Towards Real-Time Cognitive State Monitoring Using an Explainable Data-Driven Deep Learning Framework on Physiological Signals

Mohammed Ibrahim Altwijri
Department of Computer Science-Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Cognitive states represent dynamic mental conditions that impact how individuals identify their environment, process information, and make decisions. These states are influenced by factors like stress levels, cognitive workload, and external stimuli, which can drastically affect task performance and safety. Therefore, precise examination of cognitive workload is critical in applications where aligning task requirements with human processing ability is vital. Over the past several years, deep learning models have excelled in cognitive state monitoring (e.g., fatigue, workload, attention, emotion) by automatically capturing complex features from physiological signals such as electroencephalography and eye-tracking, outperforming standard techniques for better human-computer interaction (HCI), personalized learning, and mental health insights, though challenges remain in data quality, labeling, and real-world generalizability. Moreover, explainable artificial intelligence techniques are incorporated widely to ensure robust and transparent model interpretability. In this study, we propose a Real-Time Cognitive State Monitoring Framework using Explainable Deep Learning (RTCSM-XDL) based on physiological signals. The proposed RTCSM-XDL model aims to achieve accurate and transparent cognitive state classification by integrating advanced data-driven learning and explainable artificial intelligence techniques. The data preprocessing stage includes data normalization, missing data handling, and categorical encoding to ensure high-quality and consistent inputs. To recognize the most useful physiological features, a hybrid feature selection strategy combining minimum redundancy maximum relevance, recursive feature elimination, and Boruta is employed. For cognitive state classification, a temporal residual gated unit with attention is designed to efficiently learn temporal dependencies and enable real-time inference. To address the interpretability complexity associated with deep learning methods, the proposed model utilizes LIME to provide human-understandable insights into model predictions. The experimental result analysis of the RTCSM-XDL technique takes place on a benchmark dataset, and the comparative results demonstrate improved performance over the existing approaches in terms of diverse metrics.**

Keywords*—Cognitive state monitoring; physiological signals; feature selection; deep learning; explainable artificial intelligence; real-time systems

I. Introduction

In today's life, Cognitive workload and mental stress are the primary problems. Cognitive load is a physical configuration, which is an evaluation of the workload allocated for an individual to execute a specific task or go through a particular defined condition [1]. This workload measurement develops an importance for work regions, also maintaining the tasks like surgeons/physicians, aeronauts and factory labours where complete care and attentiveness are need for during the tasks. Cognitive load measurement is a task-based operation that cannot be associated with behavioural changes. Stress may be described as a condition of the brain where cognition becomes repressed, and execution declines over time. A fast change in the work situation may require a human's body to modify mentally and physically to complete as before. Our brain has to execute several incoming signals simultaneously to give a motor and sensory combination that can finally increase the cognitive load [2]. A quantity of cognitive load learned is subjective and based on the fast alteration in the situation and the individual's capability to manage the change. Cognitive Workload measurement has attained substantial significance over the past ten years. It has developed into general research, which requires the research environment's intellectual excellence to comprehend the neuroscience behind it. The load measurement might be executed utilizing physiological analysis, task execution indices, and subjective examination.

Recently, physiological measures have held a crucial role in the perception of cognitive load measurement. Physiological cognitive load measurement was mainly accomplished by utilizing the Electroencephalography (EEG) activities [3]. EEG has developed as an essential technology in neurobiology, providing real-time insights into brain activity with non-intrusive recording of neural electrical signals. Particularly, EEG is among the most extensively employed signals, because of its higher temporal convenience, cheapness, resolution, and security. Recent improvements in EEG technology have allowed more accurate and relevant data gathering, enabling researchers to discover minor changes in brain activity over various cognitive conditions. With the advancement of deep learning (DL), machine learning (ML) architectures have emerged. These methods are a subdomain of artificial intelligence (AI) techniques. The application of DL methods that depend on EEG signals is rising, thanks to the increasing accessibility of huge EEG datasets [4]. DL allows end-to-end learning from actual data, thus overcoming the limitation of processing higher-dimensional volumes of data faced by standard ML methods. In the domain of EEG data processing, DL was utilized to develop and extend existing techniques, reducing the requirement for domain-related measures and feature extraction pipelines.

Related to CNN and RNN that are the most wide-spread employed frameworks for classification of sequential data, transformers have presented a high capability to deal with a wide range of dependencies and identify patterns in data sequences, also using more understandable decision-making procedures [5].

To accurately and efficiently monitor cognitive states in real time using physiological signals, this study presents a Real-Time Cognitive State Monitoring Framework using Explainable Deep Learning (RTCSM-XDL). The proposed RTCSM-XDL model aims to improve classification accuracy and transparency by incorporating a data preprocessing pipeline, a hybrid feature selection (FS) approach, an attention-based deep learning classifier, and explainable artificial intelligence techniques. The performance assessment of the RTCSM-XDL model is carried out on benchmark physiological datasets, and the results are examined using different performance measures. The major contributions of this manuscript are given below:

- Present a hybrid FS strategy that integrates minimum redundancy maximum relevance (mRMR), recursive feature elimination (RFE), and Boruta to identify the most informative physiological features, reduce dimensionality, and enhance classification performance.
- Develop a temporal residual gated unit network with attention (TRGU-Att), which is a combination of a temporal convolutional network (TCN), a residual gated recurrent unit (ResGRU), and multi-head attention (MHA) to efficiently capture temporal dependencies in physiological signals, allowing real-time cognitive state classification.
- Employ LIME to provide human-understandable explanations of model predictions, thus improving interpretability and trust in the proposed model.
- Perform extensive experimental evaluations and comparative analyses on the benchmark dataset, demonstrating that the proposed RTCSM-XDL model performed well compared to existing techniques across different performance metrics.

II. Review of Related Studies

This section briefly reviews prior related works on cognitive state monitoring using physiological signals. Scheutz et al. [6] introduced a machine learning (ML) architecture for analysing these questions and concentrated particularly on utilizing neurophysiological and physiological conditions. Specifically, they defined a multi-task communication setting employed to attain a complete multimodal procedure that provides the basis for numerous convolutional state-of-the-art ML systems regarding their efficiency in inferring systemic cognitive states. Li et al. [7] presented a multi-model physiological signal feature presentation fusion method that depends on multi-level attention (PSFMMA). This method is intended for combining the multi-model physiological signals to detect students' cognitive conditions with better accuracy and consistency. Consequently, it produces signal presentation vectors through more semantic feature extraction and increases significant attributes with design-emphasis systems. Panwar et aal [8] presented a DL architecture for cognitive state analysis employing Electroencephalogram (EEG) neural connectivity. This method 3 EEG data via computer-based tasks: the Visual Search Task, Continuous execution Task, and Dual N-Back Task, respectively, intended to stimulate varying stages of cognitive attention, fatigue, and workload, individually. By overcoming the actual restrictions like task-specific models and small sample sizes, this study improves the applicability and generalizability of EEG-driven cognitive analysis.

Li et al. [9] proposed an analysis of the impact of feelings on cognitive control that depends on the arousal-valence method. This method is critical for increasing VR communication, responsive design, and self-correcting VR/AR applications. In [10], a psychological emotion state decoding model integrating singular spectrum examination and entropy measure is developed that depends on EEG signals. The above-mentioned technique is advanced to enable the cognitive computer system to observe emotions. Moreover, the major systems are incapable of actively understanding emotional intention or the human psychological cognitive state. Vogl et al. [11] intended to enhance and estimate support vector machine (SVM) learning methods for forecasting cognitive workload (CWL) that depend on physical information. The purposes for making strong binary classifiers, increasing these to multi-class techniques.

Liu and Zhu [12] presented functional connectivity patterns among EEG as well as fMRI methods. They used a complete method that combines dynamic and static analysis of fMRI-EEG data, capable of uncovering distinctive relationship states and describing their sequential variations. The outcome revealed a modular structure along with the intrinsic connectivity network of the mind, underlining the important parts of sensory methods and evasion mode network. Chen et al. [13] presented a brain neuronal communication with electroencephalography signals to establish relations in brain areas over several subjects for estimating real-life cognitive conditions. Specifically, a virtual reality-dependent flight environment is structured with multiple fixed viewpoints. These outcomes are inferred through various viewpoints, containing cross- and within-subject analyses for gender- and task-wise highlighting of brain network interactions. Table I highlights the comparative analysis of cognitive state monitoring, including their goals, models, outcomes, and research gaps.

TABLE I. A LITERATURE REVIEW OF COGNITIVE STATE MONITORING

References	Years	Goals	Used Models	Outcomes	Research gap
Scheutz et al. [6]	2024	To classify human cognitive states using multimodal physiological signals in mixed-initiative tasks.	ML model	Achieves better recognition.	Physiological signals alone provide modest accuracy, indicating the need for contextual information to improve cognitive state classification.
Li et al. [7]	2025	To identify learners' cognitive states using multimodal physiological signals and multi-level attention.	DL model	accuracy of 96.32 ± 0.32% and 99.15 ± 0.12%.	Based on unimodal physiological data, which are sensitive to noise and fail to fully capture learners' cognitive states.
Panwar et al. [8]	2024	To assess workload, attention, and fatigue using EEG-based DL models.	CNN, GCN, and LSTM models.	Accuracy of 96.53% for workload, 98.40% for attention, and 97.86% for fatigue.	The methods are limited by task specificity and small datasets, restricting broader applicability.
Li et al. [9]	2025	To classify cognitive control levels in VR using physiological signals and emotional arousal.	SVM, KNN, and DL methods.	Accuracy of 84.52%.	To neglect emotional arousal, limiting understanding of its effect on cognitive control in VR/AR contexts.
Huangfu and Cheng, [10]	2025	To enable AI to perceive and recognize human emotions using EEG signals.	SVM, LSTM, and RNN techniques.	Accuracy of 64.15 %.	Most AI systems cannot actively perceive or interpret human cognitive and emotional states, limiting generalization.
Vogl et al. [11]	2025	To predict cognitive workload in real time using individualized SVM models.	CWL and SVM models.	Accuracy of 70.5% and 80.4%.	The approach cannot dynamically or individually predict workload, limiting adaptive interventions.
Liu and Zhu [12]	2024	To investigate dynamic EEG–fMRI connectivity to reveal cognitive state patterns.	DL model	Higher detection of cognitive states.	Studies provide limited insight into the transient dynamics of brain connectivity over short timescales.
Chen et al. [13]	2024	To assess real-time cognitive states using EEG-based brain connectivity.	ML and DL techniques.	Accuracy of 95.83%.	Existing cognitive state evaluation methods lack timeliness, generalization, and interpretability across subjects.

III. SYSTEM ARCHITECTURE

This study follows a multi-faceted approach for real-time cognitive state monitoring on physiological signals. The benchmark physiological signals dataset serves as the foundation for model development and evaluation. Fig. 1 showcases the end-to-end architecture of the RTCSM-XDL approach. The MEFAR dataset will first be pre-processed to guarantee its quality and usability. This involves normalizing data, handling missing values, and encoding categorical values. Additionally, FS will be performed to reduce the dimensionality of the dataset and enhance the model efficiency. mRMR, RFE, and Boruta are employed to pick the pertinent attributes for classifying cognitive states. Besides, the TRGU-Att model is leveraged for classification, which captures temporal dependencies and enables real-time inference. To ensure model interpretability, the proposed model utilizes LIME-based explainability for providing human-understandable insights into model predictions. Finally, the proposed system is evaluated using a benchmark dataset under diverse evaluation measures.

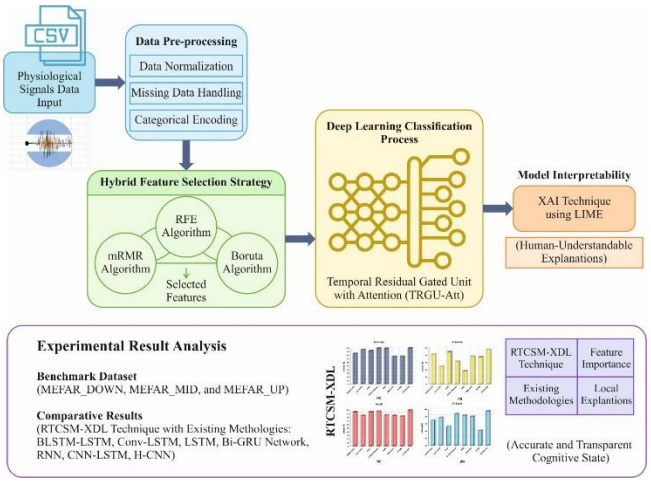


Fig. 1. End-to-end architecture of the RTCSM-XDL approach.

A. Data Preprocessing

The data preprocessing step encompasses data normalization, missing data handling, and categorical encoding to ensure high-quality and consistent inputs.

1) Missing value handling: To avoid bias and protect model strength because of insufficient data, it can be systematically recognized and handled missing values. Assume that the dataset is $D = \{(x)^i, (y)^i\}_{i=1}^{N}$. Misplaced accesses in feature matrices $X \in R^{N \times d}$ are processed as given below:

Numerical Attributes: Missing values have been imputed using the median of the individual feature: $b_j^{(i)} \leftarrow median(\{b_j^{(k)} : b_j^{(k)} \neq NaN\})$ for all $b_j^{(i)} = NaN, j \in N$. Now, N represents sets of mathematical feature indices.

Categorical Attributes: Missing values are completed by applying the approach (more extensive value) for every column:

$b_j^{(i)} \leftarrow mode(\{b_j^{(k)} : b_j^{(k)} \neq NaN\})$ for all $b_j^{(i)} = NaN, j \in C$.

Here, C represents the types of definite feature index. Later imputation, the database has been verified to confirm the mathematical outcomes $\sum_{i=1}^{N} \sum_{j=1}^{d} 1\{b_j^{(i)} = NaN\} = 0$.

2) Feature scaling: To improve convergence and numerical stability in training the model, each continuous feature can be normalized by applying Z-score normalization; its calculation is given by:

$$z_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{\sigma_j} \quad (1)$$

where, μ_j and σ_j describes the mean and standard deviation of the feature χ_j.

3) Categorical encoding: Categorical values are modified into numerical formats by utilizing label encoding. All categories at a column $\chi_j \in C$ are mapped to an exclusive integer: Encoder $(v) = k, v \in Domain(x_j)$. The mapping maintains the class individuality that is required for tree-based systems without presenting the sparseness of one-hot encoding. To ensure a reliable encoding process, every categorical value has been clearly mapped into a string class previously; it uses the LabelEncoder

B. Hybrid Feature Selection Strategy

For the feature reduction process, the proposed model deploys a hybrid FS approach integrating mRMR, RFE, and Boruta to identify the most informative physiological features.

1) mRMR: Features with a greater relationship to the target variables have been selected through mRMR [14], and their redundancy can also be reduced. The major intention of this study is to optimize relevancy of gene with output whereas decreasing its redundancy. The genes are selected and minimized from the score metrics by employing the mutual information difference (MID) as a main process. By applying similar feature counts, it can be predicted that the mRMR feature sets have higher generalization features and representation of the target phenotypes. The similar boundaries are proficiently enclosed by a small gathering of mRMR features rather than by a greater conventional set of features. Eq. (2) indicates the MID function.

$$mRMR(A_i) = Relevance(A_i, B) - \frac{1}{|S|} \sum_{A_j \in S} Redundancy(A_i, Y_j) \quad (2)$$

Here A_i denotes the genes from the datasets and B denotes the target variable.

Now, Redundancy $(A, B) = I(A_i, B)$ and Relevance $(A_i, B) =$ mutual information, $I(A_i, B)$

The primary feature is chosen by applying the method given in Eq. (3),

$$MaxV_I, V_I = \frac{1}{|S|} \sum_{i \varepsilon S} I(A_i, B) \quad (3)$$

Other features have been chosen in a progressive manner, and previous selected features remain in the feature sets. Algorithm 1 summarizes the detailed step-by-step procedure comprised in the mRMR method.

Algorithm-1: mRMR technique

Step1: Compute the MI among specific features and outcomes

Step2: Categorize the outcomes with respect to the particular variables

Step3: Eliminate the features with 0, the MI from the subsets of features that can be preferred with the help of dual-way search, and maintain the attributes with the maximum correlations

Step4: Select the weakly interrelated features for classifier validation while waiting for the outcomes

will be satisfied

Step5: Calculate mutual information entropy (MIQ) and MID for chosen feature subclasses

Step6: Sort the MID and MIQ and validate the attributes with the highest values

Step7: Repeat steps 3 to 6 until the outcomes are satisfied

Step8: Add the weight component to MID or MIQ

Step9: Determine the integration of the features and also match the optimum classification indicators over numerous sequences of computation and validation of the classifiers

Step10: Choose the optimum feature sets acquired from the two search techniques.

2) RFE: RFE is a technique that recognizes the major features from the datasets for prediction of the target variables. This technique is utilized a FS wrapper method, utilizing a different ML method at its core, contained by Recursive Feature Elimination (RFE) to choose the appropriate features. RFE is examined the training dataset for identifying the subsets of features, beginning with accessible features and systematically

eliminating them while the target quantity of features has been attained or while waiting for correctness enhancements. In conclusion, intended to improve prediction accuracy and decrease classifier execution time. The binary FS method includes the assessment of the GDM dataset, described as an instance space $= \{a_1, a_2, \dots, a_n\}$, a feature set $H = \{f_1, f_2, \dots, f_r\}$ and a target label sets $y = \{b_0, b_1\}$. This dataset is standardized in Eq. (4) as given by:

$$F(A, B) = \{(A_i, B_i) | X_i \in R^n, B_i \in \{b_0, b_1, \}\}_{i=1}^{k} \quad (4)$$

where, $b_i \in \{b_0 = 0, b_1 = 1\}^t$ indicates the binary class labels, and every individual sample $A_i = \{a_1, a_2, \dots, a_n\} \in R^n$ matches a data point in the dataset. When a data point x_i can be related to the label b and $b_{ij} = 1$, or else $b_{ij} = 0$. Thus, the matrix $A = \{a_1, a_2, \dots, a_n\}^T \in R^n$ encloses entire samples, whereas $B = \{, b_1\}^T \in \{0,1\}^{n*1}$ signifies the matrix of matching output labels. The particular steps in the RFE method have been exhibited in Algorithm 2.

Algorithm-2: RFE approach

Step1: Initialization of the classifier with features pre-defined by the mRMR method as the baseline data subsets

Step2: Measure feature significance and apply average data expansion

Step3: Assess the classification correctness by using the cross-validation method

Step4: Remove the low-graded features from the existing subsets to produce reviewed feature subsets

Step5: Find the implication of every feature in the restructured feature subsets

Step6: Reevaluate the classification accuracy of the upgraded feature subsets

Step7: Repeat steps 4 to 6 once the features are not removed

Step8: Select the feature subsets that generate the maximal classification precision under the sets of K different features.

3) Boruta: It is a wrapper-based FS technique that can be developed for identifying each related variable by repeatedly assessing their significance via a random forest (RF) method. Different conventional FS techniques that can be removed the significant features but faintly correlated parameter, Boruta maintains features that provides effectively for classification. The major score I_j for every feature j can be measured as given by:

$$I_j = \frac{1}{T} \sum_{t=1}^{T} G_j^{(t)} \quad (5)$$

Now, $G_j^{(t)}$ indicates the Gini significance of the feature j under the t^{th} tree of the random forest algorithm. T represents the overall counts of trees from the forest.

In comparison with Boruta, all the main feature scores are compared with shadow features (arbitrarily shuffled replications of original features). A feature can be represented when it is related to it:

$$I_j > \max(I_{shadow}) \quad (6)$$

Here, $\max(I_{shadow})$ denotes the maximum significance rate between shadow features.

During this study, Boruta has been used over 300 iterations, permitting features with minimum 85% position to be taken. The technique is assigned a major threshold and is also iteratively associated with actual features and randomly disorganized shadow features. The variable quantity reliably exceeded the shadow features and then measured the importance for classification. This confirmed that no possibly valuable features have been impulsively removed, creating a perfect selection for investigative FS in higher-dimensional information.

C. Deep Learning-Based Cognitive State Classification

The TRGU-Att model is developed to effectively capture temporal dependencies and enable real-time inference.

1) TCN: TCN is a CNN framework involved in temporal data processing. With the help of integrating causal, dilated, and residual convolutional networks (ResNet), the TCN [15] keeps the benefits of effective corresponding computation of CNNs. Causal convolution becomes a crucial element of TCN, and its main concept is to ensure that the output relies on the time t and the input immediately follows the causation of temporal data.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad (7)$$

$$M = \begin{pmatrix} s_{ij} & s_{ij+1} & s_{ij+2} \\ s_{i+1j} & s_{i+1j+1} & s_{i+1j+2} \\ s_{i+2j} & s_{i+2j+1} & s_{i+2j+2} \end{pmatrix} \quad (8)$$

$$T_{ij} = \sum_{m=1}^{3} \sum_{n=1}^{3} a_{mnn} s_{i+m-1j+n-1} \quad (9)$$

Here, the dimension of m is reliable and independent of feature size, convolution kernel dimension of mxn For example, the matrix of $m = 3$ indicates the width of convolutional kernel, and $n = 3$ refers to the length of the convolutional kernel, and the convolution matrix h has extensive to confirm the convolutional processes comprises entire features. n can be varied, and generally an odd number higher than 3.

During normal dilated convolution processes, the dimension of convolution kernel cannot be changed, as well as the variety of data recognized by each convolution. At dilated convolution, the convolution kernel can be extended by the all the area, thus a greater number of data are removed with no improvement in the quantity of computation and parameters, and the lost data in the middle blank area may be accompanied by the succeeding convolution, Therefore, it is not individual missing the data as well as also redundant data in normal convolution processes must be excellently filtered.

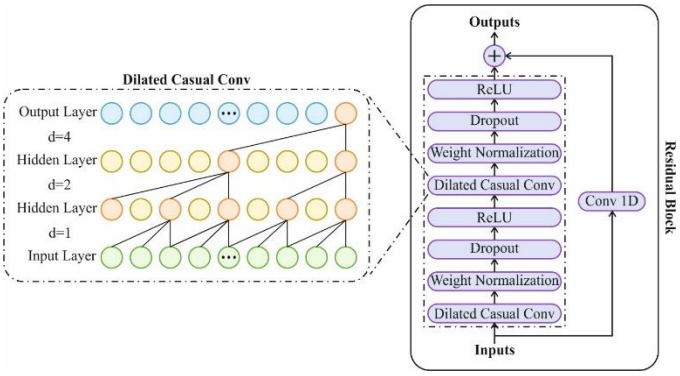


Fig. 2. Model diagram of TCN.

$$y_{ij} = \sum_{m=1}^{3}\sum_{n=1}^{3} a_{mn}\, x_{i+m-1j+d\times(n-1)} \tag{10}$$

Where the expansion co-efficient d signifies expansion dimension of convolutional kernels, and modifies size of convolution kernel by using the dramatic increase of the expansion co-efficient, then the model may be proficiently obtained the multi-scaled with local and global features. The expansion co-efficient $d = 1$, The dimension of convolution kernels has 3 columns, and there is no variance between the standard and dilated convolutions. Moreover, if the expansion co-efficient $d = 4$, the dimension of the convolution kernel will reach 9 columns; only the 1st, 5th, and 9th columns contribute to the convolution process, and the receptive fields are more extended. Fig. 2 represents the model diagram of TCN.

2) ResGRU: At this study, it can be improved the GRU and present a ResNet to develop a ResGRU architecture. The ResNet changes the standard network using learnable features of every time step to learn the variance features among every time series. Assume the standard network inputs are denoted as x and also the output represented as $f(x)$, next to present the ResNet, the output will be calculated as $H(x) = f(x) + x$. Consider the overall duration of the time step is indicated as n, the reset gate weight matrix indicates W_z, the preceding instant can be categorized by h_{t-1}, input during the existing time x_t, bias refers b_z, update gate weighted matrix denoted as W_r, bias indicates b_r, candidate phase weighted matrix is W_h, bias denotes b_h, and computation stages are represented by:

Measure reset gate:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{11}$$

The reset gate handles amounts of the past data impacts the existing candidate stage; if σ shows $sigmod$, the activation function can be represented by:

$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{12}$$

Calculation of update gate:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{13}$$

Find the fusion proportion of novel and past states data.

Calculation of candidate state:

$$h_t = tanh(W_h \cdot [r_t * h_{t-1}, x_t] + b_h) \tag{14}$$

$*$ means the Hadamard product, multiplication of the matching components.

Calculation of the final state:

$$h_t = (1 - z_t) * h_{t-1} + z_t * \hat{h}_t + h_{t-1} \tag{15}$$

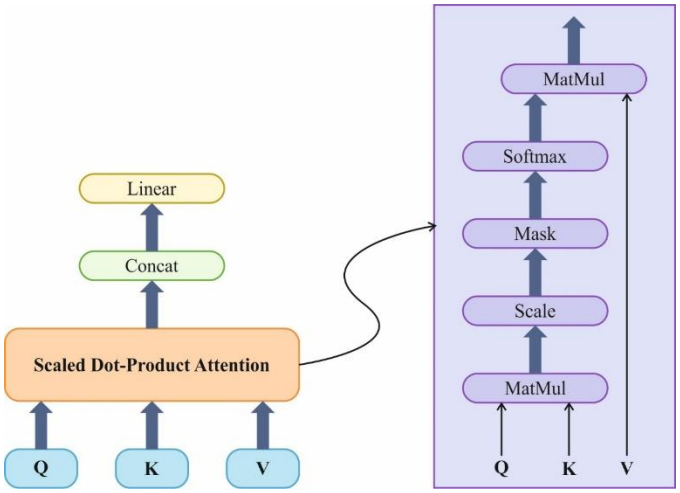


Fig. 3. Structure of MHA.

3) MHA: With the help of parallelizing numerous sets of attention heads, this method allows the system to simultaneously highlights various representations subset of the input sequences that majorly increases the model capability. Fig. 3 denotes the structure of MHA. During the tracking process, MHA may instantaneously capture spatio-temporal multi-dimensional features, and its effectiveness will be greater when compared to conventional frameworks.

Set matrix $x \in R^{n\times d_{model}}$; sequence length denoted as n, and model size referred to as d_{model}. The mathematical stages have been represented by:

Linear projection: every attention head h for learning distinct weight matrices, generally $d_k = d_v = d_{model}/H, W_h^Q \in R^{d_{model}\times d_k}, W_h^K \in R^{d_{model}\times d_k}, W_h^V \in R^{d_{model}\times d_v}$. The mathematical expression of Q, K, V can be given by:

$$Q_h = XW_h^Q \tag{16}$$

$$K_h = XiV_h^K \tag{17}$$

$$V_h = XiV_h^V \tag{18}$$

Parallel computing: calculate individual head h, attention weight and output, wherever the scaled parameter $\sqrt{d_k}$. Avoid the dot product being too large because the gradient will vanish; softmax normalization has been executed to produce a weight matrix. The mathematical equation must be denoted as:

$$Attention(Q_h, K_h, V_h) = softmax\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right)V \tag{19}$$

Result fusion: outputs of all heads can be merged and linearly converted, and then the computational expression is given by:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \dots, head) \quad (20)$$

The expression $head_n = Attention(Q_h, K_h, V_h)$, $W^0 \in R^{Hd_0 \times d_{model}}$ is the output weight matrix. Later entire equations have been combined and provided as given below:

$$MultiHead(X) = Concat(softmax\left(\frac{XW_h^Q (XW_h^K)^T}{\sqrt{d_k}}\right) XW_h^V)_{h=1}^{H} W^0 \quad (21)$$

D. Explainable AI Analysis using LIME

To overcome the interpretability challenges associated with DL models, the proposed model utilizes LIME [16] to offer human-understandable insights into model predictions. A black-box model classifier f is trained to convert the training data into the probabilities of individual labels, with the mathematical factors used to increase the possibility of the accurate label. LIME is a technique that offers local explanations g for a particular sample a depends on the trained black-box model f (for example, a neural network (NN) classifier). It can be denoted as a reduction complexity as given below:

$$\xi(a) = \text{argmin}_{g \in G}\, \mathcal{L}(f, g, \mathcal{W}_a) + \Omega(g), \quad (22)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{s,s' \in S} \pi_a(s)(f(s) - g(s'))^2, \quad (23)$$

Here, $\mathcal{L}$ determines how adjacent the explanation, g represents the predictable black-box method f. Ω signifies the complexity of g. The term G denotes the set of explainable models. s and s' describes instances produced from the distribution of all the datasets, and S indicates the types of these instances, and $\cdot\, x$: Samples are described. Fig. 4 demonstrates the general diagram of XAI.

f: Black-box model

g: Locally interpretable model

s, s': Produced instances

π_a: Kernel function

The kernel function parameter (π_a) can be represented by:

$$\pi_a(s) = \exp\left(-\frac{D(a,s)^2}{\sigma^2}\right), \quad (24)$$

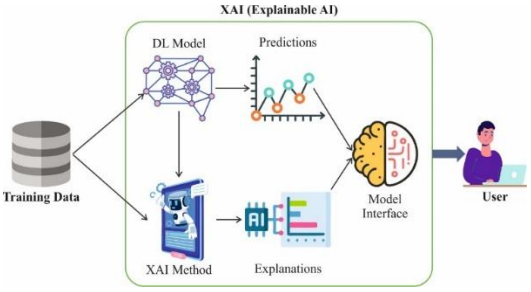


Fig. 4. General diagram of XAI.

IV. Findings and Interpretation

This section presents the dataset used, the evaluation metric, and the performance analysis of the proposed RTCSM-XDL method using benchmark Physiological Signals Based Mental Fatigue Analysis and Recognition dataset.

A. Dataset Used

In this subsection, the proposed model uses the MEFAR Dataset [17]. The processed datasets are called MEFAR_DOWN, MEFAR_MID, and MEFAR_UP, respectively. The database is formed exactly for analyzing mental fatigue depending on physiological signals. The database involves both raw and processed data. Raw data relates to the actual recorded signals, while processed data normally includes signal preprocessing models like artifact removal, feature extraction, and filtering.

TABLE II. Details of Dataset

Classes	Labels	Number of Instances		
		MEFAR_DOWN (1 HZ)	MEFAR_MID (32 Hz)	MEFAR_UP (64 Hz)
Fatigue Condition_Postive	Class 1	14464	25000	25000
Fatigue Condition_Negative	Class 0	13106	25000	25000
Total Instances		**27570**	**50000**	**50000**

The MEFAR_DOWN database contains 17 features, all of which are chosen. The MEFAR_MID database covers 17 features, but only 15 features are selected. The MEFAR_UP database includes a total of 17 features; only 16 features are selected. The database encompasses 27570 samples in the MEFAR_DOWN database, 50000 samples in the MEFAR_MID database, and 50000 samples in the MEFAR_UP database, as portrayed in Table II.

B. Experimental Evaluation Metrics

To calculate the performance of the developed model. The metrics are used to determine the values and are also described in detail as follows:

1) Accuracy: This metric evaluates how accurate the classifier is. It is mainly the ratio of accurate classifications to the total number of samples.

$$Accuracy = \frac{NumberofCorrectClassification}{Tota1NumberofC1assification} \quad (25)$$

2) Precision: A system's precision must be computed as the proportion of correctly categorized positive conditions to all predicted positive values.

$$Precision = \frac{TruePositives}{TruePositives + Fa1sePositives} \quad (26)$$

3) Recall: The ratio of properly categorized positive outcomes to the overall definite positives is recall.

$$Recall = \frac{TruePositives}{TruePositives + Fa1seNegatives} \quad (27)$$

4) Fl -SCORE: This is a measurement for comprehending the robustness and precision of a model.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (28)$$

5) Confusion matrix: The confusion matrix is an assessment matrix that is used to compute the false positive (FP), false negative (FN), true positive (TP), and true negative (TN), and supports analyzing the efficiency of a system.

6) ROC-AUC: A graphical representation that shows how indicates a model may distinguish between classes, named as ROC curve. It can plot the True Positive Rate (TPR) correlated with the False Positive Rate (FPR) through diverse threshold values, whereas a superior PR and lower FPR indicate higher performance of the model. A model indicator's overall efficiency is the AUC; higher AUC values are represented superior performance.

C. Detailed Result Analysis

This subsection provides a detailed analysis of experimental outcomes, highlighting classification performance, comparative analysis, and EigenGrad-CAM results. Fig. 5 shows the correlation matrix of RTCSM-XDL method on MEFAR_DOWN dataset that suggests most features have very lower correlation with together. This matrix displays that the features are largely independent and don't overlap much in information. Only certain feature pairs reveal mild correlations because of related physiological signals. Overall, the database comprises different and informative features appropriate for analysis.

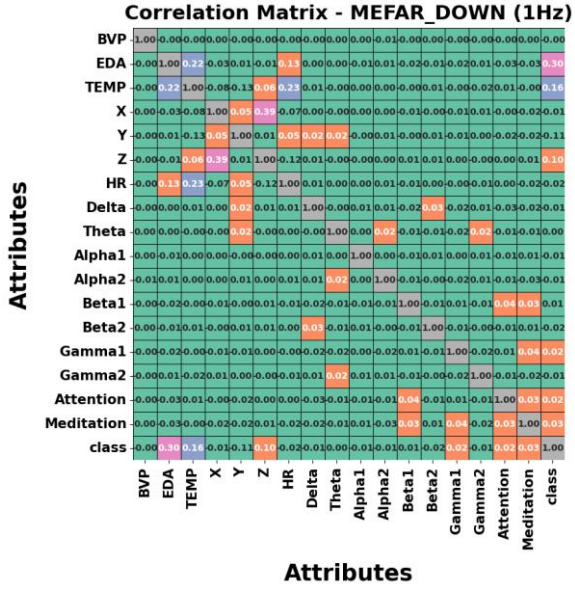


Fig. 5. Correlation matrix of RTCSM-XDL method on MEFAR_DOWN dataset.

Table III shows the confusion matrices and classifier outcomes of the RTCSM-XDL approach on the MEFAR_DOWN dataset. The simulation outcome values indicate that the RTCSM-XDL technique properly categorized the instances. According to 70% TRAIP, the RTCSM-XDL technique offers an average of $accur_y$ of 98.04%, $preci_n$ of 98.00%, $recal_l$ of 98.04%, $F1_{score}$ of 98.02%, and AUC_{score} of 98.04%, correspondingly. At same time, with 30% TESTP, the RTCSM-XDL approach offers an average of $accur_y$ of 98.13%, $preci_n$ of 98.10%, $recal_l$ of 98.13%, $F1_{score}$ of 98.11%, and AUC_{score} of 98.13%.

TABLE III. CONFUSION MATRICES AND CLASSIFIER OUTCOMES OF RTCSM-XDL METHOD ON MEFAR_DOWN DATASET

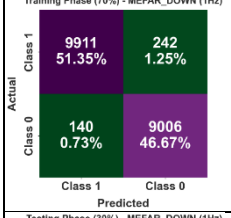
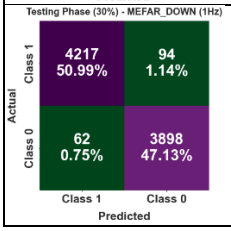
Confusion Matrix	Class	Accuracy	Precision	Recall	F1-Score	AUC Score
	Class 1	97.62	98.61	97.62	98.11	98.04
	Class 0	98.47	97.38	98.47	97.92	98.04
	Average	**98.04**	**98.00**	**98.04**	**98.02**	**98.04**
	Class 1	97.82	98.55	97.82	98.18	98.13
	Class 0	98.43	97.65	98.43	98.04	98.13
	Average	**98.13**	**98.10**	**98.13**	**98.11**	**98.13**

Fig. 6 establishes the classifier results of the RTCSM-XDL method on the MEFAR_DOWN dataset. Fig. 6(a) shows the accuracy study of the RTCSM-XDL system. The figure shows that the RTCSM-XDL model achieves rising values over growing epochs. Likewise, the increased validation over training reveals that the RTCSM-XDL system learns successfully on the test dataset. Also, Fig. 6(b) depicts the loss examination of the RTCSM-XDL. The outcome suggests that the RTCSM-XDL model reaches similar values of training and validation loss. It is observed that the RTCSM-XDL approach learns effectively on the test dataset. In Fig. 6(c), the ROC curve of the RTCSM-XDL technology is examined. The experimentation outcome point toward that the RTCSM-XDL technique accomplishes increased ROC outcomes overall classes, signifying an important ability to differentiate the classes.

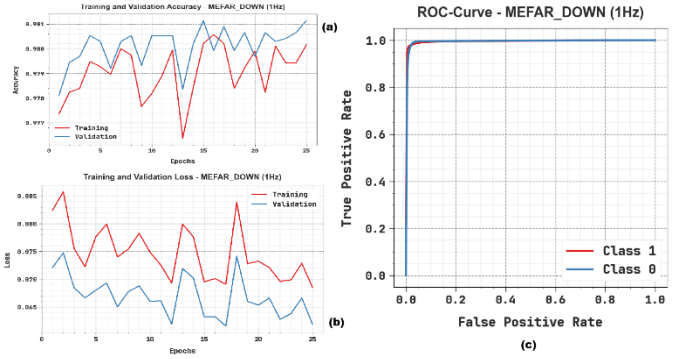


Fig. 6. RTCSM-XDL method (a) $Accur_y$, (b) Loss and (c) ROC curves on the MEFAR_DOWN dataset.

Fig. 7 presents the correlation matrix of the RTCSM-XDL method on the MEFAR_MID dataset, showing that most features have very low correlation values, implying smaller

redundancy. Only certain moderate correlations appear among related signals like accelerometer axes and EDA–TEMP. EEG band features stay largely independent of motion and physiological signals. Overall, this implies a well-distributed feature set appropriate for strong and unbiased model learning.

Table IV depicts the confusion matrices and classifier outcomes of the RTCSM-XDL approach on the MEFAR_MID dataset. The outcome implies that RTCSM-XDL system precisely recognized the instances. Under 70% TRAIP, the RTCSM-XDL system gives an average of $accur_y$ of 97.02%, $preci_n$ of 97.04%, $recal_l$ of 97.02%, $F1_{score}$ of 97.02%, and AUC_{score} of 97.02%, correspondingly. At the same time, with 30% TESTP, the RTCSM-XDL model offers an average of $accur_y$ of 97.11%, $preci_n$ of 97.14%, $recal_l$ of 97.11%, $F1_{score}$ of 97.11%, and AUC_{score} of 97.11%, correspondingly.

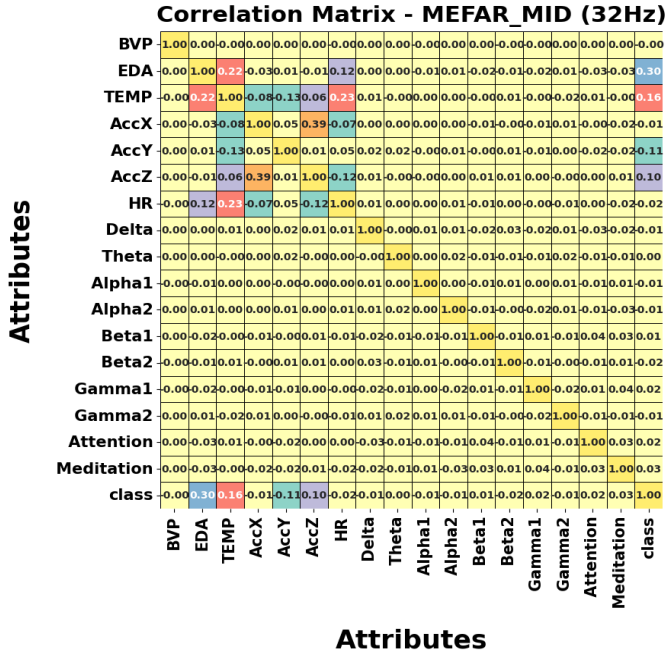


Fig. 7. Correlation matrix of RTCSM-XDL method on MEFAR_MID dataset.

TABLE IV. CONFUSION MATRICES AND CLASSIFIER OUTCOMES OF RTCSM-XDL METHOD ON MEFAR_MID DATASET

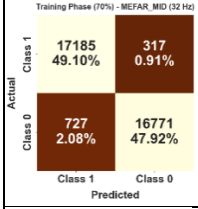
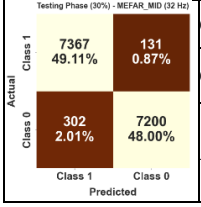
Confusion Matrix	Class	$Accur_y$	$Preci_n$	$Recal_l$	$F1_{score}$	AUC_{Score}
	Class 1	98.19	95.94	98.19	97.05	97.02
	Class 0	95.85	98.14	95.85	96.98	97.02
	Average	**97.02**	**97.04**	**97.02**	**97.02**	**97.02**
	Class 1	98.25	96.06	98.25	97.15	97.11
	Class 0	95.97	98.21	95.97	97.08	97.11
	Average	**97.11**	**97.14**	**97.11**	**97.11**	**97.11**

Fig. 8 portrays the classifier outcomes of the RTCSM-XDL model on the MEFAR_MID dataset. Fig. 8(a) shows the accuracy analysis of the RTCSM-XDL method. The figure reveals that the RTCSM-XDL system attains increased values over rising epochs. Further, the growing validation over training shows that the RTCSM-XDL method acquires efficacy. At last, Fig. 8(b) proves the loss examination of RTCSM-XDL model. The outcome implies that the RTCSM-XDL methodology achieves similar values of training and validation loss. In Fig. 8(c), the ROC curve of the RTCSM-XDL model is investigated. The simulation outcome values suggest that the RTCSM-XDL approach obtains higher ROC outcomes for overall classes, representing considerable capacity to distinguish the classes.

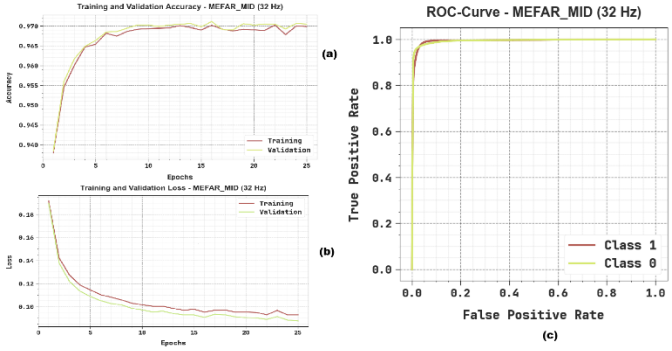


Fig. 8. RTCSM-XDL method (a) $Accur_y$, (b) Loss and (c) ROC curves on the MEFAR_MID dataset.

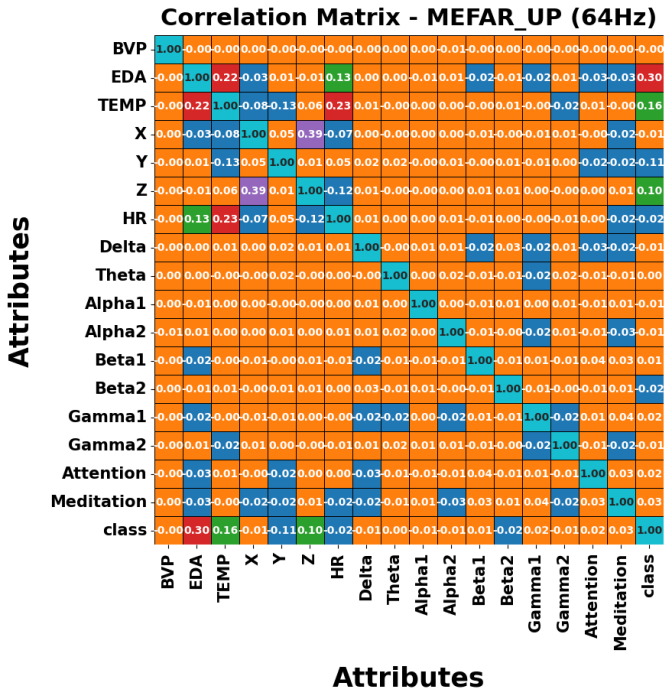


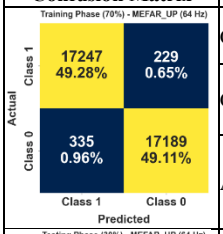
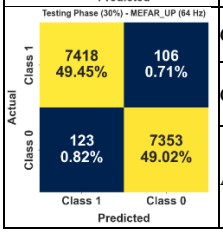
Fig. 9. Correlation matrix of the RTCSM-XDL method on the MEFAR_UP dataset.

Fig. 9 represents the correlation matrix of RTCSM-XDL method on the MEFAR_UP dataset exhibits that most features have very minimal correlation values, demonstrating lower dependency among them. Some moderate correlations appear among related signals, namely accelerometer axes and

physiological measurements such as EDA and TEMP. EEG frequency bands stay largely independent, implying decreased redundancy through spectral attributes. Overall, the dataset upholds better feature diversity that helps efficient and unbiased technique learning.

Table V displays the confusion matrices and classifier outcomes of the RTCSM-XDL technique on the MEFAR_UP dataset. The outcome implies that the RTCSM-XDL method correctly categorized the instances. With 70% TRAIP, the RTCSM-XDL model obtains an average of $accur_y$ of 98.39%, $preci_n$ of 98.39%, $recal_l$ of 98.39%, $F1_{score}$ of 98.39%, and AUC_{score} of 98.39%, correspondingly. At the same time, with 30% TESTP, the RTCSM-XDL model offers an average of $accur_y$ of 98.47%, $preci_n$ of 98.47%, $recal_l$ of 98.47%, $F1_{score}$ of 98.47%, and AUC_{score} of 98.47%, correspondingly.

TABLE V. CONFUSION MATRICES AND CLASSIFIER OUTCOMES OF RTCSM-XDL METHOD ON MEFAR_UP DATASET

Confusion Matrix	Class	$Accur_y$	$Preci_n$	$Recal_l$	$F1_{score}$	AUC_{Score}
	Class 1	98.69	98.09	98.69	98.39	98.39
	Class 0	98.09	98.69	98.09	98.39	98.39
	Average	**98.39**	**98.39**	**98.39**	**98.39**	**98.39**
	Class 1	98.59	98.37	98.59	98.48	98.47
	Class 0	98.35	98.58	98.35	98.47	98.47
	Average	**98.47**	**98.47**	**98.47**	**98.47**	**98.47**

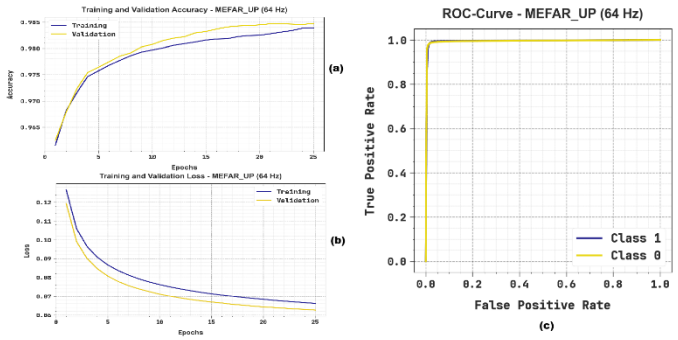


Fig. 10. RTCSM-XDL method (a) $Accur_y$, (b) Loss and (c) ROC curves on the MEFAR_UP dataset.

Fig. 10 represents the classifier results of RTCSM-XDL approach on the MEFAR_UP dataset. Fig. 10(a) describes the accuracy investigation of the RTCSM-XDL method. The figure shows that the RTCSM-XDL model realizes superior values over rising epochs. Also, the rising validation over training displays that the RTCSM-XDL system learns effectively on the test dataset. Lastly, Fig. 10(b) shows the loss examination of the RTCSM-XDL methodology. The outcome suggests that the RTCSM-XDL model achieves similar values of training and validation loss. In Fig. 10(c), the ROC curve of the RTCSM-XDL model is studied. The experimentation outcome indicated that the RTCSM-XDL technique realizes greater ROC outcomes of every class, indicating considerable ability to distinguish the classes.

Table VI and Fig. 11 portray the performance evaluation of the RTCSM-XDL approach with existing models [18]. The simulation results indicate that the method has shown better performance. The result indicated that the CNN-LSTM and H-CNN methodologies had inefficient performance with $accur_y$ of 75.00% and 75.21%, whereas the BLSTM-LSTM technique attains somewhat maximal performance with $accur_y$ of 82.57%, $preci_n$ of 95.23%, $recal_l$ of 93.62% and $F1_{score}$ of 92.17%, respectively. In addition, the LSTM method obtains slightly superior performance with $accur_y$ of 90.80%, $preci_n$ of 96.95%, $recal_l$ of 93.82% and $F1_{score}$ of 87.92%, respectively. Also, the Conv-LSTM method achieves significant performance with $accur_y$ of 93.75%, $preci_n$ of 86.94%, $recal_l$ of 83.01% and $F1_{score}$ of 93.94%, respectively. Furthermore, the Bi-GRU Network and RNN models reach performance values close to $accur_y$, $preci_n$, $recal_l$ and $F1_{score}$ Values. Ultimately, the proposed RTCSM-XDL approach achieves assuring results with $accur_y$ of 98.47%, $preci_n$ of 98.47%, $recal_l$ of 98.47%, and $F1_{score}$ of 98.47% when compared to other models, respectively

TABLE VI. COMPARISON OUTCOMES OF RTCSM-XDL METHOD WITH EXISTING MODELS

Classifier	$Accur_y$	$Preci_n$	$Recal_l$	$F1_{score}$
BLSTM-LSTM	82.57	95.23	93.62	92.17
Conv-LSTM	93.75	86.94	83.01	93.94
LSTM	90.80	96.95	93.82	87.92
Bi-GRU Network	98.00	90.15	95.26	96.47
RNN	97.71	83.78	85.10	95.33
CNN-LSTM	75.00	93.87	83.06	94.66
H-CNN	75.21	93.50	81.73	84.97
RTCSM-XDL	98.47	98.47	98.47	98.47

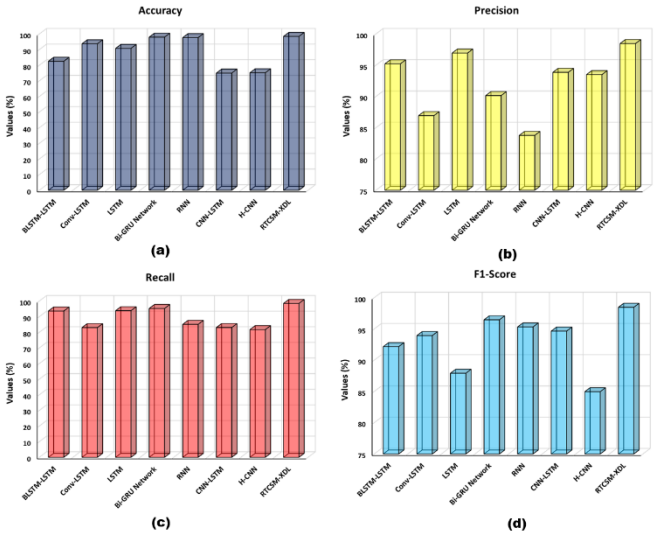


Fig. 11. Comparative outcomes of the RTCSM-XDL method (a) $accur_y$, (b) $preci_n$, (c) $recal_l$ and (d) $F1_{score}$.

Fig. 12 reveals LIME feature importance explanations for six individual test cases. Every subplot corresponds to one case and explains a single model prediction. Green bars imply features that support the predicted label, while red bars oppose it. The bar length signifies the strength of every feature's influence on the prediction. The probability displays the model's confidence in its prediction. The explanation fit specifies how well LIME approximates the local model.

Fig. 13 presents LIME explanations for the first six test cases of the hold-out dataset. All subplots explains one prediction by emphasizing influential features. Green bars denote features that support the predicted class, while red bars contradict it. The bar length shows the strength of every feature's contribution. The probability reveals the model's confidence in its prediction. The explanation fit displays how precisely LIME approximates the local model.

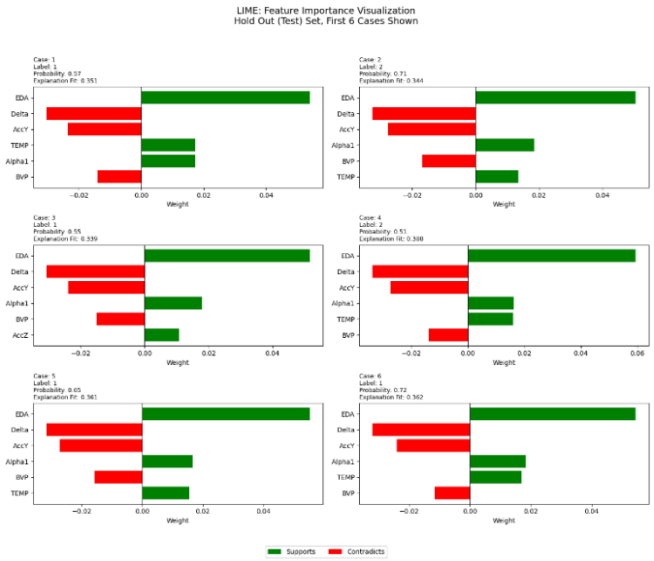


Fig. 12. LIME-based global and local interpretability analysis of the MEFAR_DOWN dataset.

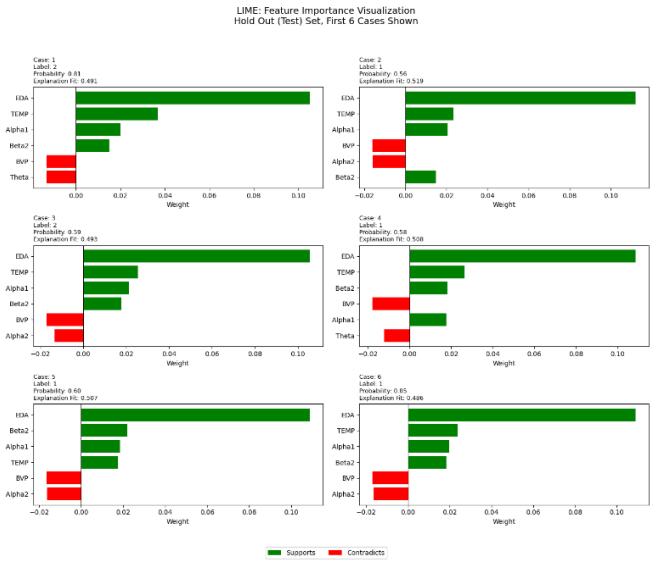


Fig. 13. LIME-based global and local interpretability analysis of MEFAR_MID dataset.

Fig. 14 depicts LIME feature importance explanations for six test cases from the hold-out dataset. Every subplot explains one prediction by displaying the most influential input features. Green bars signify features that support the predicted class, while red bars oppose it. The length of every bar signifies the strength of the feature's contribution. The prediction probability exhibits the model's confidence for all case. The explanation fit reflects how well LIME approximates the local model.

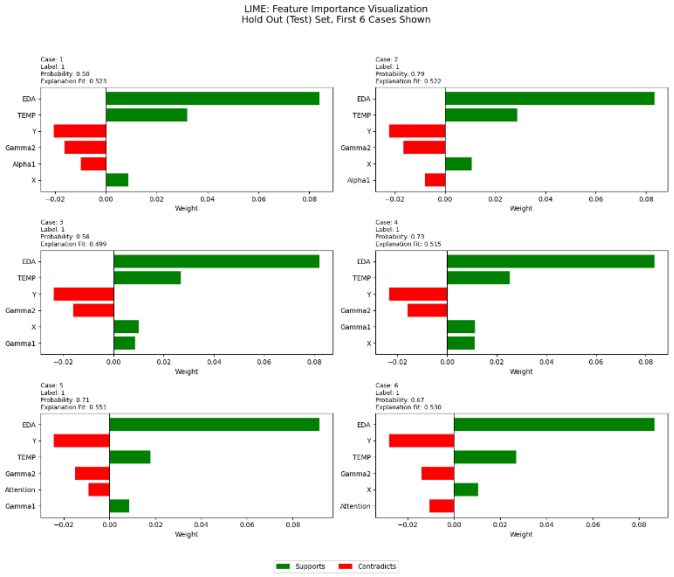


Fig. 14. LIME-based global and local interpretability analysis of MEFAR_UP dataset.

V. Conclusion

This study has presented the RTCSM-XDL model to attain robust and transparent cognitive state classification based on physiological signals by incorporating advanced data-driven learning and XAI methods. The data preprocessing step comprises data normalization, missing data handling, and categorical encoding to guarantee high-quality and consistent inputs. For FS, mRMR, RFE, and boruta is utilized to select the most informative physiological features. TRGU-Att model is developed to proficiently capture temporal dependencies and enable real-time inference. To overcome the interpretability challenges associated with DL architectures, the proposed technique employs LIME to offer human-understandable insights into model predictions. Experimental result of the proposed RTCSM-XDL technique are conducted on benchmark dataset and the comparison outcomes pointed out the superiority of the proposed technique across existing approaches in terms of different metrics.

Data Availability Statement: The data that support the findings of this study are openly available at https://data.mendeley.com/datasets/z3g26tphnv/3, reference number [17].

References

[1] S. R. Jeevakala and H. Ramasangu, "Classification of Cognitive States using Task-Specific Connectivity Features," Eng. Technol. Appl. Sci. Res., vol. 13, no. 3, pp. 10675–10679, Jun. 2023.

[2] V. A. Maksimenko, A. E. Hramov, N. S. Frolov, A. Lüttjohann, V. O. Nedaivozov, V. V. Grubov, et al., "Increasing human performance by

sharing cognitive load using brain-to-brain interface," Front. Neurosci., vol. 12, p. 949, Dec. 13, 2018.

[3] L. Monish and S. G. Shaila, "A Hybrid Fuzzy CNN–LSTM Approach for Emotion Recognition from EEG–ECG Physiological Signals," Eng. Technol. Appl. Sci. Res., vol. 15, no. 6, pp. 30405–30411, Dec. 2025.

[4] V. R. Kokitkar and A. Ghuli, "A Multi-Scale Inverted Spatial-Temporal Network for EEG-Based Emotion Recognition," Eng. Technol. Appl. Sci. Res., vol. 16, no. 3, pp. 36482–36489, Jun. 2026.

[5] N. Jamil and A. N. Belkacem, "Advancing real-time remote learning: A novel paradigm for cognitive enhancement using EEG and eye-tracking analytics," IEEE Access, vol. 12, pp. 93116–93132, 2024.

[6] M. Scheutz, S. Aeron, A. Aygun, J. P. de Ruiter, S. Fantini, C. Fernandez, et al., "Estimating systemic cognitive states from a mixture of physiological and brain signals," Top. Cogn. Sci., vol. 16, no. 3, pp. 485–526, Jul. 2024.

[7] Y. Li, Y. Li, X. He, J. Fang, C. Zhou, and C. Liu, "Learner's cognitive state recognition based on multimodal physiological signal fusion," Appl. Intell., vol. 55, no. 2, p. 127, 2025.

[8] N. Panwar, V. Pandey, and P. P. Roy, "Eeg-cognet: A deep learning framework for cognitive state assessment using eeg brain connectivity," Biomed. Signal Process. Control, vol. 98, p. 106770, 2024.

[9] [9] M. Li, J. Pan, Y. Li, Y. Gao, H. Qin, and Y. Shen, "Multimodal physiological analysis of impact of emotion on cognitive control in VR," IEEE Trans. Vis. Comput. Graph., vol. 30, no. 5, pp. 2044–2054, May 2024.

[10] B. Huangfu and W. Cheng, "Cognitive computing method based on decoding psychological emotional states," International Journal of Cognitive Computing in Engineering, vol. 6, pp. 32–43, 2025.

[11] J. Vogl, K. O'Brien, and P. St Onge, "One size does not fit all: A support vector machine exploration of multiclass cognitive state classifications using physiological measures," Front. Neuroergon., vol. 6, p. 1566431, Jun. 18 2025.

[12] G. Liu and B. Zhu, 2024. Dynamic EEG-fMRI mapping: Revealing the relationship between brain connectivity and cognitive state. arXiv preprint arXiv:2411.19922.

[13] J. Chen, A. Chen, B. Jiang, M. S. Obaidat, N. Li, and X. Zhang, 2024. Cross-subject Brain Functional Connectivity Analysis for Multi-task Cognitive State Evaluation. arXiv preprint arXiv:2408.15018.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[15] S. Bai, J. Z. Kolter, and V. Koltun, 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[17] https://data.mendeley.com/datasets/z3g26tphnv/3

[18] M. A. Afzal, Z. Gu, B. Afzal, and S. U. Bukhari, "Cognitive workload classification in industry 5.0 applications: Electroencephalography-based bi-directional gated network approach," Electronics (Basel), vol. 12, no. 19, p. 4008, 2023.