Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 7, No. 1, pp. 102-112, January 2025; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v7i1.60

RESEARCH PAPER

OPEN ACCES

Implementation of Vision Transformer for Early Detection of Autism Based on EEG Signal Heatmap Visualization

Aufa Rafiki[1], Melinda Melinda[1], Maulisa Oktiana[1], Ernita Dewi Meutia[1], Afnan Afnan[1], Muliyadi Muliyadi[2], and Lailatul Qadri Zakaria[3]

[1] Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia
[2] Department of Doctoral Engineering Study Program, Universitas Syiah Kuala, Banda Aceh, Indonesia
[3] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

ABSTRACT

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder characterized by difficulties in social interaction, communication, and repetitive behavioral patterns. Early detection of ASD is crucial for improving the quality of life of affected individuals and alleviating the burden on their families. This study proposes a computer-aided diagnostic system for ASD by applying a pre-trained Vision Transformer (ViT-B/16) architecture to EEG signal data obtained from King Abdul Aziz University. The dataset comprises EEG recordings from 16 subjects (8 normal and 8 ASD) that have undergone preprocessing—including filtering using the Discrete Wavelet Transform (DWT), segmentation (windowing), and conversion into heatmap representations—and were subsequently partitioned into training, validation, and testing subsets. The ViT model was trained for 100 epochs with a batch size of 16, using the AdamW optimizer and the CrossEntropy loss function, while two learning rate configurations (0.0001 and 0.00001) were evaluated; the best-performing weights were selected based on the lowest validation loss. Test results indicate that the model trained with a learning rate of 0.00001 achieved a testing accuracy of 99.53%, accompanied by excellent precision, specificity, recall, and f1-score, thereby demonstrating strong generalization capabilities and minimal overfitting. Future research is recommended to incorporate locally sourced datasets and to further customize the ViT architecture through comprehensive hyperparameter tuning, with the aim of developing a mobile application to support clinical ASD diagnosis.

PAPER HISTORY

Received Dec. 10, 2024
Revised Jan. 20, 2025
Accepted Feb. 10, 2025
Published Feb. 24, 2025

KEYWORDS

EEG;
ASD;
DWT;
Heatmap Visualization;
Vision Transformer

CONTACT:

[1]Melinda Melinda
melinda@usk.ac.id

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder characterized by a broad range of difficulties in social interaction, interpersonal communication, and sensory processing. The disorder is also marked by restricted, repetitive, and stereotypical patterns of interest and behavior [1], [2]. Such symptoms can significantly disrupt behavior, language, communication, and social interaction, thereby posing considerable challenges to the learning process [3]. Moreover, ASD is frequently associated with hyperactivity, which negatively impacts daily activities and diminishes the quality of life for both affected individuals and their families [4], [5].

Early detection of ASD plays a crucial role, as timely intervention can have a profoundly positive impact on developmental outcomes—particularly in enhancing communication skills and social interaction. An accurate and early diagnosis not only paves the way for improving the quality of life of individuals with ASD but also helps mitigate the burden on their families [6]. However, a primary challenge in diagnosing ASD lies in the heterogeneity of its symptoms . In other words, each individual with ASD may exhibit a highly variable combination of symptoms. For instance, some children may experience substantial delays in language development and verbal communication, whereas others—despite possessing adequate speech capabilities—may demonstrate difficulties with nonverbal communication, such as limited eye contact, diminished facial expressiveness, or challenges in interpreting social cues. Additionally, repetitive behaviors and sensory sensitivities may present with differing intensities among individuals. This variability complicates the uniform application of conventional screening tools and diagnostic methods [7].

Recent advancements in Artificial Intelligence (AI) have driven significant innovations in the development of advanced medical diagnostic systems, particularly

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 7, No. 1, pp. 102-112, January 2025; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v7i1.60

concerning Autism Spectrum Disorders (ASD). Traditional diagnostic methods for ASD primarily rely on behavioral assessments and clinical observations, which, although valuable, can be subjective and time-intensive. Moreover, the rapid progression of deep learning techniques within AI over recent years has fundamentally altered the landscape of medical data analysis [8]. Deep learning methodologies have been extensively applied to neuroimaging modalities to discern patterns associated with ASD. These AI-driven models analyze intricate brain structures and functions, enabling the detection of subtle anomalies that might not be evident through conventional assessments. For instance, research has demonstrated the effectiveness of AI in analyzing neuroimaging data to differentiate individuals with ASD from neurotypical controls, thereby enhancing diagnostic precision and efficiency [9].

Several studies have examined the application of deep learning methods for detecting ASD by visualizing EEG signals as heatmaps and spectrograms. For example, one study using the ResNet101 approach successfully classified ASD and normal subjects based on heatmap images, achieving a training accuracy of 99.3% and a validation accuracy of 98.9% [10]. Another study that applied CNN to spectrogram images reported a training accuracy of 99.15% [11]. Although CNNs have demonstrated impressive performance, their inherent reliance on local feature extraction restricts their ability to capture long-range dependencies and global contextual information within EEG signals. This limitation is significant when addressing the complex and variable nature of ASD symptoms, which may be more effectively characterized by considering global patterns across the entire signal [12].

In contrast, the Vision Transformer (ViT) architecture leverages self-attention mechanisms to capture global inter-patch relationships, thereby providing a more comprehensive analysis of EEG signals. Recent research has highlighted that ViT-based models can achieve superior performance by integrating global context, which is essential for detecting the nuanced patterns associated with ASD [13]. For instance, a study published in Healthcare (2023) [14] reported that the Vision Transformer achieved an accuracy of 99.06%, a precision of 99.06%, a recall of 99.14%, and an F1-score of 99.1% in its classification tasks—demonstrating its significant potential to outperform traditional CNN-based approaches in certain diagnostic contexts.

Building on these insights, the principal contribution of this study is the use of the Vision Transformer (ViT) architecture for classifying EEG signals to detect ASD — marking it as the first study to employ ViT in this context. As illustrated in Fig. 1, raw EEG signals are first filtered using the Discrete Wavelet Transform (DWT), then segmented into overlapping windows with a 50% overlap, and subsequently converted into heatmap images. These

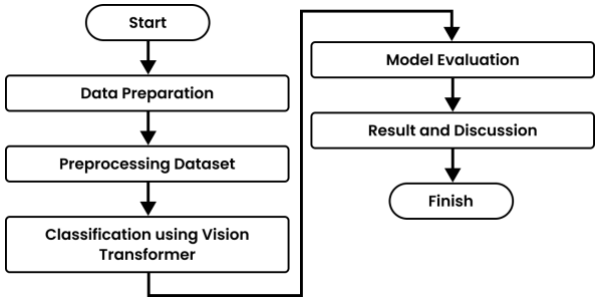


Fig. 1. Workflow of This Study.

preprocessed heatmaps are finally classified using ViT. The model is trained with standardized parameters—100 epochs, a batch size of 16, the AdamW optimizer, cross-entropy as the loss function, and the best-performing weights—while two different learning rate values are compared to evaluate performance.

In summary, this study aims to address the following research objectives:

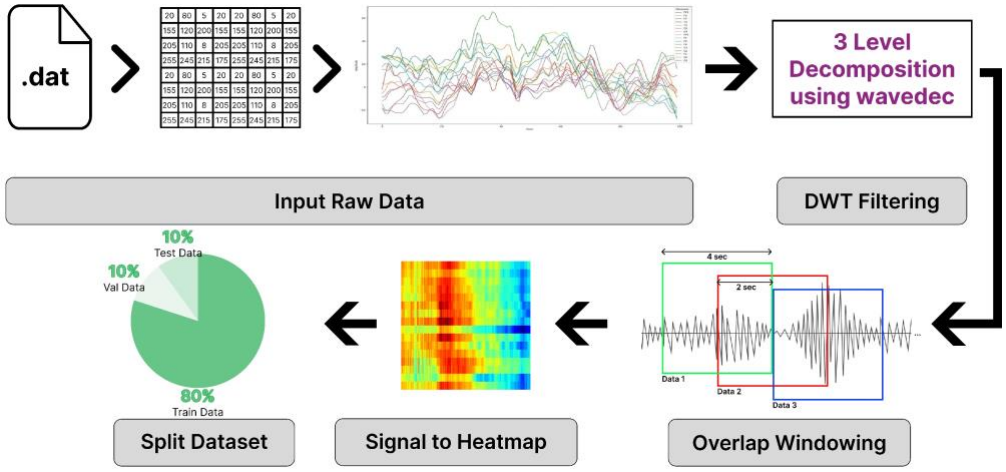


Fig. 2. Preprocessing processes.

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

- To determine the optimal learning rate for training the Vision Transformer using performance metrics such as accuracy, specificity, precision, recall, and F1-score.
- To compare the training accuracy of the Vision Transformer with methods used in previous studies.

This study is organized as follows: Section II outlines the dataset used, the proposed methodology, and the training and testing schemes applied. Section III presents the outcomes from each method as well as the accuracy results of the Vision Transformer. Section IV discusses the interpretation of the results, comparisons with other studies, and the limitations of the research. Finally, Section V concludes by summarizing the objectives, key findings, and future research directions.

2. MATERIALS AND METHOD

A. Dataset

This proposed method is evaluated using a dataset from King Abdulaziz University (KAU) Hospital in Jeddah, Saudi Arabia [15] , which has also been used in previous studies [11][16][17]. The dataset is publicly available and can be obtained by submitting a formal email request to the dataset owner, Dr. Mohammed Jaffer Alhaddad via email address, as described in [18]. In our study, we followed the same procedure to acquire the dataset while ensuring the anonymity of participants by not publishing any personal identification information. The dataset comprises EEG recordings from sixteen subjects, including eight ASD subjects (all boys, aged 10–16 years, with a total signal duration of 4104.2 second) and eight control subjects (all boys, aged 9–16 years, with a total signal duration of 4534.9 second) with no history of neurological disorders. The EEG signals were recorded using Ag/AgCl electrodes with a g.tec EEG cap, g.tec USB amplifiers, and BCI2000 software, with subjects in a relaxed state to obtain artifact-free data. Recordings were conducted using 16 channels following the international 10–20 system. A band-pass filter (0.1–60 Hz) and a notch filter (60 Hz) were applied during recording to filter the data, and all signals were digitized at a 256 Hz sampling rate [15][11]. The dataset is stored in a ".dat" format, which organizes the data in a matrix structure to ensure compatibility with various modern analytical methods and to support experiments requiring well-organized data structures [10].

B. Preprocessing

The dataset preprocessing is performed in several stages: importing the EEG data, cleaning the EEG data using Discrete Wavelet Transform (DWT), windowing, converting the EEG into heatmap visualizations, and splitting the dataset into three portions. The overall process is depicted in Fig. 2.

a) Input Raw Data

At the initial stage of preprocessing, the EEG data is imported using a specialized library such as BCI2kReader, which facilitates the retrieval of EEG signals into the Python programming environment [10]. Additionally, to support accurate analysis, it is essential to access signal-related information—such as the sampling frequency, the number and names of the recorded channels, and other pertinent details—through the 'parameters' attribute of the imported data entity.

b) DWT Filtering

After data preparation, the next step involves cleaning the EEG signal by applying the Discrete Wavelet Transform (DWT) method. One of the key techniques employed is Discrete Wavelet Transform (DWT), which is well-suited for EEG signals due to its ability to analyze both time and frequency characteristics simultaneously. In this study, the Daubechies 4 (db4) wavelet was chosen because of its smoothness and compact support, which makes it particularly effective in capturing transient components in EEG signals. Furthermore, db4 has been widely used in EEG analysis due to its ability to preserve signal characteristics while minimizing distortion [19].

The decomposition process was carried out up to the third level using the wavedec function from the pywt library, following findings from previous studies that suggest three-level decomposition effectively captures key EEG frequency bands without excessive data loss. At each level, the signal is decomposed into approximation (A) and detail (D) coefficients. [20]. Specifically:

- Level 1: A1 (0–128 Hz) and D1 (128–256 Hz)
- Level 2: A2 (0–64 Hz) and D2 (64–128 Hz)
- Level 3: A3 (0–32 Hz) and D3 (32–64 Hz)

Since EEG signals are often contaminated with noise from external sources (e.g., electromagnetic interference) and physiological artifacts (e.g., eye movements), a soft thresholding technique was applied to the detail coefficients (D1, D2, and D3) to suppress noise while retaining meaningful signal components. The threshold value was empirically set to 10, based on prior research that demonstrated effective noise reduction while preserving signal integrity in EEG applications [21].

Finally, the cleaned signal was reconstructed using the inverse DWT (IDWT), where all processed coefficients—including the approximation coefficient A3 and the thresholded detail coefficients (D3, D2, and D1)—were recombined to generate a denoised EEG signal. This preprocessing step ensures that the signal retains critical neurological information, enhancing its suitability for subsequent feature extraction and classification in autism detection tasks [22], [23].

c) Windowing

At this stage, the cleaned EEG signal is segmented into intervals of 4 seconds in duration, corresponding to 1024 amplitude samples at a sampling frequency of 256 Hz [17]. To augment the dataset while preserving the continuity of temporal information, an overlapping windowing technique is applied, wherein a 2-second overlap (i.e., 50% overlap) is introduced between consecutive segments. It is important to note that the first

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

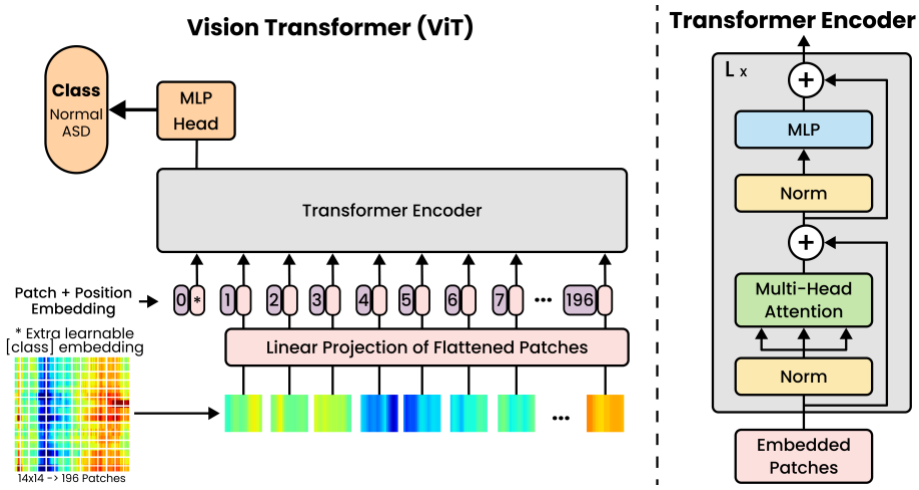


Fig. 3. Architecture ViT Overview.

4 seconds of each EEG signal file are omitted, as they typically contain artifacts or noise resulting from electrode placement, and any final segment shorter than 4 seconds is also eliminated, as it does not meet the required duration criteria [24]. In implementing this segmentation process, Python's slicing technique is employed, as the EEG signal is initially converted into an array format, enabling efficient extraction of segments. The slicing operation follows the format array[start:end], where start determines the beginning of each 4-second segment and is initialized at 1024 samples to exclude the first 4 seconds of data. The end index is then calculated as start + 1024, ensuring that each segment maintains the required duration. By applying this structured segmentation approach, the dataset is effectively expanded while preserving essential temporal characteristics, which is crucial for machine learning models that require a large and well-structured dataset, such as the architecture employed in this study [13][25].

d) Signal to Heatmap

Following the windowing process, the cleaned and segmented EEG signal is converted into a visual format in the form of a heatmap. This transformation aims to convert numerical EEG data into an easily interpretable visual representation, thereby enabling researchers and healthcare professionals to rapidly identify differences in intensity and patterns of brain activity. Consequently, subtle variations that might be concealed in the raw data become clearly discernible through color gradients on the heatmap [26]. The heatmap is generated using the pcolormesh function from the matplotlib library, which creates a plot with a consistent figure size of 5×5 inches to ensure uniform resolution across all heatmaps. This process utilizes the 'jet' colormap to produce a gradient ranging from blue to red, representing amplitude values between -120 and 120 with the adjustment of the vmin and vmax parameters, and is ultimately saved as a PNG-formatted image.

e) Split Dataset

In the final stage of preprocessing, the dataset is divided into three parts: 80% for training, 10% for validation, and 10% for testing. This partitioning method has been successfully applied in previous studies that utilized architectures similar to the one employed in this study [27]. For each class (e.g., ASD and NORMAL), the available files are first randomized using the random.shuffle function to ensure an unbiased data distribution. This approach enables the model to be trained optimally and evaluated accurately using data that it has not encountered before. The method has proven to be effective, straightforward, and well-suited for small- to medium-sized datasets, such as the one used in this study [28][29][30].

A strict separation between the test data and the model development process was implemented to ensure the integrity of the model evaluation by organizing them into separate folders. The test dataset, consisting of 427 images, was completely isolated throughout the entire model training process. It was not used for model development or hyperparameter tuning.

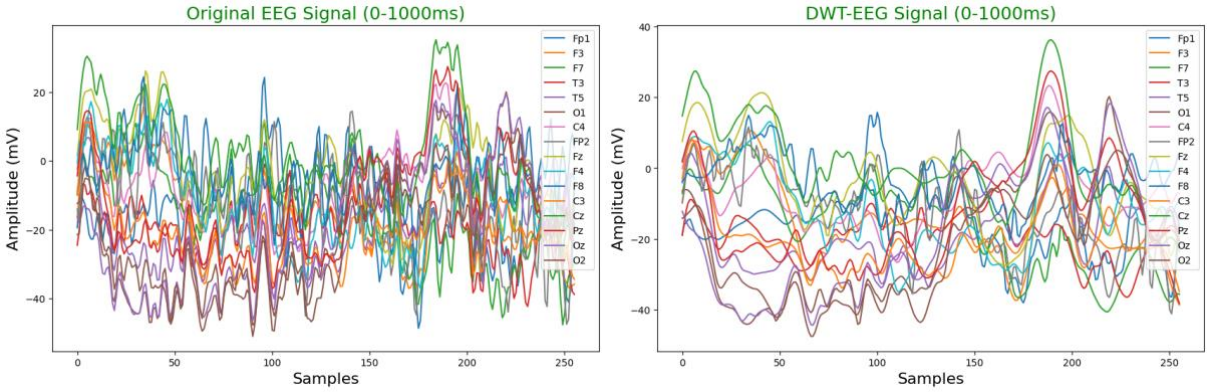
Only the training data (80%) and validation data (10%) were utilized during the training and refinement phases of the model. Measures were taken to prevent any test data leakage into the training or validation stages, ensuring that model performance is evaluated on entirely unseen data.

C. Classification using Vision Transformer (ViT)

This study implements the pre-trained Vision Transformer (ViT-B/16) architecture from the torchvision library for the classification of EEG images. ViT was selected for its ability to achieve competitive performance in image recognition tasks, including unconventional data such as EEG representations [31]. An overview of the model architecture is provided in Fig. 3. The input image, with a resolution of 224×224 pixels and three color channels (RGB), is segmented into a sequence of patches

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 7, No. 1, pp. 102-112, January 2025; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v7i1.60



(a) (b)

Fig. 4. EEG Signal Visualization, (a) Original Signal, (b) DWT Signal.

measuring 16×16 pixels, resulting in 196 patches per image. Each patch is then transformed into a vector with dimensions defined by $N \times (p^2 \cdot C)$, where $N = 196$, $p = 16$, and $C = 3$. Positional encoding is subsequently added to these vectors to preserve spatial relationships among the patches [13][32], addressing the inherent limitation of ViT in implicitly capturing relative positional information.

Following patch embedding, the resultant sequence is processed by a Transformer Encoder composed of Multi-Head Self-Attention (MHSA) layers and Feedforward Neural Networks (FFN). In contrast to Convolutional Neural Networks (CNNs), which rely on convolutional kernels for local feature extraction, ViT captures global inter-patch relationships through self-attention, thereby facilitating the identification of complex patterns within the images [13][32]. A class token is appended to the beginning of the patch sequence to serve as a global representation of the image. This token is processed alongside the patch embeddings by the encoder, and its final output is used to predict the class via an MLP Head (fully connected layer) [33].

In addition to the image patches, the class token—serving as the comprehensive representation of the entire image—is integrated within the Transformer Encoder. The final value of this token is employed as the principal representation for class prediction, as demonstrated in several related studies [33].

The training procedure leverages pre-trained ViT weights to address the challenge of large dataset requirements and to mitigate the risk of overfitting. During training, both accuracy and loss on the training and validation datasets are continuously monitored. Similar methodologies have been applied in prior research to ensure effective learning without overfitting [34]. Before training, the hyperparameters were carefully determined to achieve optimal performance. The selected configuration included a batch size of 16, CrossEntropyLoss as the loss function, and a total of 100 training epochs. The AdamW optimizer was chosen due to its improved weight decay handling, which helps prevent overfitting and enhances generalization performance compared to standard Adam. Additionally, two learning rate values (0.0001 and 0.00001) were tested to evaluate their impact on training stability and model performance. The best learning rate was determined based on the results obtained from the confusion matrix and its five derived metrics, ensuring that the final model achieved the highest classification performance. The best-performing weights were selected based on the lowest validation loss to enhance generalization.

D. Evaluation Of Metrics

The final evaluation of the model is conducted on the test data using a confusion matrix along with its derived evaluation metrics. The confusion matrix is an $n \times n$ matrix (where n represents the number of classes) that summarizes the model's performance by categorizing predictions into four components:

- True Positive (TP): The number of ASD cases correctly identified as ASD.
- True Negative (TN): The number of normal cases correctly identified as normal.
- False Positive (FP): The number of normal cases incorrectly classified as ASD (false alarms).
- False Negative (FN): The number of ASD cases incorrectly classified as normal (missed detections).

From the confusion matrix, five evaluation metrics are computed to assess model performance:

a) Accuracy

Accuracy measures the overall correctness of the model's predictions by calculating the proportion of correctly classified instances (both ASD and normal) over the total number of cases (Eq. (1)):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \times 100 \quad (1)$$

b) Specificity

Specificity evaluates the model's ability to correctly classify normal cases, minimizing false alarms (Eq. (2)):

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100 \quad (2)$$

c) Recall

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

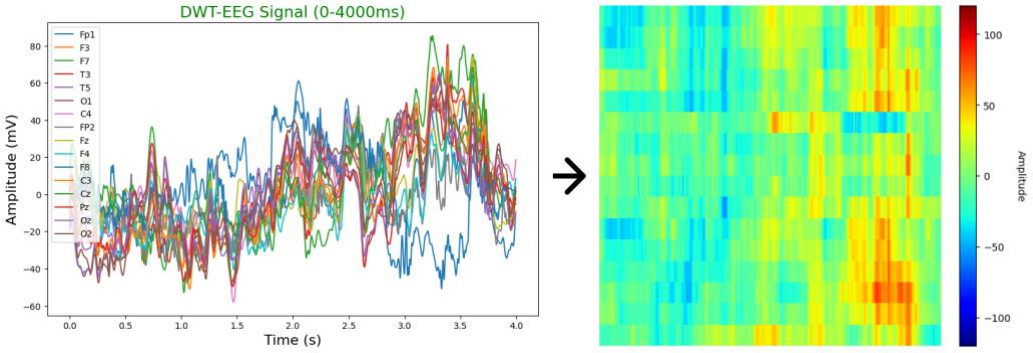


Fig. 5. DWT Signal to Heatmap.

Recall, also referred to as sensitivity, quantifies the model's ability to correctly identify ASD cases (Eq. (3)):

$$\text{Recall} = \frac{TP}{TP+FN} \times 100 \quad (3)$$

d) Precision

Precision measures how many of the predicted ASD cases are actually ASD (Eq. (4)):

$$\text{Precision} = \frac{TP}{TP+FP} \times 100 \quad (4)$$

e) F1-Score

The F1-score is the harmonic mean of recall and precision, balancing both metrics ((Eq. (5)):

$$\text{F1-Score} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (5)$$

The choice of these metrics is particularly relevant for ASD detection due to the potential consequences of misclassification. By incorporating these metrics, the study ensures a comprehensive evaluation of the model's performance in detecting ASD from EEG images, considering both false negatives and false positives.

Table 1. Information about Dataset

Parameter	Value
Sampling Rate	256 Hz
Number of Channels	16
Name of Channels	Fp1, F3, F7, T3, T5, O1, C4, FP2, Fz, F4, F8, C3, Cz, Pz, Oz, O2

3. RESULTS

A. Data Preprocessing Result

a) Input Raw Data Result

The process of reading the Electroencephalogram (EEG) signal data produces an output in the form of a two-dimensional array representing brain activity. The first dimension corresponds to the channels, while the second dimension indicates the amplitude values at each time point. Fig. 4(a) illustrates 16 EEG signal channels within a sample range from 1 to 256, with each channel rendered in a distinct color using the Matplotlib library in the Python programming language. Additionally, important information regarding the acquired EEG signals, obtained from the 'parameters' attribute, is presented in Table 1.

b) DWT Filter Result

At this stage, the signal is processed to remove noise and artifacts without reducing the number of samples or altering the amplitude. This is achieved by filtering the detail components using a thresholding technique rather than removing them. If the input signal consists of n samples, the output of this process will also consist of n samples, as illustrated in Fig. 4(b).

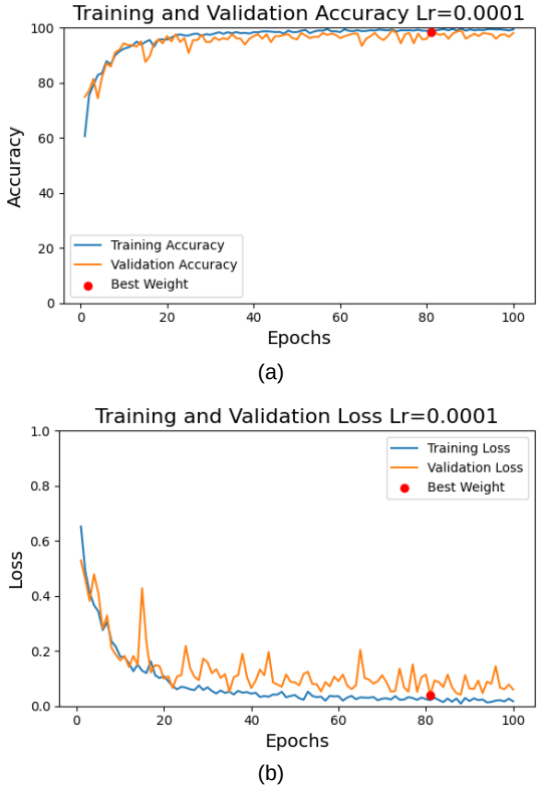


Fig. 6. Training and Validation Performance at Lr=0.0001. (a) Accuracy Curves, (b) Loss Curves.

c) Windowing Result

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 7, No. 1, pp. 102-112, January 2025; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v7i1.60

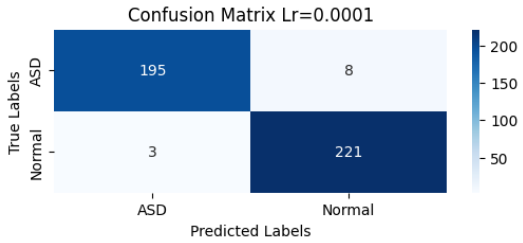
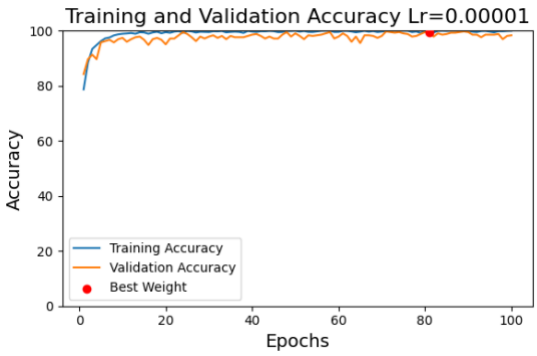
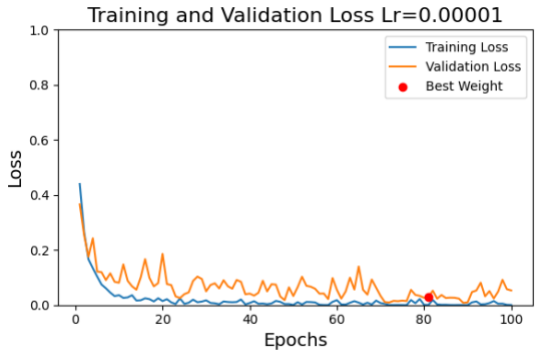


Fig. 7. Confusion Matrix for a learning rate of 0.0001.



(a)



(b)

Fig. 8. Training and Validation Performance at Lr=0.00001. (a) Accuracy Curves, (b) Loss Curves.

At this stage, the signal, which was segmented into 4-second slices with a 2-second overlap, resulted in a total of 4,262 segments, representing a substantial increase relative to the initial 16 datasets.

d) Signal to Heatmap Result

The output of this stage is a heatmap image that converts the amplitude values of the signal from each segment—obtained from the windowing stage into corresponding color intensities, as illustrated in Fig. 5. Specifically, the heatmap visualization is composed of 16 rows representing the number of channels in the EEG signal and 1024 columns depicting the amplitude values for each data segment, thereby yielding an image analogous to a matrix in which numerical values are replaced by variations in color. It is observed that, at the onset of the signal, the amplitude is relatively low, resulting in a predominance of blue and green hues; conversely, toward the end of the signal period, the amplitude increases, which gives rise to yellowish-reddish tones. Furthermore, in the heatmap representation, the data from the first channel is positioned at the bottom of the image, while the 16th channel is displayed at the top.

Table 2. Comparison of Evaluation Metrics (in %) for Two Different Learning Rates in the Proposed Model.

Metrics	Lr=0.0001	Lr=0.00001
Accuracy	97.42%	99.53%
Precision	98.48%	100%
Specificity	98.66%	100%
Recall	96.06%	99.01%
F1-Score	97.26%	99.50%

e) Split Dataset Result

At this stage, a total of 3,409 images were collected for the training dataset, comprising 1,791 images categorized as normal and 1,618 images classified as ASD. Subsequently, the validation set consisted of 426 images, divided into 224 normal images and 202 ASD images. The testing set, on the other hand, comprised 427 images, with 224 images allocated to the normal class and 203 images to the ASD class.

B. Model Training and Evaluation Results

At this stage, two models were obtained with the best weights derived from different learning rate configurations. Next, a confusion matrix was employed to provide a detailed depiction of the classification results between the two primary classes, namely positive (ASD) and negative (normal), in order to evaluate the model's performance in identifying autism.

a) Learning Rate = 0.0001

Fig. 6 illustrates the model's performance during the training process with a learning rate of 0.0001, where the accuracy curve is presented in Fig. 6(a) and the loss curve in Fig. 6(b). As shown in Fig. 6(a), training accuracy increases rapidly during the first 20 epochs before stabilizing at 99.12% upon the final weight update. Similarly, validation accuracy follows a consistent upward trend, reaching 98.59% at the last weight update. The close alignment between the training and validation accuracy curves indicates that the model generalizes well without signs of overfitting. Meanwhile, Fig. 6(b) displays the loss curves, providing further insights into the model's optimization process. The training loss continuously decreases as the number of epochs increases, ultimately reaching 0.0213 at the final weight update. The validation loss also remains stable, with a final value of 0.0403. The absence of a significant divergence between training and validation loss further reinforces the model's ability to generalize effectively. Throughout the training process, the model's weights were updated 15 times, signifying 15 instances of performance improvement based on reductions in validation loss. The optimal weights were

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

obtained at epoch 81, where the validation loss reached its lowest point, as indicated by the red marker in both subfigures. The confusion matrix for the test data (see Fig. 7), where the positive class represents ASD, shows a TP of 195, FP of 3, TN of 221, and FN of 8.

b) Learning Rate = 0.00001

After training with a learning rate of 0.00001, the model effectively identified complex patterns in the data. Analysis was conducted through accuracy graphs, loss graphs, and the evaluation of the confusion matrix. The results presented in Fig. 8 indicate that the training accuracy reached 100% and the validation accuracy 99.77%, suggesting that the model did not exhibit overfitting. Both the training loss and validation loss graphs displayed a consistent decline over several

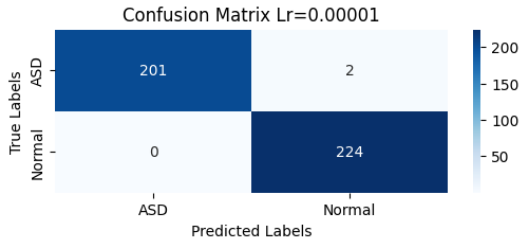


Fig. 9. Confusion Matrix for a learning rate of 0.00001.

epochs, with the training loss reaching 0.0001 and the validation loss 0.0078. Notably, the best-performing weights were obtained at epoch 89, based on the lowest validation loss. The confusion matrix for the test data (see Fig. 9), where the positive class represents ASD, shows a TP of 201, FP of 0, TN of 224, and FN of 2. Table 2 presents the evaluation metrics, providing a comparative analysis between the two learning rate values. Specifically, Table 2 details the metrics—accuracy, precision, recall, specificity, and F1-Score—which are computed in accordance with Equations 1, 2, 3, 4, and 5. The values for false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) are derived from the testing data evaluation, as represented by the confusion matrices shown in Fig. 7 for a learning rate of 0.0001 and in Fig. 9 for a learning rate of 0.00001. The findings indicate that the learning rate of 0.00001 outperformed the configuration with a learning rate of 0.0001 across all evaluated metrics.

4. DISCUSSION

As highlighted in Fig. 6 and Fig. 8, the model training results provide comprehensive insights into its performance throughout the training process. The blue line, representing training accuracy, indicates the extent to which the model correctly predicts the data at each epoch. A consistent increase in this line signifies that the model is effectively learning from the training data. In contrast, the orange line, which reflects validation accuracy, demonstrates the model's ability to generalize to data that has not been used during training. Ideally, simultaneous increases in both lines indicate that the model is not merely memorizing the training data but is also capable of recognizing the underlying patterns more broadly. However, as highlighted, if training accuracy continues to rise while validation accuracy stagnates or declines, this may indicate overfitting, whereby the model becomes overly focused on the training data and loses its ability to generalize to new data.

Fig. 6 displays the performance of the model with a learning rate of 0.0001. In this graph, the blue line exhibits high stability and approaches 100%, indicating that the model has achieved near-perfect accuracy on the training data, thus effectively optimizing its predictions on the training dataset. Meanwhile, the orange line shows validation accuracy ranging from 90% to nearly 100%, suggesting that the model also performs very well on the validation data. The loss graphs support these observations, with the training loss consistently decreasing toward zero and the validation loss remaining within a range of approximately 0.2 to slightly above 0.1. Therefore, as highlighted by these results, this model can be considered free from indications of overfitting.

Fig. 8 illustrates the training performance of the model with a learning rate of 0.00001. The graph displays a pattern similar to that of the previous model, with the blue line remaining stable and reaching 100%, reflecting optimal training accuracy. The orange line also indicates high validation accuracy, above 95% and approaching 100%, signifying an excellent generalization capability. The loss graphs further corroborate these findings by showing that the training loss approaches zero while the validation loss remains low. Therefore, as highlighted by these results, this model can be considered free from indications of overfitting.

The classification results on the test data for each learning rate configuration are presented in the confusion matrices shown in Fig. 7 and Fig. 9. In Fig. 7, which corresponds to the model trained with a learning rate of 0.0001, the model achieved an accuracy of 97.42% with a misclassification rate of 2.58%. The detailed performance metrics include a precision of 98.48%, a specificity of 98.66%, a recall of 96.06%, and an f1-score of 97.26%. Meanwhile, Fig. 9 illustrates the confusion matrix for the model trained with a learning rate of 0.00001. This configuration demonstrated superior performance, achieving an accuracy of 99.53% with a misclassification rate of only 0.47%. Additionally, the model attained a precision of 100%, specificity of 100%,

Table 3. Comparison of Classification Performance in Previous Studies

Reference	Method	Test Accuracy
[9]	ResNet-101	98.9%
[10]	Customized CNN	99.15%
This Study	ViT B-16	99.53%

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

recall of 99.01%, and an f1-score of 99.50%. Based on the evaluation of all performance parameters, it can be concluded that a learning rate of 0.00001 represents the optimal configuration in this study. This learning rate enables smoother and more stable weight updates during training, allowing the model to learn more effectively.

Table 3 presents a comparative analysis of the results obtained in this study with those of previous research utilizing EEG-based image classification to differentiate between children with and without autism. A prior study employing the ResNet-101 architecture [10] reported a test accuracy of 98.9%, while another study implementing a conventional CNN architecture [11] achieved an accuracy of 99.15%. In contrast, this study attained an accuracy of 99.53% using the Vision Transformer (ViT) B-16 architecture.

The superior performance of the ViT B-16 model can be attributed to its ability to capture long-range dependencies and global patterns more effectively than traditional CNNs, which primarily rely on localized feature extraction through convolutional operations. CNNs have demonstrated strong performance in EEG-based classification; however, their constrained receptive fields may limit their ability to learn complex spatial relationships within EEG-derived images. In contrast, the self-attention mechanism inherent in Vision Transformers facilitates a more comprehensive contextual understanding of EEG signals, potentially contributing to enhanced classification accuracy. These findings suggest the potential of transformer-based architectures for EEG analysis and highlight their promise in advancing more precise diagnostic tools for autism spectrum disorder. However, further validation is required to confirm these advantages across larger and more diverse datasets.

The implications of this study are significant for noninvasive autism detection and the medical field, particularly in the early detection of autism through EEG signal analysis using the Vision Transformer model. The high precision of this system minimizes the risk of false-positive diagnoses, thereby reducing unnecessary psychological and financial burdens on patients and their families. However, the successful deployment of this model in clinical settings requires further validation through independent datasets and collaboration with medical professionals. Moreover, our findings contribute to the broader field of ASD diagnostics by demonstrating that transformer-based architectures can more effectively capture complex EEG patterns compared to traditional CNNs. Future research should focus on expanding the dataset with larger and more diverse samples, as well as exploring the integration of additional diagnostic modalities, to enhance the generalizability and clinical applicability of these promising results.

Despite these promising results, this study has several limitations. The relatively small dataset may affect the model's generalization capability when applied in real-world scenarios. Additionally, the dataset used in this study originates from King Abdul Aziz University, Saudi Arabia, which may not fully reflect the characteristics of populations from other regions, such as Aceh Province, Indonesia. EEG signals exhibit variations influenced by demographic, cultural, and genetic factors, which may impact the model's ability to generalize beyond the training data. Therefore, future research should focus on expanding the dataset with a more diverse and representative sample to enhance the robustness and reliability of the model in clinical applications.

By addressing these limitations and conducting further research, the proposed approach can contribute to the advancement of AI-assisted diagnostic tools, particularly by improving accessibility and accuracy in autism detection across diverse populations. Future studies should prioritize the integration of locally sourced datasets to enhance the model's generalizability and facilitate direct local implementation. Additionally, comprehensive hyperparameter tuning is also recommended to refine the Vision Transformer model for improved accuracy and efficiency. Furthermore, the development of a mobile application incorporating the optimized model could provide a practical, user-friendly tool to assist clinicians in autism diagnosis, bridging the gap between research advancements and real-world clinical applications.

5. CONCLUSION

This study utilizes a dataset from King Abdul Aziz University [18] to train a model employing the Vision Transformer (VIT B-16) architecture. In the training process, parameters used include 100 epochs, a batch size of 16, the AdamW optimizer, and the CrossEntropyLoss function. Two learning rate values were tested to determine the optimal configuration. In addition, the best-performing weights obtained during training were preserved based on the lowest validation loss to ensure optimal generalization. The test results indicated that the model trained with a learning rate of 0.00001 achieved an accuracy of 99.53%, whereas the model with a learning rate of 0.0001 attained an accuracy of 97.42%.

Future research may be directed towards incorporating locally sourced datasets to enable direct local implementation and thereby enhance the practical utility of the study. Additionally, further investigation into the customization of the Vision Transformer architecture through comprehensive hyperparameter tuning is recommended, with the aim of optimizing model accuracy and performance. Moreover, the development of a mobile application integrating the refined model is proposed, which would serve as a user-friendly tool to support clinical decision-making in the diagnosis of autism.

ACKNOWLEDGE

We would like to acknowledge Universitas Syiah Kuala and to all parties that have contributed to this work.

REFERENCES

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

[1] A. Miranda, C. Berenguer, I. Baixauli, and B. Roselló, "Childhood language skills as predictors of social, adaptive and behavior outcomes of adolescents with autism spectrum disorder," *Res Autism Spectr Disord*, vol. 103, May 2023, doi: 10.1016/j.rasd.2023.102143.

[2] E. C. McCanlies *et al.*, "Parental occupational exposure to solvents and autism spectrum disorder: An exploratory look at gene-environment interactions," *Environ Res*, vol. 228, Jul. 2023, doi: 10.1016/j.envres.2023.115769.

[3] Z. Wu, "Challenges Encountered by Children with Autism Spectrum Disorder: from the perspective of academic performances and education service providers," 2022.

[4] M. E. Andonovski and G. S. Antonarakis, "Autism spectrum disorder and dentoalveolar trauma: A systematic review and meta-analysis," *J Stomatol Oral Maxillofac Surg*, vol. 123, no. 6, pp. e858–e864, Nov. 2022, doi: 10.1016/j.jormas.2022.06.026.

[5] M. Romero-González *et al.*, "EEG abnormalities and clinical phenotypes in pre-school children with autism spectrum disorder," *Epilepsy and Behavior*, vol. 129, Apr. 2022, doi: 10.1016/j.yebeh.2022.108619.

[6] L. K. Koegel, R. L. Koegel, K. Ashbaugh, and J. Bradshaw, "The importance of early identification and intervention for children with or at risk for autism spectrum disorders," *Int J Speech Lang Pathol*, vol. 16, no. 1, pp. 50–56, Feb. 2014, doi: 10.3109/17549507.2013.861511.

[7] A. M. Grabrucker, "Autism Spectrum Disorders."

[8] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *Int J Multimed Inf Retr*, vol. 11, no. 1, pp. 19–38, Mar. 2022, doi: 10.1007/s13735-021-00218-1.

[9] P. Moridian *et al.*, "Automatic autism spectrum disorder detection using artificial intelligence methods with MRI neuroimaging: A review."

[10] M. Melinda, F. Arnia, A. Yafi, N. Afny, C. Andryani, and K. A. Enriko, "Design and Implementation of Mobile Application for CNN-Based EEG Identification of Autism Spectrum Disorder," vol. 14, no. 1, 2024.

[11] M. N. A. Tawhid, S. Siuly, H. Wang, F. Whittaker, K. Wang, and Y. Zhang, "A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from EEG," *PLoS One*, vol. 16, no. 6 June, Jun. 2021, doi: 10.1371/journal.pone.0253094.

[12] H. Yan, V. Mubonanyikuzo, T. E. Komolafe, L. Zhou, T. Wu, and N. Wang, "Hybrid-RViT: Hybridizing ResNet-50 and Vision Transformer for Enhanced Alzheimer's disease detection," *PLoS One*, vol. 20, no. 2, p. e0318998, Feb. 2025, doi: 10.1371/journal.pone.0318998.

[13] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[14] M. F. Almufareh, S. Tehsin, M. Humayun, and S. Kausar, "Artificial Cognition for Detection of Mental Disability: A Vision Transformer Approach for Alzheimer's Disease," *Healthcare (Switzerland)*, vol. 11, no. 20, Oct. 2023, doi: 10.3390/healthcare11202763.

[15] M. J. Alhaddad *et al.*, "Diagnosis Autism by Fisher Linear Discriminant Analysis FLDA via EEG," 2012.

[16] M. Melinda, M. Oktiana, Y. Yunidar, N. Hasna Nabila, I. Ketut, and A. Enriko, "International Journal on Informatics Visualization journal homepage : www.joiv.org/index.php/joiv International Journal on Informatics Visualization Classification of EEG Signal using Independent Component Analysis and Discrete Wavelet Transform based on Linear Discriminant Analysis." [Online]. Available: https://malhaddad.kau.edu.sa/Pages-BCI-Datasets-En.aspx,

[17] M. Melinda, F. H. Juwono, I. K. A. Enriko, M. Oktiana, S. Mulyani, and K. Saddami, "Application Of Continuous Wavelet Transform And Support Vector Machine For Autism Spectrum Disorder Electroencephalography Signal Classification," *Radioelectronic and Computer Systems*, no. 3(107), pp. 73–90, 2023, doi: 10.32620/reks.2023.3.07.

[18] Dr. Mohammed Jaffer Alhaddad, "BCI Datasets at King AbdulAziz University." Accessed: Feb. 11, 2025. [Online]. Available: https://malhaddad.kau.edu.sa/Pages-BCI-Datasets-En.aspx

[19] M. Lathifa, A. Sarna, M. R. Hossain, and M. A. Islam, "Comparative Analysis of STFT and Wavelet Transform in Time-Frequency Analysis of Non-Stationary Signals," *International Journal of Novel Research in Engineering and Science*, vol. 11, pp. 72–78, 2024, doi: 10.5281/zenodo.11230337.

[20] N. Ji, L. Ma, H. Dong, and X. Zhang, "EEG signals feature extraction based on DWT and EMD combined with approximate entropy," *Brain Sci*, vol. 9, no. 8, Aug. 2019, doi: 10.3390/brainsci9080201.

[21] N. McCallan *et al.*, "Seizure Classification of EEG based on Wavelet Signal Denoising Using a Novel Channel Selection Algorithm," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.00852

[22] S. Chatterjee, R. S. Thakur, R. N. Yadav, L. Gupta, and D. K. Raghuvanshi, "Review of noise removal techniques in ECG signals," Dec. 01, 2020, *Institution of Engineering and Technology*. doi: 10.1049/iet-spr.2020.0104.

[23] S. Phadikar, N. Sinha, and R. Ghosh, "Automatic Eyeblink Artifact Removal from EEG Signal Using Wavelet Transform with Heuristically Optimized Threshold," *IEEE J Biomed Health Inform*, vol. 25, no. 2, pp. 475–484, Feb. 2021, doi: 10.1109/JBHI.2020.2995235.

[24] M. A. Maria, M. A. H. Akhand, A. B. M. A. Hossain, M. A. S. Kamal, and K. Yamada, "A Comparative Study on Prominent Connectivity Features for Emotion Recognition From EEG," *IEEE Access*, vol. 11, pp. 37809–37831, 2023, doi: 10.1109/ACCESS.2023.3264845.

[25] L. Cao *et al.*, "A Novel Deep Learning Method Based on an Overlapping Time Window Strategy for Brain–Computer Interface-Based Stroke Rehabilitation," *Brain Sci*, vol. 12, no. 11, Nov. 2022, doi: 10.3390/brainsci12111502.

[26] T. Xu, Y. Zhou, Z. Hou, and W. Zhang, "Decode Brain System: A Dynamic Adaptive Convolutional Quorum Voting Approach for Variable-Length EEG Data," *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/6929546.

[27] M. Waseem Sabir, M. Farhan, N. S. Almalki, M. M. Alnfiai, and G. A. Sampedro, "FibroVit—Vision transformer-based framework for detection and classification of pulmonary fibrosis from chest CT images," *Front Med (Lausanne)*, vol. 10, 2023, doi: 10.3389/fmed.2023.1282200.

[28] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," Dec. 2020, doi: 10.1080/00401706.2021.1921037.

[29] T. Pronk, D. Molenaar, R. W. Wiers, and J. Murre, "Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment," 1948, doi: 10.3758/s13423-021-01948-3/Published.

[30] K. M. Kahloot and P. Ekler, "Algorithmic Splitting: A Method for Dataset Preparation," *IEEE Access*, vol. 9, pp. 125229–125237, 2021, doi: 10.1109/ACCESS.2021.3110745.

[31] K. Al-hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: a tutorial and survey," Dec. 01, 2023, *Springer*. doi: 10.1186/s42492-023-00140-9.

[32] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," May 01, 2023, *MDPI*. doi: 10.3390/app13095521.

[33] M. A. Mulkey, H. Huang, T. Albanese, S. Kim, and B. Yang, "Supervised deep learning with vision transformer predicts delirium using limited lead EEG," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-35004-y.

[34] D. Wang, J. Lian, H. Cheng, and Y. Zhou, "Music-evoked emotions classification using vision transformer in EEG signals," *Front Psychol*, vol. 15, 2024, doi: 10.3389/fpsyg.2024.1275142.

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia

AUTHOR BIOGRAPHY

Aufa Rafiki was born on April 20, 2003, in Banda Aceh. He is a student at Department of Electrical and Computer Engineering, Universitas Syiah Kuala. His undergraduate studies focus on multimedia technology, and his research explores EEG signals. He actively participates in class and continues to develop his knowledge in his field. In addition to his studies, he has experience as a teaching assistant and a programming lab assistant from his second to fourth semester. Enrolled in the 2021 cohort, he is committed to expanding his expertise and gaining practical experience. His academic journey reflects his dedication to both theoretical and applied learning, preparing him to contribute to technological advancements in his field. He can be contacted at aufarafi21@mhs.usk.ac.id.

Melinda was born in Bireuen, Aceh, on June 10, 1979. She received a B.Eng degree from the Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh in 2002. She completed her master's degree at the Faculty of Electrical Department, University of Southampton, United Kingdom, with a concentration in field study of Radio Frequency Communication Systems in 2009. She has already completed her Doctoral degree at the Department of Electrical Engineering, Engineering Faculty of Universitas Indonesia in February 2018. She has been with the Department of Electrical Engineering, Faculty of Engineering, Universitas Syiah Kuala since 2002. She is also a member of IEEE. Her research interests include multimedia signal processing and fluctuation processing. She can be contacted at email: melinda@usk.ac.id.

Maulisa Oktiana received B.Eng. degree in electrical engineering from Universitas Syiah Kuala (USK) in 2013. She received her Ph.D. degree in Electrical and Computer Engineering from Universitas Syiah Kuala in 2020. She was awarded a scholarship from the Ministry of Research, Technology, and Higher Education, the Republic of Indonesia, under the Scheme of Pendidikan Magister menuju Doktor untuk Sarjana Unggul (PMDSU). She visited Chiba University as an exchange student in November 2018. She is currently a lecturer in the Electrical and Computer Engineering Department at Universitas Syiah Kuala. Her research interests are image processing, biometric and pattern recognition. She can be contacted at maulisaoktiana@usk.ac.id.

Dr. Afnan Afnan is an Assistant Professor in Electrical and Computer Engineering at Universitas Syiah Kuala, Indonesia. Previously, she was an Adjunct Lecturer at the University of Illinois Chicago (UIC), USA. She holds a Bachelor's in Computer Engineering from Institut Teknologi Sepuluh Nopember (ITS), Indonesia, a Master's in Management of Technology from the University of Melbourne, Australia, a Master's in Computing from the University of Manchester, UK, and a Ph.D. in Management Information Systems from UIC. Her research focuses on health IT, technology consequences, healthcare networks, and machine learning. Recognized with prestigious awards like the Chevening and Fulbright Scholarships, she has presented at AMCIS and ICIS conferences. She can be contacted at afnan@usk.ac.id.

Ernita Dewi Meutia earned her Bachelor's degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, in 1991. She obtained a Postgraduate Diploma in Electronics and IT from the University of Birmingham, UK, in 1996. In 2009, she completed her Master's degree in Telecommunication Engineering at the University of Arkansas, USA, under a Fulbright Scholarship. She has been a faculty member at Universitas Syiah Kuala since 1992. Her research interests include telecommunication engineering and deep learning, and she has published several papers in these areas. She is actively involved in academic and research activities. She can be contacted at ernita.dmeutia@usk.ac.id.

Mulyadi was born in Punteut, Lhokseumawe, on October 28, 1976. He earned his Bachelor's degree in Applied Science from Institut Teknologi Sepuluh Nopember, Surabaya, majoring in Electrical Engineering - Information Technology. He then pursued his Postgraduate studies in Electrical Engineering at Universitas Sumatera Utara, Medan. Currently, he serves as a faculty member at Politeknik Negeri Lhokseumawe (PNL). Throughout his career, he has held various professional roles at PNL, including teaching staff, practitioner, and expert consultant for both government and non-government institutions. His research interests focus on Electrical Engineering and Information Engineering, and he actively engages in research and community service. He can be contacted at mulyadi@pnl.ac.id

Dr. Lailatul Qadri Zakaria earned her Ph.D. from the University of Southampton, U.K. She is currently a Senior Lecturer at the Centre of Artificial Intelligence (CAIT), Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM). She is also a member of the Asian Language Processing (ASLAN) research group. Her research interests include natural language processing (NLP), computational linguistics, and semantic web technologies. She has contributed to various studies on text analysis and machine learning for NLP. With extensive academic experience, she actively participates in scientific publications, research collaborations, and mentoring students in the field of artificial intelligence and language technology. She can be contacted at lailatul.qadri@ukm.edu.my

Corresponding author: Melinda, melinda@usk.ac.id, Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Darussalam, Banda Aceh, 23111, Aceh, Indonesia