Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 9356452, 12 pages
https://doi.org/10.1155/2021/9356452

Research Article

Employing Multimodal Machine Learning for Stress Detection

Rahee Walambe,[1,2] **Pranav Nayak**,[1] **Ashmit Bhardwaj**,[1] **and Ketan Kotecha**[1,2]

[1]*Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 411215, India*
[2]*Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune 411215, India*

Correspondence should be addressed to Rahee Walambe; rahee.walambe@sitpune.edu.in and Ketan Kotecha; drketankotecha@gmail.com

Received 28 September 2021; Revised 13 October 2021; Accepted 15 October 2021; Published 28 October 2021

Academic Editor: Deepak Kumar Jain

In the current information age, the human lifestyle has become more knowledge-oriented, leading to sedentary employment. This has given rise to a number of health and mental disorders. Mental wellness is one of the most neglected, however crucial, aspects of today's fast-paced world. Mental health issues can, both directly and indirectly, affect other sections of human physiology and impede an individual's day-to-day activities and performance. However, identifying the stress and finding the stress trend for an individual that may lead to serious mental ailments is challenging and involves multiple factors. Such identification can be achieved accurately by fusing these multiple modalities (due to various factors) arising from a person's behavioral patterns. Specific techniques are identified in the literature for this purpose; however, very few machine learning-based methods are proposed for such multimodal fusion tasks. In this work, a multimodal AI-based framework is proposed to monitor a person's working behavior and stress levels. We propose a methodology for efficiently detecting stress due to workload by concatenating heterogeneous raw sensor data streams (e.g., face expressions, posture, heart rate, and computer interaction). This data can be securely stored and analyzed to understand and discover personalized unique behavioral patterns leading to mental strain and fatigue. The contribution of this work is twofold: firstly, proposing a multimodal AI-based strategy for fusion to detect stress and its level and, secondly, identifying a stress pattern over a period of time. We were able to achieve 96.09% accuracy on the test set in stress detection and classification. Further, we were able to reduce the stress scale prediction model loss to 0.036 using these modalities. This work can prove important for the community at large, specifically those working sedentary jobs, to monitor and identify stress levels, especially in current times of COVID-19.

1. Introduction

The increase in the percentage of socioeconomic category of knowledge workers has been on the phenomenal rise worldwide in the last few decades. In 2019, this socioeconomic category surpassed 1 billion people, accounting for more than 30% of the world's total employed population [1]. With the majority of the world's working hours spent at a desk, employees' mental health becomes the most crucial issue. Workplaces are becoming more demanding than ever, requiring employees to deliver excellent results in the shortest amount of time. While this may appear acceptable at first, the health of the individuals suffers throughout prolonged working life, resulting in stress, worry, reduced productivity, and, in the worst-case scenario, depression. Workplaces have attempted to make employees' jobs more manageable but only so much can be done. To that end, it is critical to maintain the record and monitor people's mental health and performance over a period of time and take appropriate action as needed. Stress detection is a multimodal fusion problem involving various modalities of the data and can be solved using multimodal AI methods. In the following sections, the detailed literature review encompassing the multimodal fusion techniques in general applications, specific application of multimodal techniques in healthcare and stress detection, and use of ML for stress detection are discussed. Various datasets that are useful for such tasks are also presented.

1.1. Multimodal Data and Relevant Applications. The concept of combining data streams from several sources to achieve an outcome seems intuitive, yet there are several obstacles to overcome. Combining data from several sources, such as sensors, has proven to be more effective in forecasting outcomes. In [2], Ngiam et al. have examined multimodal deep learning. A novel approach to applying deep learning to different modalities like audio and video is proposed, and cross-modality feature learning is reported. This paper presents a method for learning enhanced representations for a single modality (e.g., video) from other modalities (e.g., audio and video) that are present throughout feature learning time. With the rise in low-powered sensors in wearable devices, the amount of data generated is enormous, but they are varied in discrete and continuous sampling rates. As a result, integrating them is challenging. In [3], Radu et al. propose a method for concatenating diverse sensor types utilizing four deep learning algorithms such as DNN and CNN. Lahat et al. [4] provide perspectives, guidelines, and ideas on multimodal data fusion approaches and their applications and techniques in multiple domains like health and biomedical and multisensory systems. In [5], Gros Dut presents an in-depth analysis of the logic underpinning data fusion and discusses data fusion and multisensor integration approaches. Narkhede et al. [6] propose a method to detect gaseous emissions using multimodal data collected from gas sensors and thermal cameras. The fused model achieved 96% accuracy on the testing set instead of 82% on LSTM applied to sensor data and 93% on CNN applied to camera images for individual modalities. Cai et al. [7] review an innovative approach of using explainable AI on multimodal deep neural networks. This not only improves predictions owing to the usage of many modalities but also deviates from a neural network's black box decision-making and gives us insight into how the model arrived at any given result. Explainability also improves the model's trustworthiness and acceptability. There are multiple application domains where multimodal AI is employed, with one of the most relevant being healthcare. Healthcare data is typically multimodal and has to be fused to obtain more meaningful outcomes.

1.2. Multimodal Data in Healthcare. The work related to multimodal data in healthcare is highly relevant. Before drawing any conclusions, medical specialists examine various images, data, and patient histories. So, if a machine learning algorithm is employed for decision-making, having a mechanism for fusing multimodalities arising from various individual modalities becomes critical, since any model is only as effective as the data it is trained on. Recently, in 2019, Cai et al. [7] tried to explore all the existing technologies and state-of-the-art methods used in the multimodal data healthcare industry. USA, China, and Canada are the top three countries at the forefront of smart healthcare. In [8], Iakovidis and Smailis propose a semantic model to mine multimodal data defined to be stored as feature spaces and easy to work with, train, and test. On similar lines, [9] Wang et al. achieved the same task implemented on top of the Hadoop framework, enabling parallelization. Collecting these datasets, let alone any model implementation on them, is a very long and tedious process. The few that are already present make them even more critical in the healthcare domain. Brain Tumor Image Segmentation Benchmark [10] is a multimodal dataset containing 3D brain MRI images used to detect brain tumors. It also contains a number of different approaches used to predict brain tumor presence and locations with accuracy scores and other metrics [10]. Cetin et al. [11] researched Schizophrenia Classification using multimodal deep learning methodologies to predict the brain disease of a patient using fMRI and magnetoencephalography (MEG). They were able to achieve 85% accuracy using these modalities and ensemble neural networks of these two individual features. Radiology Objects in COntext (ROCO) [12] is a multimodal dataset to recognize the interaction of visual features and semantic links in radiological pictures. The goal is achieved by obtaining all image-caption pairings from PubMed Central, an open-access biomedical literature database, because captions represent visual content in its semantic context. Computer Tomography, Ultrasound, X-ray, Fluoroscopy, Positron Emission Tomography, Mammography, Magnetic Resonance Imaging, and Angiography are among the medical imaging modalities included in the ROCO collection. Alzheimer's Disease Neuroimaging Initiative (ADNI) [13] was collected to describe cross-sectional and longitudinal clinical assessments in healthy people, people with MCI, and people with mild Alzheimer's disease such that neuroimaging and chemical biomarker measurements may be evaluated. One of the exciting and most challenging areas in healthcare is stress detection. There are multiple approaches for stress detection using machine learning and artificial intelligence.

1.3. Stress Detection Using Machine Learning. For stress detection, typically, questionnaires are created with the help of domain experts such as clinicians and psychologists. Such questionnaires are often used in research in the field of psychology to get insight into general working experiences and behavioral analysis of the participants. In areas where computing or the use of AI algorithms can be applied, the most commonly used modality is Electrocardiogram (ECG). However, using a single sensor modality to detect stress has certain limitations, such as less accuracy and more false positives/negatives. Research from various fields shows the usage of different modalities and the potential use of sensors for estimating stress, mental and affective states, and the context in which they appear. Multiple modalities representing physical, neurophysiological, and computer interactions and so forth can be fused to generate better outcomes. Since wearable sensors are getting more affordable and can be readily integrated into generic devices, such data can be generated and collected with ease. Koldijk et al. [14] have combined several modalities in a unique dataset with features like computer interactions, facial expressions, postures coordinates, and body sensors. In [15], Ahuja et al. have collected data of university students using a questionnaire and assigning certain weightage to these questions

and then using various machine learning algorithms for predicting stress, with support vector machines yielding the highest accuracy (85%). In [16], Smets et al. have compared various machine learning algorithms for detecting mental stress on physiological responses in a controlled environment. They recruited 20 participants, conducted stress tests, and recorded data from two physiological sensors, wireless electrocardiography (ECG) sensor and NeXus-10 MKII, to measure galvanic skin response (GSR). In [17], Wijsman et al. measured physiological signals and features like skin conductance, Electrocardiogram, respiration, and surface electromyogram (sEMG) of the upper trapezius muscle wearable systems to predict stress in an office-like environment and reached an average accuracy rate of 74.5%. In [18], Healey and Picard collected and analyzed physiological data of real-world driving tasks to determine stress levels. They found out that, in most cases, driver's physiological data, that is, heart rate metrics, skin conductance, and so forth, are closely correlated with driver stress level. In [19], Can et al. have tried to perform continuous stress detection using unobtrusive wearable devices like Samsung S series devices and Empatica E4. They collected data from participants of an algorithmic programming contest for evaluation. An accuracy of 84% was achieved with Samsung S devices and multilayer perceptron, yielding the highest accuracy.

In [20], Mohd et al. aimed to present a novel approach for mental stress detection by using thermal imaging of the subject's face. They found a correlation between stress and blood flow in the face and have developed an automatic thermal face, Supraorbital, Periorbital, Maxillary, and Nostril Detection to estimate the person's internal state. In this work, we have considered the problem of identifying the stress of an individual based on various distinct different modalities using machine learning techniques. We considered the SWELL-KW dataset [14] for experimentation and demonstration of our multimodal fusion techniques.

1.4. Multimodal Datasets for Stress Detection. Hence, in this work, we propose a strategy using multimodal artificial intelligence to classify mental stress and identify the scale of the stress. A dataset consisting of multimodal data called SWELL-KW [14] is used to validate and demonstrate our framework. This dataset is collected through the standard devices around any working individual to sense various modalities and utilize them for fusion using multimodal AI. SWELL-KW [14] is a powerful resource for accurately measuring sedentary jobs' work-associated mental stress. Most datasets regarding the stress monitoring domain have a single modality. However, the SWELL-KW dataset comprises four distinct modalities that, when combined, can be extremely useful for diagnosing stress and reliably predicting stress. The data was gathered as part of the SWELL Project by Kraaij et al. [21] and made publicly available in 2017. Since then, many forms of techniques have been applied to achieve state-of-the-art results in predicting stress based on the available modalities. The majority of stress detection research has focused on heart rate variability and related features as the data pairs well with related datasets like WESAD [22] and DREAMER [23]. Sriramprakash et al. (2017) [24] implemented an SVM classifier with RBF kernel to achieve 92.75% accuracy by employing only physiological signals available in the dataset. They also examined the individual features and their importance in predicting stress and concluded that the first stress indicator is galvanic skin response and heart rate. In [25], Nkurikiyeyezu et al. have validated their model on SWELL-KW, which was trained using Advanced Trail Making Test [26], and achieved an accuracy of 99.25% using physiological data. In [27], Koldijk et al. focused on ranking the modalities to be more correlated to the prediction of stress and mental effort. The conclusion was that posture and facial expressions yielded the most valuable information. In [28], Koldijk et al. showed us the visualizations of different modalities of the SWELL-KW dataset for better insights. SWELL-KW is a powerful resource for accurately measuring the work-related mental stress of sedentary jobs.

This work employs an artificial neural network (ANN) for feature extraction and early and late fusion-based techniques for multimodal data fusion, considering all four modalities. This approach is unique and has not been explored. In summary, the contribution of this work is threefold:

(1) Implementing early and late fusion using machine learning to predict whether a person is stressed or not, given four specific modalities: computer interactions, body posture, facial features, and heart rate variability

(2) Applying transfer learning on early fusion features from the stress model to predict the NASA-TLX score, which predicts the stress level on a scale of 0 to 100

(3) Providing a method to save the data from monitoring a person's mental state as the task load increases across the specific timeline

This paper is organized as follows: Section 2 discusses the dataset employed and the methods applied for the multimodal fusion. Section 3 discusses the pipeline and workflow of the model. Section 4 includes the findings and analysis of all of the predictions and assessment measures for each. Section 5 provides limitations of this work and the scope for future improvements. Section 6 presents the conclusion.

2. Materials and Methods

2.1. Dataset. To demonstrate our approach of multimodal fusion using machine learning, the SWELL knowledge work (SWELL-KW) dataset [14] is used. This dataset was first presented in 2014 at the 16th ACM International Conference on Multimodal Interaction by Koldijk et al. [14]. It is available in a publicly accessible repository [29]. The dataset was collected as a part of the research project wherein 25 subjects that performed either traditional intelligence work or sedentary occupations. Making presentations, writing reports, reading emails, and researching information were

all part of the experience. The participants' working environments were often exploited by the researchers who exposed the subjects with stress-inducing stimuli such as e-mail interruptions and time constraints. A total of 25 participants' data was produced. There were eight females and seventeen males in this study, with an average age of 25, comprising Delft University of Technology students and TNO (the Netherlands Organization for applied scientific research) interns. Since they were workforce-ready, they had experience with large volumes of data and operating computers. Computer logs captured facial expression from camera recordings and body postures data points from a Kinect 3D sensor [30] and heart rate variability and and skin conductance from sensors connected to the participant's body were all included in the dataset. The dataset contains raw, preprocessed, and features extracted data, all readily available to work with. Validated questionnaires were administered to participants to assess their subjective interactions with task load, needed mental commitment, mood during these activities, and perceived stress. Participants were advised not to smoke or drink any caffeinated beverages three or four hours before the experiment because these are potential confounders. The experiment was classified into three blocks for the various stress conditions, with each session lasting approximately one hour. The dataset contains 3000+ examples that were used to train individual models. The ground truth labeling of whether the person was in a stressed state or not was provided along with the modalities' numerical values. Modalities collected are shown in Table 1.

2.1.1. Computer Logging. The researchers used a background application on the users' computers. The application was a key-logging uLog [31] (version 3.2.5, by Noldus IT) for logging users' computer interactions. Table 2 shows some examples of computer logging data.

2.1.2. Facial Features. A high-resolution USB camera was deployed to record the participants' faces and upper bodies. The specifications of the camera were iDS uEye UI-1490RE, 1152×768. To preserve the privacy of participants, no videos were made public. Researchers used a proprietary software called Noldus FaceReader [32] to interpret the data presented. Using deep learning and computer vision, this program analyzes facial expressions in real time. It provides details in the form of txt-logs containing information about facial expressions and emotions. FaceReader [32] app has over 30 functions, including head orientations, facial expressions, action units, and emotions. Table 3 shows some examples of facial expressions data.

2.1.3. Body Posture. A per-time frame analysis of the participant's body orientation was included in the datasets. They acquire the coordinates of all the joints by fitting the Kinect [30] skeletal model in this manner. These CSV files contain all of the coordinates necessary to determine angles between upper-body joints and bones. The dataset also includes upper-body bone orientations with timestamps relative to the x-, y-, and z-axes. Over 90 characteristics were included in the posture data. Table 4 provides few samples of observations on body posture.

2.1.4. Body Sensors. The ECG was recorded using a Mobi unit (TMSI [33]) with self-adhesive electrodes. The recording software Portilab2 [34] was created with some preprocessing. Skin conductance was measured using Mobi and finger electrodes.

Out of these 3000 + samples, due to failure in capturing a reading at any particular moment for *all* modalities, only 956 examples that reported all three modalities correctly were eventually used to train the final model on 70-30 train-test split in this work. The archive contains over 900 documents, containing both raw and structured records. Since some of the functions were not labeled and had many missed values, we had to merge and preprocess several files with each modality. We closely examined these files before selecting and sorting different files and merging them to create a single data file. Python was used for all of the data preprocessing activities.

2.2. Methods

2.2.1. Artificial Neural Network. A neural network [35] is a layer-by-layer connection of neurons which attempts to replicate the functioning of the human brain. The first layer of a neural network is the input layer, and the last is the predicted output layer. The hidden layers between the input and output layers take the output of the last layer's neurons as input and return some output after a mathematical calculation. Each layer is added in a sequential order, with the previous layer's output serving as the input for the next layer.

2.2.2. Dropout. Some selected neurons are disregarded during the training process and are not included in the computation of the output or in the backpropagation [36]. Since each neuron is trained on a specific collection of examples, this helps us avoid overfitting. The neurons are dropped out to differ in each epoch and are chosen at random. Here, a dropout [37] rate of 0.5 is used, which means that 50% of neurons would be ignored at each step.

2.2.3. Activation Functions. The representation power of a deep NN is due to its nonlinear activation functions.

Sigmoid. Sigmoid activation is implemented at the output layer. The focus is on biclass classification (stressed or not). Hence, sigmoid activation is the best fit, since it predicts the probability as an output.

The mathematical formula for this is shown as follows:

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{1}$$

ReLU: Rectified Linear Unit is an activation function that increases linearly for positive inputs and outputs zero for

Table 1: Sample data details.

Type	Available raw and preprocessed data	Available features
Computer interactions	uLog output and parsed selection of data	Mouse (3), keyboard (7), applications (2)
Facial expressions	FaceReader output and parsed data	Head orientation (3), facial movements (10), action units (19), emotion (8)
Body postures	Joint coordinates extracted with Kinect SDK and angles of the upper body	Distance (1), joint angles (10), bone orientations (3×11) (as well as the study of the above for the amount of movement (44))
Physiology	Data from Mobi	Heart rate (variability) (2), skin conductance (1)

Table 2: Sample examples for computer interaction data.

Mouse activity	Left clicked	Right clicked	Double clicked	Wheel	Char ratio	Error key ratio
0.0125	12	7	0	1	0.603774	0.216216
0.401786	10	5	0	0	0.5	0.181818
0.034188	0	0	0	0	0.7	0
0.233333	34	12	0	0	0.605263	0.068966
0.179916	37	17	0	2	0.875	0

Table 3: Sample examples for facial expressions data.

Squality	Sneutral	Shappy	Ssad	Sangry	Ssurprised	Sscared
0.944941	0.968862	0.023946	0.0013	0.016315	0.002024	0.001087
0.930303	0.88457	0.076952	0.001144	0.017392	0.002032	0.000651
0.933104	0.931965	0.031468	0.000371	0.023774	0.001722	0.001756
0.904466	0.806947	0.105516	0.006459	0.009809	0.001563	0.000441
0.929025	0.951412	0.028358	0.001095	0.01813	0.001309	0.003466

Table 4: Sample examples for body posture data.

Avg. depth	Left shoulder angle avg.	Right shoulder angle avg.	Lean angle avg.
2102.597393	−116.055931	115.017758	92.340895
2099.725525	−116.301605	115.986636	92.083385
2102.365778	−115.963089	114.073054	92.38169
2104.116968	−115.62963	113.972465	92.428905
2105.284007	−116.359699	111.538437	92.461537

negative inputs. The formula for the ReLU function is seen in equation (2). ReLU is used for the model's hidden layers.

$$f(x) = \max(0, x). \tag{2}$$

2.2.4. Loss Functions. Binary cross entropy: The final layer generates output that is compared to the ground reality, and a loss function is used to quantify the error, which is then back-propagated [28] to train individual neurons' formulae for improved results. The formula for binary cross entropy [38] is as shown in equation (3), where $\widehat{y}_i$ is the i-th value predicted by the model, y_i is the corresponding actual value, and the output size is the number of scalar values in the model output.

$$\begin{aligned} \text{loss} = -\frac{1}{\text{output size}} \sum_{i=1}^{\text{output size}} & Y_i * \log \widehat{y}_i \\ & + (1 - y_i) * \log(1 - \widehat{y}_i). \end{aligned} \tag{3}$$

Root mean squared error: on regression model predictions, Root Mean Square Error [39] is the most suitable evaluation metric. The RMSE is computed as shown in equation (4), where N is the number of examples, $\widehat{x}_i$ is the value predicted by the model, and x_i is the actual value or observation.

$$\text{loss} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \widehat{x}_i)^2}{N}}. \tag{4}$$

2.2.5. Early Fusion. Early fusion techniques incorporate various modalities by constructing a joint representation of input features. The final prediction can be expressed as seen in equation (5), where concatenation indicates concurrently represented modality features. Since only one model is used, the training procedure is simple. It often requires highly engineered and preprocessed features from several modalities in order for them to align well or have similar meanings [40].

$$p = h(v_1, v_2, \ldots\ldots v_m), \tag{5}$$

where h is used to denote a single model, v stands for input features from multiple modalities, and P is final prediction. In layman's terms, early fusion occurs as the modalities are merged or features are mapped *before* attempting to classify them.

2.2.6. Late Fusion. Late fusion employs a fusion technique to combine decision values from individual modalities [40]. Assume that model h_i is used on modality i ($i = 1, \ldots, M$); the final prediction is shown as follows:

$$p = F(h_1(v_1), h_2(v_2) \ldots\ldots h_m(v_m)). \tag{6}$$

The late fusion method admits the employment of several models on various modalities, providing greater flexibility. Because the predictions are created independently, it is easier to deal with a missing modality. In layman's terms, late fusion entails classifying outcomes through individual modalities before integrating the model predictions to characterize the final production.

2.2.7. NASA-TLX. The Nasa Task Load Index [41] is a measurement of the workload of any particular job. It was developed over three years by NASA's Ames Research Center's Human Performance Group, which used more than 40 laboratory simulations. It considers all aspects of a job, including mental need, physical demand, temporal demand, efficiency, effort, and frustration. NASA-TLX scores range from 0 to 100, with 0 representing rest mode or no work requirement or effort at all and 100 representing a task that requires complete efforts, both mental and physical.

3. Model

Our system design maps three individual neural network architecture components to predict status based on body orientation, facial expression, and keystroke dynamics. Vanilla Neural Network is employed as an individual network for all three modalities as the data was present in numeric form. The final layer of each modality neural network can then be linked with other neural networks to form an ensemble neural network architecture [42]. Our model flow uses a Hard Parameter Multitask Learning, wherein the model has common layers that split into task-specific layers further. This simply indicated that the feature maps are used to transform a large number of individual modalities' features into a small number of each and used that as an input to our ensemble neural network. The model hyperparameters were selected after experimenting with different combinations, and the model which outperformed in training and testing phases has been explained below. The model was trained using cross-validation [43] to prevent overfitting. The system was trained on approximately 3,000 examples over 200 epochs. Each epoch took, on average, one minute to train and the model as a whole took 16 hours: 10 hours to train individual models and 3 hours each to train the stress classifier and NASA-TLX regressor.

3.1. Stress Classifier. Each of the individual neural network layers has been equipped with a dropout layer. This helps us prevent overfitting the data as several columns might contain irrelevant information or might not be as useful as others. ReLU activation function is used in hidden layers as it demonstrated the best results in all our findings. Sigmoid activation is used for the output layer, giving us the probability of whether a person is stressed or not (biclass prediction). In the case of the keystroke dynamics neural network, the model's number of parameters was not too high, so a simple neural network with one hidden layer was enough to produce good results. However, in the case of models such as facial expressions and body posture, there was a need for a more complex neural network due to the higher number of features in each modality. For skin conductance and heart rate variability metric, two features and one feature are present, respectively. Hence, different neural network architecture was not necessary. Instead, when the output for the individual models was generated, these three features (namely, skin conductance (2) and heart rate variability (1)) were provided as the input to our next and final neural network, which was also equipped with a dropout layer and used ReLU and sigmoid layer on the hidden and output layer, respectively. All the neural network models were trained on the binary cross entropy loss function. The results for both early fusion and late fusion were compared.

In early fusion, output of the last hidden layer is used as an input to the combination neural network. Figure 1 depicts the construction of individual architectures and the use of the early fusion technique to predict stress and NASA-TLX score. In the late fusion prediction, the probabilities of individual models for stress are provided as an input to the final neural network. Figure 2 shows the architecture for the late fusion technique to classify whether a person is stressed or not.

3.2. NASA-TLX Regression Model. The same feature map used to predict stress with early fusion was a good predictor of NASA-TLX scores. With this, it can be observed that the feature maps our original input as an indicator of stress. Similar to the stress detection, a network with 2 hidden layers was designed from the output of individual neural networks to make an ensemble neural network regression model. RMS Error was used as the loss function for the same. This is a transfer learning solution carried out as the same model feature map was used for a prediction of different but related entity. NASA-TLX predictions were also carried out with late fusion model but the results were not significant. This can be attributed to the fact that late fusion model had only a few features which were not able to scale the features to an extent that early fusion could.

4. Results and Discussion

The prediction and analysis tasks were divided into two streams: prediction of NASA Task Load Index using

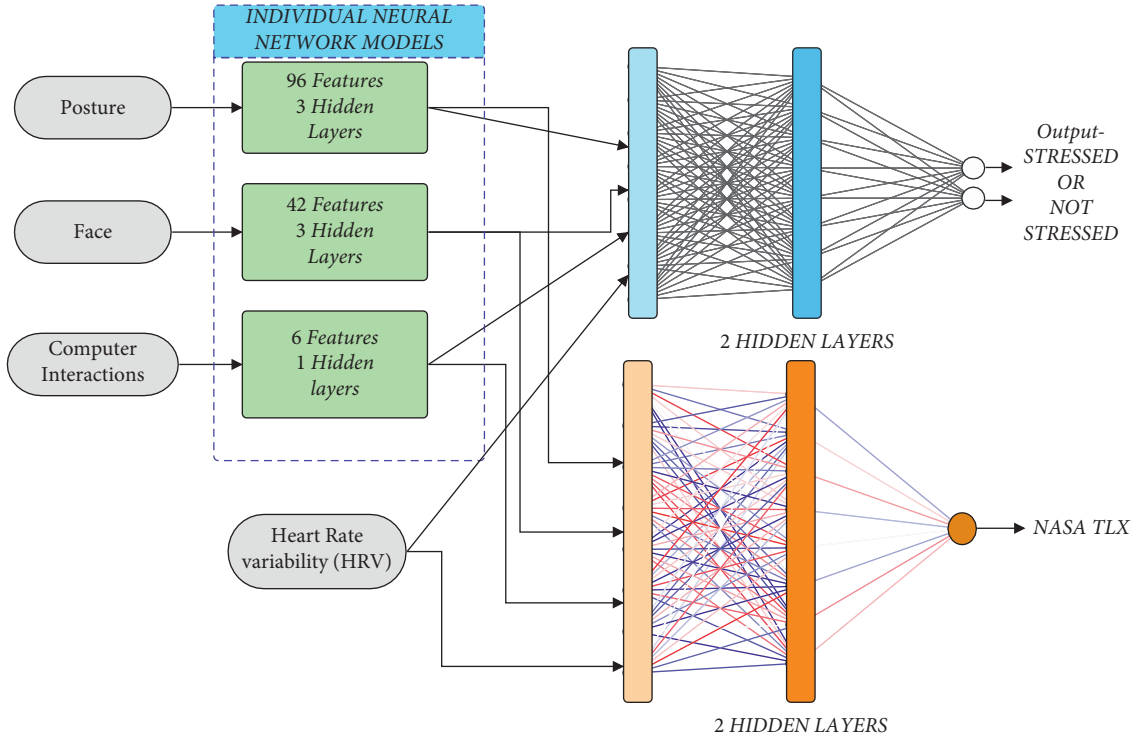

Figure 1: Model architecture and workflow for early fusion and NASA-TLX prediction.

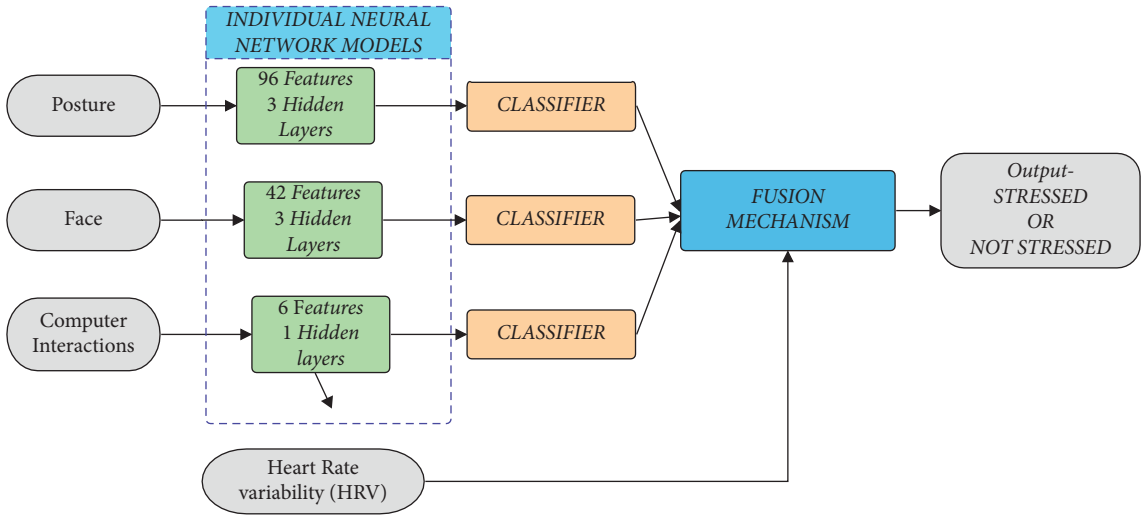

Figure 2: Model architecture and workflow for the late fusion technique.

regression model and predicting whether the user is stressed or not using neural networks classification.

4.1. NASA-TLX. We achieved an RMSE of 0.047 on the training set and 0.036 on the test set in neural network predictions. The better model performance on the test set can be attributed to the dropout layer, which is at its optimum capacity only during the test phase. Figure 3 shows how loss varies with increasing epochs during the training phase.

Note that the NASA-TLX score scale is 0–100, so an average loss of 0.036 on this scale is minimal.

4.2. Stress Detection. The metrics of the individual stress classification model and the ensemble neural network architecture can be found in Table 5. As seen, body posture is the best indicator of stress, giving an accuracy of 77%.

Both early fusion and late fusion techniques were used on these three models to form the main neural network for final predictions. These early and late fusion outputs were

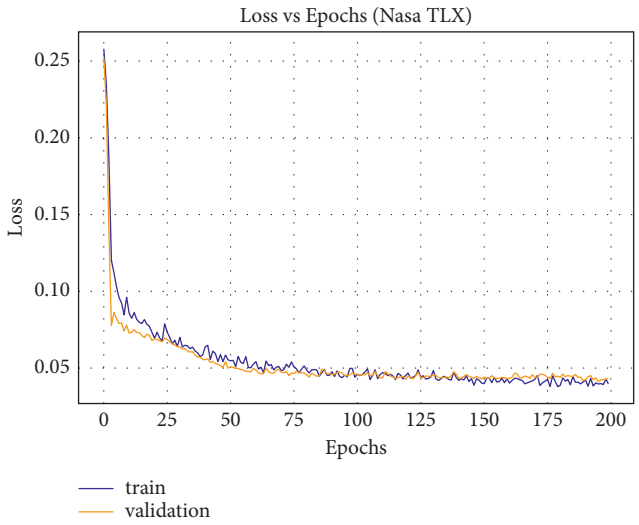

Figure 3: Loss versus epochs: NASA-TLX model.

Table 5: Evaluation metrics for individual models.

Modality	Accuracy (%)	Precision	Recall	F1 score
Body posture	77.56	84.45	76.01	78.03
Facial expressions	74.05	82.05	74.02	71.02
Keystroke dynamics	71.33	72.45	68.02	71.01

also added with three heart rate variability and skin conductance features. Figure 4 shows the training and validation accuracy plots for both late and early fusion models demonstrating the superiority in accuracy and early convergence of the early fusion model. Figure 5 shows the loss charts for both early fusion and late fusion. It can be seen that the loss keeps decreasing with increase in epochs and for early fusion the loss is even smaller as compared to the late fusion model. Figures 6 and 7 present the confusion matrices for both the respective models. The false positives and false negatives for the early fusion model are comparatively lesser than those for the late fusion model proving the better performance of the early fusion model. Figure 8 demonstrates the residual plot for the predictions made by NASA-TLX regression model on the test set. As clearly visible, most of the predictions lie within ±0.5 with only a few outliers going out of ±2 range. Note that the score is in the range of 0–100, so a decimal error is relatively affordable. Figure 9 shows the ROC curve demonstrating the classifier performance at every threshold.

Table 6 shows the evaluation metrics for each of the final models using early fusion and late fusion, respectively.

4.3. Comparison with Other Works Using the SWELL-KW Dataset. A number of approaches for detecting the stress are reported on SWELL-KW Dataset. These approaches employed subset of the available modalities and their accuracy scores. In [26], similar stress detection experiment on different database and using different metrics for stress measurement are carried out. Hence, the results from [26] are not included. However, it can be seen that using the

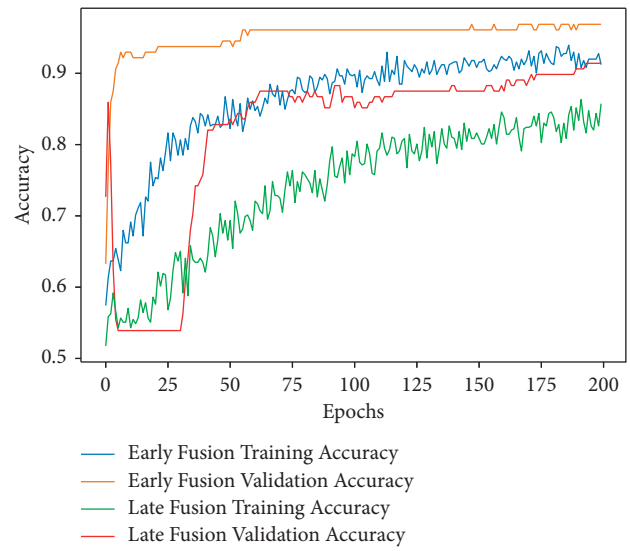

Figure 4: Accuracy chart for early fusion and late fusion.

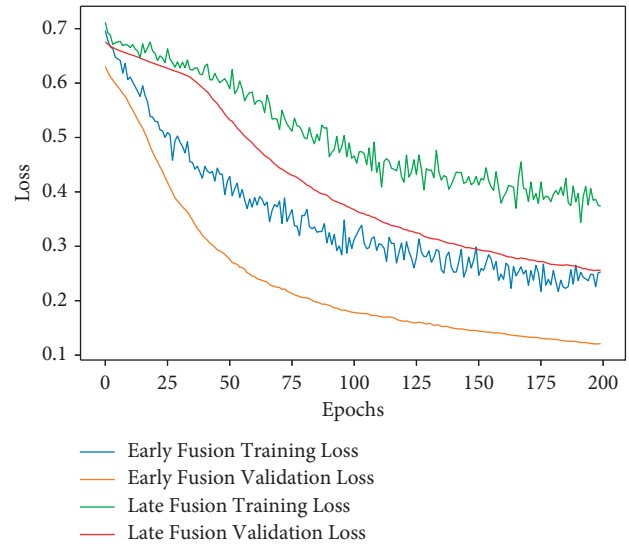

Figure 5: Loss plot for early fusion and late fusion.

subset of modalities and implementing the various machine learning models such as SVM [24, 27], Fast-GRNN [21], and Active Bayesian Learning [44] underperform our model with early fusion. Table 7 summarizes the research models used for the same dataset by other researchers, their modalities used, and accuracies achieved. Our model outperforms all of the present models, achieving a state-of-the-art accuracy score, which can be credited to the use of multimodal fusion techniques with an ensemble neural network model. With all these evaluations, it becomes evident that early fusion performs better than late fusion technique. We can attribute this to the fact that many features of individual modalities are better mapped by early fusion, which ameliorates our final result. Finally, the real-time prediction is demonstrated in Figure 8. Figure 10 consists of the plot showing "orange" whenever the state of the user is stressed

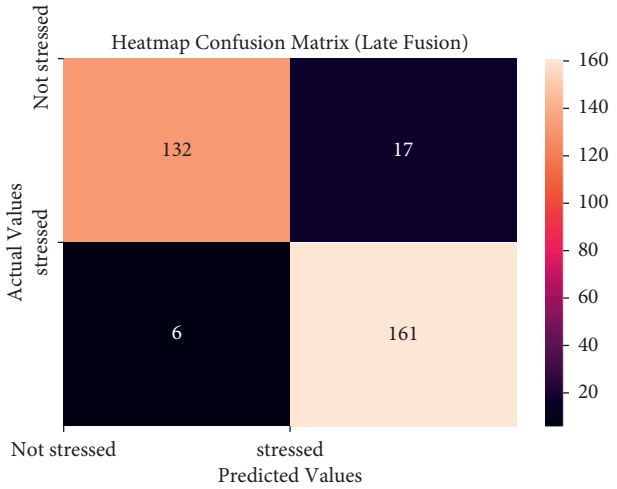

Figure 6: Confusion matrix for late fusion.

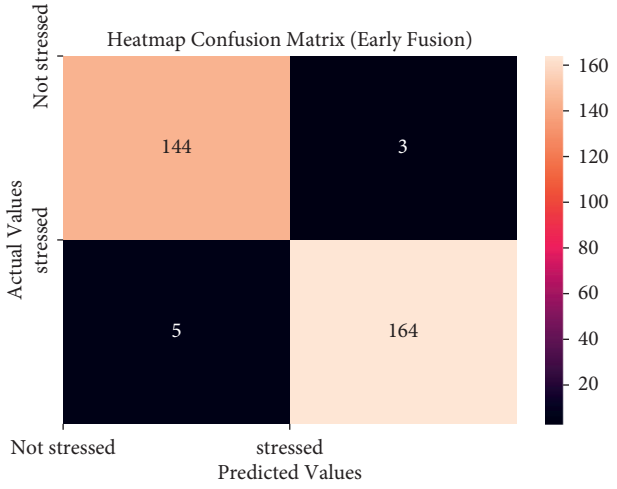

Figure 7: Confusion matrix for early fusion.

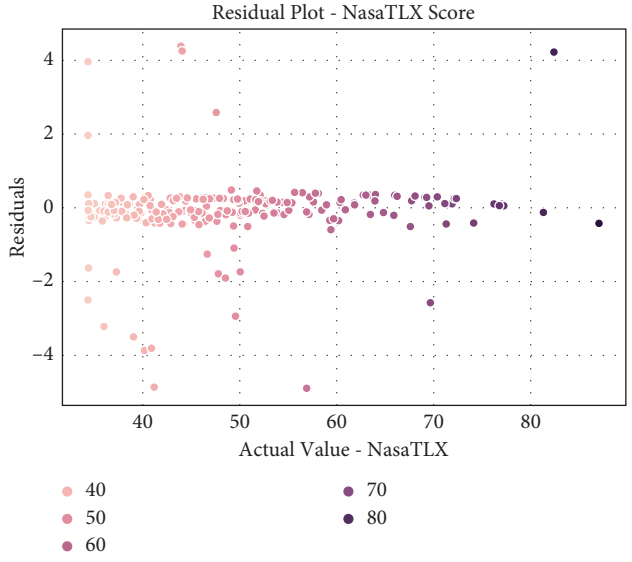

Figure 8: Residual plot for NASA-TLX prediction model.

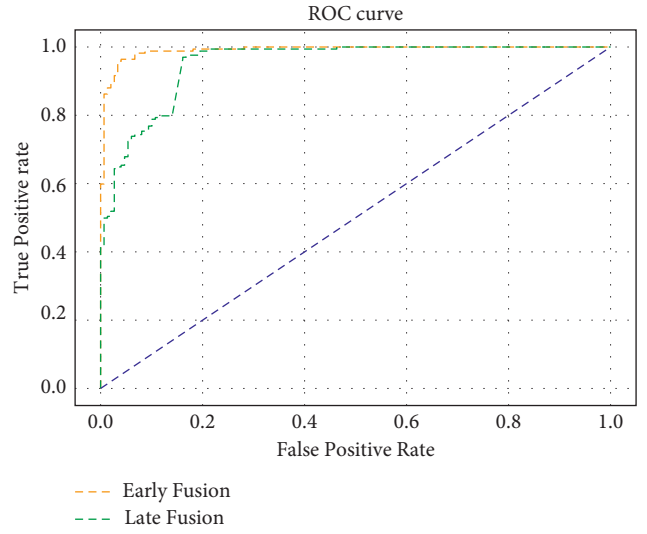

Figure 9: ROC curve for the early fusion and late fusion models.

Table 6: A comparison of metrics for early versus late fusion models.

Fused model	Accuracy (%)	Precision	Recall	F1 score
Late fusion	90.45	0.91	0.90	0.90
Early fusion	96.67	0.95	0.95	0.95

and "blue" whenever they are relaxed. The graph clearly shows that when workload increases, the person starts to get stressed when pursued too long. Some orange dots in the middle might be false indicators of stress, so the administrator monitoring this might not notify the user if the stress levels are not high and persistent for long periods. Figure 9 shows the ROC curve demonstrating the classifier performance at every threshold, which clearly indicates that early fusion has a higher area under curve and, hence, better predictions, which is consistent with other evaluation metrics.

5. Limitations

This research provides a high accuracy for stress classification using multimodal AI data fusion techniques, but there are a few limitations. The model works best with availability of more modalities. It may work with lesser modalities but the idea is to have samples from as many modalities representing stress as possible leading to better accuracies and more importantly lesser false positives or false negatives. The state of our model, as it stands, needs all the input parameters to produce the results, so future work may include an extension of this research on multimodal colearning where the study can be carried out to understand the robustness of the multimodal model in the absence of one or more modalities at test/train time. This will benefit the users who cannot provide all the modalities.

Table 7: Other research works on SWELL-KW dataset, the modalities used, and accuracy scores.

Research model	Modalities used	Accuracy (%)
SVM classifier with RBF kernel [24]	Heart rate variability and physiological data	92.75
Fast-GRNN [44]	Heart rate variability and physiological data	87.87
Support Vector Machine [27]	Heart rate variability, computer interactions, body posture, and facial features	90
Active Bayesian Learning [45]	Heart rate variability and physiological data	91.92
Our model	Heart rate variability, computer interactions, body posture, and facial features (with early fusion)	96.67

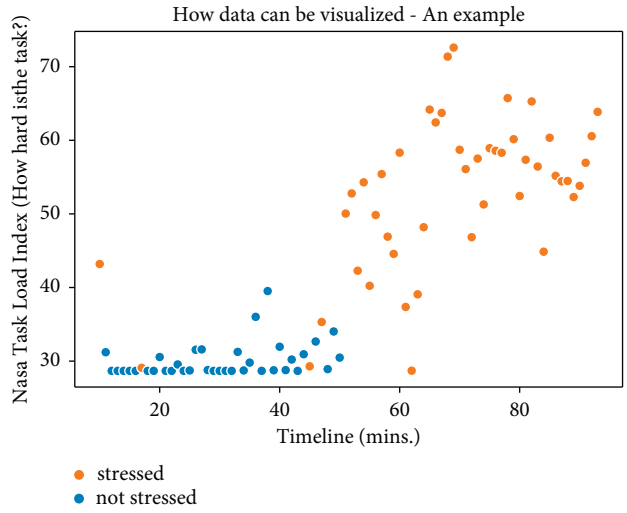

Figure 10: Timeline showing a person's state as the task load increases.

6. Conclusions

This paper investigates new ways to leverage the SWELL-KW dataset to predict stress levels and task load based on the modalities provided in the dataset. We used different multimodal fusion algorithms for the predictions and evaluated them to compare and report the best one. The early fusion model showed the best results on stress classification. It also showed better results than multiple linear regression models for predictions of task load. Finally, we showed how the data could be stored according to the timeline, which clearly shows that a prolonged increase in task load leads to stress. The input modalities can be easily replicated using simple resources around any knowledge worker who makes this set easy to use in any environment. The factor of stress affecting any person should not be ignored for too long as it causes health issues, both mental and physical. Furthermore, using the power of artificial intelligence in healthcare that goes beyond our supervision will go a long way [46].

Data Availability

The SWELL-KW dataset can be accessed at [14].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was funded by the Symbiosis International University, Pune, India, under the research support fund.

References

[1] S. Ricard, *The Year of the Knowledge Worker*, Forbes Technology Council, MA, USA, 2020, https://www.forbes.com/sites/forbestechcouncil/2020/12/10/the-year-of-the-knowledge-worker/?sh=c58dfc47fbba.

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning ICML*, Bellevue, Washington, July 2011.

[3] V. Radu, C. Tong, S. Bhattacharya et al., "Multimodal deep learning for activity and context recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–27, 2018.

[4] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[5] X. E. Gros Dut, "Data Fusion – A Review," , 1997.

[6] P. Narkhede, R. Walambe, S. Mandaokar, P. Chandel, K. Kotecha, and G. Ghinea, "Gas detection and identification using multimodal artificial intelligence based sensor fusion," *Applied System Innovation*, vol. 4, no. 1, 2021.

[7] Q. Cai, H. Wang, Z. Li, and X. Liu, "A survey on multimodal data-driven smart healthcare systems: approaches and applications," *IEEE Access*, vol. 7, Article ID 133583, 2019.

[8] D. Iakovidis and C. Smailis, "A semantic model for multimodal data mining in healthcare information systems," *Studies in Health Technology and Informatics*, vol. 180, pp. 574–578, 2012.

[9] F. Wang, V. Ercegovac, T. S. Mahmood et al., "Large-scale multimodal mining for healthcare with mapreduce," in *Proceedings of the 1st ACM International Health Informatics Symposium*, pp. 479–483, ACM, Washington, DC, USA, November 2010.

[10] B. H. Menze, A. Jakab, S. Bauer et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[11] M. S. Cetin, J. M. Houck, B. Rashid et al., "Multimodal classification of schizophrenia patients with MEG and fMRI data using static and dynamic connectivity measures," *Frontiers in Neuroscience*, vol. 10, 2016.

[12] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in context (ROCO): a multimodal image dataset," in *Intravascular Imaging and Computer-Assisted Stenting and Large-Scale Annotation of Biomedical Data and*

Expert Label Synthesis, D. Stoyanov, Ed., Springer, Cham, Switzerland, pp. 180–189, 2018.

[13] S. G. Mueller, M. W. Weiner, L. J. Thal et al., "Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI)," *Alzheimer's and Dementia*, vol. 1, no. 1, pp. 55–66, 2005.

[14] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th international conference on multimodal interaction*, pp. 291–298, ACM, Istanbul, Turkey, November 2014.

[15] R. Ahuja and A. Banga, "Mental stress detection in university students using machine learning algorithms," *Procedia Computer Science*, vol. 152, pp. 349–353, 2019.

[16] E. Smets, P. Casale, U. Großekathöfer et al., "Comparison of machine learning techniques for psychophysiological stress detection," in *Proceedings of the International Symposium on Pervasive Computing Paradigms for Mental Health*, pp. 13–22, Springer, Barcelona, Spain Cham, April 2016.

[17] J. Wijsman, B. Grundlehner, H. Liu, J. Penders, and H. Hermens, "Wearable physiological sensors reflect mental stress state in office-like situations," in *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 600–605, IEEE, Geneva, Switzerland, September 2013.

[18] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[19] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, 2019.

[20] M. N. H. Mohd, M. Kashima, K. Sato, and M. Watanabe, "Mental stress recognition based on non-invasive and non-contact measurement from stereo thermal and visible sensors," *International Journal of Affective Engineering*, vol. 14, no. 1, pp. 9–17, 2015.

[21] W. Kraaij, S. Verberne, S. Koldijk et al., "Personalized support for well-being at work: an overview of the SWELL project," *User Modeling and User-Adapted Interaction*, vol. 30, no. 3, pp. 413–446, 2020.

[22] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 400–408, ACM, Boulder, Colorado, USA, October 2018.

[23] S. Katsigiannis and N. Ramzan, "DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[24] S. Sriramprakash, V. D. Prasanna, and O. V. R. Murthy, "Stress detection in working people," *Procedia computer science*, vol. 115, pp. 359–366, 2017.

[25] K. Nkurikiyeyezu, K. Shoji, A. Yokokubo, and G. Lopez, "Thermal comfort and stress recognition in office environment," in *Proceedings of the 12th Conference on Health Informatics*, pp. 256–263, Prague, Czech Republic, February 2019.

[26] K. Mizuno and Y. Watanabe, "Utility of an advanced trail-making test as a neuropsychological tool for an objective evaluation of work efficiency during mental fatigue," in *Fatigue Science for Human Health*, pp. 47–54, Springer, Tokyo, Japan, 2008.

[27] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Transactions on affective computing*, vol. 9, no. 2, pp. 227–239, 2016.

[28] S. Koldijk, J. Bernard, T. Ruppert, J. Kohlhammer, M. Neerincx, and W. Kraaij, "Visual analytics of work behavior data insights on individual differences," Edited by E. Bertini, J. Kennedy, and E. Puppo, Eds., The Eurographics Association, 2015.

[29] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:58624/tab/2.

[30] "uLog by Noldus," https://www.noldus.com/observer-xt/modules.

[31] "Kinect - 3D Sight with Kinect," https://docs.microsoft.com/en-us/archive/msdn-magazine/2012/november/kinect-3d-sight-with-kinect.

[32] "FaceReader by Noldus," https://www.noldus.com/facereader.

[33] "TMSi Ecg Tracking," https://www.tmsi.com/products/saga32-64/.

[34] "Mobi PortiLab 2," http://www.mobihealth.org/html/internal/trials/software/pl2.html.

[35] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[36] Y. LeCun, D. Touresky, G. Hinton, and T. Sejnowski, "A theoretical framework for backpropagation," in *Proceedings of the 1988 Connectionist Models Summer School*, D. Touretzky, G. Hinton, and T. Sejnowski, Eds., vol. 1, pp. 21–28, Morgan Kaufmann, CMU, Pittsburg, PA, 1988.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] G. E. Nasr, E. A. Badr, and C. Joun, "Cross entropy error function in neural networks: forecasting gasoline demand," in *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, pp. 381–384, FL, USA, May 2002.

[39] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: empirical comparisons," *International Journal of Forecasting*, vol. 8, no. 1, pp. 69–80, 1992.

[40] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," 2018, https://arxiv.org/abs/1805.11730.

[41] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load Index): results of empirical and theoretical research," *Advances in Psychology*, vol. 52, pp. 139–183, 1988.

[42] Z. H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial Intelligence*, vol. 137, no. 1-2, pp. 239–263, 2002.

[43] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence Ijcai*, vol. 14, no. 2, pp. 1137–1145, Montreal Quebec Canada, August 1995.

[44] A. Ragav, N. H. Krishna, N. Narayanan, K. Thelly, and V. Vijayaraghavan, "Scalable deep learning for stress and affect detection on resource-constrained devices," in *Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1585–1592, IEEE, Boca Raton, FL, USA, December. 2019.

[45] A. Ragav and G. K. Gudur, "Bayesian active learning for wearable stress and affect detection," 2020, https://arxiv.org/abs/2012.02702.

[46] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, vol. 9, 2021.