in Figure 13, the highest test accuracy of 65.9% (validation accuracy 63.1%) is achieved by combining fNIRS and EEG statistical features with the NASA-TLX target using neural networks, highlighting the significance of EEG statistical measures [32] and both linear [61,62] and nonlinear fNIRS features [45]. Overall, most predictor sets show higher accuracy when NASA-TLX is used as the target.

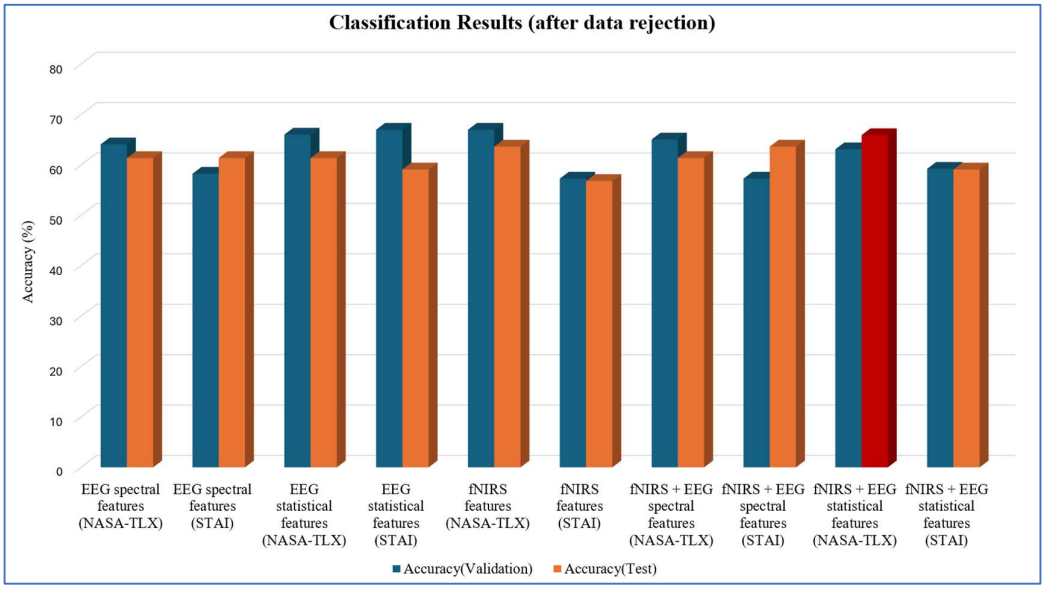


Figure 13. Classification results for the dataset after data rejection. NASA-TLX and STAI scores are used as labels. The red bar illustrates the best-performing predictor set.

 doi:10.20944/preprints202604.0485.v1

After applying the data-selection criteria, validation accuracies decrease slightly compared to the unrejected dataset, but test accuracies become significantly more consistent. For example, the fNIRS + EEG spectral feature set (target: STAI score) shows a sharp drop from ~70% validation accuracy to ~55% on the test set, indicating overfitting. In contrast, the fNIRS + EEG statistical feature set (target: NASA-TLX score) achieves a consistent test accuracy of ~66%, closely matching its validation accuracy and demonstrating improved robustness. This pattern aligns with the principle that cleaner datasets may yield lower validation performance yet generalise better [63], as noise removal reduces overfitting and helps models learn meaningful patterns [64]. Consequently, although accuracy decreases slightly after data rejection, the resulting models are more reliable and generalisable.

Further refinements, removing the initial five seconds of each EEG and fNIRS recording and filling missing values using averaging, improve performance. Using combined fNIRS and EEG statistical features with STAI as the target, SVM achieves a test accuracy of 79.55% (validation accuracy: 66.02%). After normalisation, the same test accuracy is maintained (validation: 64.08%), suggesting enhanced generalisation to unseen data. The significant difference between validation and test accuracy likely reflects that the validation set is not entirely representative of the underlying data distribution, or that the test set aligned better with the features used by the classifier.

3.3. Results After Feature Selection and Ensemble Approach

As outlined in Section 2.9, multiple feature selection algorithms are used to retain the most relevant features while rejecting redundant ones, for both regression and classification approaches. After applying the machine learning algorithms, an ensemble model is utilised to further reduce bias and improve robustness. The final model, chosen based on its optimal performance, is used to predict mental stress.

3.3.1. Regression

For the regression approach, the combined fNIRS and EEG spectral feature set with STAI score as the target has shown the best performance. Two feature selection techniques, ReliefF and the F-test, are applied to identify the most significant features. Table 4 displays the number of features retained, the number of machine learning approaches used to apply the ensemble model and the resulting test RMSE values.

Using ReliefF, with 0 as the threshold, has retained 213 of 284 features, producing the lowest test RMSE of 10.86 when combined with the ensemble model. The F-test, with the 10th-percentile threshold, has kept 211 features and achieved a higher RMSE of 11.32 when employed with the ensemble model. As ReliefF performed best, as shown in Table 4, it is used for channel selection. The channel selection criterion followed here is that if more than 4 features in a channel are rejected, the entire channel is removed. This led to the exclusion of one EEG channel and six fNIRS channels, leaving 210 features and yielding a test RMSE of 11.19.

Table 4. Feature selection algorithms applied to the selected feature set (fNIRS and EEG spectral features).

Feature Selection algorithm	Number of features	Number of classifiers for Ensemble Learning Approach	RMSE (STAI score)
Without Feature Selection	284	3	10.9745
ReliefF	**213**	**3**	**10.856**
FTest	211	6	11.3269
After Channel Selection using ReliefF	210	5	11.1882

Without feature selection, the baseline RMSE is slightly below 11. Applying ReliefF reduces the RMSE to about 10.85, giving the best performance among all methods. In contrast, the F-test produces

 doi:10.20944/preprints202604.0485.v1

the highest RMSE at around 11.35, indicating weaker performance. After channel selection using ReliefF, the RMSE rises to approximately roughly 11.2, still better than the F-test but not as strong as using ReliefF alone.

ReliefF outperforms the other feature-selection methods because it effectively handles complex, high-dimensional, and noisy multimodal data, capturing nonlinear feature interactions that statistical tests like the F-test cannot [65]. Its robustness to noise and redundancy improves generalisability and reduces RMSE, achieving the best value of about 10.85. The strongest regression performance is obtained by applying ReliefF followed by an ensemble model using the combined fNIRS and EEG statistical feature set, which is then used to quantify factory workers' mental stress.

3.3.2. Classification

Similar to the regression analysis, the combined fNIRS and EEG statistical feature set, with the STAI score as the target, is used for feature selection and ensemble modelling. Four feature selection techniques (Kruskal–Wallis, ANOVA, Chi^2, and ReliefF) are applied to select the most informative features. Table 5 compares the results after applying the feature selection algorithms followed by an ensemble model.

For Kruskal–Wallis, features scoring below 10% of the highest rank are eliminated, leaving 316 of 417 features. ReliefF has used a threshold of 0, while ANOVA and Chi^2 are tested with two thresholds: 10% of the highest rank and the 10th percentile. As shown in Table 5, the highest accuracy (84.1%) is achieved using Chi^2 with the 10% rank threshold. The selected Chi^2 features are then utilised for channel selection, applying the same rule as in the regression analysis; channels with more than four rejected features are removed. This resulted in retaining 351 of 417 features, with only one EEG channel and six fNIRS channels discarded. Using these selected channels, the final ensemble model has achieved a test accuracy of 77.27%.

Table 5. Feature selection algorithms applied to the selected feature set (fNIRS and EEG statistical features).

Feature Selection algorithm	Number of features	Number of classifiers for Ensemble Learning Approach	Accuracy (%)
Without Feature Selection	417	3	81.81
Kruskal Wallis	316	7	79.54
ReliefF	197	5	75
ANOVA (10% of the highest rank value as threshold for exclusion)	274	5	79.54
ANOVA (10th percentile for exclusion)	267	7	79.54
Chi^2 (10% of the highest rank value as threshold for exclusion)	**338**	**7**	**84.1**
Chi^2 (10th percentile for exclusion)	331	7	79.54
After channel selection using Chi^2 (10% of the maximum value as a threshold for exclusion)	351	7	77.27

Without feature selection, the ensemble classifier achieves a baseline accuracy of about 81%. Using Chi^2 with a 10% rank threshold increases accuracy to approximately 85%, demonstrating its effectiveness in identifying relevant features and removing redundancy. ANOVA and Kruskal–Wallis yield similar accuracies of around 80%, while ReliefF performs the worst at roughly 75%. After

 doi:10.20944/preprints202604.0485.v1

channel selection based on the best Chi² configuration, test accuracy decreases to about 78%, suggesting that some relevant features may have been eliminated.

Chi² is particularly effective for categorical outcomes because it evaluates how strongly each feature's distribution correlates with discrete class separation [66]. In this experiment, it successfully identifies distinct EEG and fNIRS patterns across stress levels. Eliminating the lowest-ranked 10% of features reduces noise and minimises overfitting, enabling the model to focus on the most informative predictors. This balance enhances overall classification accuracy by retaining relevant features while discarding redundant ones.

3.4. Digital Solution Development

An Android application, *Stress Sense*, was developed to support the factory workers' stress assessment experiment by enabling supervisors and managers to monitor employees' mental states. The app is built using data collected during the experiment and allows stress scores from all nine task episodes to be aggregated for each participant. An average stress score is then calculated and displayed. These scores are mapped onto a visual stress scale, as shown in Figure 14, integrated into the app's interface using predictions from the ensemble model based on the best-performing regression approach combining fNIRS and EEG spectral features, with the STAI score as the target.

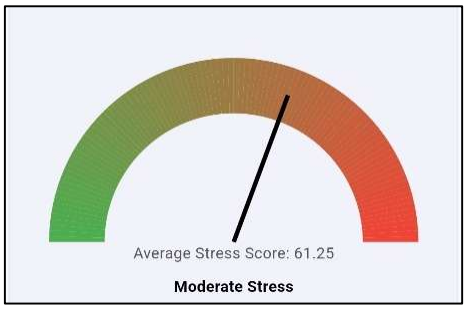
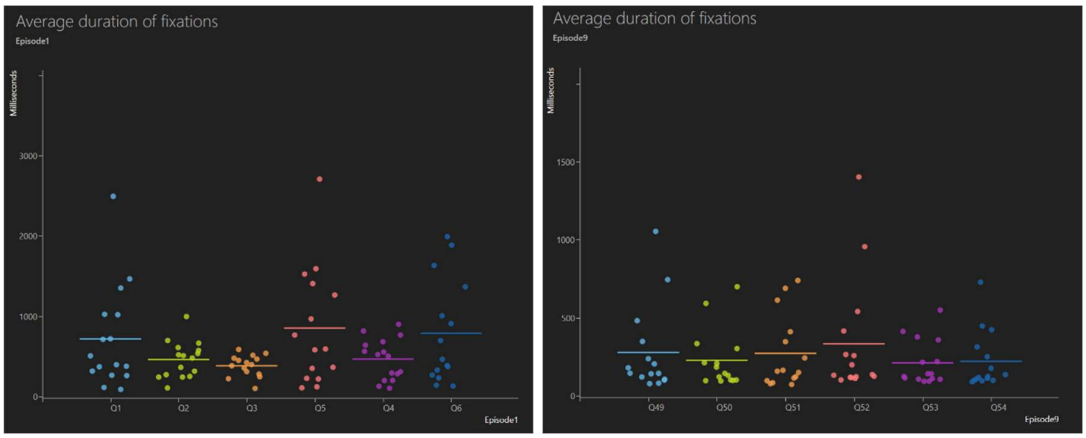


Figure 14. Stress scale displayed in the *Stress Sense* user interface.

3.5. Gaze Tracking

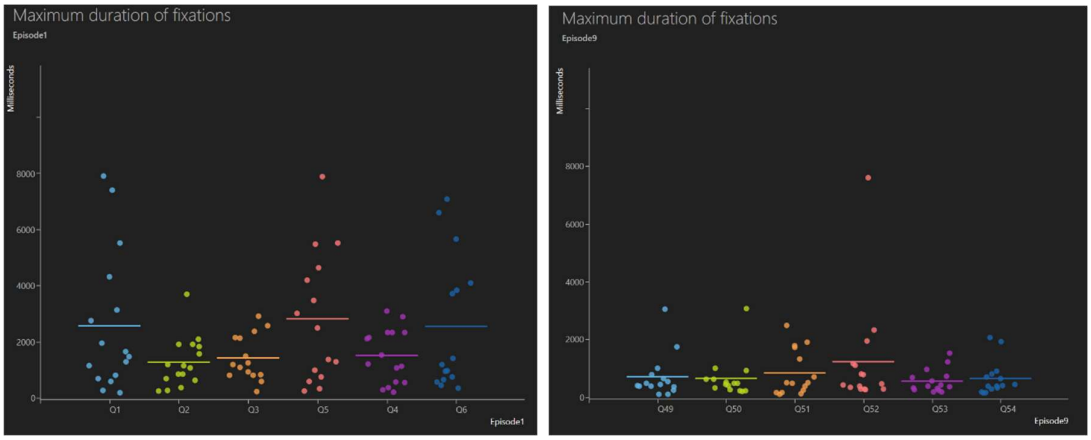
Several eye-tracking parameters are explored in this study, including fixation duration, pupil diameter, visit duration, and number of visits. Results show that average durations of fixation, maximum duration of fixation, and average time of a visit are highest in the first episode and steadily decline by episode 9, while the number of visits increases, as can be seen in Figures 15–19. Average pupil diameter is larger in low-light conditions (5–6 mm) and smaller in high-light conditions (4–5 mm). Each episode contains six cognitive questions, represented as Q1–Q6 in episode 1 and Q49–Q54 in episode 9 across the figures.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2026 doi:10.20944/preprints202604.0485.v1



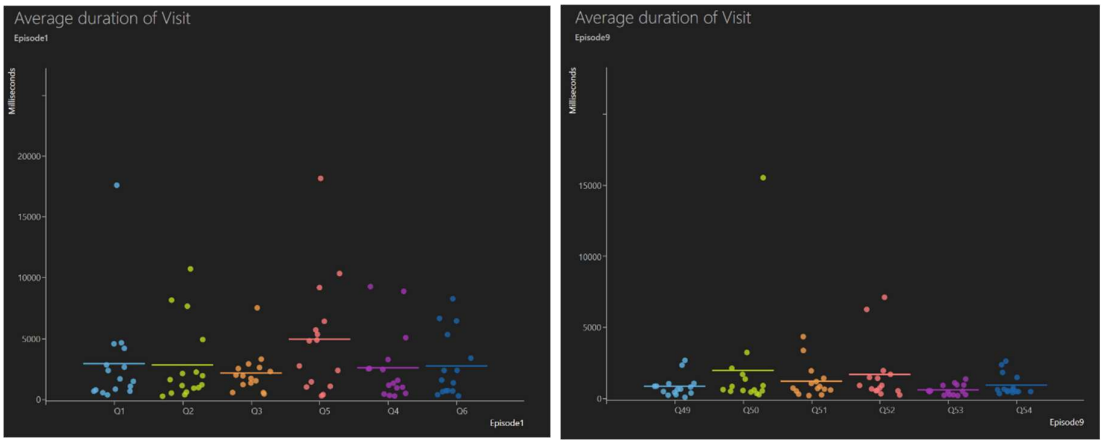
(a) (b)

Figure 15. Average duration of fixation (a) Episode 1, (b) Episode 9.



(a) (b)

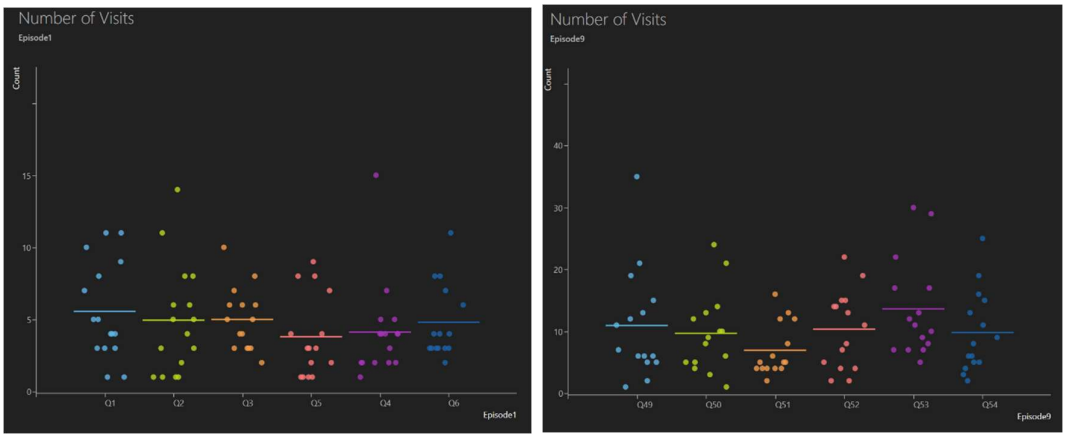
Figure 16. Maximum duration of fixation (a) Episode 1, (b) Episode 9.



(a) (b)

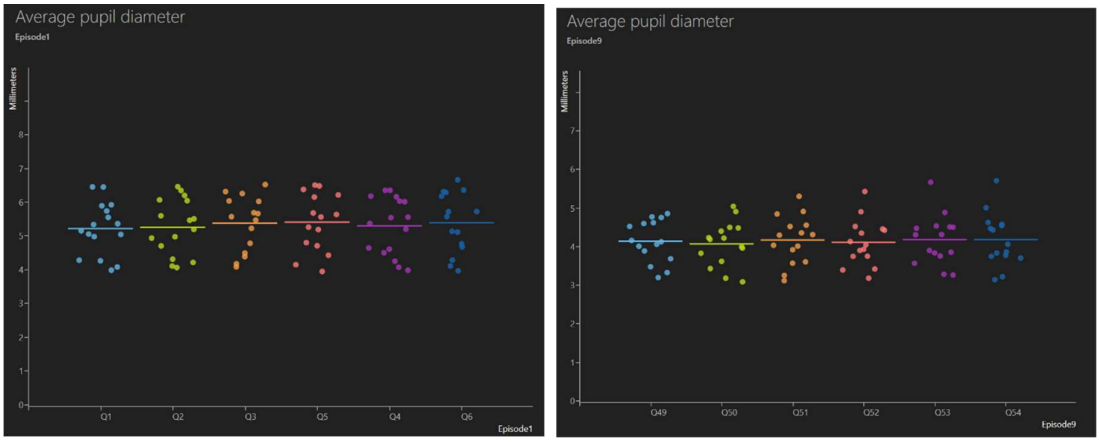
Figure 17. Average duration of visit (a) Episode 1, (b) Episode 9.

 doi:10.20944/preprints202604.0485.v1



(a) (b)

Figure 18. Number of visits (a) Episode 1, (b) Episode 9.

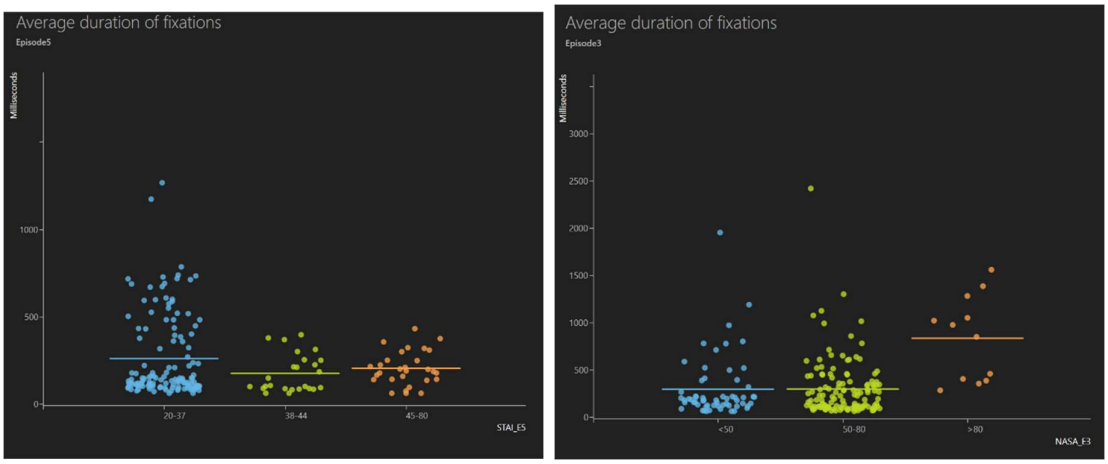


(a) (b)

Figure 19. Average pupil diameter (a) Episode 1, (b) Episode 9.

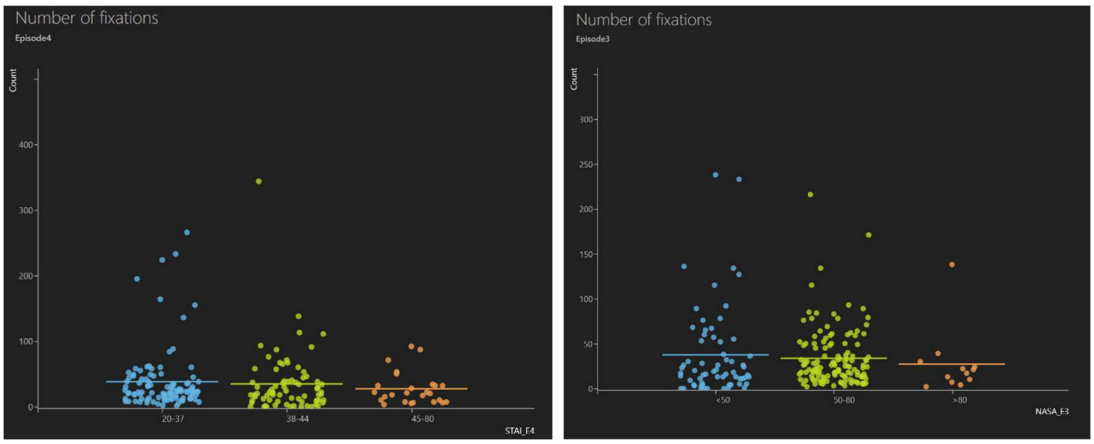
Fixation duration is higher for low and high STAI scores but lower for medium scores, and it generally increases with NASA-TLX across episodes, as shown in Figure 20. The number of fixations has decreased as both NASA-TLX and STAI scores have risen, as shown in Figure 21. Average pupil diameter has also declined with higher NASA-TLX and STAI scores, as indicated in Figure 22. Additionally, Figure 23 shows that fewer fixations have occurred for incorrect answers compared with correct ones.

 doi:10.20944/preprints202604.0485.v1



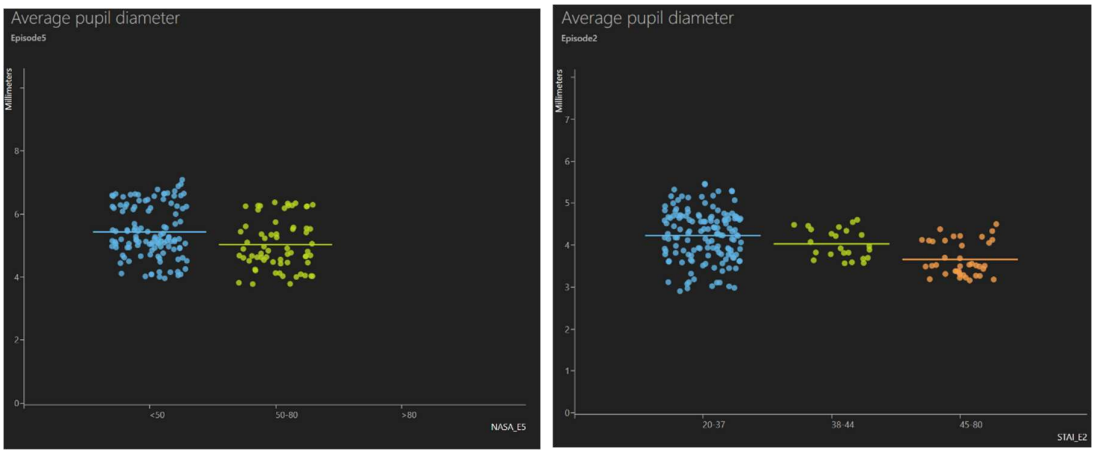
(a) (b)

Figure 20. Average duration of fixation changing with (a) STAI score, (b) NASA-TLX score.



(a) (b)

Figure 21. Number of fixations changing with (a) STAI score, (b) NASA-TLX score.



(a) (b)

Figure 22. Average pupil diameter changing with (a) STAI score, (b) NASA-TLX score.

 doi:10.20944/preprints202604.0485.v1

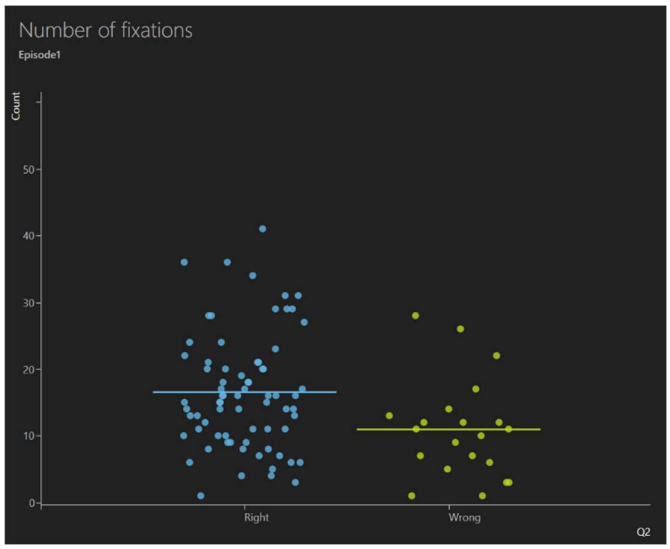


Figure 23. Number of fixations changes with the error rate.

This research reveals clear patterns in visual attention and cognitive load. Longer fixation durations and visit times in the first episode indicate stronger initial cognitive effort, which diminishes as the task becomes more familiar, suggesting a learning effect. Pupil diameter variations between low- and high-light settings match expected physiological responses. The increased number of visits after the first episode indicates a shift in the visual search method. Associations between fixation metrics and STAI/NASA-TLX scores indicate that higher workload leads to longer but fewer fixations [67]. Fewer fixations on incorrect answers than on correct ones suggest different information-processing strategies.

3.6. Subjective Measures

3.6.1. STAI Score

STAI scores vary across episodes with changes in task and ambient variables (Figure 24). Noise and light show clear effects in the first four episodes. Episode 7 shows the highest score due to the most challenging conditions, i.e. high task speed and complexity, elevated noise, and low light, while episode 2 records the lowest score under favourable settings. Episodes 8 and 9 show comparable scores, indicating that replacing the robot with a human has minimal variance. The decreased stress trend in later episodes suggests a learning effect, as familiarity reduces mental tension.

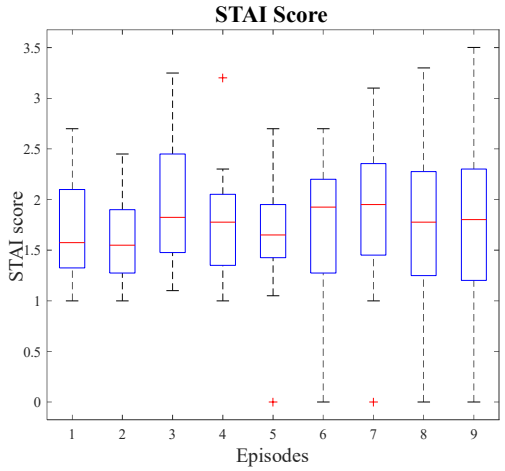


Figure 24. STAI Score for all task episodes for the factory workers' stress assessment experiment.

 doi:10.20944/preprints202604.0485.v1

3.6.2. NASA-TLX Score

Figure 25 shows that the NASA-TLX scores for all episodes vary significantly depending on the task and the ambient conditions.

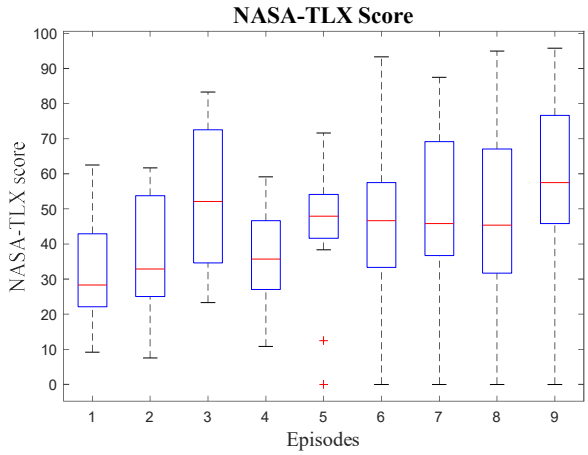


Figure 25. NASA-TLX rating for all task episodes for the factory workers' stress assessment experiment.

The effects of noise and light are evident in the first four episodes, where workload perceptions fluctuate. Episode 9 shows the highest NASA-TLX score, indicating increased mental stress by the end of the experiment, while episode 1 records the lowest score, reflecting minimal cognitive load at the start. Overall, the NASA-TLX scores steadily increase across episodes, with early episodes, particularly episode 1, representing a low-strain baseline. Although noise and light influence stress in the initial episodes, the progressive increase highlights the combined impact of prolonged task engagement and growing fatigue on cognitive stress.

3.7. Behavioural Measures

3.7.1. Error Rate

Figure 26 shows that error rates vary across episodes with shifts in environmental conditions. In the first four episodes, the combined influence of noise and light is evident. The highest error rate occurs in episode 3 when high noise is introduced, whereas episodes 4 and 8 record the lowest error rates under high light. Overall, noise spikes increase errors, while higher light levels help reduce mistakes, highlighting the sensitivity of error rates to sudden environmental changes.

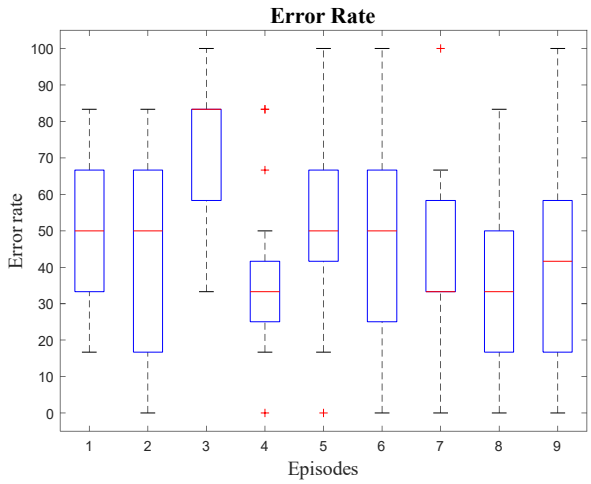


Figure 26. Error rate for all task episodes for the factory workers' stress assessment experiment.

 doi:10.20944/preprints202604.0485.v1

For the secondary task, episode 5 shows the highest error rate in Figure 27, highlighting the added difficulty introduced by this task. Episode 8 has the lowest error rate because the participants received the fewest damaged packs, while episode 9, where the most damaged packs were provided, shows more spread.

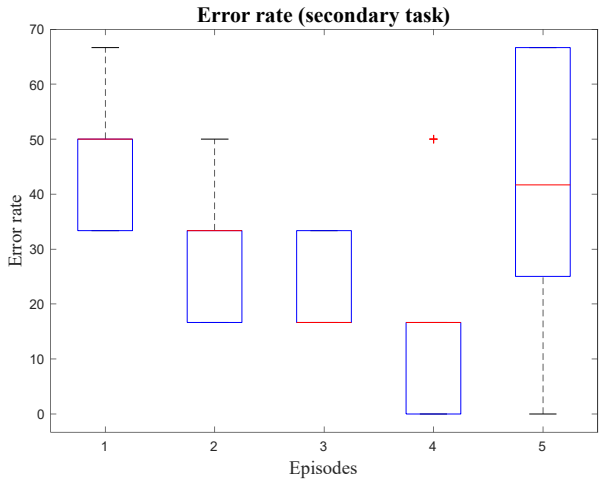


Figure 27. Error rate for task episodes, including the secondary task for the factory workers' stress assessment experiment.

4. Limitations and Future Work

Although this study shows strong potential for cognitive stress evaluation using neuroimaging, physiological data, and machine learning, several limitations must be acknowledged. The proposed framework does not operate in real time, limiting its use in dynamic, stressful environments, and currently depends on offline analysis followed by stress score prediction. The experiment has been conducted in a controlled simulation to study baseline stress patterns; it may not fully capture real workplace variability, such as unpredictability, environmental noise, or individual differences. The dataset used is relatively small, increasing the risk of overfitting [68]. The framework has only been tested with factory workers, restricting its robustness and generalisability across other occupational settings.

Future research can contribute to strengthen the stress-assessment framework by enabling real-time monitoring through an enhanced pipeline that integrates multimodal data with adaptive algorithms. Extending the framework to different occupational sectors would help evaluate its flexibility and improve its generalisability. Larger and more diverse datasets are also needed to extend the findings across various demographics, vocations and stress conditions. Improving physiological data quality remains important, as some artefacts could not be fully removed in this study; future studies should consider more accurate sensors with enhanced connectivity with the participant's body and experimental designs that minimise movement-related noise and artefacts. Overall, the framework can be advanced by using higher-quality and larger datasets, and adaptive real-time methods to broaden its application across a wider range of occupational contexts.

5. Conclusions

This study presents and validates a novel stress-quantification framework for factory workers, using machine learning to predict cognitive stress as a measurable and quantified entity. The experimental setup combines IQ-style cognitive questions with a pick-and-place activity in an HRC environment. The results demonstrate the effectiveness of integrating EEG and fNIRS features with feature selection and ensemble learning. Chi² has produced the highest classification accuracy, while ReliefF achieved the best regression performance, underscoring the importance of effective feature

 doi:10.20944/preprints202604.0485.v1

selection for robust modelling. Ensemble techniques further improved predictive accuracy, yielding the best results for both regression (RMSE: 10.86) and classification (accuracy: 84.1%), with the STAI score outperforming NASA-TLX as the target measure. An Android app-based digital solution has been developed to provide a visual tool to interpret mental stress in the form of a stress scale. Gaze-tracking analysis showed that fixation duration, number of fixations, visit duration, and pupil diameter varied with changes in process parameters. The average duration of fixation, number of fixations, and average pupil diameter are all affected by changes in subjective measures as well as the error rate. Changes in the levels of process parameters have an impact on NASA-TLX and STAI scores. Subjective variables have shown an elevated mental stress under adverse conditions. Overall, human-machine interaction, combined with ambient variables such as low light (50 to 60 lux) and high noise (above 80dB), raises the mental stress of a human participant.

This study advances occupational stress research by demonstrating that a multimodal stress-monitoring framework, combined with machine learning, can deliver meaningful insights and implementable solutions. The framework provides a solid foundation for future work, including real-time stress evaluation, long-term stress tracking, and broader application across diverse work environments. Overall, the novel framework presented in this study is an important step toward safer, healthier, and more sustainable workplaces, enabling proactive stress detection, intelligent decision-making, and timely interventions.

Author Contributions: Conceptualisation, A.K. and A.O.; Data curation, A.A and Z.Z.; Formal analysis, A.A., Z.Z., A.K. and A.O.; Funding acquisition, A.K., A.O. and P.B.; Investigation, A.A., A.O. and Z.Z.; Methodology, A.A., A.K., A.O. and Z.Z.; Project administration, Z.Z., A.K. and P.B.; Resources, A.K. and P.B.; Software, A.A. and Z.Z.; Supervision, A.O., Z.Z., A.K. and P.B.; Validation, A.A. and Z.Z.; Visualisation, A.A, Roles/Writing - original draft, A.A.; Writing - review & editing, A.O., A.K., P.B. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This research has received funding from PepsiCo International Limited.

Declaration of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Funding: This research was funded by PepsiCo International Limited.

Ethics Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Human Invasive Ethics Committee of the School of Science and Technology, Nottingham Trent University (Application ID: 1754129 and date of approval: 13/09/2023).

Informed Consent Statement: Informed consent was obtained for experimentation with human subjects from all subjects involved in the study.

Data Availability Statement: The code for the Android app, *Stress Sense*, is publicly available on the GitHub site: https://github.com/ekumamatthew/stressSense, https://github.com/ekumamatthew/stress-bee.

References

1. Othman, U., & Yang, E. (2023). Human–robot collaborations in smart manufacturing environments: review and outlook. *Sensors*, *23*(12), 5663.
2. Ajoudani, A., Zanchettin, A. M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., & Khatib, O. (2018). Progress and prospects of the human–robot collaboration. *Autonomous robots*, *42*(5), 957-975.
3. De Simone, V., Di Pasquale, V., Giubileo, V., & Miranda, S. (2022). Human-Robot Collaboration: An analysis of worker's performance. *Procedia Computer Science*, *200*, 1540-1549.
4. Sun, Y., Jeelani, I., & Gheisari, M. (2023). Safe human-robot collaboration in construction: A conceptual perspective. Journal of safety research, 86, 39-51.

 doi:10.20944/preprints202604.0485.v1

5. Rohwer, E., Flöther, J. C., Harth, V., & Mache, S. (2022). Overcoming the "dark side" of technology—a scoping review on preventing and coping with work-related technostress. *International journal of environmental research and public health*, *19*(6), 3625.
6. Vagaš, M., Galajdová, A., & Šimšík, D. (2020, October). Techniques for secure automated operation with cobots participation. In *2020 21th International Carpathian Control Conference (ICCC)* (pp. 1-4). IEEE.
7. Patil, S., Vasu, V., & Srinadh, K. V. S. (2023). Advances and perspectives in collaborative robotics: a review of key technologies and emerging trends. *Discover Mechanical Engineering*, *2*(1), 13.
8. Liu, Y., Habibnezhad, M., & Jebelli, H. (2021). Brainwave-driven human-robot collaboration in construction. *Automation in Construction*, *124*, 103556.
9. Zakeri, Z., Omurtag, A., Breedon, P., Hilliard, G., & Khalid, A. (2021, September). Studying mental stress factor in occupational safety in the context of the smart factory. In *Proceedings of the 31st European Safety and Reliability Conference, ESREL* (pp. 92-99).
10. Blandino, G. (2023). How to measure stress in smart and intelligent manufacturing systems: A systematic review. *Systems*, *11*(4), 167.
11. Hijry, H., Naqvi, S. M. R., Javed, K., Albalawi, O. H., Olawoyin, R., Varnier, C., & Zerhouni, N. (2024). Real time worker stress prediction in a smart factory assembly line. *IEEE Access*, *12*, 116238-116249.
12. Gualtieri, L., Palomba, I., Wehrle, E. J., & Vidoni, R. (2020). The opportunities and challenges of SME manufacturing automation: safety and ergonomics in human–robot collaboration. *Industry 4.0 for SMEs: Challenges, opportunities and requirements*, 105-144.
13. Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of health and social behavior*, 385-396.
14. Braarud, P. Ø. (2020). An efficient screening technique for acceptable mental workload based on the NASA Task Load Index—development and application to control room validation. *International Journal of Industrial Ergonomics*, *76*, 102904.
15. Legler, F., Trezl, J., Langer, D., Bernhagen, M., Dettmann, A., & Bullinger, A. C. (2023). Emotional experience in human–robot collaboration: suitability of virtual reality scenarios to study interactions beyond safety restrictions. *Robotics*, *12*(6), 168.
16. Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human factors*, *57*(1), 125-143.
17. Braarud, P. Ø., Bodal, T., Hulsund, J. E., Louka, M. N., Nihlwing, C., Nystad, E., ... & Wingstedt, E. (2021). An investigation of speech features, plant system Alarms, and operator–system interaction for the classification of operator cognitive workload during dynamic work. *Human factors*, *63*(5), 736-756.
18. Upasani, S., Srinivasan, D., Zhu, Q., Du, J., & Leonessa, A. (2024). Eye-tracking in physical human–robot interaction: Mental workload and performance prediction. *Human factors*, *66*(8), 2104-2119.
19. Jaiswal, A. (2023). An Intelligent Multi-Modal Framework Towards Assessing Human Cognition.
20. Aylward, J., & Robinson, O. J. (2017). Towards an emotional 'stress test': a reliable, non-subjective cognitive measure of anxious responding. *Scientific reports*, *7*(1), 40094.
21. Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., & Yang, X. (2018). A review of emotion recognition using physiological signals. *Sensors*, *18*(7), 2074.
22. Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2009). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine*, *14*(2), 410-417.
23. González Ramírez, M. L., García Vázquez, J. P., Rodríguez, M. D., Padilla-López, L. A., Galindo-Aldana, G. M., & Cuevas-González, D. (2023, August). Wearables for stress management: a scoping review. In *Healthcare* (Vol. 11, No. 17, p. 2369). MDPI.
24. Tiwari, R., Kumar, R., Malik, S., Raj, T., & Kumar, P. (2021). Analysis of heart rate variability and implication of different factors on heart rate variability. *Current cardiology reviews*, *17*(5), 74-83.
25. Jerčić, P., Sennersten, C., & Lindley, C. (2020). Modeling cognitive load and physiological arousal through pupil diameter and heart rate. *Multimedia Tools and Applications*, *79*(5), 3145-3159.
26. Kumar, A., Sharma, K., & Sharma, A. (2021). Hierarchical deep neural network for mental stress state detection using IoT based biomarkers. *Pattern Recognition Letters*, *145*, 81-87.

 doi:10.20944/preprints202604.0485.v1

27. Gomes, N., Pato, M., Lourenco, A. R., & Datia, N. (2023). A survey on wearable sensors for mental health monitoring. *Sensors*, *23*(3), 1330.
28. Zhang, P., Li, F., Zhao, R., Zhou, R., Du, L., Zhao, Z., ... & Fang, Z. (2021). Real-time psychological stress detection according to ECG using deep learning. *Applied Sciences*, *11*(9), 3838.
29. Zhang, J., Yin, H., Zhang, J., Yang, G., Qin, J., & He, L. (2022). Real-time mental stress detection using multimodality expressions with a deep learning framework. *Frontiers in Neuroscience*, *16*, 947168.
30. Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., Depaepe, F. & Van den Noortgate, W. (2021). Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work*, *23*(3), 567-585.
31. Ayres, P., Lee, J. Y., Paas, F., & Van Merrienboer, J. J. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in psychology*, *12*, 702538.
32. Katmah, R., Al-Shargie, F., Tariq, U., Babiloni, F., Al-Mughairbi, F., & Al-Nashash, H. (2021). A review on mental stress assessment methods using EEG signals. *Sensors*, *21*(15), 5043.
33. Causse, M., Lepron, E., Mandrick, K., Peysakhovich, V., Berry, I., Callan, D., & Rémy, F. (2022). Facing successfully high mental workload and stressors: An fMRI study. *Human Brain Mapping*, *43*(3), 1011-1031.
34. Molina-Rodríguez, S., Hidalgo-Munoz, A. R., Ibáñez-Ballesteros, J., & Tabernero, C. (2023). Stress estimation by the prefrontal cortex asymmetry: Study on fNIRS signals. *Journal of Affective Disorders*, *325*, 151-157.
35. HSE, "Lighting at Work Health," *Health and Safety Executive*, vol. 38, p. 47, 1987
36. "Regulations - HSE." Accessed: Aug. 01, 2025. [Online]. Available: https://www.hse.gov.uk/noise/regulations.htm
37. Jin, K., Rubio-Solis, A., Naik, R., Onyeogulu, T., Islam, A., Khan, S., ... & Mylonas, G. (2022). Identification of cognitive workload during surgical tasks with multimodal deep learning. *arXiv preprint arXiv:2209.06208*.
38. Argyle, E. M., Marinescu, A., Wilson, M. L., Lawson, G., & Sharples, S. (2021). Physiological indicators of task demand, fatigue, and cognition in future digital manufacturing environments. *International Journal of Human-Computer Studies*, *145*, 102522.
39. Onkhar, V., Dodou, D., & De Winter, J. C. F. (2024). Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions. *Behavior Research Methods*, *56*(5), 4221-4238.
40. Zheng, W. L., & Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, *7*(3), 162-175.
41. Bonilauri, A., Sangiuliano Intra, F., Baselli, G., & Baglio, F. (2021). Assessment of fNIRS signal processing pipelines: towards clinical applications. *Applied Sciences*, *12*(1), 316.
42. "GitHub - farhadabedinzadeh/AutomaticEEGSignalPreprocessingAndLinearNonlinearFeatureExtraction." Accessed: Aug. 01, 2025. [Online]. Available: https://github.com/farhadabedinzadeh/AutomaticEEGSignalPreprocessingAndLinearNonlinearFeatureExtraction
43. Hasan, M. J., & Kim, J. M. (2019). A hybrid feature pool-based emotional stress state detection algorithm using EEG signals. *Brain sciences*, *9*(12), 376.
44. Shon, D., Im, K., Park, J. H., Lim, D. S., Jang, B., & Kim, J. M. (2018). Emotional stress state detection using genetic algorithm-based feature selection on EEG signals. *International Journal of environmental research and public health*, *15*(11), 2461.
45. Arefi, S. R., Setarehdan, S. K., & MOTIE, N. A. (2018). Classification of mental stress levels by analyzing fNIRS signal using linear and non-linear features.
46. Pedroni, A., Bahreini, A., & Langer, N. (2019). Automagic: Standardized preprocessing of big EEG data. *NeuroImage*, *200*, 460-473.
47. Arora, I., Gambhir, J., & Kaur, T. (2021). Data normalisation-based solar irradiance forecasting using artificial neural networks. *Arabian Journal for Science and Engineering*, *46*(2), 1333-1343.
48. Singh, D., & Singh, B. (2022). Feature wise normalization: An effective way of normalizing data. *Pattern Recognition*, *122*, 108307.

 doi:10.20944/preprints202604.0485.v1

49. Ramadhana, H., Nasution, H., & Absah, Y. (2021). Mental Workload Analysis Using NASA-TLX Method at Bank XYZ-Medan Balaikota Consumer Loan Unit. International Journal of Research and Review, 8(12), 622-626.
50. Fauquet-Alekhine, P., Rouillac, L., Bertoni, J., & Granry, J. C. (2016). Heart rate vs stress indicator for short term mental stress. British journal of medicine and medical research, 17(7), 1-11.
51. Reyes, O., Morell, C., & Ventura, S. (2015). Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context. *Neurocomputing*, *161*, 168-182.
52. Galvao, R. K. H., Araujo, M. C. U., Fragoso, W. D., Silva, E. C., Jose, G. E., Soares, S. F. C., & Paiva, H. M. (2008). A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. *Chemometrics and intelligent laboratory systems*, *92*(1), 83-91.
53. McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology*, 1-1.
54. Elssied, N. O. F., Ibrahim, O., & Osman, A. H. (2014). Research article a novel feature selection based on one-way ANOVA F-test for E-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, *7*(3), 625-638.
55. Jin, X., Xu, A., Bie, R., & Guo, P. (2006, April). Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In *International workshop on data mining for biomedical applications* (pp. 106-115). Berlin, Heidelberg: Springer Berlin Heidelberg.
56. Spolaôr, N., Cherman, E. A., Monard, M. C., & Lee, H. D. (2013, October). ReliefF for multi-label feature selection. In *2013 Brazilian Conference on Intelligent Systems* (pp. 6-11). IEEE.
57. P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare 2023, Vol. 11, Page 1808*, vol. 11, no. 12, p. 1808, Jun. 2023, doi: 10.3390/HEALTHCARE11121808.
58. Webb, G. I., & Zheng, Z. (2004). Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. *IEEE transactions on knowledge and data engineering*, *16*(8), 980-991.
59. Urigüen, J. A., & Garcia-Zapirain, B. (2015). EEG artifact removal—state-of-the-art and guidelines. *Journal of neural engineering*, *12*(3), 031001.
60. Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109-139). Cham: Springer International Publishing.
61. Paulmurugan, K., Vijayaragavan, V., Ghosh, S., Padmanabhan, P., & Gulyás, B. (2021). Brain–computer interfacing using functional near-infrared spectroscopy (fNIRS). *Biosensors*, *11*(10), 389.
62. Hong, K. S., Khan, M. J., & Hong, M. J. (2018). Feature extraction and classification methods for hybrid fNIRS-EEG brain-computer interfaces. *Frontiers in human neuroscience*, *12*, 246.
63. Yang, S., Xie, Z., Peng, H., Xu, M., Sun, M., & Li, P. (2022). Dataset pruning: Reducing training data by examining generalization influence. *arXiv preprint arXiv:2205.09329*.
64. Mwangi, B., Tian, T. S., & Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229-244.
65. Pupo, O. G. R., Morell, C., & Soto, S. V. (2013, November). ReliefF-ML: an extension of ReliefF algorithm to multi-label learning. In *Iberoamerican Congress on Pattern Recognition* (pp. 528-535). Berlin, Heidelberg: Springer Berlin Heidelberg.
66. Peker, N., & Kubat, C. (2021). Application of Chi-square discretization algorithms to ensemble classification methods. *Expert Systems with Applications*, *185*, 115540.
67. Liu, J. C., Li, K. A., Yeh, S. L., & Chien, S. Y. (2022). Assessing perceptual load and cognitive load by fixation-related information of eye movements. *Sensors*, *22*(3), 1187.
68. Ghasemzadeh, H., Hillman, R. E., & Mehta, D. D. (2024). Toward generalizable machine learning models in speech, language, and hearing sciences: Estimating sample size and reducing overfitting. *Journal of Speech, Language, and Hearing Research*, *67*(3), 753-781.