IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 September 2022, accepted 5 October 2022, date of publication 12 October 2022, date of current version 19 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3213996

RESEARCH ARTICLE

Decoding the User's Movements Preparation From EEG Signals Using Vision Transformer Architecture

MAGED S. AL-QURAISHI[1,2], (Member, IEEE), IRRAIVAN ELAMVAZUTHI[1], (Senior Member, IEEE), TONG BOON TANG[1], (Senior Member, IEEE), MUHAMMAD S. AL-QURISHI[3], (Member, IEEE), SYED HASAN ADIL[4], (Member, IEEE), MANSOOR EBRAHIM[4], (Senior Member, IEEE), AND ALBERTO BORBONI[5], (Member, IEEE)

[1]Smart Assistive and Rehabilitative Technology (SMART) Research Group, Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, Bandar, Seri Iskandar 32610, Malaysia
[2]Faculty of Engineering, Thamar University, Dhamar, Yemen
[3]Research and Innovation Division, Research Department, Elm Company, Riyadh 12382, Saudi Arabia
[4]Faculty of Engineering, Sciences and Technology, Iqra University, Karachi 75500, Pakistan
[5]Mechanical and Industrial Engineering Department, Universita degli Studi di Brescia, 25123 Brescia, Italy

Corresponding author: Irraivan Elamvazuthi (irraivan_elamvazuthi@utp.edu.my)

This work was supported by the International Joint Research Project under Grant 015ME0-225.

ABSTRACT Electroencephalography (EEG) signals have a major impact on how well assistive rehabilitation devices work. These signals have become a common technique in recent studies to investigate human motion functions and behaviors. However, incorporating EEG signals to investigate motor planning or movement intention could benefit all patients who can plan motion but are unable to execute it. In this paper, the movement planning of the lower limb was investigated using EEG signal and bilateral movements were employed, including dorsiflexion and plantar flexion of the right and left ankle joint movements. The proposed system uses Continuous Wavelet Transform (CWT) to generate a time–frequency (TF) map of each EEG signal in the motor cortex and then uses the extracted images as input to a deep learning model for classification. Deep Learning (DL) models are created based on vision transformer architecture (ViT) which is the state-of-the-art of image classification and also the proposed models were compared with residual neural network (ResNet). The proposed technique reveals a significant classification performance for the multiclass problem ($p < 0.0001$) where the classification accuracy was 97.33 ± 1.86 % and the F score, recall and precision were 97.32 ± 1.88 %, 97.30 ± 1.90 % and 97.36 ± 1.81 % respectively. These results show that DL is a promising technique that can be applied to investigate the user's movements intention from EEG signals and highlight the potential of the proposed model for the development of future brain-machine interface (BMI) for neurorehabilitation purposes.

INDEX TERMS Continuous wavelet transform, deep learning, electroencephalography, motor-related cortical potentials, vision transformers architecture.

I. INTRODUCTION

In the fields of neural rehabilitation and human-robot interaction HRI, electroencephalography (EEG) is one of the most commonly used physiological signals [1]. EEG can propagate commands from the brain without involving potentially weakened physical neural pathways (such as peripheral nerves and muscles), which is why both healthy and disabled people can use it. Numerous methods for decoding the motion intentions using EEG in a brain-machine interface (BMI) have been investigated (actual, attempted, or imagined) [2], [3], [4], [5]. To generate input for a closed

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

loop device, the user's movement intentions (real, attempted, or imagined) must be detected from the cortical signals within a short latency. Motion preparation detection is useful whenever control of a device or avatar is desired, e.g., in a rehabilitation program. Therefore, developing a model with high movement recognition accuracy is crucial to ensure smooth and effective control techniques depending on the estimation of the user's movement. Movement-related cortical potentials (MRCPs) are associated with both executed and imagined motor tasks and reflect the preparatory processes directly related to motor execution [6], [7]. MRCPs are slow EEG changes that occur between 1.5 to 2 seconds before actual movement onset and are correlated with movement planning and execution. Many researchers have attempted to predict limb activity using MRCPs and sensorimotor rhythms (SMR). G. R. Muller-Putz et al. [8], studied event-triggered EEG changes in paraplegic patients to determine the intention of foot movement. Researchers recently discovered that noninvasive EEG could decode lower extremity movements. This study also demonstrated the feasibility of an EEG-based BMI that could help paralyzed people regain mobility. To discriminate between different types of brain activities, T. Noda et al. [9], devised a new classification technique. They use the covariance matrices of the captured EEG signals as decoder inputs. In their study, EEG signals were used to detect the subject's walking intention and allow them to control the movements of the exoskeleton. In addition, fatigue and effort levels were constantly tracked. Finally, the gait motion state was decoded using a classification paradigm based on Sparse Discriminant Analysis (SDA). The groups of healthy and disabled subjects had decoding accuracies of 84.44 $\pm$ 14.56 % and 77.61 $\pm$ 14.72 %, respectively. Another approach based on Spiking Neurons was applied to classify the motor imagery tasks (rest, left hand, right hand, foot and tongue movements) from EEG signals [10]. Besides, M. Antelis et al. [11], developed a dendrite morphological neural network (DMNN) to classify the voluntary movements during motor execution and motor imagery tasks using the EEG signals. The results depicted that the DMNN obtained 80% decoding accuracy for the motor execution and 77% for imagery.On the other hand, EEG signals can be integrated with other brain waves, such as function near infrared fNIRS, to enhance the recognition accuracy [12]. for instance, M. Khan et al. [13], proposed a model based on EEG and fNIRS to classify the finger tipping task and the mean accuracy of the proposed method was 86.0%.

DL models have recently achieved considerable success in image, video, speech, and text recognition tasks, with numerous studies demonstrating their potential applications [2], [14], [15]. Moreover, in BMI applications, DL algorithms can facilitate the creation of more sophisticated analytic systems than conventional machine learning techniques. Over the previous several years, numerous researchers have used deep learning to create more sophisticated BMI systems and have achieved impressive results [16], [17]. In addition, DL has been used in standard EEG-based BMI systems such as P300, along with concepts such as steady-state visual evoked potentials (SSVEP), motor imagery (MI), and passive BMI applications such as workload and emotion recognition [18].

The motor areas responsible for lower limb movements in an adult human are somatotopically located nearby (right leg, left leg, and foot) [19]. The mesial surface of both brain hemispheres generates ipsilateral potentials for foot movement that overlap at the midline region and are usually deep enough to be classified at the surface [20]. Therefore, the classification of different lower limb movements is challenging with the current level of noninvasive technology [21]. Since the motor regions that enable the foot and knee movements are adjacent, the challenge is more remarkable for lower limbs than for upper limbs. Finally, the foot area in the motor cortex is located near the central area. Another factor that should be considered is that most previous works using EEG signals to decode lower and upper limb movements focused on differentiating between the limbs, i.e., left and right arm or leg. Only a limited number of studies have investigated intra-limb movements based on EEG signals.

To this end, we use standard ViT architecture with a custom configuration of hyper-parameters to fit our purpose. Standard ViT represents an early implementation of attention-based methods for visual tasks. It provided irrefutable evidence that architectures of this type can successfully process images with accuracy comparable to top CNNs on the classification task. Starting from this foundation, we enhance it by modifying the architecture and adding a residual connection from the embedding layer so that the model can capture more information about the image. A vision Transformer was used to classify the image obtained from the EEG signals. As far as we know, this is the first study to deploy a Transformer model for such a task. Finally, we used the so-called Twins model, which contains two architectures. The first combines a Pyramid Vision Transformer (PVT) and conditional positional embedding. The second is the Twins-SVT model, which is based on a spatially-separable Vision Transformer which can consistently provide a good trade-off between computational demands and the accuracy of predictions. This is a consequence of the altered attention mechanism, which can better suit the nature of visual tasks. In addition, we fine-tuned an intense pre-trained model (ResNet150) and evaluated its results against the two transformer models. Therefore, this work developed an approach based on DL to decode the user's motor preparation for lower limb movements. The models detect the intention of the ankle joint movements, i.e. dorsiflexion and plantar flexion. These movements are crucial for maintaining basic walking positions and postures, so the correct detection of intent can be crucial to interpreting signals acquired during lower limb motor preparation.

II. RELATED WORK

Despite the use of DL in EEG-based BMI systems such as (P300, steady-state visual evoked potentials (SSVEP), motor imagery (MI), and passive BCI (for emotion and workload

recognition)), motor preparatory EEG signals have rarely been used in DL. Moreover, most of the related work on upper limb movements focuses on lower limb movements. Recognizing motor preparation is helpful whenever the goal is controlling peripheral devices such as assistive rehabilitation robotics. The following paragraphs summarize state of the art in the motor imagination and DL technique. The essential aim of previous studies is to employ the different DL algorithms to boost the detection and recognition of brain waves during different tasks.

Y. R. Tabar et al. [22], Investigated a convolution neural network (CNN) and stacked auto-encoders (SAE) to classify EEG motor imagery signals during left- and right-hand movements. Combined features, including time, frequency and spatial information, were extracted from the EEG signal. The integrated features were classified using the combination of CNN and SA. The outcomes revealed the improved performance of the classification accuracy. For the same task, Z. Tang et al. [23], proposed a CNN model to perform feature extraction and classification for a single trial motor imagery EEG. Then the authors compared their outcomes with three machine learning approaches with different feature extraction methods, including; AR, CSP and power with SVM. The result depicted that the combination of spatial-temporal with CNN outperformed the other conventional techniques.

J. Yang et al. [24], presented a deep fusion feature learning based on LSTM and CNN to overcome the problem of the conventional deep learning networks to generate Spatio-temporal representation concurrently and the dynamic correlation for the motor imagery signal. Moreover, they applied discrete wavelet transformation decomposition to obtain the spectral information of the EEG signals. J. Xue et al. [25], proposed a feature extraction method by implementing a multifrequency brain network with CSP from the EEG signal during the motor imagery movements. Then CNN model was developed to classify the MI task. On the other hand, a study reported by I. Majidov et al. [26] incorporates two feature extraction techniques, including CSP and Riemannian geometry feature extraction to extract the features from recorded EEG signal during the left- and right-hand imagination movements. Furthermore, a feature selection algorithm was employed to remove the redundant feature based on the particle swarm optimization method. Thereafter, the processed data were fed to the CNN to map the two imagined movements.

The graph-based hierarchical attention model (G- HAM) was introduced by D. Zhang et al. [27], and uses a graph structure to characterize the spatial information of EEG signals and a hierarchical attention mechanism to focus on both the most discriminative time periods and EEG channels. Using time series of EEG signal, G. Zhang et al. [28], proposed LSTM with an attention mechanism to decode the actual movements of the left and right hand. The authors conducted two classification schemes, including intra-subject and cross-subject. Few studies have attempted to identify EEG-based intentional movement before movement execution compared with the number of studies on EEG obtained during movement execution or movement imagination. N. Mammone et al. [18], investigated the motor planning activity based on EEG signals to decode motor preparation phases. Data were collected from 61 EEG channels during unilateral arm movements such as elbow flexion/extension, forearm pronation/supination, and hand open/close. The authors implemented 21 binary classifications, 15 for pre-movements vs another pre-movements epoch and 6 for those related to premovement vs rest epochs. The proposed approach generates a time-frequency (TF) map of each source signal in the motor cortex for each epoch using beamforming and Continuous Wavelet Transform (CWT), then embeds all maps in a volume and feeds them into a Deep CNN. The suggested approach achieved an average accuracy of 90.3 % in distinguishing pre-movement from resting and 62.47 % in distinguishing pre-movement vs pre-movement. Although CNNs have been used widely in almost all previous work, Transformer models that rely on self-attention to track long-distance relationships have been used in Natural Language Processing with impressive results. In recent years, this architectural blueprint has been regarded as a worthy replacement for convolution-based models even in the field of computer vision, mainly for tasks like image classification, image creation and enhancement, object detection, scene segmentation, video processing and 3D processing [29], [30]. What makes the Transformer superior to other architectures, such as CNN or LSTM, is that it doesn't rely on inductive bias and instead uses global attention to interpret long-distance connections. Meanwhile, local weights can be dynamically aggregated based on the relationships discovered between tokens belonging to the same local window, which is the opposite approach to the one employed by CNN, which relies on fixed weights for spatially proximate pixels. On the other hand, adaptive weight aggregation allows the networks to perform better with tasks that require recognition [31].

III. MATERIALS AND METHODS

Fig 1 demonstrates the general framework of this work, and the pipeline started with data collection and experimental setup using EEG signals. Then the signals underwent prepossessing to remove the unwanted signals. MRCP was evaluated from processed EEG signals to investigate and extract the motor preparation period. The time-frequency was evaluated using continuous wavelet transform CWT. Furthermore, the movement onset detection was evaluated using an EMG signal. Next, the transformed EEG signal is passed to the deep learning structure for recognition. A paired t-test was used in this work to evaluate the significance level of the proposed modalities.

A. EXPERIMENTAL SETUP

This study included twenty healthy right-handed participants (aged 27.9 ± 2.9 years). This study was endorsed by Monash University Human Research Ethics Committee (MUHREC).

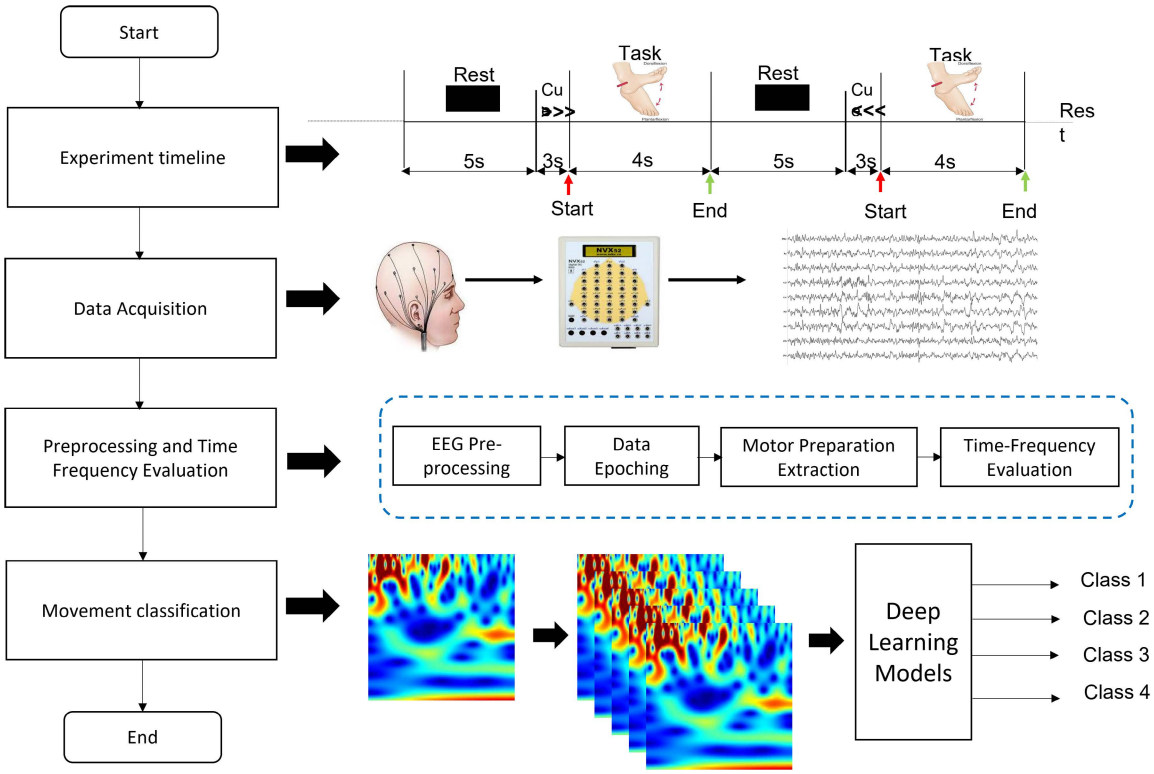

FIGURE 1. General framework of the proposed technique.

The experiment was carried out in compliance with the Helsinki Declarations, and all volunteers provided written informed consent. Before data collection, the procedure of the experiment was explained to the participants. The focus of this study was on two movements of the ankle joint. These are dorsiflexion and plantarflexion (DF and PF). The DF is a movement that minimizes the angle between the foot and the shank. The PF, on the other hand, is a movement that maximizes the angle between the shank and the foot, as if the foot were pressing on the gas pedal of a car. These movements were selected because they are essential for maintaining proper walking position and posture [32]. The tibialis anterior and gastrocnemius lateralis muscles were selected for this study because they play a dominant role in the execution of dorsiflexion and plantar flexion. To ensure maximum range of motion of the ankle joints, each subject sat in a comfortable chair with the legs not touching the floor, as described in Fig 2.

To provide visual guidance for the movement task, the monitor was placed approximately 1 meter in front of the patient. The task proceeded as follows: the subject was asked to move the ankle joint dorsiflexion and maintain the contraction for three seconds, then repeat the same movements until the number of trials, T = 30, was reached. Between each trial, there was a time of rest, and the plantar flexion movement of the ankle joint was performed in the same way. This experiment was conducted for the right ankle joint.

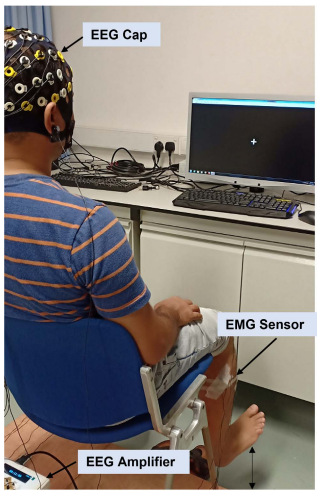

FIGURE 2. Subject position during the data acquisition process.

B. EEG DATA ACQUISITION

EEG signals were recorded from 21 channels (Fz, FC3, FC1, FCz, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, Cp5, Cp3, Cp1, Cpz, Cp2, Cp4, Cp6, Pz) using Ag/AgCl electrodes and MCScap; all channels were positioned according to the international 10-10 standard. The ground electrode was placed between Fz and Fpz, and the reference electrode was placed on the left and right earlobes. The signal was amplified and sampled at 2 kHz using NVX52 (MKS cooperation Inc,

Russia). Before recording the EEG signal, some measures were taken, such as checking the placement of the EEG cap and ensuring that the electrode impedance was less than 5K ohms, which can be achieved by placing a conductive electrogel between the EEG electrode and the scalp. The EEG signals were first filtered with a finite impulse response (FIR) bandpass filter (0.05-40Hz). After that, a segmented data stream was developed (with a duration of 7s: 4s prior and 3s after the movement's onset). Segmented data were subjected to independent component analysis (ICA) to remove visible artefacts such as eye movements, heart signals, and muscle contractions. These artefacts were removed from the ICA components. Next, the remaining components were projected back to build EEG signal-free form artifacts. To detect the movement's onset, EMG signals were recorded from two shank muscles, Tibialis Anterior TA and Gastrocnemius Lateralis. Surface EMG for Non-Invasive Muscle Evaluation (SENIAM, seniam.org) guidelines were used to position the EMG electrodes. The muscle belly was palpated to determine the best location for the electrode, which was then placed along the main fibre course [33]; moreover, the subjects were encouraged to perform maximal voluntary contractions to validate the positioning. For more details on the data collection procedure, readers were invited to refer to our previous work [34].

C. MOTOR RELATED CORTICAL POTENTIALS

MRCPs are defined as a slow negative potential recorded in EEG prior to the movement execution. MRCPs have been categorized into two segments: the first segment is the readiness potential (RP) begins 1.5 to 1 s before the movement onset and was observed throughout the whole pre-supplementary motor area. The second segment is the motor potential (MP) associated with movement execution. To MRCP, the processed EEG signals were filtered using a second-order Butterworth filter at (0.5,4) Hz. The EEG signal was epoched into a 6s long segment from -4 to 2 s concerning the movement onset.

D. TIME FREQUENCY REPRESENTATION

The processed EEG signals were represented in the time-frequency domain by evaluating the CWT. The Morelet wavelet was employed as a wavelet mother function in this work. The minimum and maximum frequencies for the complex Morelet wavelet convolution were set to 0.5 Hz and 40 Hz, respectively. The wavelet cycle was set to 5 cycles, and the number of frequencies was set to 30. After evaluating the TF map using CWT, the TF maps were converted to RGB images to feed them to the DL architecture. The wavelet mother used allowed us to span the target range under investigation (0.5–40 Hz) with high resolution. This range contains the five primary brain waves of general interest (delta, theta, alpha, beta, and gamma), including MRCP and SMR, which are significant to movement analysis.

E. DEEP LEARNING MODEL ARCHITECTURE

Our proposed deep learning model is created based on vision transformer architecture [35], which is the state-of-the-art of image classification. Transformer architecture was first introduced in the seminal work of Vaswani et al. [36], and has since been applied to many different problems such as EEG person identification [37], seizure prediction [38], hand movement recognition based on Electromyography signals (EMG) [39] and Visual stimulus classification [40], however, it has been implemented mostly in the Natural Language Processing field. Its main advantage is that relationships between every two tokens in a sequence can be tracked and analyzed. By analogy, in image recognition, the model would have to track relationships between every two pixels in an image, which is extremely computationally demanding. In this section, three modified ViT, as described in Fig.3 were employed and the outcomes of these three models were compared with the ResNet model.

1) VISION TRANSFORMER

A new deep learning tool takes a known model and adjusts it to a different input, using visual information instead of text. The original Transformer model was taken as the foundation of the new solution in an almost identical form, which saved a lot of effort on model design. However, the model was altered to process two-dimensional sequences while retaining its ability to analyze contextual relations. The new model was named ViT (short for Visual Transformer), and its primary purpose is to correctly classify visual images, similarly to how language models can solve linguistic problems. The transformation of input into a 2D sequence corresponding to a patch of pixels from the image is performed by reshaping the images as follows: for each image: $x \in \mathbb{R}^{H\times W\times C}$ will be in the shape of $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$ with the resolution of the original image (H, W), the resolution of the isolated patch (P, P), the number of channels C and the effective length of the input sequence $N = \frac{H\times W}{P^2}$ taken into account. In order to create linear projections that are suitable for model training, all patches were flattened to D dimensions, where D is the constant latent vector size, and in terms of language, the model is 768 lengths. Patch embedding, which is formed in this way, is fed into the algorithm during the training stage.

The Transformer model consists of a stack of layers with attached multi-head attention mechanisms, with the output of one layer serving as the input for the next one. Ultimately, information is passed along with the attention weights to a classification layer where the decision about the particular patch is made. In our model, we did not use the original idea mentioned in ViT as illustrated in Fig 3a. Still, instead, we used another idea found in the field of Natural Language Processing called "Residual Attention Layer Transformer" or RealFormer in short form. This model is almost identical to the original Transformer and consists of a stack of encoder-decoder layers. The difference here is that it uses residual multi-head instead of the standard multi-head, as shown in

Fig 3b. Each layer contains a residual multi-head attention mechanism and calculates attention scores that are passed to the next layer. The output of all attention heads is concatenated and linearly projected into an attention matrix, which includes raw attention scores for all patches. In the standard Transformer, the multi-head self-attention (MSA) can be calculated by the following Equation 1:

$$MSA(Q, K, V) = \text{concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h) \times W^o \quad (1)$$

where headi is calculated as

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$$

the Q, K and V are the three values that produce the attention score. Both Q and K have the dimension d_k whereas V has a dimension d_v. These three matrices can be gotten by the following Equations 2,3,4.

$$Q_i = x_i \times W^Q \quad (2)$$
$$K_i = x_i \times W^K \quad (3)$$
$$V_i = x_i \times W^V \quad (4)$$

where W is the weighted matrix that projects the attention parameters into new space to extract the essential features from each patch. Next, the attention score of the Q and K are normalized by the Softmax function and then passed to a scaled dot-product operation along with the V to produce the final attention score as shown in the following Equation 5:

$$\text{Attention}(Q_i, K_i, V_i) = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \cdot V_i \quad (5)$$

In our model, the normalization is performed before Layer Norms are inserted, but with added skip edges that create a connection between the attention mechanisms in layers positioned next to each other as follows.

$$\text{Residual Multi Head}(Q, K, V, \text{Prev}) = \text{concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h) \times W^0 \quad (6)$$

where $head_i$ is calculated as

$$\text{head}_i = \text{ResidualAttention}(Q_i, K_i, V_i, \text{Prev}_i)$$

This is accomplished by sending an additional piece of data, the attention score before Softmax activation, to the attention heads in the current layer. This input parameter is described as residual attention, and its weighted sum is calculated using the same formula as for the normal attention scores with a slight differences as follows:

$$\text{ResidualAttention}(Q_i, K_i, V_i, \text{Prev}_i) = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}} + \text{Prev}_i\right) \cdot V_i \quad (7)$$

where $Prev_i$ takes the shape of (N, N). The point of this procedure is to create a direct connection between attention modules in separate layers, thus strengthening the predictive capacities of the entire model. The classification head that serves for this purpose is added to a multilayer perceptron (MLP) with a hidden layer in the pre-training stage and a linear layer in the training stage. The MLP consists of two layers and can be described as follows:

$$z_0 = \left[x_{\text{class}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] \times \text{where } E \in \mathbb{R}^{(P^2 \cdot C) \times D} \quad (8)$$

And by adding the positional embedding:

$$z_0 = \left[x_{\text{class}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{pos} \text{ where } E_{pos} \in \mathbb{R}^{(N+1) \times D}$$
$$z'_l = LN(\text{ResidualMultiHead}(z_{l-1})) + z_{l-1} \quad (9)$$

where l is the number of layers and it can be represented as $l = 1 \ldots L$and z_l is the output of layer l.

$$z_l = LN\left(MLP\left(z'_l\right)\right) + z'_l$$
$$y = LN\left(z_L^0\right) \quad (10)$$

The output y of the encoder works as image representation such as a sentence contextualized embedding.

2) MODIFIED VISION TRANSFORMER (TWINS)

Vision transformers are highly effective tools for completing many complex image analysis tasks, but they are inherently computationally demanding and difficult to implement. The main reason for the complexity is the way the self-attention mechanism is constantly re-calculated and, in particular, how the algorithm handles the spatial division of the image. Improvement of this procedure would make the visual transformer architecture more cost-efficient, which can dramatically impact the practical value of this type of deep learning network. The authors are aware of the previous attempts to simplify self-attention by introducing sub-sampling, and they expand this idea by adding some innovative elements [30]. They propose a spatial redesign of the self-attention mechanism where sub-samples are grouped, and positional encoding is deployed. Depending on the grouping criterion, they develop two algorithms – one with locally-grouped self-attention and another with global grouping. The globally based model was named Twins- Pyramid Vision transformer based on condition position encoding (Twins-PCPVT), and it inserts conditional positional encoding generators (PEG) after the first encoder block as illustrated in Fig 3c. The locally based model was named Twins-SVT, which aims to reduce complexity by creating sub-samples that can be analyzed separately. The authors introduce the concept of spatially separable self-attention attention, which is better suited for visual tasks and consists of globally sub-sampled attention and locally grouped attention. In other words, a 2D feature map is first divided into several local windows in which self-attention can be calculated easily but with low generality. To address this issue, the sub-sampling function based on separable convolutions is introduced, serving to summarize

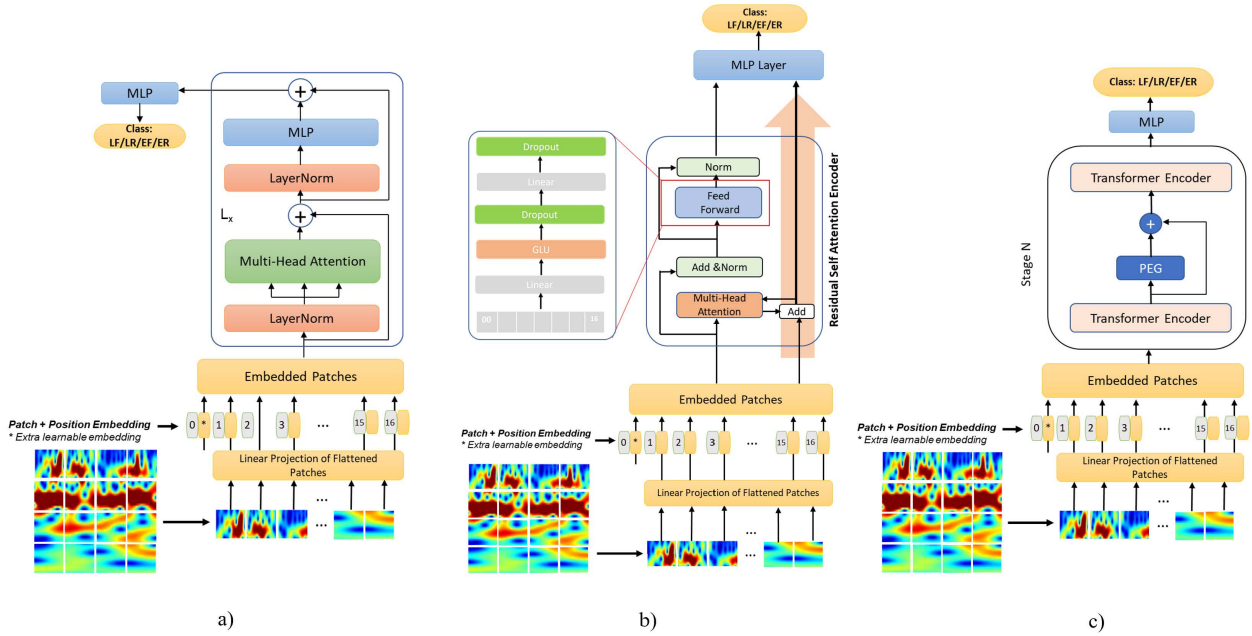

FIGURE 3. Vision Transformer models architecture a) ViT b) ResViT c) TWINS.

the information within windows and provide communication across those windows. This reduces the complexity of the model to the following:

$$O(mnHWd) = O((H2W2d)/k1k1 + k1k1HWd)$$

Both described models are simple to implement and more computationally friendly than alternative configurations of visual transformer architecture. Their parameters, such as the number of layers and hidden dimensions, number of heads, expansion ratio etc. are intentionally different in order to study the impact of such factors on model performance.

3) DEEP RESIDUAL LEARNING

The number of stacked layers within the network architecture, often referred to as network depth, impacts the ability of machine learning systems to complete various tasks (including image recognition) with a high level of accuracy, with deeper networks producing better results in general. However, training and optimization of deep networks are associated with several problems, namely vanishing gradients and decaying training efficiency. This significantly reduces the practical applicability of such systems and makes their pre-training more computationally expensive and time-consuming.

In response to the common difficulties with the optimization of deep architectures, the authors propose a solution based on the inclusion of a shallower model within the deep network. Thanks to the integration of features across the network and residual learning, the efficiency of the deep network can be increased in this scenario. To accomplish that, the authors introduced shortcut connections, which can be used for identity mapping. In this way, a building block consisting of several layers can be constructed using the formula:

$$Y = F(x, \{W_i\} + x) \tag{11}$$

where x and Y are the output vectors for the included layers, and the function represents residual learning. This formula provides a way to transfer inputs to additional layers through residual learning without introducing new parameters that could complicate training. If the dimensions of layers within a building block are not identical, linear projection is used to equalize them. In terms of architecture, the authors started from a convolutional network design with a global pooling layer and fully connected layer, then inserted the shortcuts and attached them to filters to create the residual network. The resulting network has fewer total filters and thus reduces computational complexity to only a fraction of that of a plain network with a comparable number of layers. Scale augmentation and color augmentation procedures were performed with the images before they were used for training, with batch normalization implemented after every convolution. Identity mapping doesn't require input padding when the layer dimensions are increased, contributing to its efficiency.

The hypothesis that identity mapping with shortcut connections can reduce the size of training error was examined by testing several different variations of deep learning networks on publicly available data sets containing a large number of images. The impact of the network depth was reversed when residual learning was introduced, with training error being lower for a deeper architecture rather than higher as with a plain network. The same trend was observed when looking at top-1 and top-5 image classification, with the proposed method outperforming several state-of-the-art models. Model accuracy was even higher when the number of layers

inside each building block was increased from 2 to 3. This configuration could reduce the top-5 error to 3.47% while still requiring less computational power than the benchmark models of the same depth. Those results confirm that residual learning alleviates some known issues with deep learning network training.

4) TRAINING STAGE

We trained our models on the large size of our dataset of 28 subjects. The settings of the hyperparameters are described in Table 1. During the training stage, the visual information from the images is transformed into linear embeddings through a process that involves isolating patches and connecting a series of such patches into input sequences. Those sequences are processed by the transformer model in the same way as token sequences in NLP. After attention weights are calculated for each patch, it becomes possible to calculate the 'attention distance' between various image elements. Meanwhile, low-level representations within each patch are preserved. Thus, the model can effectively learn about the global distribution of similarities within the sequence and capture latent connections between distant tokens. Due to the fact that all self-attention layers are global, the proposed model displays far lower levels of image-specific inductive bias than any other neural network model, such as CNN or RNN. This occurs because the starting positions are blindly chosen during initialization and all spatial relations have to be learned based on input processing. The settings of the hyper-parameters During the training stage of the twins model are described in Table2.

TABLE 1. Visual transformer hyperparameters.

Parameter	Value
Learning rate	$3e^{-5}$
Dropout pro	.1
Batch size	32
Number of epochs	50
Hidden layer size	128
No. heads	8
Depth	12
Gamma	0.7
Seed	42
Image size	224
Patch size	16
Num classes	4
Optimizer	Adam
Loss	Cross Entropy Loss

IV. RESULTS

A. RESULT OF MRCP ANALYSIS

Fig 4. shows the average MRCP from −4 to 2 s with respect to the movement onset at the Cz area during the PF and DF movements. In both movements, the negative deflection was observed before the movement's onset and peaked immediately after the onset of the movement where the actual movement started. During the PF movement, the MRCP peaked at 0.16 s after the onset of the movement with a maximum

TABLE 2. Twins transformer hyper-parameters.

Parameter	Value
Stage 1:	
embedded dimension	64
patch size	4
local patch size	7
global attention key	7
depth	1
Stage 2:	
embedded dimension	128
patch size	2
local patch size	7
global attention key	7
depth	1
Stage3:	
embedded dimension	256
patch size	2
local patch size	7
global attention key	7
depth	5
Stage 3:	
embedded dimension	512
patch size	2
local patch size	7
global attention key	7
depth	4

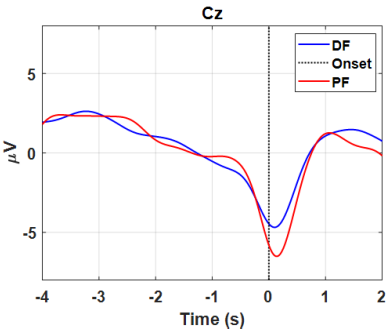

FIGURE 4. Average MRCPs at the Cz area during DF and PF movements.

peak (MP) −6.47 μV. While during the DF movements, the negative deflection peaked at 0.12 s with −4.68 μV MP. On the other hand, Fig 5. illustrates the average MRCP for the EEG electrodes at the supplementary motor area SMA and dorsal primary motor area PMAdr during the DF movement of the right ankle joint. The results show the large negative deflection in the midline region (Cpz, Cz, and FCz). Since RP could represent the motor preparation time, only this interval was considered here. It can be seen that RP in Cz and C1 starts 2 seconds before the onset of the movement. However, in other channels such as FCz and FC1, it occurs later and starts about 1.5 to 1 second before movement onset. Therefore, the duration of 1.5s before the start of the motion was chosen for the TF mapping and classification phase.

B. TIME FREQUENCY ANALYSIS RESULTS

The average TF plot of the EEG channels on the right and left motor cortex areas during the movement of the right ankle is shown in Fig 6. The yellow color represents an increase

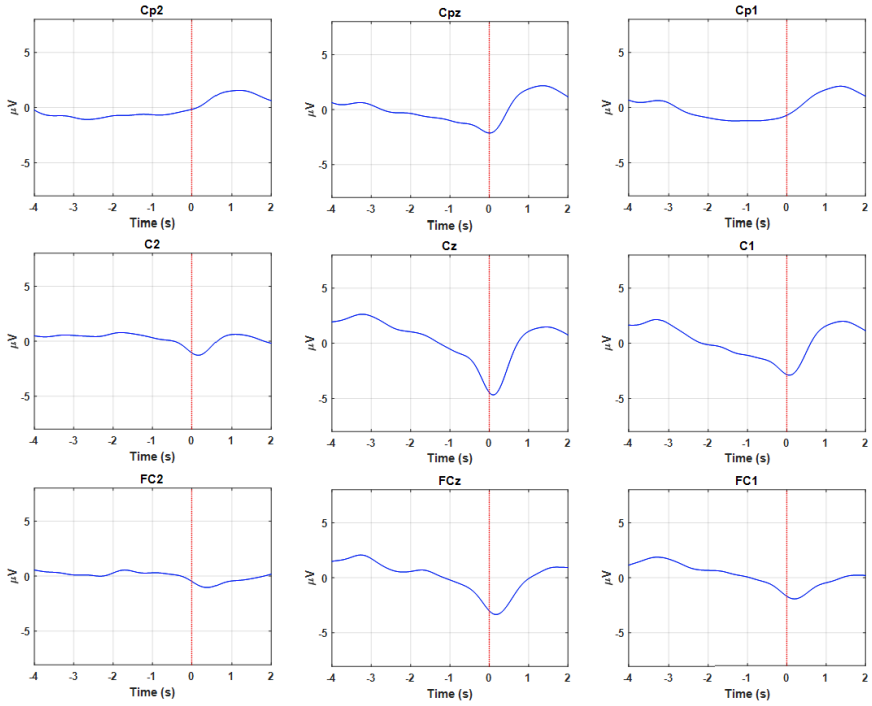

FIGURE 5. Average MRCPs for the EEG electrodes over the motor cortex area.

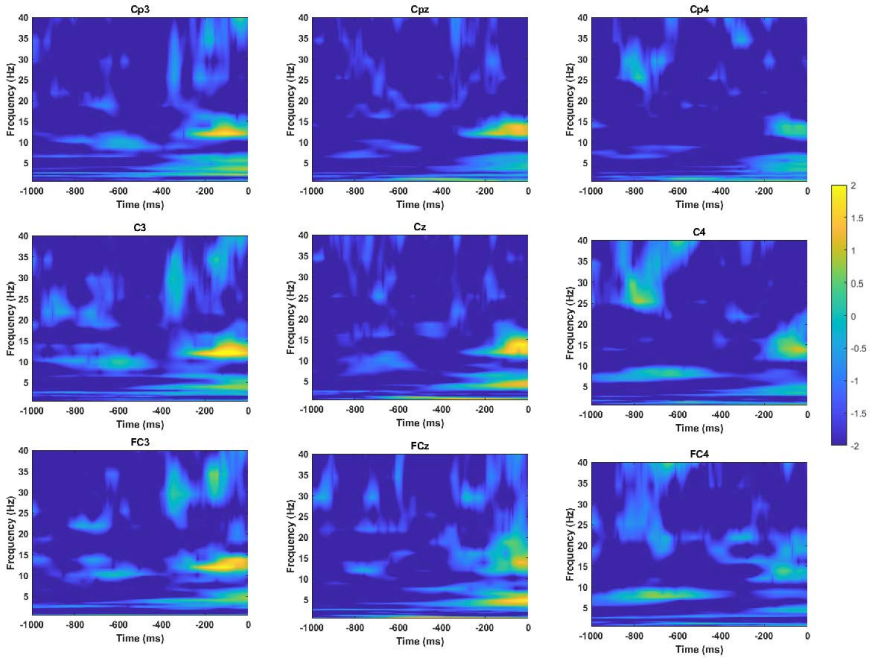

FIGURE 6. Time-Frequency mapping for the motor cortex electrodes during the DF movement for Subject 5.

of power in the delta and lower beta bands (1- 4)Hz and (12 - 18) Hz respectively, which is known as event-related synchronization ERS. Additionally, Fig 6. shows that the ERS is most pronounced in the left primary motor cortex (Cp3, C3 and FC3) and central line represented by (Cpz, Cz and FCz).

C. CLASSIFICATION RESULTS

In this stage the interval of 1s proceeding the movements onset was extracted. This is because this duration reflects the motor preparation. As investigated before using MRCP signals the RP is more prominent at 1s prior the movements.

TABLE 3. Classification performance results using ViT.

Subjects	F1_Score	Precision	Recall	Accuracy
1	0.8972	0.8963	0.8985	0.8988
2	0.9325	0.9327	0.9325	0.9328
4	0.8397	0.8391	0.8413	0.8416
5	0.9301	0.9306	0.9302	0.93
6	0.8476	0.8477	0.8483	0.8478
7	0.9329	0.9333	0.9330	0.932
8	0.9155	0.9157	0.9157	0.9169
9	0.6082	0.6146	0.6054	0.6031
10	0.9678	0.9689	0.9677	0.9683
11	0.9177	0.9189	0.9175	0.9179
12	0.9468	0.9473	0.9469	0.9470
13	0.9052	0.9054	0.9055	0.9047
14	0.9108	0.9130	0.9102	0.9100
15	0.9207	0.9213	0.9207	0.9206
16	0.9365	0.9375	0.9374	0.9365
17	0.8905	0.8899	0.8918	0.8904
18	0.9311	0.9320	0.9310	0.9312
19	0.9130	0.9136	0.9136	0.9126
20	0.9084	0.9101	0.9074	0.9076
21	0.9397	0.9397	0.9400	0.9401
22	0.9445	0.9445	0.9446	0.9444
23	0.9344	0.9330	0.9365	0.9345
24	0.8631	0.8632	0.8636	0.8654
25	0.9171	0.9185	0.9170	0.9188
26	0.9127	0.9134	0.9128	0.9126
27	0.8182	0.8183	0.8184	0.8205
28	0.9394	0.9413	0.9383	0.9393
Average	**0.8980**	**0.8988**	**0.8983**	**0.8982**
Std	**0.0669**	**0.0661**	**0.0673**	**0.0675**

TABLE 4. Classification performance results using ResViT.

Subject	F1	precision	recall	accuracy
1	0.939	0.9403	0.9403	0.9387
2	0.968	0.9679	0.9679	0.9689
3	0.8811	0.8821	0.8821	0.8821
4	0.9284	0.9294	0.9294	0.93
5	0.913	0.9135	0.9135	0.9148
6	0.9509	0.9518	0.9518	0.9501
7	0.8958	0.8971	0.8971	0.8965
8	0.9445	0.9452	0.9452	0.9441
9	0.9728	0.9729	0.9729	0.9733
10	0.9421	0.9432	0.9432	0.9414
11	0.9631	0.9636	0.9636	0.9634
12	0.9128	0.9145	0.9145	0.9125
13	0.9638	0.9642	0.9642	0.9635
14	0.9333	0.9341	0.9341	0.933
15	0.9634	0.9636	0.9636	0.9637
16	0.9439	0.9433	0.9433	0.9447
17	0.9546	0.9552	0.9552	0.9548
18	0.9628	0.9631	0.9631	0.963
19	0.9176	0.9174	0.9174	0.9181
20	0.9685	0.9696	0.9696	0.9681
21	0.9552	0.9552	0.9552	0.9554
22	0.9171	0.9172	0.9172	0.9178
23	0.8811	0.883	0.883	0.8805
24	0.9335	0.9345	0.9345	0.9348
25	0.918	0.9233	0.9233	0.9193
26	0.8703	0.8705	0.8705	0.8706
27	0.9518	0.9522	0.9522	0.9518
28	0.8929	0.9022	0.9022	0.8939
Average	**0.9335**	**0.9346**	**0.9346**	**0.9339**
Std	**0.029**	**0.0284**	**0.0284**	**0.0289**

TABLE 5. Classification performance results using ResNet.

Subject	F1_Score	Precision	Recall	Accuracy
1	0.9685	0.9683	0.9687	0.9682
2	0.9768	0.9764	0.9763	0.9762
3	0.9484	0.9484	0.9484	0.9486
4	0.9407	0.9435	0.9408	0.942
5	0.9450	0.9476	0.9443	0.9446
6	0.9778	0.9774	0.9789	0.978
7	0.9523	0.9522	0.9525	0.9525
8	0.9622	0.9627	0.9628	0.9623
9	0.9960	0.9957	0.9963	0.9960
10	0.9726	0.9724	0.9729	0.9722
11	0.9891	0.9894	0.9889	0.9900
12	0.9506	0.9514	0.9503	0.9503
13	0.9467	0.9492	0.9460	0.9464
14	0.9397	0.9457	0.9389	0.9365
15	0.9922	0.9920	0.9925	0.9920
16	0.9772	0.9775	0.9770	0.9774
17	0.9801	0.9802	0.9803	0.9801
18	0.9487	0.9490	0.9491	0.9484
19	0.9661	0.9654	0.9673	0.9664
20	0.9901	0.9903	0.9901	0.9902
21	0.9879	0.9879	0.9881	0.9880
22	0.9702	0.9708	0.9699	0.9705
23	0.9341	0.9387	0.9337	0.9347
24	0.9507	0.9510	0.9521	0.9509
25	0.9342	0.9354	0.9339	0.9345
26	0.9297	0.9314	0.9300	0.9308
27	0.9818	0.9816	0.9825	0.9822
28	0.8804	0.8832	0.8798	0.8799
Average	**0.9603**	**0.9612**	**0.9604**	**0.9603**
Std	**0.0248**	**0.0238**	**0.0250**	**0.0249**

Therefore, the TF maps were evaluated over the motor cortex area and the resultant scalogram diagram for each 1s epoch was converted to RGB image. Then these converted images were fed as an input to the proposed DL model. Furthermore, Three ViT models were evaluated to assess the powerful of the proposed deep learning method and the results of these models also compared with ResNet model. Among the three ViTs model, TWINS showed the highest classification performance metrics. Moreover, the TWINS also outperforms the ResNet model. Table 4 to Table 9 depicts the individual classification performance measures for the proposed ViTs models in addition to the ResNet technique. Amonge the DL models,TWINS technique reveals a significant classification performance measures for the multi-class problem including RDF, LDF, RPF and LPF. Where the classification accuracy is 97.33 $\pm$ 1.86 % and the F-score 97.32 $\pm$ 1.88 %.

To demonstrate the significant of the classification performance measures improvements of the motor preparation during lower limb movement, paired t-test was employed between ViT and other machine learning classifiers, Table 3 illustrates the different p values. Where the classification performance measures using the proposed ViT method was compared with the other DL models as shown in Fig 7 and Fig 8.

V. DISCUSSION

EEG signals are commonly used to interpret motoric actions and predict movement, but they are poorly suited to facilitate recognition of lower limb movement. For this reason, this study explored whether the inclusion of EMG data and fusion

TABLE 6. Classification performance results using TWINS.

Subject	F1_Score	Precision	Recall	Accuracy
1	0.9733	0.9732	0.9736	0.9735
2	0.9871	0.9863	0.9882	0.9868
3	0.9705	0.9714	0.9700	0.9709
4	0.9892	0.9888	0.9897	0.9893
5	0.9593	0.9598	0.9590	0.9604
6	0.9845	0.9846	0.9847	0.984
7	0.9811	0.9811	0.9812	0.9815
8	0.9712	0.9712	0.9713	0.9708
9	0.9975	0.9977	0.9972	0.9973
10	0.9737	0.9743	0.9733	0.9735
11	0.9972	0.9972	0.9972	0.9973
12	0.9683	0.9682	0.9686	0.9682
13	0.9788	0.9790	0.9787	0.9788
14	0.9600	0.9603	0.9600	0.9603
15	0.9923	0.9923	0.9923	0.9920
16	0.9786	0.9779	0.9779	0.9780
17	0.9788	0.9791	0.9786	0.9788
18	0.9603	0.9603	0.9608	0.9603
19	0.9766	0.9771	0.9764	0.9762
20	0.9947	0.9947	0.9949	0.9947
21	0.9844	0.9845	0.9844	0.9841
22	0.9787	0.9788	0.9787	0.9790
23	0.9251	0.9312	0.9230	0.9261
24	0.9510	0.9513	0.9513	0.9528
25	0.9788	0.9795	0.9784	0.9788
26	0.9579	0.9605	0.9561	0.9577
27	0.9843	0.9850	0.9837	0.9841
28	0.9161	0.9168	0.9159	0.9160
Average	**0.9732**	**0.9736**	**0.9730**	**0.9733**
Std	**0.0188**	**0.0181**	**0.0190**	**0.0186**

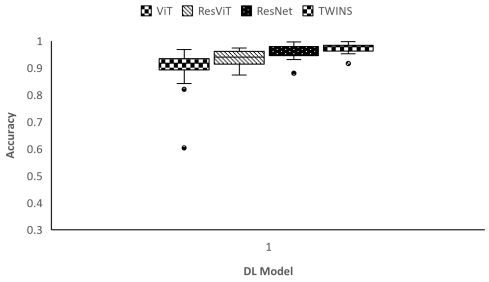

FIGURE 7. Comparison of the classification accuracy.

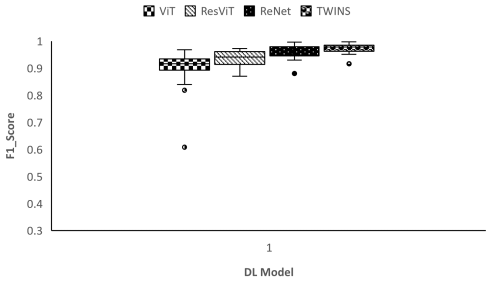

FIGURE 8. Comparison of the F1_Score.

of input from those two sources could improve the prediction accuracy and make the construction of BCI devices for foot rehabilitation possible. This work aims to recognize the motor preparation of the user based on the EEG signal during the movements of the ankle joint. MRCPs represent brain activity

TABLE 7. *p*_value of the classification perfromance measures for the ViT and ResNet models in comparison with TWINS model.

	Classification performance measures			
DL models	F1_Score	Precision	Recall	Accuracy
ViT	4.007E-07	3.572E-05	4.767E-07	4.917E-07
ResNet	4.227E-05	3.572E-05	6.584E-05	3.886E-05

changes associated with movement in the time domain [41]. The Bereitschaftspotential (BP) or readiness potential (RP) represents the motor preparation stage of movement and is thought to be produced by the supplementary motor area (SMA) [42], motor cortex, and cingulate gyrus [43]. The analysis of MRCPs in this work is consistent with this approach, where the negative deflection appeared around 2 s before the movement's onset on the SMA, and the large negative deflection appeared in the Cz area. The motor potential (MP) is a late subcomponent of the MRCPs that is thought to be produced partly by afferents stimulated by movement and by the under-lying motor cortex [2]. For both ankle joint movements, MP during the PF movement is higher than that during the DF, where the MRCP peaked at 0.16 s after the movement's onset with maximum peak (MP) $-6.47\ \mu$V during the PF. While during the DF movements, the negative deflection peaked at 0.12 s with $-4.68\ \mu$V MP. The MRCPs' negativity amplitude can be related to the amount of energy needed for the movement, whereas the MRCPs' onset period is defined as the time spent planning and preparing the movement [44]. Also, it can be noted from the MRCP analysis during the movement execution, the motor cortex area was activated bilaterally. Although there is a negative deflection at the ipsilateral area in C2 and contralateral area C1, the MP of the C1 is higher than that in the C2 area. The MP in the C1 area was $-2.87\ \mu$V, while the MP value at the C2 was $-1.27\ \mu$V; therefore, there is a significant difference in the MP amplitude in both areas ($p < 0.001$). Several studies utilized MRCPs for the movement's intention detection and recognition [45], [46], [47], [48], [49]. According to [50], self-directed grasping movements of the upper limbs can be detected with an accuracy of about 80% using MRCP correlates before the movement. Furthermore, in a recent related study, MRCP features were used to predict foot torque movement on a single trial [50]. Depending on the wavelet and the classification process, they achieve a classification accuracy of about 84.2 %. Recent research, which focused at detecting pre-movement states from MRCP correlations when executing ankle dorsiflexions, shows an 82.5% performance for movement execution [51]. On the other hand, according to the time-frequency mapping and alpha beta ERD data, there is a bilateral control phenomenon in movement execution. Alpha ERS was most pronounced during the movement intention or preparation phase, indicating that brain excitability has a contralateral function in the pre-movement phase [1]. In the current study, alpha oscillations in the brain's central part represent the neural populations' synchronous activities. SMR in alpha and beta oscillation have been utilized in

recent research to detect the movements intention and movements execution, [52], [53], [54], [55]. However, studies that used SMR for movement's detection showed lower detection accuracy compared to that studies used MRCPs as reported in [56]. MRCP and ERD feature varied in lateralization phenomena, but it was apparent that both contralateral and ipsilateral motor cortices were engaged in motor preparation tasks. Not only for bilateral but also unilateral movements, neural networks within and between hemispheres are needed to coordinate motor functions [47].

In addition, In this paper, we have developed some enhancements to the standard ViT by introducing a Residual connection. In addition, we have used ResNet and a recent implementation of Transformer architecture called Twins, where the models were found to deliver more astute predictions than the comparable Transformer variations and the ResNet. The ability to consistently provide a favourable ratio between computational demands and accuracy of predictions confirms that the altered attention mechanism revealed that both conditional positional encoding and the SSSA mechanism could better suit the nature of visual tasks. This approach brings tangible improvements over any existing alternative forms of vision Transformer in model accuracy and training efficiency. All DL models were tested and compared against each other on the image classification task. Our proposed ResidualViT outperformed the standard ViT, with almost 4% higher accuracy; however, its performance was inferior to ResNet. For this reason, we used Twins, which was even more accurate than ResNet on image classification task, with accuracy margins reaching as high as 1.3%; and outperforming standard ViT and ResidualViT by.5% and 3.9%, respectively. Overall, the results indicate that using Twins retains excellent generalization ability and broad contextual awareness, and the number of parameters that must be accounted for is significantly reduced. These encouraging results also highlight the potential of the EEG signals with a deep learning-based ViT approach for accelerating the development of a BMI for movement rehabilitation in the future. Additionally, the developed model might encourage the development of bio-robotics assistive devices that enhance human movement and improve quality of life.

VI. CONCLUDING REMARKS

A. CONCLUSION

Movement recognition based on EEG signals today significantly influences neuroscience research. This work investigated and implemented the motor preparation phase based on EEG signals for lower limb movement recognition. Four movements of the right and left ankle joints were involved in this study, including right and left dorsiflexion and plantar flexion. The time-frequency (TF) map of each EEG signal in the motor cortex is generated using the Continuous Wavelet Transform (CWT). The obtained images are then fed into deep-learning models for classification. The proposed deep learning models are based on the vision transformer architecture (ViT). The findings of this study demonstrate the effectiveness of the deep learning approach based on EEG signals for the development of future BMI for lower limb rehabilitation.

B. FUTURE DIRECTIONS

The proposed approach successfully recognized the actual ankle joint movements. Nevertheless, its real-time ability to classify those movements remains to be tested. Furthermore, more studies should be carried out to cover both actual and imagined movements. Besides, more rigorous research needs to be performed to incorporate the results of this study into clinical practice.

ACKNOWLEDGMENT

The authors would like to thank the sponsors for their support.

REFERENCES

[1] H. Li, G. Huang, Q. Lin, J.-L. Zhao, W.-L.-A. Lo, Y.-R. Mao, L. Chen, Z.-G. Zhang, D.-F. Huang, and L. Li, "Combining movement-related cortical potentials and event-related desynchronization to study movement preparation and execution," *Frontiers Neurol.*, vol. 9, p. 822, Oct. 2018.

[2] B. Gudiño-Mendoza, G. Sanchez-Ante, and J. M. Antelis, "Detecting the intention to move upper limbs from electroencephalographic brain signals," *Comput. Math. Methods Med.*, vol. 2016, pp. 1–11, Apr. 2016.

[3] Y. He, D. Eguren, J. M. Azorín, R. G. Grossman, T. P. Luu, and J. L. Contreras-Vidal, "Brain–machine interfaces for controlling lower-limb powered robotic systems," *J. Neural Eng.*, vol. 15, no. 2, Apr. 2018, Art. no. 021004.

[4] K. Lee, D. Liu, L. Perroud, R. Chavarriaga, and J. D. R. Millán, "A brain-controlled exoskeleton with cascaded event-related desynchronization classifiers," *Robot. Auton. Syst.*, vol. 90, pp. 15–23, Apr. 2017.

[5] R. Ron-Angevin, F. Velasco-Álvarez, Á. Fernández-Rodríguez, A. Díaz-Estrella, M. J. Blanca-Mena, and F. J. Vizcaíno-Martín, "Brain-computer interface application: Auditory serial interface to control a two-class motor-imagery-based wheelchair," *J. NeuroEng. Rehabil.*, vol. 14, no. 1, pp. 1–16, Dec. 2017.

[6] O. F. D. Nascimento, K. D. Nielsen, and M. Voigt, "Movement-related parameters modulate cortical activity during imaginary isometric plantarflexions," *Exp. Brain Res.*, vol. 171, no. 1, pp. 78–90, May 2006.

[7] O. F. D. Nascimento and D. Farina, "Movement-related cortical potentials allow discrimination of rate of torque development in imaginary isometric plantar flexion," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 11, pp. 2675–2678, Nov. 2008.

[8] G. R. Müller-Putz, D. Zimmermann, B. Graimann, K. Nestinger, G. Korisek, and G. Pfurtscheller, "Event-related beta EEG-changes during passive and attempted foot movements in paraplegic patients," *Brain Res.*, vol. 1137, pp. 84–91, Mar. 2007.

[9] T. Noda, N. Sugimoto, J. Furukawa, M.-A. Sato, S.-H. Hyon, and J. Morimoto, "Brain-controlled exoskeleton robot for BMI rehabilitation," in *Proc. 12th IEEE-RAS Int. Conf. Hum. Robots (Humanoids)*, Nov. 2012, pp. 21–27.

[10] C. D. Virgilio G., J. H. Sossa A., J. M. Antelis, and L. E. Falcón, "Spiking neural networks applied to the classification of motor tasks in EEG signals," *Neural Netw.*, vol. 122, pp. 130–143, Feb. 2020.

[11] J. M. Antelis, B. Gudiño-Mendoza, L. E. Falcón, G. Sanchez-Ante, and H. Sossa, "Dendrite morphological neural networks for motor task recognition from electroencephalographic signals," *Biomed. Signal Process. Control*, vol. 44, pp. 12–24, Jul. 2018.

[12] M. S. Al-Quraishi, I. Elamvazuthi, T. B. Tang, M. Al-Qurishi, S. H. Adil, and M. Ebrahim, "Bimodal data fusion of simultaneous measurements of EEG and fNIRS during lower limb movements," *Brain Sci.*, vol. 11, no. 6, p. 713, May 2021.

[13] M. J. Khan, U. Ghafoor, and K.-S. Hong, "Early detection of hemodynamic responses using EEG: A hybrid EEG-fNIRS study," *Frontiers Hum. Neurosci.*, vol. 12, p. 479, Nov. 2018.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–12.
[15] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
[16] J. Li, C. Li, and A. Cichocki, "Canonical polyadic decomposition with auxiliary information for brain–computer interface," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 263–271, Jan. 2017.
[17] J. Li, Z. Struzik, L. Zhang, and A. Cichocki, "Feature learning from incomplete EEG with denoising autoencoder," *Neurocomputing*, vol. 165, pp. 23–31, Oct. 2015.
[18] N. Mammone, C. Ieracitano, and F. C. Morabito, "A deep CNN approach to decode motor preparation of upper limbs from time–frequency maps of EEG signals at source level," *Neural Netw.*, vol. 124, pp. 357–372, Apr. 2020.
[19] J. D. Meier, T. N. Aflalo, S. Kastner, and M. S. A. Graziano, "Complex organization of human primary motor cortex: A high-resolution fMRI study," *J. Neurophysiol.*, vol. 100, no. 4, pp. 1800–1812, Oct. 2008.
[20] D. Liu, W. Chen, K. Lee, R. Chavarriaga, F. Iwane, M. Bouri, Z. Pei, and J. D. R. Millán, "EEG-based lower-limb movement onset decoding: Continuous classification and asynchronous detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1626–1635, Aug. 2018.
[21] D. G. E. Robertson, G. E. Caldwell, J. Hamill, G. Kamen, and S. Whittlesey, *Research Methods in Biomechanics*. Champaign, IL, USA: Human Kinetics, 2013.
[22] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, 2016, Art. no. 016003.
[23] Z. Tang, C. Li, and S. Sun, "Single-trial EEG classification of motor imagery using deep convolutional neural networks," *Optik*, vol. 130, pp. 11–18, Feb. 2017.
[24] J. Yang, S. Yao, and J. Wang, "Deep fusion feature learning network for MI-EEG classification," *IEEE Access*, vol. 6, pp. 79050–79059, 2018.
[25] J. Xue, F. Ren, X. Sun, M. Yin, J. Wu, C. Ma, and Z. Gao, "A multifrequency brain network-based deep learning framework for motor imagery decoding," *Neural Plasticity*, vol. 2020, pp. 1–11, Dec. 2020.
[26] I. Majidov and T. Whangbo, "Efficient classification of motor imagery electroencephalography signals using deep learning methods," *Sensors*, vol. 19, no. 7, p. 1736, Apr. 2019.
[27] D. Zhang, L. Yao, K. Chen, S. Wang, P. D. Haghighi, and C. Sullivan, "A graph-based hierarchical attention model for movement intention detection from EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2247–2253, Nov. 2019.
[28] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, and A. Etemad, "Classification of hand movements from EEG using a deep attention-based LSTM network," *IEEE Sensors J.*, vol. 20, no. 6, pp. 3113–3122, Mar. 2020.
[29] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," 2021, *arXiv:2105.07581*.
[30] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–12.
[31] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.
[32] A. Mai and S. Commuri, "Intelligent control of a prosthetic ankle joint using gait recognition," *Control Eng. Pract.*, vol. 49, pp. 1–13, Apr. 2016.
[33] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of recommendations for SEMG sensors and sensor placement procedures," *J. Electromyogr. Kinesiol.*, vol. 10, no. 5, pp. 361–374, 2000.
[34] M. S. Al-Quraishi, I. Elamvazuthi, T. B. Tang, M. Al-Qurishi, S. Parasuraman, and A. Borboni, "Multimodal fusion approach based on EEG and EMG signals for lower limb movement recognition," *IEEE Sensors J.*, vol. 21, no. 24, pp. 27640–27650, Dec. 2021.
[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[37] Y. Du, Y. Xu, X. Wang, L. Liu, and P. Ma, "EEG temporal–spatial transformer for person identification," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, Aug. 2022.
[38] X. Zhang and H. Li, "Patient-specific seizure prediction from scalp EEG using vision transformer," in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Mar. 2022, pp. 1663–1667.
[39] R. V. Godoy, G. J. G. Lahr, A. Dwivedi, T. J. S. Reis, P. H. Polegato, M. Becker, G. A. P. Caurin, and M. Liarokapis, "Electromyography-based, robust hand motion classification employing temporal multi-channel vision transformers," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10200–10207, Oct. 2022.
[40] S. Bagchi and D. R. Bathula, "EEG-ConvTransformer for single-trial EEG-based visual stimulus classification," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108757.
[41] J. G. Colebatch, "Bereitschaftspotential and movement-related potentials: Origin, significance, and application in disorders of human movement," *Movement Disorders*, vol. 22, no. 5, pp. 601–610, Apr. 2007.
[42] S. Yazawa, A. Ikeda, T. Kunieda, S. Ohara, T. Mima, T. Nagamine, W. Taki, J. Kimura, T. Hori, and H. Shibasaki, "Human presupplementary motor area is active before voluntary movement: Subdural recording of bereitschaftspotential from medial frontal cortex," *Exp. Brain Res.*, vol. 131, no. 2, pp. 165–177, Mar. 2000.
[43] A. Shakeel, M. S. Navid, M. N. Anwar, S. Mazhar, M. Jochumsen, and I. K. Niazi, "A review of techniques for detection of movement intention using movement-related cortical potentials," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–13, Oct. 2015.
[44] D. J. Wright, P. S. Holmes, and D. Smith, "Using the movement-related cortical potential to study motor skill learning," *J. Motor Behav.*, vol. 43, no. 3, pp. 193–201, 2011.
[45] Y. Sato, M. Fukuda, M. Oishi, and Y. Fujii, "Movement-related cortical activation with voluntary pinch task: Simultaneous monitoring of near-infrared spectroscopy signals and movement-related cortical potentials," *Proc. SPIE*, vol. 17, no. 7, 2012, Art. no. 076011.
[46] G. Garipelli, R. Chavarriaga, and J. D. R. Millán, "Single trial analysis of slow cortical potentials: A study on anticipation related potentials," *J. Neural Eng.*, vol. 10, no. 3, Jun. 2013, Art. no. 036014.
[47] I. K. Niazi, N. Jiang, M. Jochumsen, J. F. Nielsen, K. Dremstrup, and D. Farina, "Detection of movement-related cortical potentials based on subject-independent training," *Med. Biol. Eng. Comput.*, vol. 51, no. 5, pp. 507–512, May 2013.
[48] M. Jochumsen, C. Rovsing, H. Rovsing, I. K. Niazi, K. Dremstrup, and E. N. Kamavuako, "Classification of hand grasp kinetics and types using movement-related cortical potentials and EEG rhythms," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–8, Aug. 2017.
[49] M. S. Mirzaee and S. Moghimi, "Detection of reaching intention using EEG signals and nonlinear dynamic system identification," *Comput. Methods Programs Biomed.*, vol. 175, pp. 151–161, Jul. 2019.
[50] E. Lew, R. Chavarriaga, S. Silvoni, and J. D. R. Millán, "Detection of self-paced reaching movement intention from EEG signals," *Frontiers Neuroeng.*, vol. 5, p. 13, Jul. 2012.
[51] I. K. Niazi, N. Jiang, O. Tiberghien, J. F. Nielsen, K. Dremstrup, and D. Farina, "Detection of movement intention from single-trial movement-related cortical potentials," *J. Neural Eng.*, vol. 8, no. 6, Oct. 2011, Art. no. 066009.
[52] C. S. Nam, Y. Jeon, Y.-J. Kim, I. Lee, and K. Park, "Movement imagery-related lateralization of event-related (de)synchronization (ERD/ERS): Motor-imagery duration effects," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 567–577, Mar. 2011.
[53] E. Formaggio, S. Masiero, A. Bosco, F. Izzi, F. Piccione, and A. D. Felice, "Quantitative EEG evaluation during robot-assisted foot movement," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1633–1640, Sep. 2017.
[54] S. Y. Gordleeva, M. Lukoyanov, S. Mineev, M. Khoruzhko, V. Mironov, A. Y. Kaplan, and V. Kazantsev, "Exoskeleton control system based on motor-imaginary brain–computer interface," *Mod. Technol. Med.*, vol. 9, no. 3, pp. 31–36, 2017.
[55] D. P. Murphy, O. Bai, A. S. Gorgey, J. Fox, W. T. Lovegreen, B. W. Burkhardt, R. Atri, J. S. Marquez, Q. Li, and D.-Y. Fei, "Electroencephalogram-based brain–computer interface and lower-limb prosthesis control: A case study," *Frontiers Neurol.*, vol. 8, p. 696, Dec. 2017.
[56] A. I. Sburlea, L. Montesano, and J. Minguez, "Continuous detection of the self-initiated walking pre-movement state from EEG correlates without session-to-session recalibration," *J. Neural Eng.*, vol. 12, no. 3, Jun. 2015, Art. no. 036007.

MAGED S. AL-QURAISHI (Member, IEEE) received the B.Sc. degree in biomedical engineering from Baghdad University, Iraq, in 2005, the M.Sc. degree in biomedical engineering from the Universiti Putra Malaysia (UPM), in 2015, and the Ph.D. degree from the Universiti Teknologi PETRONAS, Malaysia, in 2021. He is currently working as a Postdoctoral Researcher at the Universiti Teknologi PETRONAS. He is also an Academic Staff at Thamar University, Yemen. His research interests include biomedical signal processing, instrumentation, machine learning, and rehabilitation robotics.

IRRAIVAN ELAMVAZUTHI (Senior Member, IEEE) received the Ph.D. degree from the Department of Automatic Control and Systems Engineering, The University of Sheffield, U.K., in 2002. He is currently an Associate Professor at the Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS (UTP), Malaysia. His research interests include control, robotics, mechatronics, power systems, and biomedical applications. He is also the Chair of the IEEE Robotics and Automation Society (Malaysia Chapter).

TONG BOON TANG (Senior Member, IEEE) received the B.Eng. degree in electronics and electrical engineering and the Ph.D. degree in intelligent sensor fusion from the University of Edinburgh, U.K., in 1999 and 2006, respectively. In 2012, he joined the Department of Electronic and Electrical Engineering, Universiti Teknologi PETRONAS, as an Associate Professor. Since 2018, he has been appointed as the Director of the Institute of Health & Analytics. He is currently the Chair of the IEEE Circuits and Systems Society (Malaysia Chapter).

MUHAMMAD S. AL-QURISHI (Member, IEEE) received the Ph.D. degree from the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia, in 2017. He was a Postdoctoral Researcher with the Chair of Pervasive and Mobile Computing (CPMC), CCIS, KSU. He is one of the Founding Members of CPMC. He is currently a Data Scientist working with the Research and Innovation Department, ELM Company. He has published several articles in refereed journals (IEEE, ACM, Springer, and Wiley). His research interests include data science, big data analysis and mining, pervasive computing, and machine learning. He received an Innovation Award for a Mobile Cloud Serious Game from KSU 2013 and the Best Ph.D. Thesis Award from CCIS, KSU, in 2018. He got the IBM Data Science Professional Certificate and the Deep Learning Certification from deep learning.ai.

SYED HASAN ADIL (Member, IEEE) received the B.S. degree in computer science from the National University of Computer and Emerging Sciences, and the M.S. and Ph.D. degrees in computer science from Iqra University, Pakistan. He is currently an Associate Professor and the Chair of the Department of Software Engineering, Iqra University, where he has been since 2009. His research interests include machine learning, optimization, parallel computing, and data science. Much of his work has been on improving the solutions of various applied problems, mainly through the application of data mining, optimization, and performance evaluation. He has given numerous professional trainings, invited talks, and tutorials.

MANSOOR EBRAHIM (Senior Member, IEEE) received the B.S. degree in computer engineering from the Sir Syed University of Engineering and Technology, Karachi, Pakistan, the M.Sc. degree in telecommunication from the Queen Mary, University of London, U.K, and the Ph.D. degree in computing from Sunway University, Malaysia. He is currently associated as an Assistant Professor with the Department of Computer Science, Iqra University, Pakistan. His current research interests include image and video analysis and compression, optimization, image cryptography, and the IoT.

ALBERTO BORBONI (Member, IEEE) was born in Brescia, Italy, in 1973. He received the master's degree in mechanical engineering and the Ph.D. degree in applied mechanics from the University of Brescia, in 1997 and 2012, respectively. He is currently an Assistant Professor of mechanics for machines and mechanical systems at the University of Brescia. He is the author of more than 100 publications. His research interests include mechanical engineering in biomedical applications, smart actuators, kinematics, and dynamics of high-speed machines, and mechanical design of industrial and biomedical systems.