This is a peer-reviewed, accepted author manuscript of the following conference paper: McGuire, N & Moshfeghi, Y 2024, On ensemble learning for mental workload classification. in G Nicosia, V Ojha, E La Malfa, et al. (eds), Machine Learning, Optimization, and Data Science: 9th International Conference, LOD 2023, Grasmere, UK, September 22–26, 2023, Revised Selected Papers, Part II. Lecture Notes in Computer Science, vol. 14506, Springer-Verlag, Cham, pp. 358-372. https://doi.org/10.1007/978-3-031-53966-4_27

# On Ensemble Learning for Mental Workload Classification

Niall McGuire and Yashar Moshfeghi

NeuraSearch Laboratory, Department of Computer and Information Sciences, of Strathclyde, Glasgow, Scotland niall.mcguire@strath.ac.uk
yashar.moshfeghi@strath.ac.uk

**Abstract.** The ability to determine a subject's Mental Work Load (MWL) has a wide range of significant applications within modern working environments. In recent years, techniques such as Electroencephalography (EEG) have come to the forefront of MWL monitoring by extracting signals from the brain that correlate strongly to the workload of a subject. To effectively classify the MWL of a subject via their EEG data, prior works have employed machine and deep learning models. These studies have primarily utilised single-learner models to perform MWL classification. However, given the significance of accurately detecting a subject's MWL for use in practical applications, steps should be taken to assess how we can increase the accuracy of these systems so that they are robust enough for use in real-world scenarios. Therefore, in this study, we investigate if the use of state-of-the-art ensemble learning strategies can improve performance over individual models. As such, we apply Bagging and Stacking ensemble techniques to the STEW dataset to classify "low", "medium", and "high" workload levels using EEG data. We also explore how different model compositions impact performance by modifying the type and quantity of models within each ensemble. The results from this study highlight that ensemble networks are capable of improving upon the accuracy of all their individual learner counterparts whilst reducing the variance of predictions, with our highest scoring model being a stacking BLSTM consisting of 8 learners, which achieved a classification accuracy of 97%.

**Keywords:** EEG · Mental Workload · Classification · Deep Learning · Ensembles.

## 1 Introduction

The ability to monitor and predict a person's mental workload levels (MWL) has a vast range of impactful applications within modern working environments, such as teaching/training to monitor the performance of a subject and how well they are handling a problem [31], safety systems to detect when a subjects mental workload levels increase to dangers levels that may increase the likelihood of accidents occurring [23, 44, 9], as well as improving performance by increasing or decreasing the workload of a subject based on their current MWL [18]. A variety

of avenues have already been explored on how to best determine the MWL of a topic, from self-review questionnaires [33] and performance measures [4, 17] as well as more direct techniques such as Functional Magnetic Resonance Imaging (fMRI) [29, 24, 28], Functional Near-Infrared Spectroscopy (fNIRS) [27], and Electroencephalography (EEG) [47, 21] which all monitor the neurophysiological features of a subject's brain.

In recent years Electroencephalography (EEG) data has been widely adopted for MWL detection. This is because EEG data is well suited for MWL detection as the signals produced by the subject's brain are strongly linked to their mental workload levels [36, 39]. EEG also allows for capturing these signals in real-time due to its high temporal resolution (millisecond scale), allowing for developing systems with lower latency and more responsiveness to a subject's mental workload levels. EEG data has already been used to detect, monitor and predict the subject's mental workload [37, 26, 1]. Recently there have been significant developments of machine/deep learning techniques designed to perform mental workload classification, which has been shown to produce high accuracy classification [47, 36].

Many of these previous studies have employed a variety of machine and deep learning models which make use of single-learner algorithms to achieve results to support their research questions, but often need to look into ways in which to obtain the optimal performance of their models. One such way is the utilisation of ensemble learning techniques. Ensemble learning consists of combining multiple individual machine/deep learning models to improve the system's overall performance [34]. The advantages of ensemble learning can include reduced variance in the models' predictions and an overall increase in accuracy score. Ensemble learning has already demonstrated its effectiveness for improving performance in various applications [49] as well as some tasks that use EEG data [51]. However, the capabilities of different ensemble learning techniques for EEG-based mental workload classification have yet to be thoroughly examined. Thus, in this study, we aim to answer the following research questions:

- **RQ1**: "*Can ensemble learning techniques improve the performance of MWL classification over that of single-learner models?*"
- **RQ2**: "*How do different ensemble techniques compare against each other?*"
- **RQ3**: "*What are the effects of differing ensemble compositions on performance?*"

To answer these research questions, we employ two popular ensemble learning techniques, these being Bagging and Stacking [34]. Each of these techniques is made up of a combination of the following state-of-the-art deep learning models used for MWL classification: Gated Recurrent neural Network (GRU), Short-Term Long Memory (LSTM), Bidirectional Gated Recurrent Neural Network (BGRU), and Bidirectional Long Short-Term Memory (BLSTM) [8, 21]. These systems are then trained and evaluated on the open-source mental workload data set STEW [25].

## 2 Background

The ability to detect, monitor and predict a subject's mental workload level/cognitive load has many potential significant applications in a variety of domains, from safety systems that consider a subject's cognitive load to performance optimisation systems [5]. As such, in recent years, it has become a rapidly evolving area of research. Mental workload (MWL) can be defined as the cognitive resources required by a subject whilst engaging in a task. In such cases, the load on the human subject can directly affect the outcome of the given study [32, 5]. Prior research has shown that when an operator's cognitive load increases, their performance may significantly suffer, resulting in lower performance and potential safety risks depending on the task [15, 10]. Thus, it is essential to understand the MWL of the subject to ensure that the job is handled appropriately. Early techniques developed to infer MWL employ subjective strategies that utilise responses gathered from the issue. Examples of these emotional techniques include the National Aeronautics and Space Administration's Task Load Index (NASA TLX) [12] and the Subjective Workload Assessment Technique (SWAT) [35]. Alternatives to subjective responses are performance measures; these estimate the MWL of a subject by evaluating their performance on various main and subtasks during multi-tasking scenarios [45]. Another potential avenue to infer the workload level of an issue is through monitoring their physiological features such as; heart rate, respiratory activity, saccadic eye movements, and brain activity [13, 40, 7]. These techniques have moved to the forefront of MWL classification as the other strategies, such as self-rating questionnaires and performance metrics, can only be feasibly measured in a post-hoc manner requiring the user's cooperation, unlike physiological features that can be monitored continuously during the on-goings of the subject's task [48].

### 2.1 Neuroscience and MWL detection

Recent works have demonstrated the effectiveness of utilising specific brain activity features to infer a subject's cognitive load directly. These studies have made use of techniques such as magnetoencephalography (MEG) [20], fMRI [30], fNIRS [27], and EEG [2, 19, 41] to extract neurophysiological features from the brain. However, EEG data is optimal for the specific task of MWL classification due to its strong correlation to the cognitive load of a subject, its portability, its unobtrusiveness, and its ability to capture data in real-time (millisecond scale).

To more accurately determine the cognitive load of a subject using EEG data, Machine Learning (ML) and Deep Learning (DL) algorithms have been developed to perform MWL classification. Works such as [37] extract the EEG from twenty subjects performing four cognitive and motor tasks consisting of arithmetic operations, finger tapping, mental rotation and lexical decisions, ranging from low to medium to high difficulty levels. The researchers used the EEG data recorded from the subjects to train a single learner ML support vector machine (SVM) to predict the current workload level of a subject from their EEG data. The results showed that they achieved classification accuracy varying from

65%-75% across subjects. Deep learning models were utilised in the works by [6], where in which the researchers employed a deep hybrid model based on a BLSTM and LSTM to perform mental workload classification when trained on the STEW EEG dataset. The results of this study outlined that the model was capable of achieving 82.57% accuracy when predicting whether the mental workload levels of a subject were "low", "medium", or "high" from their EEG data.

### 2.2 Ensemble Learning for MWL detection

Prior works that have employed ML and DL algorithms to perform their classification have primarily limited themselves to a single learner model to fulfil the requirements of their study. However, suppose these systems are to be applied to real-world applications, where accurately determining the MWL of a subject may have significant impacts on the task that the operator is engaging with. In that case, it is vital that these systems must be high performing and consistent in their MWL predictions. A potential solution to this would be using an ensemble network, which combines more than one ML and DL model into a single learner to reduce the variance of the predictions and overall improve the model's accuracy. Some recent works have made use of ensemble learning techniques during their MWL classification tasks, such as [46], where the researchers employed a bagging ensemble of stacked denoising autoencoders to perform the binary classification of two mental workload levels of "low" and "high" from the EEG data gathered across eight subjects. Similarly, in the study [38], the researchers examine the performance of several heterogeneous ML and DL algorithms to perform binary classification "low", and "high" mental workload conditions using the EEG data gathered from eight subjects. Although these studies have examined and demonstrated the capabilities of ensemble learning for MWL classification, they have yet to extensively compare the effects of different ensemble variables, such as bagging vs stacking, mixed vs heterogeneous models, and how the number of ensemble members affects overall performance.

## 3 Methodology

The main steps of this study are EEG signal acquisition, artefact removal, feature engineering, model evaluation and EEG music classification. Details of each step will be described in the following sections.

### 3.1 Data set

For mental workload classification, we utilised the open-source Simultaneous Task EEG Workload (STEW) dataset [16]. The STEW data set creates a multitasking mental workload environment by employing the study's single-session simultaneous capacity (SIMKAP) experiment [3]. The dataset contains the EEG data recorded from 45 subjects using a 14-electrode system sampled at 128 Hz.

The EEG signals were recorded from the subjects while they took part in two experimental conditions, the resting and testing conditions. During the resting state, the participants had to sit in a chair and not perform any tasks or challenges for 3 minutes whilst their EEG data was recorded. Within the testing conditions, the subjects were tasked to participate in the SIMKAP activity whilst their EEG data was recorded. For the EEG data recorded during the SIMKAP test, the last 3 minutes of the recording is considered to be the workload condition. The first and last 15 seconds of the data from each recording were excluded to reduce the effects of any between-task activities; this resulted in each recording being 2.5 minutes long. After every segment of the experiment conditions, each subject was asked to rate their perceived MWL from a 9-point rating scale of 1 to 9. This was done as a subjective validation that the subject did experience an increase in workload whilst taking part in the SIMKAP test when compared to the resting condition. The rating ranges denote that 1-3 can be considered as a low-level workload, 4-6 as a moderate workload, and 7-9 as a high-level workload.

### 3.2 EEG Preprocessing

The final data produced when recording of EEG signals can be strongly affected by the environment and electrical activities of the participant. EEG signals are commonly polluted by noise and artefacts, which are unwanted signals that can affect the raw EEG data and alter signals of interest for our task [42]. The removal of noise and artefacts is a critical step in cleaning/processing EEG data. Prior studies have demonstrated the benefits of these procedures on the overall performance of classification, specifically for ML and DL models [16]. Early techniques employed by neuroscientists to perform pre-processing involved the manual or visual analysis of the recorded EEG signals. Although these procedures were very effective in identifying and removing the majority of artefacts from EEG data, they are particularly time-consuming and require expert knowledge of the field. These procedures may also suffer from the possibility of bias when having human operators examine and remove artefacts [43].

As such, within this study, we employ automatic systems to perform these procedures, speeding up the time to process and removing the possibility of human bias interfering with the processed EEG data. To fulfil our pre-processing, we use the MNE-python library [11]. Firstly, we apply a bandpass filter to the data within a threshold of 0.5-40 Hz, EEG data within this frequency band has been shown to strongly correlate with activation frequencies of the brain, and thus we can consider anything out with this threshold to be noise and artefacts caused by electrical interference or subject muscle activity.

### 3.3 EEG Feature Extraction

After processing and cleaning the EEG signals, we extracted features from the data. Feature extraction of EEG data is important as it can provide the ML and DL models with more meaningful data and gives the models the best chance of

understanding the relationships between the data and their associated MWL. The Features extracted are detailed below:

1. **Morphological features:** the number of peaks, average non-linear energy and the curve length were extracted. Features and three morphological features were extracted in two different ways. Firstly, they were calculated at four frequency bands theta (4–8 Hz), alpha (8–13 Hz), beta (13–25 Hz), and low gamma (25–40 Hz).
2. **Statistical features:** skewness, mean, and kurtosis were computed to identify the distribution of the signal.[21] was calculated to describe time-varying processes
3. **non-linear features:** The approximate entropy (ApEn) and Hurst exponent (H) were used to quantify the unpredictability of fluctuations over a time series as well as measure the self-similarity of the time series, respectively.

All of the features above were extracted from each EEG channel using a time window of 1 second over the recorded EEG data. This resulted in 148 samples for each feature generated per trial a subject took part in.

### 3.4 Models

When examining previous studies that attempt to infer MWLs from EEG data, it can be seen that ML and DL models are commonly employed to perform this classification. Some of the models made use of by these prior studies include the Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Bidirectional Gated Recurrent Unit (BGRU), and the Bidirectional Long Short-Term Memory (BLSTM). These models have seen extensive use in both general EEG and MWL classification due to their effectiveness when learning sequential data with long-term dependencies [50]. Within this study, we will use each of these individual learner models to compose each of our ensemble learners.

**Table 1.** Deep learning model architectures

| Model | Layers/Nodes |
|---|---|
| GRU | G128-G64-G40-D32(D3) |
| BGRU | BG128-G64-G32-D16(D3) |
| LSTM | L128-L64-L40-D32(D3) |
| BLSTM | BL128-L64-L32-D16(D3) |

The architectures for each model can be seen in Table 1. G, BG, L, BL, and D correspond to GRU, BGRU, LSTM, BLSTM, and Dense layer, respectively. For example, BL128-L64-L32-D16(D3) details that there is a BLSTM layer with 128 units, an LSTM layer with 64 units, followed by a Dense-layer with 16 units, and lastly, another Dense-layer with three units (corresponding to the three levels of mental workload we are attempting to predict). SoftMax activation was

implemented in the last layer. This work implemented a dropout rate of 0.2, and the Adam optimiser was used within each of the deep-learning models. Additionally, early stopping was utilised to alleviate the potential issue of overfitting. Furthermore, we limited the number of epochs for training to 200.

### 3.5 Ensemble Strategies

As this study focuses on using ensemble learning techniques to perform MWL classification, we employed two popular ensemble strategies, Bagging and Stacking.

**Bagging** (short for Bootstrap Aggregating) has two main components in creating an ensemble. Firstly bootstrapping; this is where for each learner (in this study, GRU, LSTM, BGRU, BLSTM) within the ensemble, the training data set is randomly re-sampled into subsets that can contain multiple instances of the same data point. These subsets are then assigned to and used to train an individual model, with the idea being that since each of the models has been trained on its own random subset of the data, the ensemble will be able to better generalise for unseen data. The second function of bagging is aggregating. When the ensemble is presented with a testing dataset, it averages the class predictions across each learner and selects the class with the highest number of "votes". This is done to improve the overall accuracy of the model's predictions and reduce its variance.

**Stacking** (short for Stacked generalisation), unlike bagging's aggregation of the individual learner's predictions, uses the prediction outputs of the unique models to train a meta-learner that can better weight the output predictions of each model in the ensemble. The first step is to split the training data into two different groups, one that is used to train each model and one that is used to test the models to obtain their prediction outputs. These prediction outputs are then passed to the meta-learner (for this study, a logistic regressor model) to be trained. Once presented with unseen data, the ensemble passes the data to each learner to obtain their prediction outputs which are then given to the meta-learner to make the final prediction.

### 3.6 Metrics

In this study, the performance of each of the models is evaluated using precision, recall, f1-score and Accuracy. For a classification task such as ours, true positive (TP) is a situation that the model correctly predicted to be positive. A false negative (FN) is when the model incorrectly predicted a case to be negative. A true negative (TN) is a case where the model correctly predicted it to be negative, and a false positive (FP) is a situation where the model incorrectly predicted it be positive. Therefore, each metric is calculated as follows:

**Precision** refers to the ratio of correct positive examples to the number of actual positive examples.

$$Precision = \frac{TP}{(TP + FP)} \tag{1}$$

**Recall** is calculated as the sum of true positives across all classes divided by the sum of true positives and false negatives across all categories.

$$Recall = \frac{TP}{(TP + FN)} \tag{2}$$

**F1-score** is the harmonic mean of the precision and recall.

$$f1score = \frac{2TP}{(2TP + FP + FN)} \tag{3}$$

**Accuracy** is the sum of the number of true positives and true negatives divided by the total number of examples.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{4}$$

### 3.7 Experiment Conditions

To effectively address each of the research questions outlined for this paper, we formulated the following experimental conditions, with the goal of predicting the low, medium, and high mental workload of subjects from EEG data.

**Experiment Scenario 1** Within this study, we make use of four single-learner baseline models to compose each of our ensembles GRU, LSTM, BGRU, and BLSTM. These models act a baseline for us to compare the performance of our ensembles against (Investigating RQ1). We employ two popular ensemble learning strategies; Bagging and Stacking (Investigating RQ2). To begin with, each of these ensemble strategies (Bagging and Stacking) has 5 variations where they are composed of different single learner models: 4-GRU-Learners, 4-LSTM-Learners, 4-BGRU-Leaners, 4-BLSTM-Learners, and a mixed model containing one of each of the models: GRU, LSTM, BGRU, and BLSTM (Investigating RQ3). This means that, in total, we have 10 separate ensemble models for training and evaluation.

**Experiment Scenario 2** To further investigate RQ3 we created the following experiment condition. This scenario follows the same training/evaluation process as the first experiment scenario. However, the only difference is the number of individual learners in each of the ensembles. Within this scenario, we increase the number of individual learners in each of the ensembles from 4 learners to 6, to 8, and lastly, 10. This was done to assess how many learners can be added to an ensemble before the performance begins to stagnate, allowing us to determine the optimal number of learners to compose an ensemble network.

### 3.8 Training Procedure

For each of the experiment scenarios, training and testing data was handled by using a 5-fold time series cross-validation method [22], where 70% of the fold was utilised for training, 10% for validation, and the remaining 20% for testing.

## 4 Results

In this section, we use a box plot to display the results for the variations of the ensemble models compared against the baseline individual learners that the ensembles are made up of. Each box plot presents five crucial pieces of information: the minimum, first, second (median), third, and maximum quarterlies [14]. Where the red dot in the box represents the mean accuracy for each model.

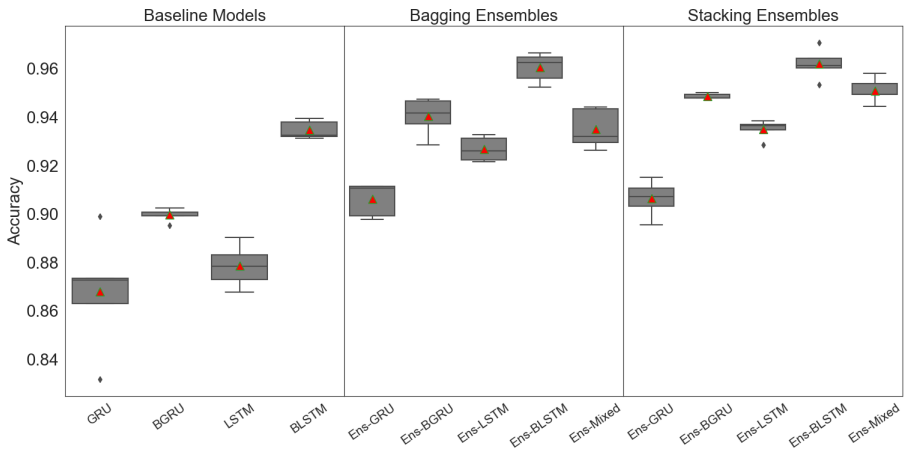



**Fig. 1.** Baseline and Ensemble Models Accuracy

### 4.1 Scenario 1

The results for experiment scenario one are displayed in figure 1. This figure contains box plots for the 5-fold cross-validation results of each model's accuracy. Within each block, we have the baseline models, the bagging models, and the stacking models. We first compare the ensemble models' results to their respective learner counterparts. It can be observed that every ensemble model for both bagging and stacking strategies has improved accuracy over each of its component learner models. On average, Bagging and Stacking improved the accuracy across each heterogeneous model by 3.7% and 4.2%, respectively. The model that saw the most significant improvement above the baseline was the Stacking LSTM, which achieved an accuracy of 93.5%, a 5.6% increase in performance

from the baseline LSTM score of 87.9%. This helps to demonstrate ensemble networks' effectiveness in improving individual learners' performance (RQ1).

Following this, we examine how the two ensemble strategies used compare against each other. On average, the Stacking models had an improved performance of 0.7% over the bagging models. When looking at the highest-performing models for both strategies, we see that the BLSTM for bagging and stacking achieves the highest accuracy scores of 96% and 96.2%, respectively, highlighting that the stacking models slightly outperform their bagging counterparts. This shows that, although very close in performance, the stacking ensemble strategy, on average, offers improved performance over that of bagging (RQ2).

Lastly, we explore how the different model compositions compare against each other. Within the bagging ensemble plots, we can see that the order of lowest to highest-performing models is as such: GRU (90.6%), LSTM (92.7%), Mixed (93.5%), BGRU (94%), and BLSTM (96%). For the stacking ensemble, the order follows GRU (90.6%), LSTM (93.5%), BGRU (94.9%), Mixed (95.1%), and BLSTM (96.2%). Both strategies follow a similar expected performance pattern that mirrors the baseline models; the only exception is the Mixed model, which reliably outperforms GRU and LSTM but differs in its ability to outperform BGRU depending on the ensemble strategy used. This is likely due to the Mixed model containing two weaker learners (GRU, LSTM) than the homogeneous BGRU and BLSTM ensembles. This implies that when given the option of composing a model of varying performance models versus a homogeneous model of strong learners, the latter of the two options will obtain greater performance (RQ3).

**Table 2.** Stacking Ensembles with Varying Learner Count

| **Model** | **Member Count** | **Precision (S.D)** | **Recall (S.D)** | **F1-Score (S.D)** | **Accuracy (S.D)** |
|---|---|---|---|---|---|
| GRU | 6 | 0.91 (0.8) | 0.91 (0.8) | 0.91 (0.8) | 0.91 (0.8) |
| GRU | 8 | 0.92 (0.3) | 0.92 (0.3) | 0.92 (0.3) | 0.92 (0.3) |
| GRU | 10 | 0.94 (0.1) | 0.94 (0.1) | 0.94 (0.1) | 0.94 (0.1) |
| BGRU | 6 | 0.95 (0.4) | 0.95 (0.4) | 0.95 (0.4) | 0.95 (0.4) |
| BGRU | 8 | 0.94 (0.8) | 0.94 (0.8) | 0.94 (0.8) | 0.94 (0.8) |
| BGRU | 10 | 0.94 (0.5) | 0.94 (0.5) | 0.94 (0.5) | 0.94 (0.5) |
| LSTM | 6 | 0.93 (0.6) | 0.93 (0.6) | 0.93 (0.6) | 0.93 (0.6) |
| LSTM | 8 | 0.93 (0.2) | 0.93 (0.2) | 0.93 (0.2) | 0.93 (0.2) |
| LSTM | 10 | 0.94 (0.3) | 0.94 (0.3) | 0.94 (0.3) | 0.94 (0.3) |
| BLSTM | 6 | 0.96 (0.3) | 0.96 (0.3) | 0.96 (0.3) | 0.96 (0.3) |
| BLSTM | 8 | 0.97 (0.5) | 0.97 (0.5) | 0.97 (0.5) | **0.97 (0.5)** |
| BLSTM | 10 | 0.96 (0.7) | 0.96 (0.7) | 0.96 (0.7) | 0.96 (0.7) |

**Table 3.** Bagging Ensembles with Varying Learner Count

| model | member count | Precision (S.D) | Recall (S.D) | F1-Score (S.D) | Accuracy (S.D) |
|---|---|---|---|---|---|
| GRU | 6 | 0.91 (0.2) | 0.91 (0.2) | 0.91 (0.2) | 0.91 (0.2) |
| GRU | 8 | 0.92 (0.4) | 0.91 (0.4) | 0.91 (0.4) | 0.91 (0.4) |
| GRU | 10 | 0.91 (0.3) | 0.91 (0.3) | 0.91 (0.3) | 0.91 (0.3) |
| BGRU | 6 | 0.94 (0.4) | 0.94 (0.4) | 0.94 (0.4) | 0.94 (0.4) |
| BGRU | 8 | 0.94 (0.5) | 0.94 (0.5) | 0.94 (0.5) | 0.94 (0.5) |
| BGRU | 10 | 0.94 (0.3) | 0.94 (0.3) | 0.94 (0.3) | 0.94 (0.3) |
| LSTM | 6 | 0.92 (0.5) | 0.92 (0.5) | 0.92 (0.5) | 0.92 (0.5) |
| LSTM | 8 | 0.93 (0.4) | 0.93 (0.4) | 0.93 (0.4) | 0.93 (0.4) |
| LSTM | 10 | 0.93 (0.6) | 0.93 (0.6) | 0.93 (0.6) | 0.93 (0.6) |
| BLSTM | 6 | 0.94 (0.7) | 0.94 (0.7) | 0.94 (0.7) | 0.94 (0.7) |
| BLSTM | 8 | 0.94 (0.3) | 0.94 (0.3) | 0.94 (0.3) | 0.94 (0.3) |
| BLSTM | 10 | 0.94 (0.5) | 0.94 (0.5) | 0.94 (0.5) | 0.94 (0.5) |

### 4.2 Scenario 2

The results for experiment scenario 2 are presented in table 2, which contains the results for the stacking ensembles, and table 3, which contains the results for the bagging ensembles. Firstly, when comparing the accuracy scores for each model across both ensemble strategies, we can observe that the bagging ensemble accuracies remain stagnant as the number of member increase from 6 to 8. Whereas it can be observed that a number of the stacking models' accuracy appears to improve, with the highest performing model being the stacking BLSTM achieving an accuracy of 97% using eight learners. These results imply that the stacking strategy appears the benefit from an increased number of learners above the base count of 4. In contrast, to the bagging strategy shows no benefits from increasing the learner count, adding further considerations when selecting optimal ensemble composition (RQ3).

## 5 Conclusion

To accurately determine a subject's Mental Work Load (MWL) level opens the door to an extensive range of beneficial applications within modern working environments. Areas such as health and safety, teaching, and resource management can benefit directly through monitoring and classifying a person's MWL. In recent years, significant progress has been made in the field of MWL detection thanks to the inclusion of direction brain monitoring techniques such as Electroencephalography (EEG). Prior studies have demonstrated the ability to train single machine/deep learning models on this data to classify MWL levels automatically. However, if these systems are to be applied in real-world scenarios, steps should be taken to explore how these models can be made as accurate and reliable as possible.

Therefore, in this study, we explored the viability of ensemble learning models to provide improved performance over single-learner models when applied to the STEW [25] MWL data set. Furthermore, we examined how different ensemble learning techniques (Bagging and Stacking) compare in performance and

explored how ensemble composition (model type and quantity) affects performance. To do this, we composed several Bagging and Stacking ensembles from the following individual learners – GRU, BGRU, LSTM, and BLSTM. Our findings demonstrated the substantial improvement of all ensemble networks over their learner counterparts, with a bagging and stacking ensemble consisting of four learners on average increasing accuracy by 3.7% and 4.2%, respectively, whilst reducing the variation in the prediction output. Moreover, our results also show that performance benefits can be seen by including more members in the ensemble when using the stacking strategy, with the highest achieving ensemble being the BLSTM with eight learners that obtained an accuracy score of 97%.

Future works that aim to perform MWL classification within the context of a real-world scenario should consider the usage of ensemble networks to improve the system's overall performance. Furthermore, the ensemble networks presented in this study can also be applied to other research areas that wish to improve their performance when performing EEG data classification.

## References


1. Aghajani, H., Garbey, M., Omurtag, A.: Measuring mental workload with eeg+ fnirs. Frontiers in human neuroscience **11**, 359 (2017)
2. Allegretti, M., Moshfeghi, Y., Hadjigeorgieva, M., Pollick, F.E., Jose, J.M., Pasi, G.: When relevance judgement is happening? an eeg-based study. In: Proceedings of the 38th international acm sigir conference on research and development in information retrieval. pp. 719–722 (2015)
3. Bratfisch, O., Hagman, E.: Simkap–simultankapazität/multi-tasking. Mödling: Schuhfried GmbH (2008)
4. Butmee, T., Lansdown, T., Walker, G.: Mental Workload and Performance Measurements in Driving Task: A Review Literature: Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics, pp. 286–294 (01 2019)
5. Cain, B.: A review of the mental workload literature (2007)
6. Chakladar, D.D., Dey, S., Roy, P.P., Dogra, D.P.: Eeg-based mental workload estimation using deep blstm-lstm network and evolutionary algorithm. Biomedical Signal Processing and Control **60**, 101989 (2020)
7. Charles, R.L., Nixon, J.: Measuring mental workload using physiological measures: A systematic review. Applied ergonomics **74**, 221–232 (2019)
8. Chen, J., Jiang, D., Zhang, Y.: A hierarchical bidirectional gru model with attention for eeg-based emotion classification. IEEE Access **7**, 118530–118540 (2019)
9. Dehais, F., Somon, B., Mullen, T., Callan, D.: A neuroergonomics approach to measure pilot's cognitive incapacitation in the real world with eeg (03 2020)
10. Deng, P.Y., Qiu, X.Y., Tang, Z., Zhang, W.M., Zhu, L.M., Ren, H., Zhou, G.R., Sheng, R.S., et al.: Detecting fatigue status of pilots based on deep learning network using eeg signals. IEEE Transactions on Cognitive and Developmental Systems **13**(3), 575–585 (2020)
11. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al.: Meg and eeg data analysis with mne-python. Frontiers in neuroscience p. 267 (2013)

12. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 50, pp. 904–908. Sage publications Sage CA: Los Angeles, CA (2006)
13. Henelius, A., Hirvonen, K., Holm, A., Korpela, J., Muller, K.: Mental workload classification using heart rate metrics. In: 2009 annual international conference of the ieee engineering in medicine and biology society. pp. 1836–1839. IEEE (2009)
14. Hofmann, T.: Collaborative filtering via gaussian probabilistic latent semantic analysis. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 259–266 (2003)
15. Hu, X., Lodewijks, G.: Detecting fatigue in car drivers and aircraft pilots by using non-invasive measures: The value of differentiation of sleepiness and mental fatigue. Journal of safety research **72**, 173–187 (2020)
16. Islam, M.K., Rastegarnia, A., Yang, Z.: Methods for artifact detection and removal from scalp eeg: A review. Neurophysiologie Clinique/Clinical Neurophysiology **46**(4-5), 287–305 (2016)
17. Jafari, M., Zaeri, F., Jafari, A., Najafabadi, A., Al-Qaisi, S., Hassanzadeh Rangi, N.: Assessment and monitoring of mental workload in subway train operations using physiological, subjective, and performance measures. Human Factors and Ergonomics in Manufacturing Service Industries **30** (01 2020). https://doi.org/10.1002/hfm.20831
18. Kandemir, C., Handley, H.A.: Work process improvement through simulation optimization of task assignment and mental workload. Computational and Mathematical Organization Theory **25**, 389–427 (2019)
19. Karameh, F.N., Dahleh, M.A.: Automated classification of eeg signals in brain tumor diagnostics. In: Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No. 00CH36334). vol. 6, pp. 4169–4173. IEEE (2000)
20. Kauppi, J.P., Kandemir, M., Saarinen, V.M., Hirvenkari, L., Parkkonen, L., Klami, A., Hari, R., Kaski, S.: Towards brain-activity-controlled information retrieval: Decoding image relevance from meg signals. NeuroImage **112**, 288–298 (2015)
21. Kingphai, K., Moshfeghi, Y.: On eeg preprocessing role in deep learning effectiveness for mental workload classification. In: International Symposium on Human Mental Workload: Models and Applications. pp. 81–98. Springer (2021)
22. Kingphai, K., Moshfeghi, Y.: On time series cross-validation for deep learning classification model of mental workload levels based on eeg signals. In: International Conference on Machine Learning, Optimization, and Data Science. pp. 402–416. Springer (2022)
23. Kingphai, X., preprocessing role in deep learning effectiveness for mental workload classification. In International Symposium on Human Mental Workload: Models, Y.M..O.E., Applications. Springer, ., Lodewijks, G.: Detecting fatigue in car drivers and aircraft pilots by using non-invasive measures: The value of differentiation of sleepiness and mental fatigue. Journal of Safety Research **72**, 173–187 (2020). https://doi.org/https://doi.org/10.1016/j.jsr.2019.12.015
24. Lim, J., chau Wu, W., Wang, J., Detre, J.A., Dinges, D.F., Rao, H.: Imaging brain fatigue from sustained mental workload: An asl perfusion study of the time-on-task effect. NeuroImage **49**(4), 3426–3435 (2010)
25. Lim, W.L., Sourina, O., Wang, L.P.: Stew: Simultaneous task eeg workload data set. IEEE Transactions on Neural Systems and Rehabilitation Engineering **26**(11), 2106–2114 (2018)
26. Lim, W.L., Sourina, O., Liu, Y., Wang, L.: Eeg-based mental workload recognition related to multitasking. In: 2015 10th International Conference on In-

formation, Communications and Signal Processing (ICICS). pp. 1–4 (2015). https://doi.org/10.1109/ICICS.2015.7459834
27. Midha, S., Maior, H.A., Wilson, M.L., Sharples, S.: Measuring mental workload variations in office work tasks using fnirs. International Journal of Human-Computer Studies **147**, 102580 (2021)
28. Moshfeghi, Y., Pinto, L.R., Pollick, F.E., Jose, J.M.: Understanding relevance: An fmri study. In: Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35. pp. 14–25. Springer (2013)
29. Moshfeghi, Y., Triantafillou, P., Pollick, F.E.: Understanding information need: An fmri study. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 335–344 (2016)
30. Moshfeghi, Y., Triantafillou, P., Pollick, F.E.: Understanding information need: An fmri study. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 335–344. SIGIR '16, Association for Computing Machinery, New York, NY, USA (2016)
31. Orru, G., Gobbo, F., O?Sullivan, D., Longo, L.: An investigation of the impact of a social constructivist teaching approach, based on trigger questions, through measures of mental workload and efficiency. In: McLaren, B.M., Reilly, R., Zvacek, S., Uhomoibhi, J. (eds.) Proceedings of the 10th International Conference on Computer Supported Education, pp. 292–302. Scitepress ? Science and Technology Publications (2018)
32. Pandey, V., Choudhary, D.K., Verma, V., Sharma, G., Singh, R., Chandra, S.: Mental workload estimation using eeg. In: 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). pp. 83–86. IEEE (2020)
33. Paxion, J., Galy, E., Berthelon, C.: Mental workload and driving. Frontiers in Psychology **5** (2014). https://doi.org/10.3389/fpsyg.2014.01344
34. Polikar, R.: Ensemble learning. Ensemble machine learning: Methods and applications pp. 1–34 (2012)
35. Reid, G.B., Nygren, T.E.: The subjective workload assessment technique: A scaling procedure for measuring mental workload. In: Advances in psychology, vol. 52, pp. 185–218. Elsevier (1988)
36. Singh, U., Ahirwal, M.K.: Mental workload classification for multitasking test using electroencephalogram signal. In: 2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES). pp. 1–6 (2021). https://doi.org/10.1109/TRIBES52498.2021.9751676
37. So, W.K., Wong, S.W., Mak, J.N., Chan, R.H.: An evaluation of mental workload with frontal eeg. PloS one **12**(4), e0174949 (2017)
38. Tao, J., Yin, Z., Liu, L., Tian, Y., Sun, Z., Zhang, J.: Individual-specific classification of mental workload levels via an ensemble heterogeneous extreme learning machine for eeg modeling. Symmetry **11**(7), 944 (2019)
39. Teplan, M.: Fundamental of eeg measurement. MEASUREMENT SCIENCE REVIEW **2** (01 2002)
40. Tokuda, S., Obinata, G., Palmer, E., Chaparro, A.: Estimation of mental workload using saccadic eye movements in a free-viewing task. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4523–4529. IEEE (2011)
41. Tzallas, A.T., Tsipouras, M.G., Fotiadis, D.I.: Epileptic seizure detection in eegs using time–frequency analysis. IEEE transactions on information technology in biomedicine **13**(5), 703–710 (2009)

42. Urigüen, J.A., Garcia-Zapirain, B.: Eeg artifact removal—state-of-the-art and guidelines. Journal of neural engineering **12**(3), 031001 (2015)
43. Vaid, S., Singh, P., Kaur, C.: Eeg signal analysis for bci interface: A review. In: 2015 fifth international conference on advanced computing & communication technologies. pp. 143–147. IEEE (2015)
44. wu, E.Q., Deng, P.Y., Qiu, X.Y., Tang, Z., Zhang, W.M., Zhu, L.M., Ren, H., Zhou, G.R., Sheng, R.S.F.: Detecting fatigue status of pilots based on deep learning network using eeg signals. IEEE Transactions on Cognitive and Developmental Systems **13**(3), 575–585 (2021). https://doi.org/10.1109/TCDS.2019.2963476
45. Xie, B., Salvendy, G.: Prediction of mental workload in single and multiple tasks environments. International journal of cognitive ergonomics **4**(3), 213–242 (2000)
46. Yang, S., Yin, Z., Wang, Y., Zhang, W., Wang, Y., Zhang, J.: Assessing cognitive mental workload via eeg signals and an ensemble deep learning classifier based on denoising autoencoders. Computers in biology and medicine **109**, 159–170 (2019)
47. Yin, Z., Zhang, J.: Cross-session classification of mental workload levels using eeg and an adaptive deep learning model. Biomedical Signal Processing and Control **33**, 30–47 (2017). https://doi.org/https://doi.org/10.1016/j.bspc.2016.11.013
48. Zarjam, P., Epps, J., Lovell, N.H.: Beyond subjective self-rating: Eeg signal classification of cognitive workload. IEEE Transactions on Autonomous Mental Development **7**(4), 301–310 (2015)
49. Zhang, C., Ma, Y.: Ensemble machine learning: methods and applications. Springer (2012)
50. Zhao, R., Yan, R., Wang, J., Mao, K.: Learning to monitor machine health with convolutional bi-directional lstm networks. Sensors **17**(2), 273 (2017)
51. Zheng, X., Chen, W., You, Y., Jiang, Y., Li, M., Zhang, T.: Ensemble deep learning for automated visual classification using eeg signals. Pattern Recognition **102**, 107147 (2020). https://doi.org/https://doi.org/10.1016/j.patcog.2019.107147