Human-Centric Intelligent Systems
https://doi.org/10.1007/s44230-024-00086-y

RESEARCH ARTICLE

An End-to-End Brain Computer Interface System for Mental Workload Estimation through Hybrid Deep Learning Model

Vipul Sharma[1] · **Mitul Kumar Ahirwal**[1]

Received: 13 February 2024 / Accepted: 30 October 2024

Abstract
In this paper, a new cascade one-dimensional convolutional neural network (1DCNN) and bidirectional long short-term memory (BLSTM) model has been developed for binary and ternary classification of mental workload (MWL). MWL assessment is important to increase the safety and efficiency in brain–computer interface (BCI) systems and professions, where multi-tasking is required. Keeping in mind the necessity of MWL assessment, a two-fold study is presented, firstly binary classification is done to classify MWL into low and high classes. Secondly, ternary classification is applied to classify MWL into low, moderate, and high classes. The cascaded1DCNN-BLSTM deep learning architecture has been developed and tested over the Simultaneous task EEG workload (STEW) dataset. Unlike recent research in MWL, handcrafted feature extraction and engineering are not done, rather end-to-end deep learning is used over 14 channel EEG signals for classification. Accuracies exceeding the previous state-of-the-art studies have been obtained. In binary and ternary classification accuracies of 96.77% and 95.36%have been achieved with sevenfold cross-validation, respectively.

Keywords Bidirectional long short-term memory · Brain–computer interface · Multivariate time series classification · Convolutional neural networks · Deep learning · Electroencephalography · Mental workload · Recurrent neural networks

Abbreviations

ATC	Air traffic control
ART	Auditory response test
ANS	Autonomic nervous system
BLSTM	Bidirectional long short-term memory
BCI	Brain-computer interface
CNN	Convolutional neural network
ECG	Electrocardiogram
EEG	Electroencephalogram
ERPs	Event related potentials
HM	Human machine
LC	Letter counting
LSTM	Long short-term memory
MWL	Mental workload
MATB	Multi-attribute task battery
NASA-TLX	NASA task load index
NFS-S2U	Need for speed-shift 2 unleashed
1DCNN	One-dimensional convolution neural network
OFS	Operator functional states
PGBM	Point-wise gated Boltzmann machines
RNN	Recurrent neural networks
RSEFNNs	Recurrent self-evolving fuzzy neural networks
STEW	Simultaneous task EEG workload
SIMKAP	Single-session simultaneous capacity
SWAT	Subjective workload assessment technique
SDBN	Switching deep-belief network
TAV	Task of attention (sound) and video stimuli
2DCNN	Two-dimensional convolutional neural network
VRT	Visual response test

✉ Mitul Kumar Ahirwal
ahirwalmitul@gmail.com

Vipul Sharma
vipuls1996@outlook.com

[1] Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, M.P. 462003, India

1 Introduction

Nowadays, almost every person suffers from mental stress either due to their lifestyle or their nature of work or profession. Mental stress may lead to mental disorders, and there is a strong correlation between the two. Psychological

Published online: 12 November 2024

and mental stress are synonymous, and are associated with anger or anxiety. These may also lead to depression, if they remain untreated for a long time. Mentally stressed conditions also affect the functionalities of autonomic nervous system (ANS). Therefore, mental stress is considered the main cause of overall degradation of a person's mental and physical health. Due to stress, a person may also lose interest in their profession and related works. Mental stress also changes attitudes toward life [1–6].

In various industries and professions, higher mental workload, mental stress, and mental pressure lead to stress-related diseases that decrease the performance and output of industries and increase the burden of medical expenses of employees. Mental workload, mental stress, and related diseases can also increase economic and social losses for the whole country. Suicides and psychiatric illnesses due to mental stress are also reported [1–6]. In universities and colleges, faculty and student's mental health also suffers due to mental stress generated from various factors as reported in [7]. Hence, precautionary measures to reduce mental stress and its proper assessment are necessary for the safety and benefit of humanity. For assessment of brain/mental state physiological signal analysis is the best way.

Electroencephalogram (EEG) signals are the most suitable physiological signal to explore the mental state of humans. Different types of experiments and simulations have been designed as protocols to understand the influence of mental workload and mental stress. Protocols and tools like visual response test (VRT), auditory response test (ART), letter counting (LC) test, stroop test, NASA task load index (NASA-TLX), mental arithmetic/calculation, NASA's multi-attribute task battery (MATB), Subjective workload assessment technique (SWAT) and single-session simultaneous capacity (SIMKAP) task were designed to observe and identify the level of alertness and mental workload. These protocols and tools are related to cognitive tasks which induce mental fatigue and change the level of the subject's alertness [8–12].

Recorded EEG data analyzed through these protocols is very useful in the proper assessment of mental workload which causes stress and other issues, and based on this timely diagnosis and treatment is possible. Recently, machine learning techniques have started being used to automate the assessment of mental workload. The major issue in these experiments and simulations is the accuracy of classification or prediction of mental workload. The accuracy of the entire protocol or model depends majorly upon factors like pre-processing of data (generally filtering), the number of classes or categories or tasks to be classified, and the classification algorithm used for training and testing. These factors are targeted by different researchers in several studies [13–19]. Among these factors, classification algorithm plays a crucial role, because, at present, methods like deep learning and specifically convolution neural networks with other models or algorithms have surpassed other classification methods that work on features extracted from EEG signals [11–13].

The common idea in most deep learning models is using the initial convolutional neural network (CNN) layers for the generation of feature maps and then the use of fully connected layers to classify these feature maps. Layers like pooling and drop-out are used to prevent over-fitting in such models. Many recent studies have used these CNN-based classification models for the classification of EEG signals [20–22]. For EEG signals that are time series data, many researchers have also applied long short-term memory (LSTM) models. In the LSTM model, generative and discriminative capabilities of recurrent neural networks (RNN) have been used. The use of RNNs allows for temporal features to be extracted, and CNNs help in extracting spatial features from the data.

2 Related Works

Some of the recent related studies are summarized in this section. Several recent study utilized deep learning concepts in the fields related to mental workload. Most of the research done focuses on handcrafted feature extraction from EEG signals. A deep 1DCNN has been used for attention classification in the stroop color test. In this, raw data, filtered data, and data in five conventional EEG bands are given for training [16]. In [17], a CNN model has been developed which can be used as a generalized model for few EEG brain–computer interface (BCI) systems working on P300 event related potentials (ERPs) with visual stimulation, neural oscillations generated for movement-related cortical potentials, and several sensorimotor rhythms generated due to real and imaginary limb movements. Another important field is driver fatigue monitoring, because of its relationship with traffic accidents. Driving simulators like "need for speed-shift 2 unleashed (NFS-S2U)" and "World Record" software are used in these experiments. Along with EEG, electrocardiogram (ECG or EKG) is recorded and used for the classification of two mental states, namely, "DROWS" and "Task of attention (sound) and video stimuli (TAV)". The models named as EEG-Conv and EEG-Conv-R that are based on deep CNN and deep residual learning concepts have been proposed in [18] for this task. In [19], a deep classifier and a deep autoencoder were used for task engagement assessment i.e., to learn and label three types of events in flight simulation. EEG and ECG signals were used to monitor the state of pilots in a 4-h flight simulation and three events were classified, namely two types of air traffic control (ATC) calls and one failure event. Point-wise gated Boltzmann machines (PGBM) have been used to classify

the mental state of subjects in task-relevant or task-irrelevant categories, where each subject underwent a working memory experiment with a set of characters [23]. Assessing operator functional states (OFS) plays an important role in safety critical human machine (HM) systems. A new switching deep-belief networks (SDBN) with adaptive-weights has been implemented for detection of separate and coupling effect of mental workload and mental fatigue across different subjects [24]. In this, the automation enhanced cabin air management system (Auto CAMS) is used as platform to simulate complex processes as control tasks for real-time HM collaboration. In [25], RNNs were used in the effective prediction of drowsiness in a high-fidelity vehicle simulator study using EEG, for driving tasks. Modification in same has been done with ensemble group-trained recurrent self-evolving fuzzy neural networks (RSEFNNs). A deep CNN-RNN model was used to predict cognitive load generated due to working memory tasks with the help of 2D azimuthal equidistant projections (AEPs) of power spectral density (PSD) features of different EEG bands [26, 27]. In most of the RNN models used, the temporal processing direction is only forward, sometimes this reduces the extent of temporal information extracted from the data. To overcome this limitation a bidirectional long short-term memory (BLSTM) model has been used for epileptic seizure classification, followed by 2DCNN and fully connected layers in [28]. The Application of BLSTM has also been reported in [29] for the classification of the mental workload from extracted EEG signal features. It involved a combination of BLSTM, and LSTM being used for the classification of mental workload during the task and no-task states. In most of the research stated above, manually extracted features like certain time and frequency domain features, linear and non-linear features, etc. are used instead of raw EEG signals. In addition to this, many feature selection techniques and evolutionary algorithms for feature selection or manual selection of features have been done. Few dedicated models like EEG-TNet, variational autoencoder and attention model, and Integrated spatio-temporal deep clustering (ISTDC) are reported in [34–36], latest studies are summarized in [37, 38].

Hence, building on the existing research, in this paper, a new cascaded 1DCNN and BLSTM model to classify mental workload in two and three classes is proposed. To the author's best knowledge, the points of novelty of this study are:

1. Most of the previous studies on the STEW dataset [14, 15] classify mental workload state between "task" and "no-task" states, but here different levels of mental workload during multitasking, i.e., "task" state, in binary as well as ternary classes were classified.
2. Before this work, a combination of CNN and BLSTM was not applied to the mental workload data used in this study. This model surpasses the current state-of-the-art models.
3. EEG signals for end-to-end deep learning were used. To the author's best knowledge, no other study on the STEW dataset has done so, instead, they have focused on handcrafted feature extraction and engineering.

Motivation: The motivation behind this research is to provide a computerized solution for mental workload analysis based on objective assessment. Subjective assessments are not reliable, due to various biases. This computerized solution is also demanding at present. Due to lack of medical professionals in the mental health field as compared to the rise in number of mental health issues caused because of excess mental workload.

Contribution: Major contributions of this work are:

1. End-to-End deep learning based classification model for mental workload classification.
2. Frame work for utilization of same database for binary as well as ternary classification of mental workload.
3. Besides the above contributions, a new learning rate modification method during the training phase of the proposed 1DCNN-BLSTM model has been also suggested.

The rest of the paper is arranged as, Sect. 2 presents the overall methodology, Sect. 3 explainsthe details of the proposed method, Sect. 4 displays the results obtained, Sect. 5 discusses the performance of the proposed model, and its comparison with recent research, and Sect. 6 gives the concluding remarks.

3 Methodology

3.1 Dataset Description

In this study, Simultaneous task EEG workload (STEW) dataset [14, 15] is used for the mental workload classification task. STEW measures the mental workload during "no task" and the workload induced by "simultaneous capacity (SIMKAP)-based multi-tasking activity". EEG recordings during SIMKAP have been analyzed in the experiments. The SIMKAP involves the subjects being given simultaneous audio-visual tasks like arithmetic, finding identical items on two separate windows, data lookup etc., and at the end of the tasks they rate their mental workload on a scale of 1–9. In STEW dataset 45 subjects' EEG recordings with their mental workload ratings during SIMKAP is provided. These ratings were binned into 2 and 3 classes respectively for binary and ternary classification. Tables 1 and 2 show the distribution of the ratings in each class.

Table 1 Frequency distribution of classes for binary classification

Class	Range of ratings	Subjects
Low workload (class 0)	4–6	20
High workload (class 1)	7–9	25

Table 2 Frequency distribution of classes for ternary classification

Class	Range of ratings	Subjects
Low workload (class 0)	4–5	13
Moderate workload (class 1)	6–7	18
High workload (class 2)	8–9	14

The EEG signals were captured with 14 electrodes, namely, AF_3, F_7, F_3, FC_5, T_7, P_7, O_1, O_2, P_8, T_8, FC_6, F_4, F_8 and AF_4, during the SIMKAP test with a sampling frequency of 128 Hz for 2.5 min. Bandpass filter with a permissible frequency range of 4–32 Hz is used to remove artefacts from the EEG recordings.

3.2 Experimental Setup

In this experiment, 45 multichannel EEG recordings have been considered, each 2.5 min long. To augment this data for deep learning models, windowing has been done over the dataset with overlapping windows of size 512 samples and shift of 128 samples. This sub-sampling is performed over 14 channels, and labels are repeated for subsample as per their original sample. This augmentation produced 6615 samples from the initial 45 subjects' data. Further, one-hot encoding for the class labels is done. Table 3 describes the shape of the dataset thus produced. For classification, 85% of data (5622 samples) were used for training purpose, and 15% (993 samples) for testing the deep learning model. In addition to this, K-fold cross-validation (CV) is also performed, after several initial experiments, fivefold and sevenfold CV are found suitable for final experiments, and to check the robustness of the results.

A deep learning model for the multivariate time series i.e., EEG signals, classification into 3 and 2 classes has been developed. The model consists of 1D convolution (1DCNN) layers followed by bidirectional long short-term memory (BLSTM) layers for feature extraction. A fully connected neural network to the output of these layers is also used for classification. The detailed structure of the layers is discussed in Sect. 3. The use of deep learning has allowed for the classification of complex multichannel EEG data without the need for handcrafted feature extraction, demonstrating the power of deep learning.

3.3 Description of Layers Used

The CNN-BLSTM model is used in the experiment for both binary and ternary classification of EEG signals, this model learns both the spatial and the temporal characteristics of multichannel EEG signals to do automated feature extraction.

3.3.1 1D Convolutional Neural Network (1D CNN)

1D CNN works based on convolution operations using kernels/filters. Several kernels of small size are passed over the data to learn local patterns from small patches of data and do feature extraction. They learn the spatial information from multivariate time series easily and are often stacked to do feature extraction from raw data.

3.3.2 Long Short-Term Memory (LSTM)

LSTM was developed by Hochreiter et al. [23] and is a special type of recurrent neural network (RNN) used for learning temporal information. RNNs are a type of neural network which utilize the previous cell's output in the current cell or state along with the sequence input. RNNs suffer from vanishing gradient problem which leads to the gradient becoming zero for long sequences. LSTM overcomes this problem and is useful to learn information from long sequences. They consist of 4 blocks, the cell state, the forget gate, the input gate, and the output gate. The cell state helps to transfer the information from earlier states to later cells solving the vanishing gradient problem. Further, the forget state learns what information should be retained or forgotten. A combination of these two helps to develop a mix of long and short-term memory.

3.3.3 Bidirectional LSTM (BLSTM)

LSTMs are traditionally unidirectional, i.e., they process the time series in only one direction from past to future. To overcome this limitation, an extension to RNNs was proposed by Schuster et al. [31–33] as a bidirectional recurrent neural network (BRNN) that can simultaneously train in the positive and negative time direction. BLSTMs are a type of BRNN that can process the data parallel in both forward and backward direction and the output of LSTMs merged to produce the final output. This bidirectional reading allows BLSTMs to learn the temporal information from the data in a better way.

Table 3 Augmented EEG data format

Array name	Array shape
EEG data	6615 × 512 × 14 samples × windowed EEG data × channels
Class labels	6615 × 2/ 6615 × 3 samples × number of classes (2 or 3)

4 Proposed Method

In this section, the proposed model is explained with description of the layers used in the model i.e., the model parameters and the hyperparameters.

The development of the proposed CNN-BLSTM model is done with the help of Python language. First of all data dimensional of input data (EEG data files) has been arranged as three dimensional file. Further, the proposed CNN-BLSTM model architecture, shown in Fig. 1, consists of two 1D CNN layers stacked with two BLSTM layers which are then followed by a dense layer and output layer for classification. The first 1D CNN layer has 32 filters, each with a kernel of size 16 and stride length 1. The output of this layer is passed through a ReLU activation function. The second layer is also a 1D CNN layer with 16 filters, each filter has a kernel size of 8 and stride length of 1 with ReLU activation function. The output of these stacked 1D CNN layers is then passed to BLSTM layers. The first BLSTM layer has 32 neurons with a tanh activation function. The output of this BLSTM layer is a sequence that is fed into another BLSTM layer with 32 neurons with tanh activation function. The second BLSTM layer generates a single vector as its output which is fed to adense layer for classification, consisting of 40 neurons and output layer having two or three neurons depending on the type of classification. A Softmaxfunctionis used at the end for the mental workload classification task. This architecture is also summarized in Table 4.

The resulting models have 51,370 and 51,411 trainable parameters respectively for binary and ternary classification tasks. The modelswere trained using stochastic gradient descent (SGD) optimizer with cross-entropy loss. Appropriate batch size is selected from the factors of the size of training data i.e., a number that evenly divides the training set. Learning rate is chosenin a specific way as described by Leslie N. Smith in [30]. Initial training started with an initial learning rate of 1e-7 and exponentially increased in each epoch using the formulae,

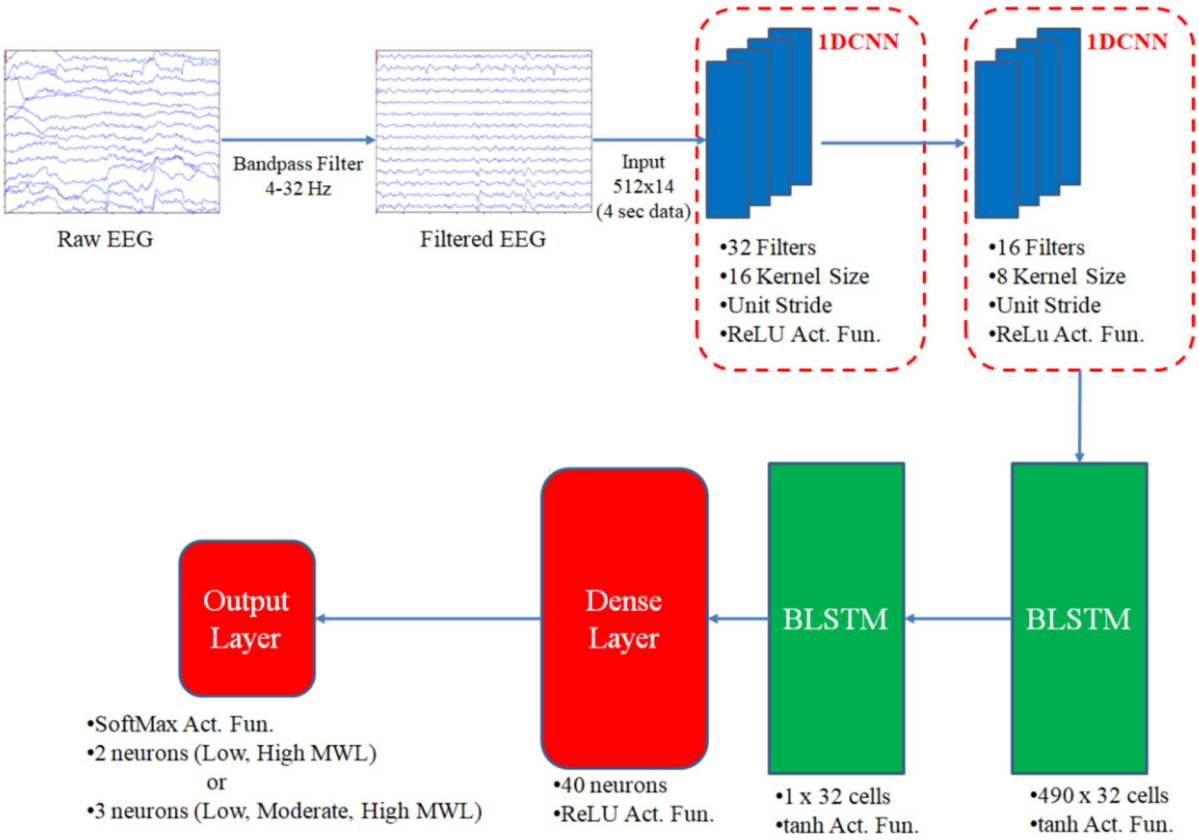

Fig. 1 The architecture of the proposed 1D CNN-BLSTM model

Table 4 Hyperparameters used for model training

Model	Number of classes	Learning rate	Momentum	Epochs	Batch size
CNN-BLSTM	2-Class	2e-3	0.9	90	245
5-fold CV	2-Class	2e-3	0.9	125	294
7-fold CV	2-Class	2e-3	0.9	115	270
CNN-BLSTM	3-Class	1e-3	0.9	200	245
5-fold CV	3-Class	1e-3	0.9	200	294
7-fold CV	3-Class	1e-3	0.9	200	270

$$Learning\ rate = 10^{-7+\frac{epoch}{40}}, \tag{1}$$

$$Learning\ rate = 10^{-7+\frac{epoch}{100}}, \tag{2}$$

where Eq. (1) was used for binary and Eq. (2) for ternary classification. After plotting the loss versus learning rate graph for each epoch, learning rate has been selected which gave the maximum decrease in loss i.e., the rate of change of loss was minimum. After fine-tuning, it was found that the hyperparameters described in Table 5 gave the fastest training and the best accuracy of each model.

Table 5 Specifications of the proposed CNN-BLSTM model

Layer	Layer configuration	Layer parameters	Number of parameters
1D CNN	Filters:32, Kernel size:16	Stride length = 1, activation = ReLU	7200
1D CNN	Filters:16, Kernel size:8	Stride length = 1, Activation = ReLU	4112
BLSTM	Memory units: 32	Activation = tanh	12,544
BLSTM	Memory units: 32	Activation = tanh	24,832
Dense layer	Neurons: 40	Activation = ReLU	2600
Dense layer	Neurons: 2/3	Activation = Softmax	82

5 Results

5.1 Model Evaluation Parameters

5.1.1 Accuracy

Accuracy is simply the fraction of correct classifications done by the model. It can be defined as,

$$Accuracy = \frac{TP + TN}{Total\ instances}, \tag{3}$$

where, *TP* means True Positive or the number of instances of positive class which are predicted correctly, and *TN* means True Negative or the number of instances of negative class which are predicted correctly.

5.2 Precision

Precision is the fraction of correct positive predictions. It can be defined as,

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

where, *FP* means False Positive or the number of instances of negative class which are predicted wrong, and *TP* is the True Positive as defined in Eq. (3).

5.2.1 Recall

Recall is the fraction of all positive instances that the model predicts correctly as positive. It can be defined as,

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

where, *FN* means False Negative or the number of instances of positive class which are predicted wrong, and *TP* is the True Positive as defined in Eq. (3).

5.2.2 F1 Score

F1 Score is defined as the harmonic mean of precision and recall and it helps to give an overall measure of the model. It is defined using Eqs. (4) and (5) as,

$$F1Score = 2 * \frac{precision * recall}{precision + recall}. \tag{6}$$

These above metrics are used to evaluate the performance of the proposed model in binary and ternary classification tasks. In results, the class weighted averages of the above metrics has been reported.

5.3 Binary Classification

In this subsection, analysis of the model trained for EEG classification into low and high Workload classes has been discussed. Model has been trained as defined in Sect. 3 using

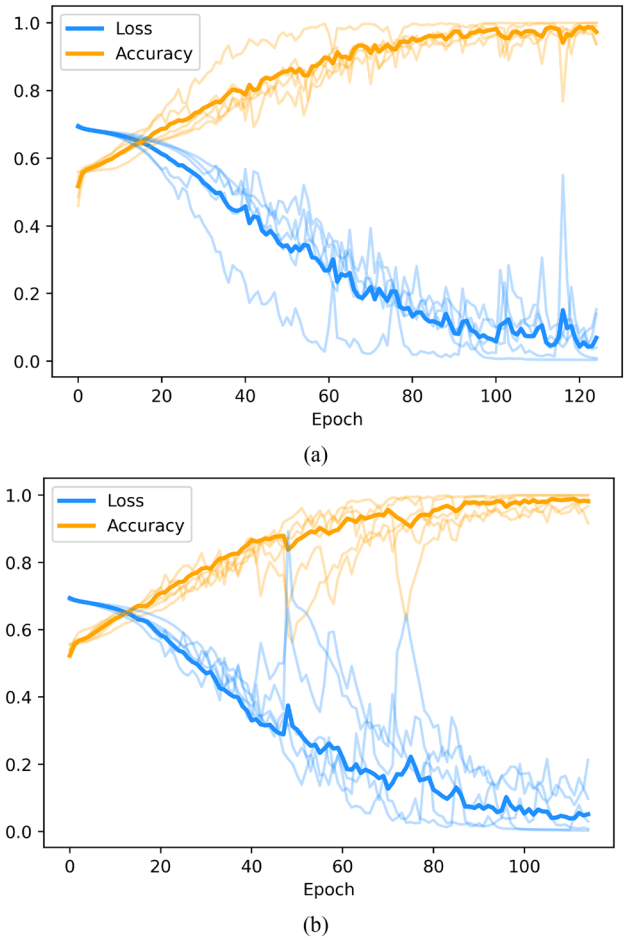

Fig. 2 Training loss and accuracy curves for **a** 5-fold CV, and **b** 7-fold CV, of the proposed model for binary classification

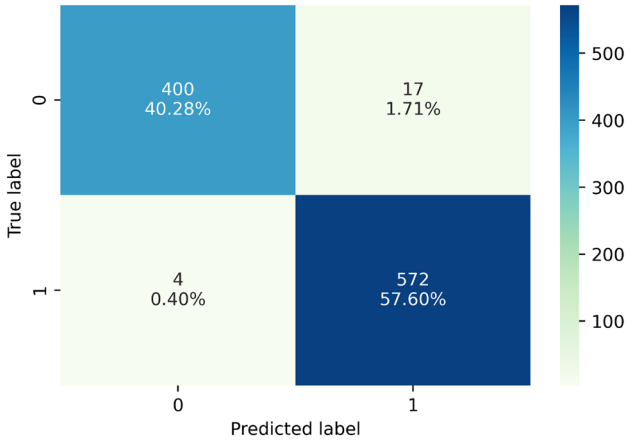

Fig. 3 Confusion matrix for binary classification

the hyperparameters mentioned in Table 5. Single model has been tested, and trained through holdout method, and fivefold and sevenfold CV as mentioned in Sect. 2. Figure 2a, b shows the training loss and accuracy for fivefold and sevenfold CV, respectively. In these figures, the bold line represents the mean of these values, and the lines in the background are the individual metrics for each fold. It has been noticed that there are sudden spikes in the loss curve during training which are quickly reduced, these arise due to a bad mini batch being randomly generated during optimization. It is also seen, that towards the end of the training, all curves stagnate and converge to around the same accuracy level which proves that the results are robust and reproducible.

Confusion matrix of the trained model on the holdout test dataset were also plotted, as shown in Fig. 3. The model performs well in separating the lowworkload EEG samples from the high workload samples. Out of 993 test samples, during the holdout method, only 21 are misclassified. The classification of Class 0 as Class 1 is slightly high which maybe related to more samples of Class 1 being available in the dataset leading to a slight class imbalance.

Table 6 describes the model's performance as measured by the model evaluation parameters discussed earlier. The CV results are stated in the format of "mean ± standard deviation" i.e., the mean and standard deviation of the metrics measured across the K-folds. The model gives an impressive accuracy of 97.89% on the test dataset when using the holdout method of training. Further, for the fivefold CV, three out of five folds had a testing accuracy greater than 97% but fold number 3 and 4 had small disturbances in their loss and accuracy curves towards the end of the training, decreasing their accuracy to 96.07% and 93.27%, respectively. This led to a slight decrease in mean accuracy to 96.54% and an increase in standard deviation for fivefold CV. Similarly for sevenfold CV, six out of seven folds had a testing accuracy greater than 96.50% but as evident from Fig. 2, the learning was unstable for some folds. Particularly, for fold number 3, there was a large spike in losswhich it could not recover from in the given epochs. It also had a drop in its accuracy in the last three epochs giving a testing accuracy of 90.89% which brought down the mean accuracy measure to 96.77%.

5.4 Ternary Classification

Same proposed model architecture as discussed in Sect. 3 with the hyperparameters for 3 class classification as mentioned in Table 5 has been used. Like binary classification, in ternary classification training was done using the holdout method, fivefold CV and sevenfold CV. Figure 4a, b shows the training loss and accuracy with the bold line representing their mean of all the folds and the lines in the background depicting each fold's measures. Like binary classification, ternary classification also shows spikes in the loss in the middle of the training for some folds caused by a bad mini

Table 6 Binary classification analysis of proposed model on simkap-based multitasking activity

Model	Accuracy (%) (train/test)	Precision (%)	Recall (%)	F1 score (%)
CNN-BLSTM	99.98/97.89	97.91	97.88	97.88
5-fold CV	97.30 ± 2.91/ 96.54 ± 1.80	96.69 ± 1.55	96.54 ± 1.80	96.55 ± 1.79
7-fold CV	98.16 ± 2.91/ 96.77 ± 2.43	96.69 ± 2.41	96.67 ± 2.43	96.67 ± 2.43

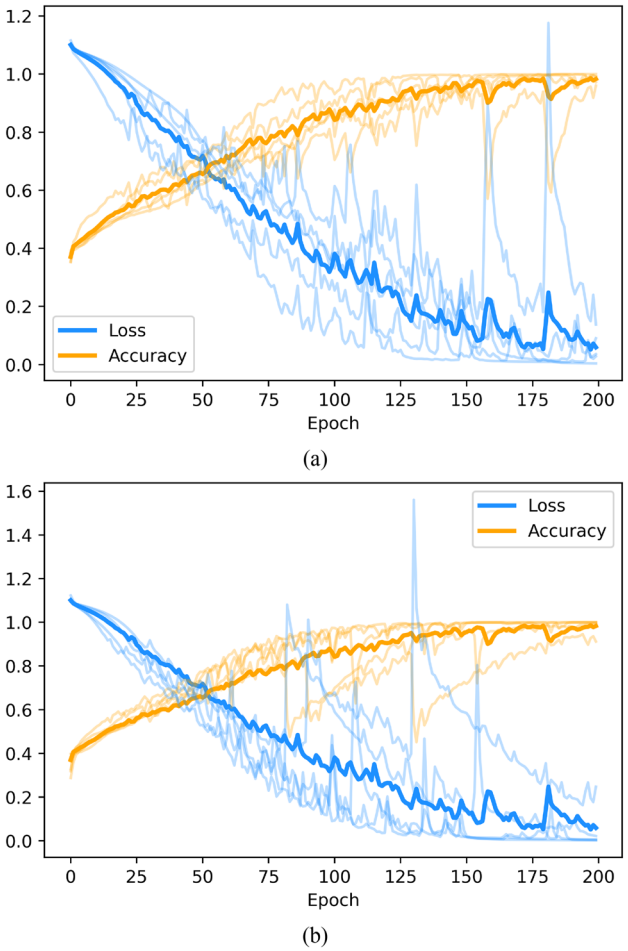

Fig. 4 Training loss and accuracy curves for the **a** 5-fold CV, and **b** 7-fold CV, of the proposed model for ternary classification

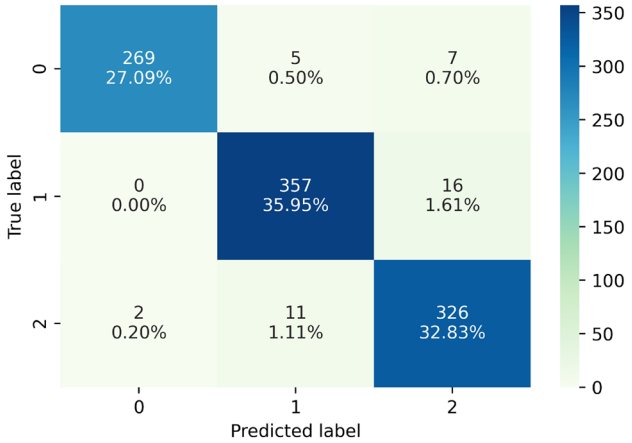

Fig. 5 Confusion matrix for ternary classification

batch being randomly generated. But at the end of training, the curves stabilize and reach the same accuracy level.

The 3-class confusion matrix of the model's performance on the holdout test dataset is shown in Fig. 5. There is only a marginal misclassification showing that the model was able to learn to distinguish low, moderate, and high workload.

The proposed model's performance and analysis using model evaluation parameters discussed earlier is shown in Table 7. Both the holdout method's results and the K-fold CV results are shown in this table. Like binary classification, the K-fold measures are reported in the format of "mean ± standard deviation". The model gives an impressive test accuracy of 95.87% with an F1 score of 95.88% for the holdout method of training. Substantial performance of the proposed model was also obtained, when judged using fivefold and sevenfold CV. For fivefold CV, despite spikes in loss during training for fold number 3, all the folds were able to learn adequately from the data giving a mean accuracy of 94.68%. Similarly for sevenfold CV, except fold number 1 and 6, the testing accuracy was more than 96% for the rest of the folds. But due to unstable learning and spikes in loss, fold number 1 and 6 had difficulty learning from the data and recovering from huge spikes in loss which resulted in their accuracy being 92.91% and 89.21%, respectively. This brought down the mean accuracy measures but was simply caused by randomness in the deep learning process and is accounted for when reporting the CV results. The CV results clearly show the robustness and generalizability of our proposed method.

6 Discussion

In this study, the impact of cognitive load during multitasking activities using EEG data which leads to mental workload was also estimated. Here, the STEW [14, 15] mental workload data for subjects during the "SIMKAP-based

Table 7 Ternary classification analysis of proposed model on simkap-based multitasking activity

Model	Accuracy (%) (train/test)	Precision (%)	Recall (%)	F1 Score (%)
CNN-BLSTM	99.72/95.87	95.93	95.87	95.88
5-fold CV	98.30 ± 1.48/ 94.68 ± 0.81	94.80 ± 0.81	94.68 ± 0.81	94.67 ± 0.80
7-fold CV	97.95 ± 3.19/ 95.36 ± 2.92	95.41 ± 2.85	95.36 ± 2.92	95.37 ± 2.90

Table 8 Comparative Analysis of Proposed Model

S.no	Study	Model	Classes/tasks	Test/stimulator/protocol	Accuracy (%)
1	Ang et al. [16]	1D CNN	2/attention, no-attention	Stroop color test	79.26
2	Zeng et al. [18]	CNN-R	2/DROWS, TAV	NFS-S2U & world record driving simulators	92.68
3	Li et al. [19]	DAEC	3/two ATC event, one failure event	4-h flight simulator	86.52
4	Yin et al. [24]	SDBN	3/low, moderate, high MWL	AutoCAMS	76.00
5	Jiao et al. [23]	PGBM	2/task relevant, task irrelevant	Character set working memory Experiment	92.37
6	Bashiva et al. [26]	1D CNN & LSTM	4/load level from 1 to 4	character set working memory experiment	91.11
7	Lim et al. [14]	Support Vector Regressor (SVR)	3/low, moderate, high MWL	SIMKAP	69.00
8	Das Chakladar et al. [29]	BLSTM & LSTM	2/task, no-task	SIMKAP	86.33
9	Das Chakladar et al. [29]	BLSTM & LSTM	3/low, moderate, high MWL	SIMKAP	82.57
10	Proposed	1DCNN & BLSTM	2/low, high MWL	SIMKAP	96.77
11	Proposed	1DCNN & BLSTM	3/low, moderate, high MWL	SIMKAP	95.36

multitasking activity" is used. A single classifier for all subjects, overcoming the subject to subject variability, which is a great challenge when using EEG data for classification has been successfully made.

The comparative analysis of the proposed model with the current state-of-the-art models has been shown in Table 8. The model architecture, the number of classes to estimate and the protocol used to induce and measure cognitive workload are mentioned in this table. The serial numbers from 1 to 6 represent recent research done on mental workload/cognitive load estimation using other testing protocols as discussed in Sect. 1. The models numbered from serial 7 to 9 represent recent research on the STEW [9, 10] dataset using handcrafted feature extraction and engineering. Serial number 10 and 11 show the proposed model's accuracy in sevenfold CV.

For the binary classification task, the current state-of-the-art model's performance is 92.68% for models made on testing protocols other than SIMKAP, and for SIMKAP testing protocol, the maximum accuracy is only 86.33%.The proposed model exceeds this performance significantly by attaining accuracy of 96.77%.

Similarly, for ternary classification, the state-of-the-art model's performance is 82.57% for SIMKAP testing protocol-based models while for other protocols this accuracy reached up to 86.52%. Again, the proposed model far exceeds these models and attains an accuracy of 95.36%. Since the dataset used is open access, the work is easily reproducible and can be extended in the future.

It is demonstrated that end-to-end deep learning can be successfully used for multi-channel EEG signals classification. Simple data preprocessing like bandpass filtering and data augmentation like windowing of data are sufficient for adapting the raw EEG data for deep learning. This study follows the recent trend of deep learning surpassing models which use handcrafted feature extraction and engineering.

The model utilizes only around 50,000 parameters which result in a fast performance and training time while other models based on deep learning have significantly more parameters. Due to the lightweight nature of the proposed model, it can easily be updated and maintained and utilized in real-time classification of mental workload. The work is focused on the "SIMKAP-based multitasking activity" part of the "STEW" dataset and not on the "no task" part which involves the subjects in a resting state. Classification of mental workload during multitasking is of more use for operator efficiency in tasks like air traffic management as compared to just being able to learn to distinguish between the "Task" and "no task" state of a subject.

The limitations of this study are that the proposed model has only been tested on the "STEW" dataset which has all the subjects of the same gender, education level and age. Further, this study is based on two and three class classification, while actual labels of mental workload levels are distributed from 4 to 9. The performance of more than three classes on the proposed model is not studied.

Implication of this work is moving one step ahead in mental workload measurement, which can helpful in taking preventive steps to reduce further problem like, anxiety, mental stress, tension, and worry, in daily life or profession-specific task.

7 Conclusion

In this paper, a new model using cascaded deep 1DCNN and BLSTM for binary and ternary classification of mental workload on the open access "STEW" dataset has been developed. The work is focused on the "SIMKAP-based multitasking activity" part of the dataset, which contained data for subjects doing multitasking activities. An end-to-end deep learning methodology that did not require any handcrafted feature extraction and engineering is used. Using only around 50,000 parameters, the proposed model achieves accuracies of 97.89% and 95.87% with the holdout method, 96.54% and 94.68% with fivefold cross-validation, and 96.77% and 95.36% with sevenfold cross-validation for binary and ternary classification, respectively, far exceeding the state-of-the-art. Finally, an end-to-end brain computer interface framework for mental workload estimation has been privied in this paper.

In the future, the proposed model is required to be evaluated on other reputed mental workload datasets, which have more diversity in subjects and recording sessions. It is also desired to explore the use of 2D convolutional neural network layer and observe wether it has a better spatial information extraction compared to 1DCNN or not.

Acknowledgements Many thanks for Maulana Azad National Institute of Technology, Bhopal, India, for providing SEED GRANT and support from Brain Computer Interface Lab at Computer Science and Engineering Department. Many thanks for STEW: Simultaneous Task EEG Workload Dataset Authors for provding access to their dataset.

Author Contributions Mitul Kumar Ahirwal conceptually designed the flow of study. Coding and implementation is done by Vipual Sharma. Both authors contributed in writing of manuscript. Mitul Kumar Ahirwal done the proof reading.

Funding This work is supported and funded by Maulana Azad National Institute of Technology, Bhopal, India, under SEED GRANT scheme.

Availability of Data and Materials Data used in this study is publically available with name "STEW: Simultaneous Task EEG Workload Dataset" on IEEE Dataport.

Declarations

Conflict of interest All authors confirm that there are no known conflicts of interest associated with this manuscript.

References

1. Othmani A, Brahem B, Haddou Y, Mustaqeem. Machine-learning-based approaches for post-traumatic stress disorder diagnosis using video and EEG sensors: a review. IEEE Sens J. 2023;23(20):24135–51. https://doi.org/10.1109/JSEN.2023.3312172.
2. Seo SH, Lee JT. Stress and EEG. Convergence and hybrid information technologies. London: IntechOpen; 2010.
3. Ryali VSSR, Bhat PS, Srivastava K. Stress in the Indian Armed Forces: how true and what to do? Med J Armed Forces India. 2011;67(3):209.
4. Wu EQ, Peng XY, Zhang CZ, Lin JX, Sheng RSF. Pilots' fatigue status recognition using deep contractive autoencoder network. IEEE Trans Instrum Meas. 2019;68(10):3907–19.
5. Seal A, Bajpai R, Agnihotri J, Yazidi A, Herrera-Viedma E, Krejcar O. DeprNet: a deep convolution neural network framework for detecting depression using EEG. IEEE Trans Instrum Meas. 2021;70:1–13.
6. Watts D, et al. Predicting treatment response using EEG in major depressive disorder: a machine-learning meta-analysis. Transl Psychiatr. 2022;12(1):332.
7. Deb S, Banu PR, Thomas S, Vardhan RV, Rao PT, Khawaja N. Depression among Indian university students and its association with perceived university academic environment, living arrangements and personal issues. Asian J Psychiatr. 2016;23:108–17. https://doi.org/10.1016/j.ajp.2016.07.010.
8. Sengupta A, et al. A multimodal system for assessing alertness levels due to cognitive loading. IEEE Trans Neural Syst Rehabil Eng. 2017;25(7):1037–46.
9. Sengupta A, Abhishek T, Aurobinda R. Analysis of cognitive fatigue using EEG parameters. In: 2017 39th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017.
10. Sengupta A et al. Analysis of loss of alertness due to cognitive fatigue using motif synchronization of EEG records. In: 2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016.
11. Singh U, Ahirwal MK. Improved classification of mental workload using one dimensional convolutional neural network and SMOTE technique. In: Proceedings of the 2023 7th international conference on graphics and signal processing. 2023.
12. Singh K, Ahirwal MK, Pandey M. Mental health monitoring using deep learning technique for early-stage depression detection. SN Comput Sci. 2023. https://doi.org/10.1007/s42979-023-02113-4.

13. Singh K, Ahirwal MK, Pandey M. Subject wise data augmentation based on balancing factor for quaternary emotion recognition through hybrid deep learning model. Biomed Signal Process Control. 2023;86:105075.
14. Lim WL, Sourina O, Wang LP. STEW: simultaneous task EEG workload data set. IEEE Trans Neural Syst Rehabil Eng. 2018;26(11):2106–14. https://doi.org/10.1109/TNSRE.2018.2872924.
15. Lim WL, Sourina O, Wang L. STEW: simultaneous task EEG workload dataset. IEEE Dataport. 2018. https://doi.org/10.21227/44r8-ya50.
16. Ang KK, Guan C. Inter-subject transfer learning with end-to-end deep convolutional neural network for EEG-based BCI. J Neural Eng. 2019;16: 026007.
17. Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: a compact convolutional network for EEG-based brain—computer interfaces. arXiv, 2016. arXiv:1611.08024
18. Zeng H, Yang C, Dai G, Qin F, Zhang J, Kong W. Classification of driver mental states by deep learning. Cogn Neurodyn. 2018;12:597–606.
19. Li F, et al. Deep models for engagement assessment with scarce label information. IEEE Trans Hum-Mach Syst. 2017;47:598–605.
20. Kose MR, Ahirwal MK, Atulkar M. Weighted ordinal connection based functional network classification for schizophrenia disease detection using EEG signal. Phys Eng Sci Med. 2023;46:1055–70. https://doi.org/10.1007/s13246-023-01273-0.
21. Singh K, Ahirwal MK, Pandey M. Quaternary classification of emotions based on electroencephalogram signals using hybrid deep learning model. J Ambient Intell Human Comput. 2023;14:2429–41. https://doi.org/10.1007/s12652-022-04495-4.
22. Kose MR, Ahirwal MK, Atulkar M. Dynamic characterization of functional brain connectivity network for mental workload condition using an effective network identifier. Int J Inf Tecnol. 2023;15:229–38. https://doi.org/10.1007/s41870-022-01151-0.
23. Jiao Z, Gao X, Wang Y, Li J, Xu H. Deep convolutional neural networks for mental load classification based on EEG data. Pattern Recogn. 2018;76:582–95.
24. Yin Z, Zhang J. Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights. Neurocomputing. 2017;260:349–66.
25. Liu YT, Lin YY, Wu SL, Chuang CH, Lin CT. Brain dynamics in predicting driving fatigue usinga recurrent self-evolving fuzzy neural network. IEEE Trans Neural Netw Learn Syst. 2016;27:347–60.
26. Bashivan P, Rish I, Yeasin M, Codella N. Learning representations from EEG with deep recurrent-convolutional neural networks. In: Proceedings of the international conference on learning representations, San Juan, Puerto Rico. 2016
27. Bashivan P, Bidelman GM, Yeasin M. Spectrotemporal dynamics of the EEG during working memoryencoding and maintenance predicts individual behavioral capacity. Eur J Neurosci. 2014;40:3774–84.
28. Thodoroff P, Pineau J, Lim A. Learning robust features using deep learning for automatic seizure detection. Mach Learn Healthc Conf. 2016;56:178–90.
29. Das Chakladar D, Dey S, Roy PP, Dogra DP. EEG-based mental workload estimationusing deep BLSTM-LSTM network and evolutionary algorithm. Biomed Signal Process Control. 2020;60:101989.
30. Smith LN. Cyclical learning rates for training neural networks, 2017. arXiv: 1506.01186
31. Hochreiter Sepp, Schmidhuber Jürgen. Long short-term memory. Neural Comput. 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
32. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process. 1997;45(11):2673–81. https://doi.org/10.1109/78.650093.
33. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. https://doi.org/10.1038/nature14539.
34. Fan C, Hu J, Huang S, Peng Y, Kwong S. EEG-TNet: an end-to-end brain computer interface framework for mental workload estimation. Front Neurosci. 2022;16: 869522.
35. Chakladar DD, Datta S, Roy PP, Prasad VA. Cognitive workload estimation using variational autoencoder and attention-based deep model. IEEE Trans Cogn Dev Syst. 2022;15(2):581–90.
36. Chakladar DD, Roy PP, Chang V. Integrated spatio-temporal deep clustering (ISTDC) for cognitive workload assessment. Biomed Signal Process Control. 2024;89: 105703.
37. Kingphai K, Moshfeghi Y. Mental workload assessment using deep learning models from EEG signals: a systematic review. IEEE Trans Cogn Dev Syst. 2024. https://doi.org/10.1109/TCDS.2024.3460750.
38. Das Chakladar D, Roy PP. Cognitive workload estimation using physiological measures: a review. Cogn Neurodyn. 2024;18(4):1445–65.