Deep Fusion of VGG19 and ConvMixer With Superlet Transform for Cognitive Load Detection

Jammisetty Yedukondalu, M. Ramesh, N. Janardhan, Sahebgoud Hanamantray Karaddi, T. Chandana Gowri, Yadavalli Murali Krishna, Lakhan Dev Sharma, *Senior Member, IEEE*, and Srinivasa Rao Karumuri

***Abstract*— The cognitive load triggers neural activity, which is crucial for understanding the brain's response to mental stress or stimuli that induce stress. Electroencephalogram (EEG) signals were collected from a mental arithmetic task (MAT), simultaneous task EEG workload datasets and segmented into 4-second intervals. These segmented signals were then transformed into images using various time-frequency conversion methods (TF), including the Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), Q Transform (QT), and Superlet Transform (SLT). The resulting TF images were fed into convolutional neural networks (CNNs), such as VGG19, ConvMixer, and a fusion of VGG19+ConvMixer. CNN models were trained using the Adam optimizer to detect cognitive load. The preprocessing involved normalization, and scaling in both phases. Among the models tested, the SLT-based TF-EEG with the fusion of the VGG19 and ConvMixer model outperformed other TF conversion methods and CNN architectures. The VGG19+ConvMixer utilizes the individual advantages of both VGG19 and ConvMixer models. It helps reduce overfitting and vanishing gradients, enhances performance with new data, improves GPU acceleration, and reduces computational cost due to its simpler architecture. However, the SLT effectively handles non-stationary data through its adaptive multi-resolution approach, making it ideal for EEG analysis. The SLT + VGG19 + ConvMixer model achieved accuracies of 97.26% on the MAT dataset and 96.04% on the STEW dataset, with all other evaluation metrics such as precision, sensitivity, specificity, F-score, MCC, Jaccard index, and Cohen's kappa exceeding 94%. These findings can enhance real-time cognitive load monitoring, benefiting areas such as personalized learning, mental health, and stress management by detecting cognitive load to improve performance and reduce stress in critical situations.**

***Index Terms*— Cognitive load, Superlet transform, time-frequency EEG, VGG19, ConvMixer.**

Received 17 April 2025; revised 27 July 2025, 24 August 2025, and 14 September 2025; accepted 1 October 2025. Date of publication 6 October 2025; date of current version 8 December 2025. *(Corresponding author: Srinivasa Rao Karumuri.)*

Jammisetty Yedukondalu is with the Department of ECE, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh 523272, India (e-mail: yedukondalu463@gmail.com).

M. Ramesh is with the School of Technology Management and Engineering, NMIMS Deemed-to-be-University, Hyderabad 509202, India (e-mail: munipalaramesh123@gmail.com).

N. Janardhan is with the Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University), Hyderabad 502329, India (e-mail: janardhan.ku@gmail.com).

Sahebgoud Hanamantray Karaddi is with the Department of ECE, Kishkinda University, Ballari, Karnataka 583120, India (e-mail: sahebgoudhkaraddi@kishkindauniversity.edu.in).

T. Chandana Gowri is with the Department of Computer Science and Engineering, SRM University-AP, Amaravati, Andhra Pradesh 522502, India (e-mail: chandanagowri.g@srmap.edu.in).

Yadavalli Murali Krishna is with the Department of ECE, QIS College of Engineering and Technology, Ongole 523272, India (e-mail: muralikrith123@gmail.com).

Lakhan Dev Sharma is with the School of Electronics Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522241, India (e-mail: devsharmalakhan@gmail.com).

Srinivasa Rao Karumuri is with the VLSI-Microelectronics Research Laboratory, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur, Andhra Pradesh 522302, India (e-mail: srinivasakarumuri@gmail.com).

Digital Object Identifier 10.1109/TCE.2025.3618009

I. Introduction

COGNITIVE load detection (CLD) refers to the process of measuring and identifying the mental effort exerted by an individual during a cognitive task. Cognitive load represents the task load of memory resources used to process information or solve problems. It seeks to evaluate this effort in real time or after the task, using various methods, including physiological signals (such as heart rate or eye movement), self-reporting and task performance metrics [1], [2], [3]. Excessive load on working memory during cognitive tasks can impair performance and make completing tasks challenging. Investigating cognitive load is crucial for preventing a range of medical and mental disorders, including seizures, depression, anxiety, and heart attacks. Brain dynamics are impacted by extended cognitive load, according to neuroscience research [3], [4], [5], [6]. CLD is crucial for enhancing learning outcomes, improving system designs, maintaining high performance in critical tasks, and fostering adaptive technology that responds to human cognitive states [7].

EEG signals are significant in cognitive load detection because they provide real-time, non-invasive insights into brain activity. EEG helps identify when an individual is experiencing high, moderate, or low cognitive load, making it valuable for applications in education, user experience design, and performance monitoring [8]. EEG signals are comprised of different frequency bands within the 0 to 100 Hz range, including delta, theta, alpha, beta, and gamma bands. Researchers can use each band to analyze and interpret brain activity in response to various tasks or mental states [9].

In recent years, a limited number of studies have explored the mental effort of operators using EEG data. Notably, Wilson [10] employed artificial neural network (ANN) for classification purposes, achieving accuracy rates of up to 90%. Wang et al. [11] introduced a multifractal analysis

of EEG data to measure mental exertion during arithmetic tasks, with a multi-channel algorithm reaching 95.87% accuracy and a one-channel algorithm achieving 84.15%. Al-Shargie et al. [12] achieved 94.79% accuracy in workload classification using EEG data and a support vector machine classifier with error-correcting output code. Malviya and Mal [13] employed DWT, CNN, and Bi-LSTM for decomposing signals and classifying stress. Zhang et al. [14] employed recurrent 3D CNNs to assess mental effort during multitasking, achieving 88.9% accuracy. Budak et al. [15] proposed an EEG-based model for driver drowsiness detection using LSTM, reaching 94.31% accuracy. Malviya and Mal [16] used, statistical importance (CIS), DWT, LDA, XGBoost and Extra Trees (ET) to achieve 97.14% accuracy in decomposition, feature extraction, and classification. Yenurkar and Mal [17] used a mayfly optimization algorithm combined with linear regression to detect COVID-19. In another study [18], random forest and XGBoost were utilized for forecasting pandemic related stress. Roy et al. [19] proposed a DWT-based CNN, BiLSTM, and GRU network model for stress classification, reaching 97.20% accuracy. Saadati and colleagues utilized a CNN for cognitive load detection using a combination of EEG and functional near-infrared spectroscopy data. They achieved an accuracy of 89%, highlighting the effectiveness of CNNs in extracting spatial features from EEG signals. Zhang et al. [20] proposed attention based multiscale spatial-temporal convolutional network for motor imagery EEG decoding and acheived an accuracy of 96.35%.

Elman Recurrent Neural Networks (RNNs) have been shown to improve epilepsy diagnosis by incorporating temporal context in cross-participant modeling of EEG data [21]. Güler et al. [22] and Übeyli [23] demonstrated that group-trained, cross-validated models reduced diagnostic error by 63% and 74%, respectively, compared to non-recurrent approaches. Another study utilized a high-fidelity vehicle simulator and recurrent neural networks to track occipital lobe activity during lane disturbance events in a simulated highway driving scenario [24]. In forecasting a sleepiness metric, ensembles of Recurrent Self-Evolving Fuzzy Neural Networks (RSEFNNs) performed marginally better than other neural network ensembles; however, a comprehensive statistical analysis was not carried out to validate these results. Despite this limitation, Liu et al. [24] demonstrated that recurrent network ensembles can achieve strong results in cross-participant, stimulus-aligned tasks. Wang et al. [25] proposed the STFT TF conversion method and a convolutional block attention module incorporated with ResNet18 and DenseNet121 for workers' fatigue detection. They reported accuracies of 94.9% and 93.9%, respectively. Huang et al. [26] presented a CWT-based TF conversion method and a convolutional neural network model with a channel and frequency attention mechanism to automatically distinguish between cognitive states. They achieved an average accuracy of 76.14%. Chanda et al. [27] proposed a Fast Fourier Transform-based TF conversion method and a CNN for detecting cognitive impairment, reporting an accuracy of 93%. Moreover, Vafaei and Hosseini [28] introduced four primary transformer architectures namely, time-series, vision, graph-based, and hybrid models and examined their variants applied to EEG tasks including motor imagery, emotion recognition, and seizure detection. Li et al. [29] developed a Transformer Neural Architecture Search (TNAS) using a multiobjective evolutionary algorithm, achieving state-of-the-art EEG emotion recognition on DEAP and DREAMER while balancing accuracy and model size. Xie et al. [30] developed Transformer-based models for motor imagery EEG, reaching 83.31% cross-individual accuracy on PhysioNet; positional embeddings boosted performance, and attention maps revealed ERD patterns, enhancing both accuracy and interpretability for BCI. Several deep learning models have been proposed for time–frequency analysis of EEG signals. EEGNet captures temporal and spatial patterns using depthwise separable convolutions [31], while ShallowNet focuses on bandpower features via shallow temporal filters [32]. BCINetV1 introduces spectral–temporal attention for improved motor imagery decoding [33], and EEGPT leverages transformer-based pretraining for generalized EEG representation across tasks [34].

The CNNs, machine learning algorithms and RNNs have served as the predominant architecture for deep learning techniques employed in computer vision applications for several years. Transformer-based architectures, like as the Vision Transformer (ViT), have recently demonstrated remarkable efficacy across several applications, frequently surpassing traditional convolutional networks, particularly in the context of large datasets. To utilize Transformers for images, their representation requires modification; directly applying self-attention layers at the per-pixel level would result in processing expenses increasing quadratically with the number of pixels in each image. The solution entails partitioning the image into several patches, linearly embedding each patch, and subsequently applying the Transformer to this assemblage of patches. A fundamental convolutional architecture known as ConvMixer was introduced to EEG cognitive load detection. It closely resembles the ViT and MLP-Mixer [35]. It employs direct patch processing, maintains uniform size and resolution of representations across all layers, halts down-sampling of representations in subsequent layers, and distinguishes between channel-wise and spatial data mixing. Unlike the ViT and MLP-Mixer, the proposed fusion model employs standard convolutions to do all these functions.

In this work, EEG data were initially preprocessed with power line notch, low-pass, and high-pass filters. The EEG data was divided and normalized using z-scores, then Images are created by several time-frequency conversion processes like STFT, Q-Transform, CWT and SLT. For CLD, a VGG19, ConvMixer and concatenated VGG19 with convMixer were employed in this article. The proposed framework combines VGG19 and ConvMixer to achieve balancing between local accuracy and global comprehension of context. VGG19's strong ability to extract spatial features and ConvMixer's ability to efficiently combine global features make this possible. This complementary integration improves the model's ability to pick up on small changes in time-frequency components derived from EEG, making it more reliable and accurate at

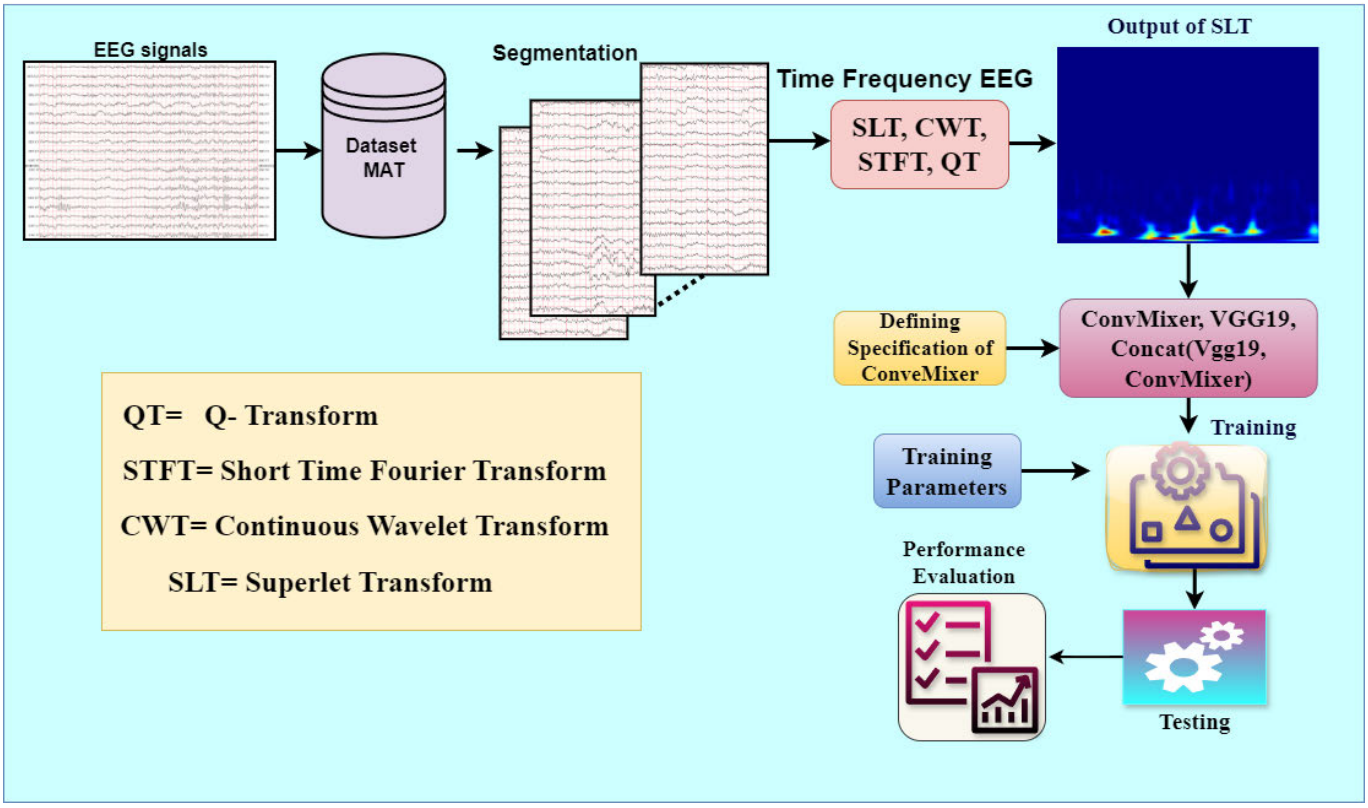


Fig. 1. Proposed method diagram for CLD.

detecting cognitive load. The main contribution of the work are as follows:

- Various time-frequency analysis models such as SLT, STFT, CWT, and QT employed for 1D to 2D EEG conversion.
- Employed VGG19, convMixer, and fusion of VGG19 and ConvMixer models to detect cognitive load.
- Applied the occlusion sensitivity (OS), Gradient class activation map (GCAM), Local interpretable model explanations (LIME) to understand the decision made by the model.
- Performance analysis done by various metrics like κ: Cohens Kappa; Sens: Sensitivity; JI: Jaccard Index; FS: f-score.

II. Dataset & Methods

In this work, the MAT and STEW publicly available datasets were employed to detect the CLD. The National Technical University of Ukraine provided a dataset comprising 36 students aged between 18 and 26 years. These students completed a MAT involving two numbers in serial subtraction. The EEG recording utilized 20 electrodes arranged in a 10-20 pattern across the scalp, despite there being 23 electrodes available. The recordings contain artifact-free EEG segments, comprising 3 minutes of relaxation with closed eyes and 1 minute of mental counting. It serves as a valuable resource about neural dynamics and cognitive processes during tasks [36]. The STEW dataset contains raw EEG data from 48 participants who participated in a multitasking workload experiment using the SIMKAP multitasking test. In addition to task-related data, the dataset includes recordings of the participants' resting brain activity taken before the test. The EEG data was collected using the Emotiv EPOC device, which features 14 channels, a sampling frequency of 128 Hz, and records for 2.5 minutes per session [37].

The proposed approach is organized into five main steps: pre-processing, segmentation, normalization, TF methods, and a combination of VGG19, ConvMixer, and their fusion (VGG19+ConvMixer). Initially, the 1D EEG signal is filtered using notch filters to eliminate noise. Following this, the data is normalized using the z-score method and then segmented into 4-sec intervals. Then various time-frequency conversion techniques are used to convert the resultant signal into an image, including STFT, QT, CWT, and SLT. A CNN network, optimized with the Adam optimizer, is employed to detect cognitive load. The proposed model was evaluated using a total of 10,080 samples from the STEW dataset and 1,242 samples from the MAT dataset, which were separately split into training, validation, and testing sets using an 8:1:1 ratio. Fig. 1 illustrates the structural flow of the proposed methodology, and which are describes the following subsections.

A. Data Normalization

Data normalization is the process of scaling individual data points to ensure consistency across a dataset, improving the performance of machine learning models and making features more comparable. It helps to bring all values into a common range, to have a mean of 0 and a standard deviation of

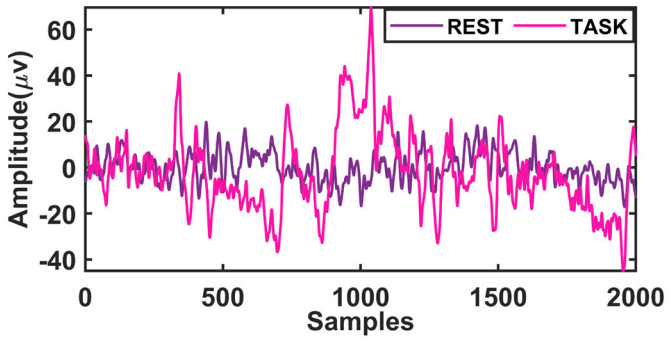


Fig. 2. 4sec segmented EEG rest and task signal.

1 (z-score normalization). Each data point is adjusted by subtracting the mean and then divided by the standard deviation. The z-score formula is mathematically represented as follows:

$$V_{norm}(n) = \frac{V_{data}(n - \bar{\mu})}{\sigma} \tag{1}$$

where, V_{norm} and V_{data} are normalized and data values. The z-score normalization was specifically conducted on a per-subject, per-channel basis. All segmented 4-second trials were utilized to calculate the mean and standard deviation individually for each EEG channel for each participant. Subsequently, normalization was performed individually for each channel by removing the subject's mean and dividing by the standard deviation. This approach reduces inter-subject variability while maintaining essential temporal and spectral features in each subject's EEG data. It also guarantees that no single topic or channel dominates the majority of the features, hence enhancing the generalizability and equity of the categorization model.

B. Time Frequency EEG

Time-Frequency EEG analysis is a technique used to examine the dynamic changes in the frequency content of EEG signals over time. Unlike traditional EEG analysis, which looks at either the time domain (changes in amplitude) or the frequency domain (distribution of power across frequency bands), time-frequency analysis integrates both perspectives, offering a more comprehensive view of brain activity. Since signals are non-stationary, their statistical characteristics change with time, time-frequency analysis effectively captures these variations, providing a more accurate representation of dynamic brain processes that may be overlooked in static time or frequency domain analysis. Hence in this article, we have employed four time frequency conversion methods: STFT, CWT, Q-Transform, and SLT. The sample 4-sec rest and task EEG signal as shown in Fig. 2 was converted 2D images using various TF techniques and which are describe in the following subsections. These methods are useful for automatic recognition of cognitive states or identifying neurological conditions.

1) STFT: It is a technique for analyzing the frequency content of non-stationary signals over time. It works by dividing the signal into overlapping time segments (windows) and performing a Fourier Transform on each segment. This produces a time-frequency representation, showing how the signal's frequency content evolves over time. The STFT of a signal $x_s(t)$ is defined as:

$$X(t, f) = \int_{-\infty}^{\infty} x_s(\tau) w(t - \tau) e^{-j2\pi f \tau} d\tau \tag{2}$$

This represents the transformation of a signal $x_s(\tau)$ using a window function $w(t - \tau)$, $e^{j2\pi f \tau}$ is the frequency content.

2) CWT: The CWT analyzes signals by breaking them into scaled and shifted wavelet representations, making it ideal for non-stationary signals as it captures both time and frequency information using localized functions. This contrasts with the Fourier Transform, which relies on sinusoidal functions and provides only frequency information without time localization. The CWT of a signal $x_s(t)$ is represented as:

$$W(p, q) = \int_{-\infty}^{\infty} x_s(t) \psi^* \left(\frac{t - q}{p} \right) dt \tag{3}$$

In this context, $\psi(t)$ represents the mother wavelet, p denotes the scale factor, q is the translation factor, and t is the time variable.

3) Q-Transform: It is a type of wavelet transform in which the wavelet is adjusted to maintain a consistent quality factor Q, ensuring that the ratio of the center frequency to bandwidth remains constant. The Q-transform is well-suited for analyzing signals with oscillatory patterns because it offers improved frequency resolution for low frequencies and enhanced time resolution for high frequencies. The Q-transform of a signal $x(t)$ is defined as:

$$X_Q(t, f) = \int_{-\infty}^{\infty} x(\tau) \psi_Q^*(t - \tau, f) d\tau \tag{4}$$

Here, $X_Q(t, f)$ represents the Q-transform of the signal in both time and frequency, while $\psi_Q(t - \tau, f)$ denotes the Q-transform wavelet.

4) SLT: The SLT is an advanced time-frequency analysis method, building on the concept of the Morlet wavelet but enhancing its resolution through an adaptive number of cycles per frequency. It is especially useful for high-resolution analysis of EEG data, capturing both precise time and frequency information. In 2021, Moca et al. [38] used SLT to create a time-frequency spectrum with outstanding resolution. To achieve a more precise and less leaky time-frequency representation, the Superlet Transform employs a set of wavelets with progressively narrower bandwidths, referred to as superlets (SL). Unlike more conventional techniques such as the CWT and STFT, the SLT does not necessitate a trade-off between frequency and temporal resolution. While STFT provides outstanding frequency resolution at high frequencies, its temporal resolution suffers as a result. On the other hand, CWT maintains a rather high temporal resolution across the spectrum, but frequency resolution becomes redundant and deteriorates with increasing frequency [38]. Due to the inherent time-frequency uncertainty, STFT and CWT are less suitable for analyzing signals with rich time-frequency content, such as brain signals. Comprehensive discussions on various time-frequency domain methods for analyzing physiological signals can be found in [39]. In the SLT method, the collaboration between short wavelets providing higher temporal

resolution and long wavelets offering superior frequency resolution aims to address the resolution limitation observed in STFT and CWT. SLT's appropriate time-frequency resolution allows for the high precision observation of fast transient oscillation. SLT is a beneficial option for examining neural signals because of its capacity to identify high-frequency bursts [38].

A superlet is a collection of Morlet wavelets that span a variety of cycles and have a fixed core frequency (f_0), as defined by Eq. (5).

$$SL_{f_0 n} = \left\{ Y_{f_0,m} | m_1, m_2, \ldots m_n \right\} \tag{5}$$

In this case, n refers to the order of the wavelet transform (SL). On the other hand, m represents the number of cycles that fluctuate within a given range, typically between 1 and n. Here, Eq. (6) defines the mother wavelet, also known as the modified Morlet wavelet, in SL.

$$Y_{f_0,m}(t) = \frac{1}{B_m \sqrt{2\pi}} e^{-\frac{t^2}{2B^2_m}} e^{i2\pi f_0 t} \tag{6}$$

B_m is time spread parameter (measured in seconds) that is specified in Eq. (7). Where f_0 and m termed as the central frequency and number of base cycles.

$$B_m = \frac{m}{\sigma_{sd} \times f_0} \tag{7}$$

B_m shows an inverse relationship with frequency and controls the wavelets' time variance. A low value of B_m corresponds to a broad frequency response, while a high value results in a narrower frequency response [40]. A wave is said to encompass m cycles within the standard deviation (σ_{sd}) of the Gaussian envelope when the value of B_m is chosen to ensure that the wavelet's oscillatory cycles fit within the Gaussian window, where the window's width is defined by the σ_{sd}.

Since it significantly affects the ability to portray different signal features using a time-frequency representation, wavelet normalization is essential. This work normalizes the wavelet with the modulus's unit integral. This normalization allows wavelets to provide a representation that is scale-free, allows for the identification of self-similar events spanning various scales [41].

The multiplicative ($m_j = j \times m_1$) or additive ($m_j = m_1 + j - 1$) rules measures the number of wavelet cycles in SL. The study uses the multiplicative relationship to find the wavelet's cycle count. As per Eq. (8), the geometric mean (GM) of the individual wavelets' responses in the SL method represents the response of SL to a signal Y.

$$\mathbb{R}[SL_{f_0 n}] = n \sqrt{\prod_{j=1}^{n} \mathbb{R}[Y_{f_0,m_j}]} \tag{8}$$

In this case, the complex convolution of the Morlet wavelet is the j^{th} wavelet's response to the signal Y and it is denoted as $\mathbb{R}[Y_{f_0,m_j}]$, which is provided by Eq. (9).

$$\mathbb{R}[Y_{f_0,m_j}] = \sqrt{2} y \times Y_{f_0,m_j} \tag{9}$$

The SL technique identifies the oscillation bursts in the signal at its central frequency, f_0. By squaring the SL magnitude, the scalogram is obtained, while the geometric mean (GM) of the response magnitudes of individual wavelets in SL is utilized to approximate the magnitude of SL [38]. Similar to CWT, SLT employs SL, whereas CWT utilises wavelet [42].In the case of SLTs of order 1, CWT can be seen as a specific instance of SLT. The study in [38] shows that, unlike the CWT, SLTs of orders greater than 1 offer a clearer, more efficient representation of signals with reduced redundancy.

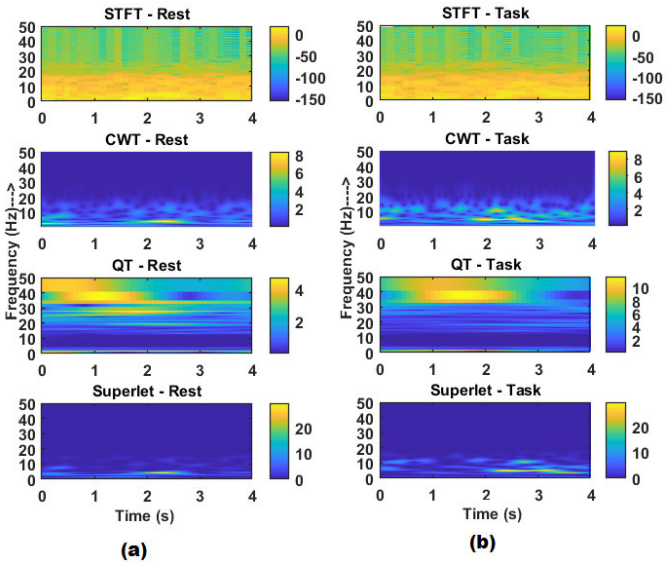


Fig. 3. Rest and task EEG TF images using STFT, CWT, Q-transform, and SLT models.

A variation of SLT called adaptive SLT (ASLT) is suggested to handle signals with a broad frequency range [38]. Adaptive SLs automatically adjust the order of SLs to the core frequency. The ASLT compensates for the wider wavelet bandwidth at higher frequencies. ASLT achieves this by employing lower orders to estimate lower frequencies and escalating the order as the frequency increases. This approach enhances the precision and improves the time-frequency resolution by adjusting the order of superlets in relation to the frequency as seen below:

$$ASL_{f0} = SL_{f_0,n} | n = a(f_0) \tag{10}$$

Here, $a(f_0)$ is an integer value function of the center frequency that increases monotonically.

$$a(f_0) = n_{min} + [(n_{max} - n_{min}).\frac{f_0 - f_{min}}{f_{max} - f_{min}}] \tag{11}$$

The smallest and largest central frequencies, f_{min} and f_{max} are are chosen as 0.1 Hz and 50 Hz, respectively. These correspond to the minimum and maximum superlet orders, $n_{min} = 1$ and $n_{max} = 30$ While ASLT is better for wide frequency ranges, SLT is appropriate for narrow band analysis.

The Fig. 3 illustrates the TF representation of the SLT, showcasing its superior representation over STFT, CWT, and Q-transform as demonstrated for the signals in Fig. 2. SLT offers exceptional resolution in both time and frequency, effectively capturing transient features and rapid changes. In contrast, STFT suffers from resolution trade-offs, CWT exhibits smearing, and the Q-transform struggles with

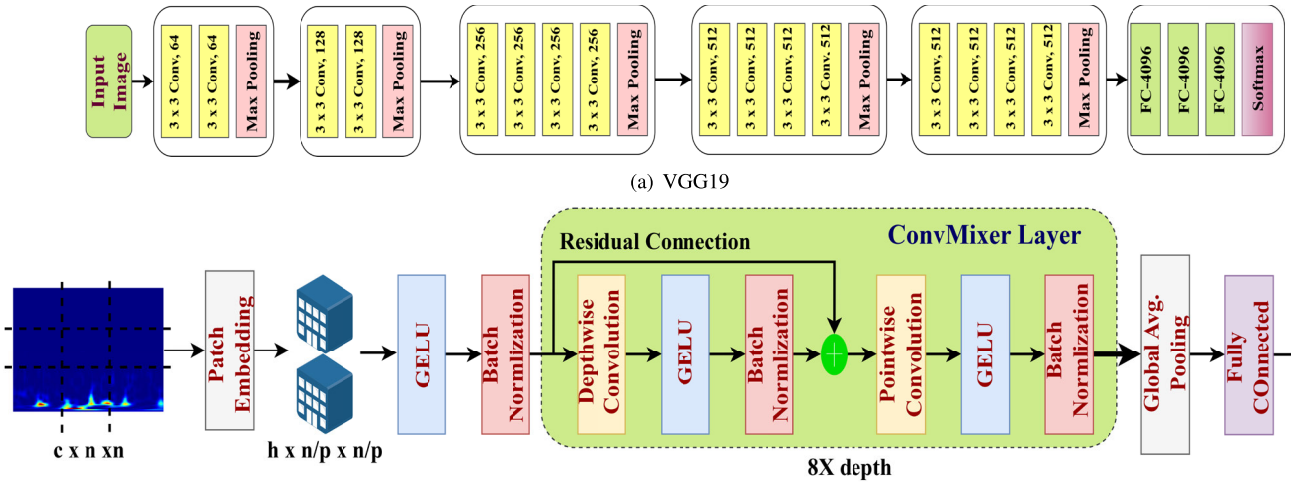


Fig. 4. Detailed architecture of the a) VGG19 and b) ConvMixer that are used for the detection.

handling complex dynamics. SLT's adaptability allows it to tailor its frequency resolution to local signal properties, making it particularly well-suited for non-stationary EEG signals. Unlike the fixed-window STFT, partially adaptive CWT, and limited Q-transform, SLT smoothly represents amplitude modulations and power variations with minimal fluctuations. Other methods, by comparison, introduce inconsistencies. The SLT enhances rest and task classification by providing high-resolution time-frequency representations of EEG signals. It effectively captures transient features, spectral power, and amplitude variations, which are critical for distinguishing rest and task states. SLT transforms 1D EEG signals into 2D time-frequency images, preserving key features. These images are then used as inputs to a classifier, such as VGG, which leverages its convolutional layers to extract hierarchical features. This integration improves the classifier's ability to identify task-specific spectral patterns and localized power changes, leading to more accurate rest and task classification.

C. Architectures of VGG19+ConvMixer

In this work the fusion VGG19 and ConvMixer were employed for classification. The VGG19 is used for its efficacy in feature extraction, efficient training durations, compact design, resilience, GPU acceleration, minimized overfitting, and enhanced performance on novel data and classification accuracy. ConvMixer employs residual or skip connections to mitigate overfitting and vanishing gradient problems inside the model. Utilizing both models, each with unique benefits and limitations, reduces the overfitting and enhances adaptability to new data. Proposed model can take the advantages of each paradigm by integrating them. This fusion also provide efficient feature extraction and fine tuning the mode. The detailed architecture of the model is described in the following sections.

1) Architecture VGG19: The Fig. 4(a) illustrates the architecture of the VGG19 CNN. VGG19 is a popular deep learning model primarily used for image classification tasks and was developed by the Visual Geometry Group (VGG) at Oxford. The model has 19 layers, which is why it's referred to as VGG19. The explanations of each layer, including the mathematical functions:

- Input Image: The network typically accepts an input image of size $224 \times 224 \times 3$, where: - 224 is the height and width of the image (pixels). - 3 represents the RGB color channels.
- Convolutional layers: Each convolutional layer in VGG19 uses a 3×3 filter with a stride of 1 and padding to preserve spatial dimensions. Convolutional layers utilize filters to analyze the input image and capture features such as edges, textures, and shapes. Mathematical representation of each convolutional operation as:

$$\text{Output}_{i,j,k} = \sum_{m,n,c} \left(\text{Input}_{i+m,j+n,c} \times \text{Filter}_{m,n,c,k}\right) + \text{Bias}_k \tag{12}$$

  Here, i, j are the spatial coordinates, k is the filter index, m, n are the filter dimensions, and c represents the input channels.
- Layers decomposition: Block 1 consists of two convolutional layers with filters of size (3×3), resulting in 64 feature maps each. It is then followed by a Max Pooling layer that reduces the spatial dimensions by half. Block 2 includes two convolutional layers with (3×3) filters, generating 128 feature maps each. Lastly, Block 3 comprises four convolutional layers with filters of (3×3), producing 256 feature maps each. Block 4 consists of four convolutional layers with (3×3) filters, each yielding 512 feature maps. In a similar manner, Block 5 comprises four convolutional layers with (3×3) filters, resulting in 512 feature maps each.
- Fully Connected Layers: which act as classifiers, are integrated into the network following the convolutional layers and max-pooling operations. Flattening: a one-dimensional vector is created from the output of the last pooling layer. FC Layers: Three fully connected layers,

each with 4096 neurons.

$$\text{Output} = \text{ReLU}(\text{Weights} \cdot \text{Input} + \text{Bias}) \quad (13)$$

This operation is repeated for each FC layer.

- Softmax Layer: It generates probabilities for each class, with the sum of probabilities equaling 1. It is employed to forecast the probability of each class in the ultimate classification task.

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (14)$$

Here, z_i is the score for class i. This architecture's use of small 3×3 filters allows for deep feature extraction while keeping computation manageable, making it effective for complex image classification tasks.

2) Architecture of ConvMixer: In ViT, the creation and representation of patches play a crucial role in enhancing the transformer's performance. The ViTs and MLP mixer networks will reduce the created patch dimensions after processing. On the other hand, ViTs and MLP mixer models perform channel-wise and spatial-wise convolutions. Therefore, this could potentially lead to an increase in the computational complexity and time required for training and testing datasets. As a result, we proposed a ConvMixer, which performs point-wise convolution while maintaining the same patch dimension throughout the image process. Therefore, this approach significantly reduces both complexity and processing time.

The structure of the ConvMixer is as shown in the Fig. 4(b). In this paper, ConvMixer_256_8 structure is used, where as 256 is the number of filters used and 8 is the depth of the model. The concept of mixing is the foundation of ConvMixer's design [35]. Specifically, we decided to blend channel locations using point-wise convolution and spatial locations using depth-wise convolution. The concept that MLPs and self-attention mechanisms can integrate information from disparate geographical locations, effectively having an infinitely broad receptive field, is a key idea derived from earlier research. As a result, we mixed remote spatial locations using convolutions with an abnormally high kernel size. After patch embedding GeLU and batch normalization layers are connected followed by eight convMixer layers are connected that consists of residual connections and point-wise convolutions with GeLU and batch normalization. The output is defined mathematically as:

$$CM_0 = BatchNorm(\sigma_{gelu}(conv_{Cin \to d}(X, S = p, k_{size} = p))) \quad (15)$$

where, CM_0 is the output of the ConvMixer layer, d is the embedding dimension, p is the patch size, S is stride, c_{in} is the input channels, and k_{size} is the kernel size. ConvMixer worked better for high kernel size in depthwise convolutions. GeLU and batch normalization are connected before and after each convolutions:

$$CM'_l = BatchNorm(\sigma_{gelu}(convDepth(CM_{l-1}))) + CM_{l-1} \quad (16)$$

$$CM_{l+1} = BatchNorm(\sigma_{gelu}(convPoint(CM'_l)) \quad (17)$$

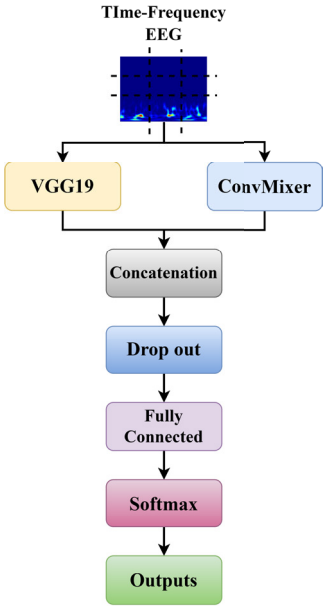


Fig. 5. Deep fusion of the VGG19 and ConvMixer using concatenation.

After convMixer layers performed average pooling to get feature vector. A fully connected (FC) layer is connected to output of the convMixer layer and followed by softmax for the classification. This architecture provide efficient feature extraction, avoid overfitting, and gives the enhanced performance to new data.

The proposed model essentially comprises an MLP-Mixer enhanced with convolutions. It functions directly on embedded patches, ensuring uniform resolution and dimensions throughout the layers. Moreover, depth-wise separable convolution differentiates between channel-wise and spatial information mixing, similar to MLP-Mixer, and incorporates comparable skip connections. The proposed framework is indeed a fully convolutional neural network. All operations of ConvMixer may be executed just using activations, batch normalization, and convolutions. Therefore, it is essentially a CNN with certain architectural hyper-parameters. The advantages of the ConvMixer_256_8 is as follows [43]:

- Simple, basic and isotropic architecture.
- Uses simple convolutional layers, batch normalization, and activation as like CNNs.
- Uses fixed-dimension of patch throughout the process.
- Less Computational complexity.

Fig. 5 shows the structure of the VGG19+convMixer. We removed the last softmax layer of both VGG19 and convMixer and then added a concatenation layer after pooling layer and followed by added dropout, a fully connected layer, and a softmax layer after concatenation. This fusion of VGG19 and convMixer achieved a compromise between local accuracy and global contextual understanding through VGG19's robust spatial feature extraction and ConvMixer's efficient global feature amalgamation. This complementing integration enhances the model's capacity to discern subtle variations in EEG-derived time-frequency components, hence augmenting its robustness and accuracy in identifying cognitive load.

TABLE I
SPECIFICATIONS AND TRAINING PARAMETERS USED TO TRAIN THE PROPOSED MODEL

Specifications	Value
Epochs	30
Input size	224
Batch size	256
Learning rate	0.0001
Weight decay factor	0.0003
Loss	Sparse_categorical_cross entropy
Optimizer	Adam
Model	ConvMixer (256 filters, depth 8)
Depth	8
Patch size	2
Activation	GELU
Total Parameters	37,558,600

In contrast to earlier studies that utilized CNNs such as AlexNet, VGG16, or ResNet for EEG classification tasks. The TF representations derived from the EEG were utilized to train the ConvMixer and VGG19 models from the ground up. This judgment was shaped by two principal elements. Initially, pretrained features may be less beneficial in this context due to the considerable divergence between EEG TF representations and natural images, which are utilized to train models like VGG16. A concatenation-based fusion technique was incorporated into the topologies of both models, substituting the softmax layers. This hampers the direct application of pretrained weights. To more effectively capture the domain-specific attributes of the EEG TF data, all layers of the model were randomly initialized and optimized during the training phase. Previous studies, [44] and [45] utilized transfer learning by transforming EEG signals into image-like time-frequency representations and optimizing the upper layers of pretrained models while preserving the initial convolutional layers. This work demonstrated that initiating the model from a baseline enabled it to acquire features precisely adapted to the EEG data properties, hence enhancing robustness and generalization.

III. PERFORMANCE ANALYSIS

This section provides the proposed model results and existing models discussion via performance metrics. The 16GB RAM, 1TB SSD with intel core i5 processor GPU and Jupiter with Python 3.0 version was used for the proposed model simulation.

A. Results

Table II presents the proposed model, VGG19, and ConvMixer results respectively, with various evolution metrics. The performance metrics reveal that VGG19 demonstrates decent accuracy (80.75%) but has room for improvement, while ConvMixer shows a significant increase in accuracy (94.6%) and better precision (96.82%), indicating enhanced reliability. The combination of VGG19 and ConvMixer achieves the highest accuracy (97.26%) and precision (98.56%), highlighting the benefits of model fusion. In terms of sensitivity, ConvMixer excels (96.32%) compared to VGG19 (82.28%), with the combined model further enhancing detection of true positives (97.96%). The Fig. 6a and 6b are kernel weights and patches of ConvMixer model. Cohen's Kappa values indicate improved agreement with the combination model (85.68%). The Jaccard Index shows a reduction in false negatives for ConvMixer (93.37%) and the combined model (96.58%). While VGG19 has lower specificity (63.72%), ConvMixer (88.16%) and the combined model (94.63%) demonstrate better performance in identifying negative cases. Finally, the F-Score reflects strong improvements across models, with VGG + ConvMixer achieving the highest score (98.26%) for MAT, showcasing exceptional sensitivity and precision balance. However, We acheived an ACC of 96.02%, JI of 93.26%, PRE of 96.05%, FS of 96.04%, SPE of 96.03%, Sens of 95.02%, κ of 85.33% for STEW dataset. The ANOVA results, with an F-statistic of 6.66 and a p-value of 0.0057, indicate that there are statistically significant differences in the performance of the three models (VGG19, ConvMixer, VGG+ConvMixer) across various evaluation metrics, such as ACC, PRE, AS, κ, JI, MCC, SPE, and FS as specified in Table II. The p-value (less than 0.05) suggests that the performance variations between the models shows a significant difference in performance. indicate that both ConvMixer and VGG+ConvMixer significantly outperform the baseline VGG19 model. However, the difference between ConvMixer and VGG+ConvMixer is not statistically significant, suggesting that both models perform comparably at a high level. The robustness of the evaluated models, 95% confidence intervals (CIs) were reported alongside the mean performance metrics. These intervals reflect the range within which the true metric values are expected to lie with high confidence, thereby indicating the consistency of model performance. The Fig. 7, 8 depicts the confusion matrix of proposed model.

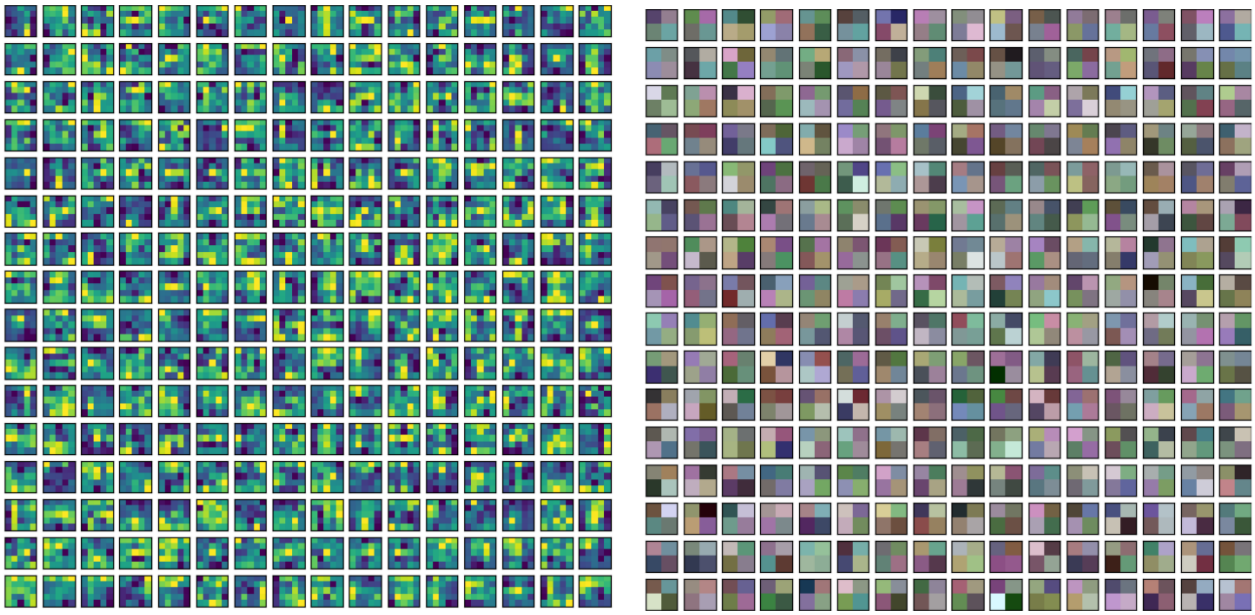
TABLE II
PERFORMANCE OF THE PROPOSED MODELS

Model	ACC	PRE	Sens	κ	JI	MCC	SPE	FS
VGG19	80.75	96.20	82.28	44.76	79.69	30.74	63.72	88.69
ConvMixer	94.60	96.82	96.32	74.55	93.37	83.91	88.16	96.57
VGG+ConvMixer (MAT)	97.26	98.56	97.96	85.68	96.58	91.83	94.63	98.26
VGG+ConvMixer (STEW)	96.02	96.05	95.02	85.33	93.26	92.11	96.03	96.04
* κ : Cohens Kappa; Sens: Sensitivity; JI: Jaccard Index; FS: f-score								

Overall, the results highlight that while VGG19 provides a solid baseline, ConvMixer significantly improves performance across all metrics. The combination of VGG19 and ConvMixer (VGG + ConvMixer) yields the best results, demonstrating that fusing models can lead to superior classification capabilities. This is particularly valuable in applications like cognitive load detection, where both precision and recall are critical. The analysis suggests that utilizing multiple model architectures can capitalize on their strengths and provide more reliable and accurate predictions.

B. Discussion

This section describes the post image analysis using deep learning heat-maps like GCAM, OS, and LIME to understand



(a) Kernel Weights (b) Patches of the Kernel

Fig. 6. ConvMixer (256 filters, depth 8) model weights and patches.

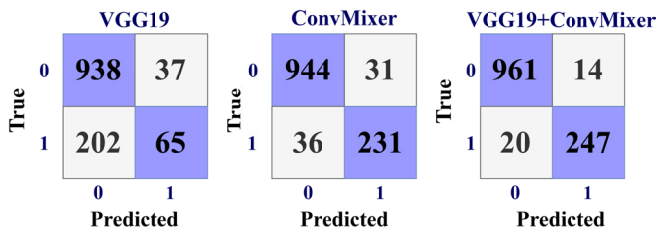


Fig. 7. Confusion matrices of the MAT dataset for VGG19, ConvMixer, and the combined VGG19+ConvMixer models.

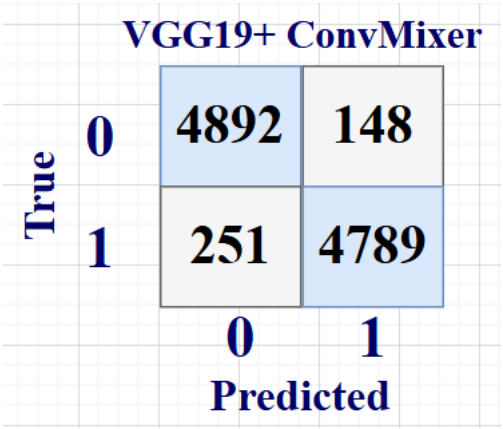


Fig. 8. STEW dataset confusion matrix of VGG19+convMixer.

the decision making by the model and comparison of the proposed model with state-of-the-art models. Fig. 9 presents the visualization of the images using GCAM, OS, and LIME to understand the decision made by the model [46].

- GCAM: Highlights the region's most influential to the model's prediction, with focused activations that align well with task-relevant features. The fusion model leverages the high resolution and adaptability of the SLT, enhancing GCAM's ability to capture transient features associated with cognitive load changes.
- LIME: Offers localized explanations by perturbing specific image regions to assess their impact on the model's output. In the fusion model, LIME captures task-relevant spectral features but exhibits slightly less localization detail compared to GCAM. This indicates the model's robustness in identifying critical regions.
- OS: Displays sensitivity by occluding parts of the input and observing the prediction shifts. The broader activations in OS heat-maps emphasize regions of importance during cognitive load detection, but with less precision compared to GCAM, highlighting the complementarity of these techniques.

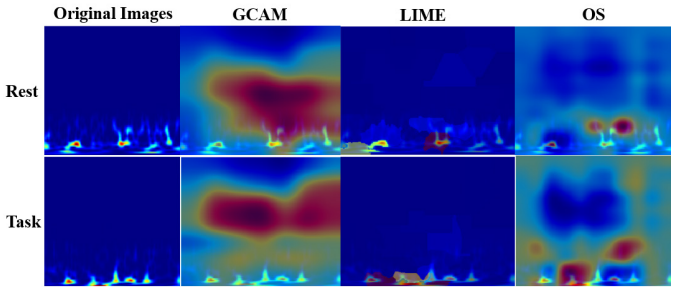


Fig. 9. Heat-maps applied to Understand the decision making by the proposed model. It describes the area that more descriminative features are extracted for decision making.

We have implemented heatmaps for rest vs. task in addition to the current qualitative visualizations. Class-specific attention patterns are evident in these visualizations, which also highlight how each interpretability technique highlights distinct but complementary areas of interest. This enhancement highlights the model's capacity to pay attention to fleeting EEG characteristics under various cognitive circumstances.

TABLE III
COMPARISON OF PREVIOUS STATE-OF-ART-METHODS WITH PROPOSED MODEL

Ref.no/year	Method	ACC	PRE	Specificity
[52]/2023	CNN	95.8	96.41	94.44
[54]/2021	LSTM	93.2	92.55	92.74
[55]/2020	GW0 + BILSTM-LSTM	86.3	86.88	70.59
[49]/2023	BiLSTM + ANN	87.1	89.99	84.61
[51]/2022	CNN + LSTM	86.2	88.88	84.19
[56]/2021	CNN + LSTM	94.8	93.10	94.55
[57]/2024	TCN+Multi-Space Deep Model	74.09	74.00	74.51
[58]/2024	1DCNN+BiLSTM	95.36	95.24	94.56
[59]/2024	SVM	88.96	88.79	88.90
[53]/2024	CNN+LSTM	83.12	83.55	84.00
Proposed	**SLT+VGG19+ConvMixer (MAT)**	**97.26**	**98.56**	**94.86**
Proposed	**SLT+VGG19+ConvMixer (STEW)**	**96.02**	**96.05**	**96.03**

These improvements show clear proof that the explanations given by the proposed fusion model are reliable and consistent, while also making the interpretability analysis deeper.

The proposed deep fusion model effectively integrates the spatial feature extraction of VGG19, the patch-based processing of ConvMixer, and the time-frequency adaptability of the SLT. This combination ensures superior identification of spectral and temporal features relevant to cognitive load. Compared to state-of-the-art models, the heat-maps generated by the fusion model exhibit sharper focus and higher clarity, particularly in distinguishing rest and task conditions. The integration of the SLT significantly enhances the model's ability to capture transient and non-stationary EEG features, making it highly effective for cognitive load detection. VGG19 alone lacks efficient spatial mixing, and ConvMixer alone lacks hierarchical feature extraction. VGG19+ConvMixer overcomes these limitations by leveraging the strengths of both, leading to better adaptability, performance, and efficiency. However, the combined model may introduce increased complexity and tuning requirements.

C. Comparison of the Proposed Model With Previous State-of-the-Art Models

This section presents a comparison between the proposed model and established deep neural network architectures. It delves into various types of deep neural networks, like stacked LSTM [47], bidirectional LSTM (BLSTM) [48], [49], CNN-LSTM [50], and stacked autoencoder [14]. The stacked LSTM architecture which has multiple LSTM layers are stacked on top of each other and each with multiple memory cells. Including more hidden layers enhances the model's depth and enables it to acquire more specific representations.The BLSTM combines outputs from two parallel LSTMs operating in both forward and backward directions. In the CNN-LSTM architecture, TCNN layers are used for feature extraction, and these features are then used for sequence prediction [51]. It looks like you want to form a sentence using the information provided. Here's a reformulation: An autoencoder model comprises two primary components: an encoder, which compresses the input data into a lower-dimensional representation, and a decoder, which reconstructs the original data from this compressed representation. One or more LSTM layers are used for the encoder and decoder functions in a stacked autoencoder with an LSTM architecture [52], [53]. Table III presents a

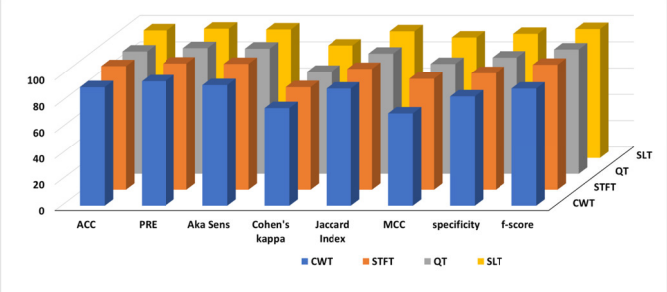


Fig. 10. Evolution metrics comparison of various TF methods with VGG19+ConvMixer.

comparative exploration of other DNN models in conjunction with the suggested model. Existing methodologies, such as CNN, LSTM, and their hybrids have shown limitations in either model complexity, overfitting to small datasets, or inadequate feature extraction capabilities for high-dimensional data. Furthermore, methods like GW0 + BiLSTM-LSTM suffer from significantly lower Specificity (70.59%), making them unsuitable for tasks requiring high sensitivity and reliability. The proposed SLT+VGG19+ConvMixer method addresses these gaps by combining the strengths of:

- Superlet Transform (SLT) for enhanced time-frequency analysis, capturing fine-grained, multi-resolution features essential for complex signal data.
- VGG19, a deep CNN architecture for robust feature extraction.
- ConvMixer, which effectively blends convolutional and mixer layers to capture both local and global patterns in the data.
- The fusion of VGG19 and ConvMixer combines the advantages of both approaches, ensuring robust spatial feature extraction and improved generalization through enhanced global-context learning.

However, we have compared the various TF techniques with the proposed CNN fusion model (VGG19+ConvMixer) achieved better performance as shown in Fig. 10. The conventional TF methods like STFT and wavelet transform face limits in resolution and require manual parameter tuning. The SLT overcomes these issues by combining multiple wavelets through a geometric mean, providing high-resolution, multi-scale representations with improved mode separation and reduced reliance on manual tuning [38], [60]. These representations are then processed by a fusion of VGG19 and ConvMixer, enabling fine-grained spatial feature extraction. As per performance analysis we conclude that the SLT based 2D EEG with fusion CNN model precise to detect the CLD rest of all TF conversion and CNN models. These innovative hybrid methods are instrumental in enhancing the analysis of EEG signals for clinical diagnoses, including identifying neurological disorders such as seizure detection, Alzheimer's detection, and abnormal EEG detection.

Table IV model presents the comparison of the proposed model with different transformer models. From the Table IV, it can be observed that proposed model achieved highest accuracy of 97.26%. The VGG19+ConvMixer model presents several improvements over Vision Transformers (ViTs) by integrating the benefits of CNNs with transformer topologies.

TABLE IV
COMPARISON OF PROPOSED MODEL WITH DIFFERENT TRANSFORMER MODELS

Methods	ACC	PRE	AS	κ	JI	MCC	SPE	FS
ViT	95.10	96.80	95.90	81.40	93.20	87.60	91.50	96.40
SVD+ViT [61]	94.86	96.05	91.72	–	–	–	89.34	93.83
Swin-T	96.50	97.70	96.10	85.20	95.90	91.70	94.00	96.40
Bio-signal Transformer [62]	83.67	97.10	–	–	–	86.40	–	85.96
Proposed (VGG19+ConvMixer)	**97.26**	**98.90**	**98.30**	**87.00**	**97.20**	**93.50**	**95.80**	**98.70**

SVD: Singular Value Decomposition; Swin-T: Swin Transformer

In contrast to ViTs, which need extensive datasets and considerable computer resources, the VGG19+ConvMixer model demonstrates greater computational efficiency and effective generalization even with little training data. By utilizing VGG19's robust local feature extraction and ConvMixer's global representation learning, the model effectively reconciles spatial hierarchy and long-range interdependence, surpassing only transformer-based methodologies. Moreover, it maintains superior interpretability in comparison to ViTs, which depend on intricate self-attention processes. The VGG19+ConvMixer model offers a more resilient, efficient, and scalable option compared to individual ViTs, rendering it exceptionally appropriate for practical applications.

IV. CONCLUSION

This paper successfully demonstrates that the proposed SLT-based TF-EEG approach, combined with the VGG19+ ConvMixer model, achieves superior accuracy and robustness in detecting cognitive load from 2D EEG data. Leveraging the SLT's multi-resolution capability makes it particularly effective for handling non-stationary EEG signals, enhancing sensitivity to variations in cognitive load when compared to TF methods like, CWT, STFT, QT. The fusion VGG19+ConvMixer architecture further strengthens the model's ability to process complex time-frequency features, achieving high accuracy (97.26%, 96.04%) and precision (98.56%, 96.06%) for EEG datasets metrics that outperform other tested methods. These findings suggest that the SLT+VGG19+ConvMixer model offers a powerful and reliable solution for CLD, with potential applications in brain-computer interfaces and cognitive monitoring systems. The future directions for CLD is:

- Using a larger dataset with a broader age range can enhance the efficiency and accuracy of cognitive load detection methods.
- RNN and GAN models can be employed to enhance classification accuracy.

REFERENCES

[1] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educ. Psychologist*, vol. 38, no. 1, pp. 63–71, Jan. 2003.

[2] J. Yedukondalu and L. D. Sharma, "Cognitive load detection using circulant singular spectrum analysis and binary Harris hawks optimization based feature selection," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104006.

[3] J. Yedukondalu, L. D. Sharma, and A. Bhattacharyya, "Cognitive load detection using adaptive/fixed-frequency empirical wavelet transform and multi-domain feature optimization," *Biomed. Signal Process. Control*, vol. 110, Dec. 2025, Art. no. 108124.

[4] C.-T. Lin, J.-T. King, J.-W. Fan, A. Appaji, and M. Prasad, "The influence of acute stress on brain dynamics during task switching activities," *IEEE Access*, vol. 6, pp. 3249–3255, 2018.

[5] I. Hassan, M. Zolezzi, H. Khalil, R. M. A. Saady, S. Pedersen, and M. E. H. Chowdhury, "Cognitive load estimation using a hybrid cluster-based unsupervised machine learning technique," *IEEE Access*, vol. 12, pp. 118785–118801, 2024.

[6] H. Jin, L. Zhu, M. Li, and V. G. Duffy, "Recognition and evaluation of mental workload in different stages of perceptual and cognitive information processing using a multimodal approach," *Ergonomics*, vol. 67, no. 3, pp. 377–397, Mar. 2024.

[7] J. Yedukondalu, D. Sharma, and L. D. Sharma, "Subject-wise cognitive load detection using time–frequency EEG and bi-LSTM," *Arabian J. Sci. Eng.*, vol. 49, no. 3, pp. 4445–4457, Mar. 2024.

[8] N. V. Thakor and S. Tong, "Advances in quantitative electroencephalogram analysis methods," *Annu. Rev. Biomed. Eng.*, vol. 6, no. 1, pp. 453–495, Aug. 2004.

[9] J. Yedukondalu and L. D. Sharma, "Cognitive load detection using ci-SSA for EEG signal decomposition and nature-inspired feature selection," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 31, no. 5, pp. 771–791, Sep. 2023.

[10] G. F. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures," *Int. J. Aviation Psychol.*, vol. 12, no. 1, pp. 3–18, Jan. 2002.

[11] Q. Wang and O. Sourina, "Real-time mental arithmetic task recognition from EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 2, pp. 225–232, Mar. 2013.

[12] F. Al-Shargie, T. B. Tang, N. Badruddin, and M. Kiguchi, "Towards multilevel mental stress assessment using SVM with ECOC: An EEG approach," *Med. Biol. Eng. Comput.*, vol. 56, no. 1, pp. 125–136, Jan. 2018.

[13] L. Malviya and S. Mal, "A novel technique for stress detection from EEG signal using hybrid deep learning model," *Neural Comput. Appl.*, vol. 34, no. 22, pp. 19819–19830, Nov. 2022.

[14] P. Zhang, X. Wang, W. Zhang, and J. Chen, "Learning spatial–spectral–temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 1, pp. 31–42, Jan. 2019.

[15] U. Budak, V. Bajaj, Y. Akbulut, O. Atila, and A. Sengur, "An effective hybrid model for EEG-based drowsiness detection," *IEEE Sensors J.*, vol. 19, no. 17, pp. 7624–7631, Sep. 2019.

[16] L. Malviya and S. Mal, "CIS feature selection based dynamic ensemble selection model for human stress detection from EEG signals," *Cluster Comput.*, vol. 26, no. 4, pp. 1–15, Aug. 2023.

[17] G. Yenurkar and S. Mal, "Future forecasting prediction of COVID-19 using hybrid deep learning algorithm," *Multimedia Tools Appl.*, vol. 82, no. 15, pp. 22497–22523, Jun. 2023.

[18] G. K. Yenurkar et al., "Multifactor data analysis to forecast an individual's severity over novel COVID-19 pandemic using extreme gradient boosting and random forest classifier algorithms," *Eng. Rep.*, vol. 5, no. 12, p. 12678, Dec. 2023.

[19] B. Roy et al., "Hybrid deep learning approach for stress detection using decomposed EEG signals," *Diagnostics*, vol. 13, no. 11, p. 1936, Jun. 2023.

[20] Y. Zhang, P. Li, L. Cheng, M. Li, and H. Li, "Attention-based multiscale spatial–temporal convolutional network for motor imagery EEG decoding," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 2423–2434, Feb. 2024.

[21] J. Elman, "Finding structure in time," *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, Jun. 1990.

[22] N. Guler, E. Ubeyli, and I. Guler, "Recurrent neural networks employing Lyapunov exponents for EEG signals classification," *Expert Syst. Appl.*, vol. 29, no. 3, pp. 506–514, Oct. 2005.

[23] E. D. Übeyli, "Analysis of EEG signals by implementing eigenvector methods/recurrent neural networks," *Digit. Signal Process.*, vol. 19, no. 1, pp. 134–143, Jan. 2009.

[24] Y.-T. Liu, Y.-Y. Lin, S.-L. Wu, C.-H. Chuang, and C.-T. Lin, "Brain dynamics in predicting driving fatigue using a recurrent self-evolving fuzzy neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 2, pp. 347–360, Feb. 2016.

[25] Y. Wang, Y. Huang, B. Gu, S. Cao, and D. Fang, "Identifying mental fatigue of construction workers using EEG and deep learning," *Autom. Construction*, vol. 151, Jul. 2023, Art. no. 104887.

[26] D. Huang et al., "Decoding subject-driven cognitive states from EEG signals for cognitive brain–computer interface," *Brain Sci.*, vol. 14, no. 5, p. 498, May 2024.

[27] C. Simfukwe, Y. C. Youn, M.-J. Kim, J. Paik, and S.-H. Han, "CNN for a regression machine learning algorithm for predicting cognitive impairment using qEEG," *Neuropsychiatric Disease Treatment*, vol. 19, pp. 851–863, Apr. 2023.

[28] E. Vafaei and M. Hosseini, "Transformers in EEG analysis: A review of architectures and applications in motor imagery, seizure, and emotion classification," *Sensors*, vol. 25, no. 5, p. 1293, Feb. 2025.

[29] C. Li, Z. Zhang, X. Zhang, G. Huang, Y. Liu, and X. Chen, "EEG-based emotion recognition via transformer neural architecture search," *IEEE Trans. Ind. Informat.*, vol. 19, no. 4, pp. 6016–6025, Apr. 2023.

[30] J. Xie et al., "A transformer-based approach combining deep learning network and spatial–temporal information for raw EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2126–2136, 2022.

[31] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[32] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[33] M. Z. Aziz, X. Yu, X. Guo, X. He, B. Huang, and Z. Fan, "BCINetV1: Integrating temporal and spectral focus through a novel convolutional attention architecture for MI EEG decoding," *Sensors*, vol. 25, no. 15, p. 4657, Jul. 2025.

[34] T. Yue et al., "BrainGPT: Unleashing the potential of EEG generalist foundation model by autoregressive pre-training," 2024, *arXiv:2410.19779*.

[35] I. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 24261–24272.

[36] I. Zyma et al., "Electroencephalograms during mental arithmetic task performance," *Data*, vol. 4, no. 1, p. 14, Jan. 2019.

[37] W. L. Lim, O. Sourina, and L. P. Wang, "STEW: Simultaneous task EEG workload data set," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 11, pp. 2106–2114, Nov. 2018.

[38] V. V. Moca, H. Bârzan, A. Nagy-Dăbâcan, and R. C. Mureşan, "Time-frequency super-resolution with superlets," *Nature Commun.*, vol. 12, no. 1, p. 337, Jan. 2021.

[39] P. K. Chaudhary, V. Gupta, and R. B. Pachori, "Fourier-bessel representation for signal processing: A review," *Digit. Signal Process.*, vol. 135, Apr. 2023, Art. no. 103938.

[40] C. Parameswariah and M. Cox, "Frequency characteristics of wavelets," *IEEE Trans. Power Del.*, vol. 17, no. 3, pp. 800–804, Mar. 2002.

[41] J. M. Lilly, "Element analysis: A wavelet-based method for analysing time-localized events in noisy time series," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 473, no. 2200, Apr. 2017, Art. no. 20160776.

[42] P. M. Tripathi, A. Kumar, M. Kumar, and R. Komaragiri, "Multilevel classification and detection of cardiac arrhythmias with high-resolution superlet transform and deep convolution neural network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[43] S. Park. (2021). *Convmixer: Patches Are All You Need? Overview and Thoughts*. [Online]. Available: https://medium.com/codex/an-overview-on-convmixer-patches-are-all-you-need-8502a8d87011

[44] M. T. Sadiq, M. Z. Aziz, A. Almogren, A. Yousaf, S. Siuly, and A. U. Rehman, "Exploiting pretrained CNN models for the development of an EEG-based robust BCI framework," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105242.

[45] S. Siuly, S. K. Khare, E. Kabir, M. T. Sadiq, and H. Wang, "An efficient Parkinson's disease detection framework: Leveraging time-frequency representation and AlexNet convolutional neural network," *Comput. Biol. Med.*, vol. 174, May 2024, Art. no. 108462.

[46] S. Hirose and T. Saitoh, "Automatic detection of concrete defects using laser ultrasonic visualization technique based on deep learning," in *Proc. IEEE Ultrason., Ferroelectr., Freq. Control Joint Symp. (UFFC-JS)*, Sep. 2024, pp. 1–3.

[47] R. Joshi, J. Ghosh, N. Kalani, and R. L. Tanna, "Assessment of stacked LSTM, bidirectional LSTM, ConvLSTM2D, and auto encoders LSTM time series regression analysis at ADITYA-U tokamak," *IEEE Trans. Plasma Sci.*, vol. 52, no. 7, pp. 2403–2409, Jul. 2024.

[48] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, Jul. 2005.

[49] G. Yoo, H. Kim, and S. Hong, "Prediction of cognitive load from electroencephalography signals using long short-term memory network," *Bioengineering*, vol. 10, no. 3, p. 361, Mar. 2023.

[50] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.

[51] N. E. Mughal et al., "EEG-fNIRS-based hybrid image construction and classification using CNN-LSTM," *Frontiers Neurorobotics*, vol. 16, Aug. 2022, Art. no. 873239.

[52] M. Saini, U. Satija, and M. D. Upadhayay, "DSCNN-CAU: Deep-learning-based mental activity classification for IoT implementation toward portable BCI," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8944–8957, May 2023.

[53] M. Safari, R. Shalbaf, S. Bagherzadeh, and A. Shalbaf, "Classification of mental workload with EEG analysis by using effective connectivity and a hybrid model of CNN and LSTM," *Comput. Methods Biomechanics Biomed. Eng.*, vol. 2024, pp. 1–15, Jul. 2024.

[54] A. Sundaresan, B. Penchina, S. Cheong, V. Grace, A. Valero-Cabré, and A. Martel, "Evaluating deep learning EEG-based mental stress classification in adolescents with autism for breathing entrainment BCI," *Brain Informat.*, vol. 8, no. 1, p. 13, Dec. 2021.

[55] D. D. Chakladar, S. Dey, P. P. Roy, and D. P. Dogra, "EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm," *Biomed. Signal Process. Control*, vol. 60, Jul. 2020, Art. no. 101989.

[56] M. Kang, S. Shin, J. Jung, and Y. T. Kim, "Classification of mental stress using CNN-LSTM algorithms with electrocardiogram signals," *J. Healthcare Eng.*, vol. 2021, pp. 1–11, Jun. 2021.

[57] H.-H. Nguyen, N. K. Iyortsuun, S. Kim, H.-J. Yang, and S.-H. Kim, "Mental workload estimation with electroencephalogram signals by combining multi-space deep models," *Biomed. Signal Process. Control*, vol. 94, Aug. 2024, Art. no. 106284.

[58] V. Sharma and M. K. Ahirwal, "An end-to-end brain computer interface system for mental workload estimation through hybrid deep learning model," *Hum.-Centric Intell. Syst.*, vol. 4, no. 4, pp. 599–609, Nov. 2024.

[59] M. Safari, R. Shalbaf, S. Bagherzadeh, and A. Shalbaf, "Classification of mental workload using brain connectivity and machine learning on electroencephalogram data," *Sci. Rep.*, vol. 14, no. 1, p. 9153, Apr. 2024.

[60] H. Bârzan, A.-M. Ichim, V. V. Moca, and R. C. Mureşan, "Time-frequency representations of brain oscillations: Which one is better?" *Frontiers Neuroinform.*, vol. 16, Apr. 2022, Art. no. 871904.

[61] D. D. Chakladar, "Vision transformer and brain connectivity patterns for estimating cognitive states," *IEEE Access*, vol. 13, pp. 74606–74616, 2025.

[62] M. M. Azizi and B. BabaAli, "Biosignals based automated driver cognitive load assessment using a pre-trained transformer," *IEEE Trans. Intell. Vehicles*, early access, Nov. 20, 2024, doi: 10.1109/TIV.2024.3501016.