IJIE

INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

Vol. 18 No. 1 (2026) 242-253
https://publisher.uthm.edu.my/ojs/index.php/ijie

Deep Learning-Based Classification of Anesthesia Depth Using EEG and Bispecral Index

Nurul Izzati Che Abdul Patah[1], Julie Roslita Rusli[1]*, Izanoordina Ahmad[1], Balqis Budiman[1], Husna Hamza[1], Wan Maziyah Ab. Halim[2], Sairul Safie[3], Rohana Sapawi[4], Sohiful Anuar Zainol Murad[5], Nor Hidayah Kahar[1], Mohd Zubir Suboh[1]

[1] *Electronics Technology Section, British Malaysian Institute, Universiti Kuala Lumpur, Batu 8, Jalan Sungai Pusu, Gombak, Selangor, 53100, MALAYSIA*

[2] *IC Design & Packaging, Semiconductor R & D, MIMOS Bhd., Technology Park Malaysia, Bukit Jalil, 57000, Kuala Lumpur, MALAYSIA*

[3] *Department of Plan Engineering, Malaysia Institute of Industrial Technology, Universiti Kuala Lumpur, Johor Baharu, Johor, 81750, MALAYSIA*

[4] *Faculty of Engineering,Universiti Malaysia Sarawak, Jln. Datuk Mohammad Musa, Kota Samarahan, 94300, MALAYSIA*

[5] *Fakulti Kejuruteraan & Teknologi Elektronik, Universiti Malaysia Perlis, Kampus Alam, Pauh Putra, 02600, Arau, Perlis, MALAYSIA*

*Corresponding Author: julie@unikl.edu.my
DOI: https://doi.org/10.30880/ijie.2026.18.01.018

Article Info

Received: 1 January 2026
Accepted: 30 March 2026
Available online: 20 April 2026

Keywords

Depth of anesthesia (DoA), electroencephalography (EEG), Bispectral Index (BIS), deep learning, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM)

Abstract

Electroencephalography (EEG) is a broadly used method designed for measuring the depth of anesthesia (DoA) due to its ability to replicate the brain's state and surgical pain. However, factors such as inaccurate assessment of DoA which can lead to unintended awareness and postoperative complications make precise monitoring difficult. Traditional methods of DoA monitoring are also time consuming, as they require continues attention to patient's vital sign during surgery. The objective of this study is to improve the classification of DoA by using deep learning techniques. We trained and evaluated Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and a combined CNN-LSTM model on a dataset of EEG signals from 30 patients undergoing general anesthesia, alongside corresponding Bispectral Index (BIS) values. The CNN model achieved moderate performance, with its best results in the General anesthesia class with the precision: 65.95%, recall: 55.99%, F1-score: 60.56%. The LSTM model achieved an improved F1-score of 85.13% for the same class. The hybrid CNN–LSTM model produced the best overall performance, achieving 65.78% accuracy, 75.00% specificity, and an F1-score of 79.36%, demonstrating the efficiency of hybrid deep learning for EEG-based DoA classification.

This is an open access article under the CC BY-NC-SA 4.0 license.

1. Introduction

The brain is the central organ responsible for regulating physiological and cognitive functions, and its activity is highly sensitive to pharmacological interventions during surgery. The administration of anesthetic agents induces characteristic changes in neural dynamics, making reliable intraoperative monitoring essential for maintaining appropriate anesthesia depth and patient safety. Electroencephalography (EEG) provides a non-invasive technique for capturing brain electrical activity with high temporal resolution through scalp-mounted electrodes. Due to its ability to reflect real-time cerebral responses to anesthetic drugs, EEG has been widely adopted in anesthesia monitoring systems. From a biomedical signal processing perspective, EEG offers rich temporal and spectral information that can be exploited for quantitative analysis, feature extraction, and the development of automated anesthesia assessment algorithms. [1], [2].

During general anesthesia, where intravenous (IV) drugs are used, continuous EEG monitoring is urged to minimize the risk of unintended awareness [3], [4]. In current clinical practice, anesthesiologists commonly evaluate the depth of anesthesia (DoA) by observing physiological parameters of an arterial blood pressure, heart rate, and oxygen saturation. However, these indicators are highly patient-dependent and can be affected by factors including age, body mass, sex, surgical conditions, and pre-existing medical status. As a result, assessing the level of consciousness based solely on vital signs remains challenging and may not reliably reflect the actual anesthetic state [5]. To address this limitation, monitoring the cerebral effects of anesthetic agents through electroencephalographic (EEG) signals has been recommended, as EEG provides a more direct representation of brain activity and helps reduce the risk of intraoperative awareness. The Bispectral Index (BIS), a widely used EEG-derived parameter, offers a quantitative measure of anesthesia depth. Nevertheless, BIS-based monitoring alone may be insufficient to fully characterize the complex and nonlinear dynamics of brain activity, mostly in the presence of inter-patient variability, multiple drug interactions, and EEG artifacts. Motivated by these challenges, this study emphasis on classifying the depth of anesthesia using EEG signals in conjunction with BIS values. By extracting informative features from EEG recordings and employing deep learning-based models, the proposed approach aims to improve the robustness and precision of DoA estimation.

In light of these challenges, there is a clear demand for accurate and computationally efficient systems capable of monitoring and estimating the depth of anesthesia (DoA) using EEG signals. Previous studies have discovered both conventional machine learning and deep learning approaches for this purpose. In [6], a machine learning framework was introduced in which time-series EEG features were extracted and classified using support vector machines and random forest models. The reported classification accuracy of 83% proven that traditional learning algorithms can effectively exploit EEG characteristics for DoA estimation. In contrast, a more advanced deep learning architecture incorporating convolutional neural networks (CNNs), bidirectional long short-term memory (LSTM) units, and an attention mechanism was presented in [5]. By enabling the model to selectively emphasize informative temporal and spatial EEG features, the attention-enhanced framework achieved improved adaptability and a higher classification accuracy of 88.7%. Additionally, a hybrid approach combining multiple EEG feature extraction methods with LSTM networks and stacked denoising autoencoders (SDAE) was proposed in [7], highlighting the potential of integrating feature learning and temporal modeling to further enhance DoA classification performance. This system significantly enhances prediction accuracy and the reliability of DoA estimation, showcasing the benefits of combining diverse feature extraction methods. Aside from that, a patient-specific decision tree classifier-based system was developed in [8] to provide accurate DoA estimation. This system emphasized individual variability by tailoring the model to specific patients, achieving a reliable classification with an accuracy of 79%. This approach showcases the potential for personalization in EEG signal processing for anesthesia depth monitoring. Additionally, in [9], a novel system was developed that utilizes a Multilayer Perceptron (MLP) classifier combined with Stationary Wavelet Transform (SWT) features to monitor DoA. This system effectively identifies the patient's consciousness level across various anesthetic agents and age groups, achieving an impressive accuracy of 96.8%. On the contrary, a different approach was explored in, where a CNN-LSTM hybrid model was designed to leverage both spatial and temporal features from EEG signals. This system can provide real-time predictions of DoA, enhancing patient safety during surgical procedures.

Although previous studies show that machine learning and deep learning have great potential for anesthesia monitoring, some important challenges still exist. Traditional machine learning relies heavily on manually crafted features, which can be time-consuming and might miss important EEG details. Advanced deep learning methods, like attention-based and hybrid models, have achieved high accuracy but often require large amounts of data, demand heavy computation, and may struggle to perform well under varying surgical conditions. Additionally, the complex nature of EEG signals such as their nonlinearity, noise, and differences between patients makes consistent depth of anesthesia (DoA) classification difficult.

2. Methodology

The methodology adopted in this study involves of three main components: the system block diagram, process flow, and flowchart. Figure 1 presents the block diagram, which illustrates the complete workflow of the proposed system. The process begins with data input, where the experimental dataset used in this work was obtained from the Department of Anesthesiology and Pain Medicine, Seoul National University College of Medicine and Seoul National University Hospital, Republic of Korea. This dataset is publicly available, with detailed documentation provided in [10], and comprises recordings from 30 patients who underwent general anesthesia. All participants were aged 18 years and above, and sevoflurane was administered as the primary anesthetic agent during the surgical procedures. EEG signal acquisition was conducted using a commercial EEG-based anesthesia monitoring device, the BIS Vista system (Medtronic, Minneapolis, MN, USA), which estimates the depth of anesthesia (DoA) through a numerical index ranging from 0, representing an isoelectric EEG, to 100, indicating a fully awake state. EEG signals were recorded using the BIS-QUATRO sensor (Aspect Medical Systems, Newton, MA, USA), consisting of four electrodes arranged into two channels and positioned on the patient's forehead. The raw EEG data were sampled at a frequency of 128 Hz, while the BIS index values were updated at a rate of one sample per second. For data preparation, the raw EEG recordings stored in the vital format were accessed using the Vital Recorder software. This software enabled the extraction and conversion of the EEG data into an Excel-compatible format for further analysis. All preprocessing, feature extraction, and classification tasks were implemented using PyCharm as the integrated development environment (IDE). The primary stages of the proposed framework include feature extraction from raw EEG signals, application of deep learning-based classification models, and performance evaluation of the system in categorizing DoA into four clinically relevant levels: Awake (BIS 80–100), Light Sedation (BIS 60–80), General Anesthesia (BIS 40–60), and Deep Anesthesia (BIS 0–40).

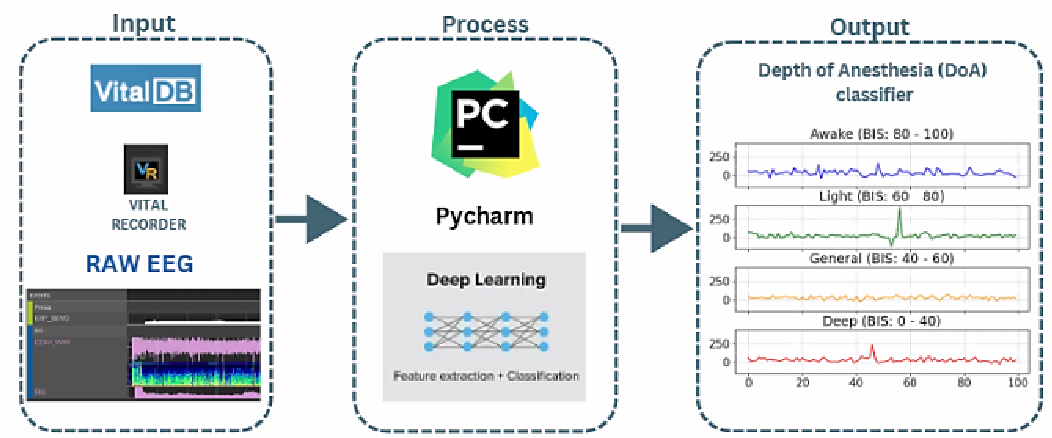


Fig. 1 *Block diagram of project*

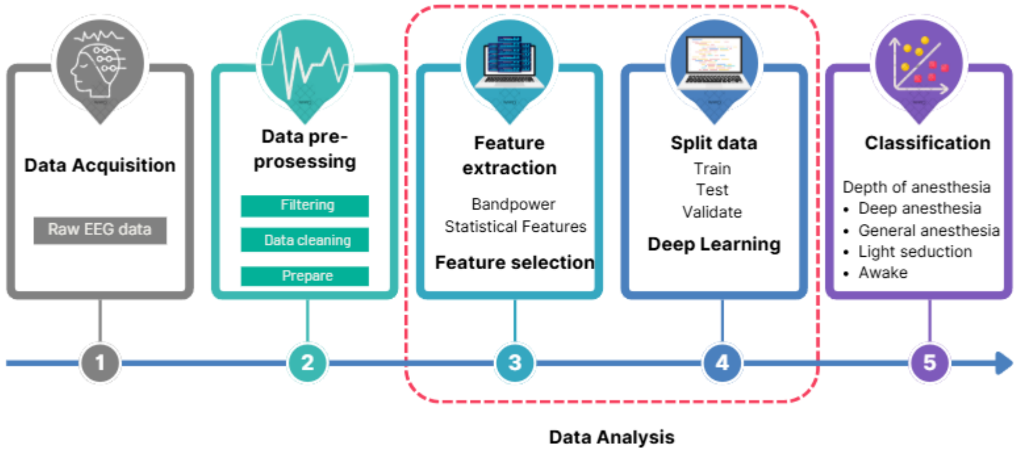


Fig. 2 *Process flow of project*

Figure 2 illustrates the detailed process flow involved in the classification of anesthesia depth using EEG signals. The framework comprises five sequential stages: data acquisition, data pre-processing, feature extraction and selection, data splitting with deep learning, and classification [11]. The details of each stage of process flow project is explained as follows:

2.1 Data Acquisition

The data acquisition process initiates the analytical pipeline by importing the EEG dataset and converting it into a structured format, such as Microsoft Excel, to facilitate ease of processing. The raw EEG signals are recorded using standard EEG acquisition systems, typically from patients undergoing different depths of anesthesia. These signals serve as the primary data source and provide rich temporal information about brain electrical activity, which is essential for evaluating anesthesia levels.

2.2 Data Pre-processing

Once the raw EEG data is acquired, pre-processing is performed to prepare it for meaningful analysis. Initially, a bandpass filter with cut-off frequencies of 0.5 Hz and 47 Hz is applied to the signal. This frequency range is selected because it captures the most relevant EEG activity associated with varying levels of anesthesia [12]. Subsequently, data cleaning procedures are carried out by removing all BIS (Bispectral Index) values of zero and any missing or NaN entries. These values are typically observed during the pre-induction phase of anesthesia and do not contain valuable information [13]. As the dataset is already structured in 1-second epochs with corresponding BIS annotations, additional segmentation is not required. This stage ensures that only clean and relevant data are passed to the next phase, improving the reliability of feature extraction and classification.

2.3 Feature Extraction and Feature Selection

Following the preprocessing stage, feature extraction is performed to distinguish the discriminative characteristics of the EEG signals. In this work, two feature extraction techniques are employed to analyze raw EEG recordings for the classification of Bispectral Index (BIS) values, which are used as quantitative indicators of the depth of anesthesia. The first technique utilizes the bandpower function, which computes the signal power within predefined EEG frequency bands. These spectral bands are known to be associated with distinct physiological and cognitive states and are therefore particularly relevant for characterizing brain activity under anesthesia. The frequency ranges considered in this study include Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–47 Hz) [13]. For instance, high power in the Delta band indicates deeper levels of anesthesia, while reduced Beta power suggests decreased cognitive activity. This function computes the band power for a specified frequency range using Welch's method to estimate the Power Spectral Density (PSD). The bandpower can be mathematically expressed as:

$$\text{Bandpower}=\int_{f2}^{f1} PSD(f)\, df \tag{1}$$

Where the $f1$ and $f2$ are the lower and upper frequency limits of the band (Delta, Theta, Alpha, Beta, Gamma). PSD (f) is the power spectral density at frequency (f). The second method involves calculating Statistical Features from the EEG signal, such as mean, variance, standard deviation, maximum, and minimum values. Statistical features are essential for understanding EEG signals, as they reveal important characteristics of the signal and offer meaningful insights into the depth of anesthesia. The following statistical measures are used in this study.

The second technique involves calculating Statistical Features, which provide complementary insights into the signal dynamics. These include the mean, variance, standard deviation, maximum, and minimum values of the EEG signal. The mean μ is calculated as:

$$\text{Mean} = \frac{1}{N}\sum_{i=1}^{N} xi \tag{2}$$

where N is the total number of samples and xi denotes each individual sample. A lower mean value may indicate reduced brain activity, which is often associated with deeper levels of anesthesia. Variance quantifies the spread of the signal values around the mean and is computed as:

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N} xi - Mean)^2 \tag{3}$$

High variance in the EEG signal may suggest increased brain activity, indicating that the patient could be in a lighter state of anesthesia. This statistic reflects the variability of the signal and is derived from variance:

$$\text{Standard Deviation} = \sqrt{Variance} \tag{4}$$

A high standard deviation indicates greater variability in brain activity, which can be relevant for assessing anesthetic depth. The maximum value of the EEG signal is determined as follows:

$$\text{Max} = max(x1,x2,\ldots,xN) \tag{5}$$

This value provides insights into peak brain activity levels during different states of consciousness. The minimum value is calculated as:

$$\text{Min} = min(x1,x2,\ldots,xN) \tag{6}$$

Understanding both maximum and minimum values aids in evaluating the overall range of brain activity.

2.4 Data Splitting

The dataset was subsequently divided into two different set of training dataset and testing datasets, where 80% of the samples were used for training process and another 20% were reserved for testing the data. The training data were utilized to adjust and optimize the parameters of the deep learning model, allowing the model to capture meaningful patterns related to different levels of anesthesia depth. Meanwhile, the testing dataset was employed to evaluate the model's generalization ability and overall classification performance on unseen data [14]. To address the issue of class imbalance, the RandomUnderSampler method from the imbalanced-learn (imblearn) library was implemented. Class imbalance arises when some classes contain significantly more samples than others, which may cause the model to favor the majority classes during training and consequently reduce prediction accuracy. The RandomUnderSampler technique mitigates this problem by randomly reducing the number of samples in the majority classes so that they match the size of the minority classes or follow a specified sampling ratio. This process retains all samples from the minority classes while balancing the distribution across categories.

In this study, RandomUnderSampler was applied to ensure that the deep learning models learn from a balanced dataset and avoid bias toward the dominant class [15]–[17]. Figure 3 shows the class distribution before and after the undersampling process. The blue bars represent the original dataset, where a significant imbalance is observed; for instance, the majority class contains 213,660 samples, whereas some minority classes contain only 6,300 samples. After applying undersampling, the class distributions (shown by the orange bars) become balanced, with each BIS category containing 6,300 samples.

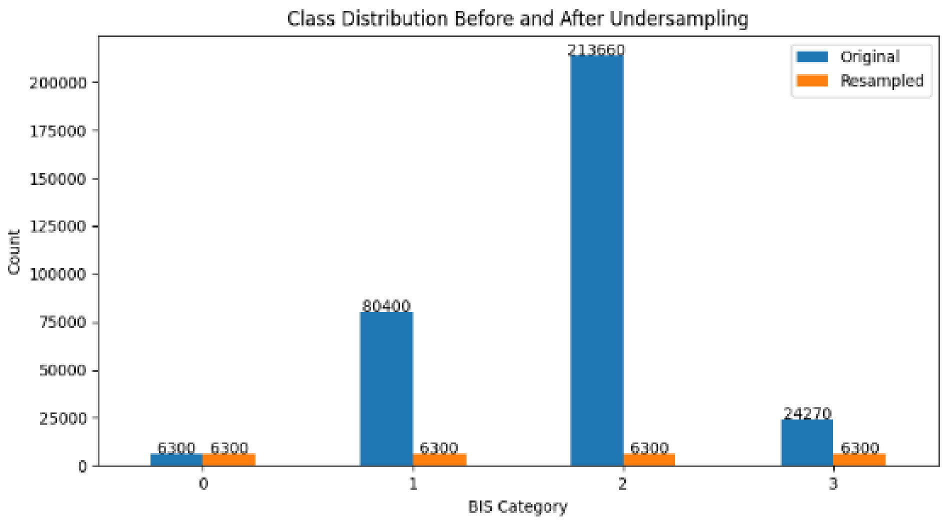


Fig. 3 *Class distribution before and after undersampling*

2.5 Deep Learning Model Development

Nexrt stage is to describes the application of deep learning models for depth of anesthesia (DoA) classification using EEG-derived features. Three architectures were investigated: a Convolutional Neural Network (CNN), a Long Short-Term Memory (LSTM) network, and a hybrid CNN–LSTM model. These models were selected to examine their ability to learn spatial patterns, temporal dependencies, and combined spatiotemporal representations from EEG signals.

2.5.1 CNN Model

A Convolutional Neural Network (CNN) was employed to perform anesthesia depth classification using extracted EEG features. While CNNs are widely recognized for image processing applications, their ability to automatically learn hierarchical representations also makes them suitable for one-dimensional time-series signals, including EEG recordings [18]–[20]. The proposed CNN architecture develop with a one-dimensional convolutional layer containing 64 filters with a kernel size of 3. This process enables the model for capturing the localized feature patterns generate from the input sequence. The acitivatgion of the Rectified Linear Unit (ReLU) function is to introduce nonlinearity, and L2 regularization is included to control model complexity and reduce overfitting. A MaxPooling1D layer is subsequently used to downsample the extracted feature maps, thereby reducing dimensionality while preserving relevant information. Batch normalization is incorporated to stabilize gradient updates and improve training convergence.The resulting feature maps are then flattened and passed to a fully connected layer with 32 neurons using ReLU activation, allowing the network to learn higher-level representations of the extracted features. To further improve generalization, a rate of 0.2 of Dropout layer is introduced before the final classification layer and the output layer uses a softmax activation function to allocate each input sample to one of the predefined anesthesia depth classes. Model training was accompanied for 50 epochs with a batch size of 32, and 20% of the training data was reserved for validation.

2.5.2 LSTM Model

To capture temporal characteristics, present in EEG signals, a Long Short-Term Memory (LSTM) network was implemented. LSTM is an advanced variant of recurrent neural networks (RNNs) designed to retain relevant information across long sequences, making it particularly effective for time-dependent signal analysis such as EEG data [21], [22]. The LSTM architecture begins with a layer containing 128 memory units, configured with return sequences=True to maintain the sequential structure of the output. L2 regularization with a coefficient of 0.001 is applied to limit model overfitting. A dropout layer (0.2) is then introduced to improve model generalization, trailed by batch regularization to stabilize the training process.

A second LSTM layer with 64 units is subsequently applied with return sequences=False, producing a fixed-length feature representation for classification. Similar regularization strategies, including L2 regularization, dropout (0.2), and batch normalization, are combined to improve training stability and prevent overfitting. The temporal representations extracted by the LSTM layers are then forwarded to a compressed layer containing 64 neurons with ReLU activation, enabling further feature refinement. Additional regularization and dropout are applied at this stage to improve robustness and decrease the risk of model overfitting.

2.5.3 CNN+LSTM Hybrid Model

To combine the advantages of spatial and temporal feature learning, a CNN–LSTM hybrid model was established. This architecture integrates convolutional layers for spatial feature extraction with LSTM layers for temporal modeling, enabling the network to capture both spatial and sequential features of EEG signals. The architecture begins with a CNN module that extracts spatial features from the bandpower and statistical attributes derived from EEG signals. These features contain important time-domain and frequency-domain information associated with varying anesthesia states [23], [24], [25]. The first stage consists of a one-dimensional convolutional layer with 32 filters and a kernel size of 3, which detects localized spatial patterns within the input features. A ReLU activation function introduces nonlinearity, while L2 regularization is applied to decrease the risk of overfitting. A MaxPooling1D layer is then used to reduce feature dimensionality while preserving relevant information. This is trailed by Batch Normalization, which stabilizes training and normalizes intermediate activations.

After spatial feature extraction, the model incorporates an LSTM layer with 16 units to capture temporal dependencies in the EEG signals. L2 regularization is applied to improve generalization, and a Dropout layer with a rate of 0.3 is included to enhance model robustness. The extracted temporal representations are subsequently passed to a dense layer with 32 neurons using ReLU activation, allowing the network to learn higher-level relationships between spatial and temporal features. An additional Dropout layer (0.3) is applied to further reduce overfitting. Finally, the softmax output layer generates probability estimates for the four predefined anesthesia depth classes. Model optimization is achieved using the Adam optimizer, while sparse categorical cross-entropy is applied as the loss function for multi-class classification.

3. Results & Discussion

Table 1 presents the classification outcomes of the CNN, LSTM, and hybrid CNN–LSTM models for determining the depth of anesthesia (DoA). The DoA is divided into four categories: Awake (0), Deep (1), General (2), and Light (3). The comparison shows variations in model effectiveness when evaluated using precision, recall, and F1-score. Overall, the hybrid CNN–LSTM model demonstrates more consistent performance across the classes, indicating its capability to capture both spatial features and temporal patterns from the EEG signals, which contributes to improved DoA classification accuracy. Precision represents the percentage of correctly predicted samples within a class, indicating the model's ability to minimize false-positive predictions. Recall, also referred to as sensitivity, measures the proportion of actual samples correctly identified by the model. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced performance indicator, particularly when class distributions are imbalanced [26], [27].

The CNN model proves its strongest performance in the General anesthesia class, achieving a precision of 65.95, recall of 55.99, and an F1-score of 60.56. However, its performance for the Awake and Deep classes remains limited, and the model does not correctly classify samples belonging to the Light class. This observation suggests that while CNN is capable of learning spatial patterns from EEG-derived inputs, it may be less effective in modeling the temporal characteristics of the signals. The LSTM model shows improved performance compared with CNN, particularly for the General anesthesia class, where it achieves a precision of 93.8, recall of 77.29, and an F1-score of 85.13. This improvement highlights the advantage of LSTM networks in capturing temporal dependencies present in sequential EEG data. Nevertheless, similar to the CNN model, the LSTM model exhibits reduced performance for the Awake and Deep classes and does not successfully identify the Light class.

From the evaluated models, the CNN–LSTM hybrid architecture achieves the most competitive overall performance by integrating spatial and temporal feature learning. For the General anesthesia class, the hybrid model achieves a recall of 100 and an F1-score of 79.36, demonstrating the effectiveness of combining convolutional feature extraction with LSTM-based temporal modeling. However, classification performance remains limited for the Awake and Deep classes, and the Light class continues to be difficult to identify. Compared with previous studies, the overall accuracy of 65.78% obtained in this work may appear lower than results reported in some earlier research. For example, accuracies of 83% using conventional machine learning techniques [6], 88.7% using CNN–BiLSTM–Attention models [5], and 96.8% using MLP–SWT methods [9] have been reported. However, these studies typically rely on larger datasets, handcrafted features, or more complex model architectures, which may reduce their practicality in real-time clinical environments.

Table 1 *Comparison of CNN, LSTM, and CNN+LSTM model performance*

DoA	CNN			LSTM			CNN+LSTM		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Awake (0)	2.05	20.99	3.74	2.57	1.623	1.99	0.00	0.00	0.00
Deep (1)	25.17	25.13	25.15	24.44	4.84	8.08	0.00	0.00	0.00
General (2)	65.95	55.99	60.56	65.72	93.8	77.29	65.78	100	79.36
Light (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OVERALL									
Accuracy	43%			63%			65.78%		
Sensitivity	25.53%			25.57%			25.00%		
Specificity	75.11%			23.69%			75.00%		

In contrast, the present study utilizes a relatively small dataset consisting of 30 patients, which naturally constrains the achievable performance. Despite this limitation, the outcomes demonstrate that the CNN–LSTM hybrid model provides measurable improvements compared with standalone CNN and LSTM models. Furthermore, its strong performance in classifying the General anesthesia state is clinically relevant, as accurate classification in this state is essential for preventing intraoperative awareness. Figure 4 presents a comparison of specificity, sensitivity, and accuracy among the three models. The CNN model achieves the highest specificity (75.11%), thoroughly followed by CNN–LSTM (75.00%), while the LSTM model exhibits considerably lower specificity (23.69%). Sensitivity values remain relatively low across all models. The LSTM model achieves the highest sensitivity (25.57%), slightly outperforming CNN (25.53%), while the CNN–LSTM model records 25.00%. In terms of overall classification accuracy, the CNN–LSTM hybrid model achieves the highest accuracy (65.78%), followed by LSTM (62.94%) and CNN (43.47%). These results indicate that integrating spatial and temporal feature learning can enhance classification performance. To further assess model effectiveness, confusion matrices and Receiver Operating Characteristic (ROC) curves are used. A confusion matrix summarizes classification outcomes by comparing actual and predicted labels, provided that insight into true positives, true negatives, false positives, and false negatives. This representation is particularly useful for analyzing classification behavior in imbalanced datasets. The ROC curve proves the relationship between the true positive rate and false positive rate across different decision thresholds. The Area Under the Curve (AUC) serves as a quantitative quantity of the model's discriminative capability [28]. An AUC value of 0.5 indicates performance equivalent to random classification, whereas values approaching 1 represent stronger discriminative performance. Together, these evaluation tools offer a comprehensive assessment of the classification models.

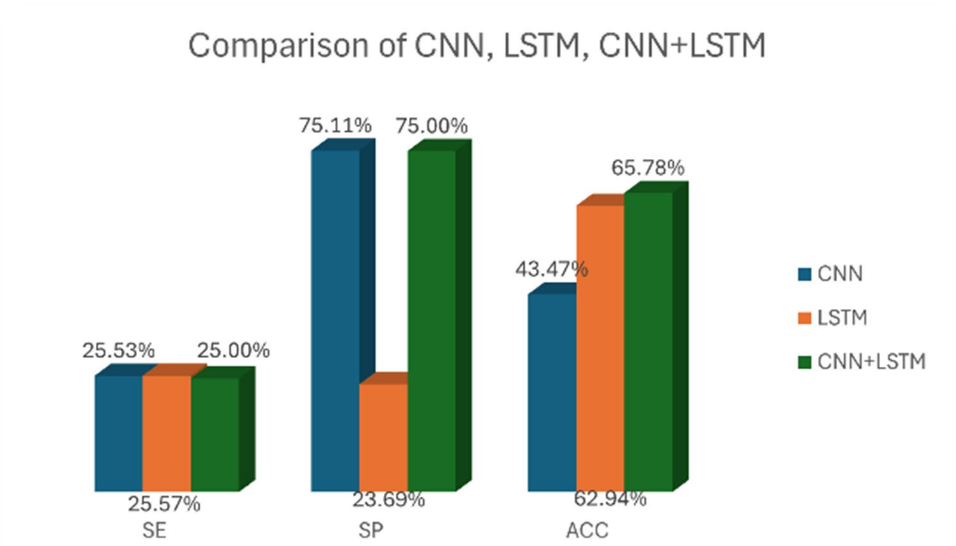
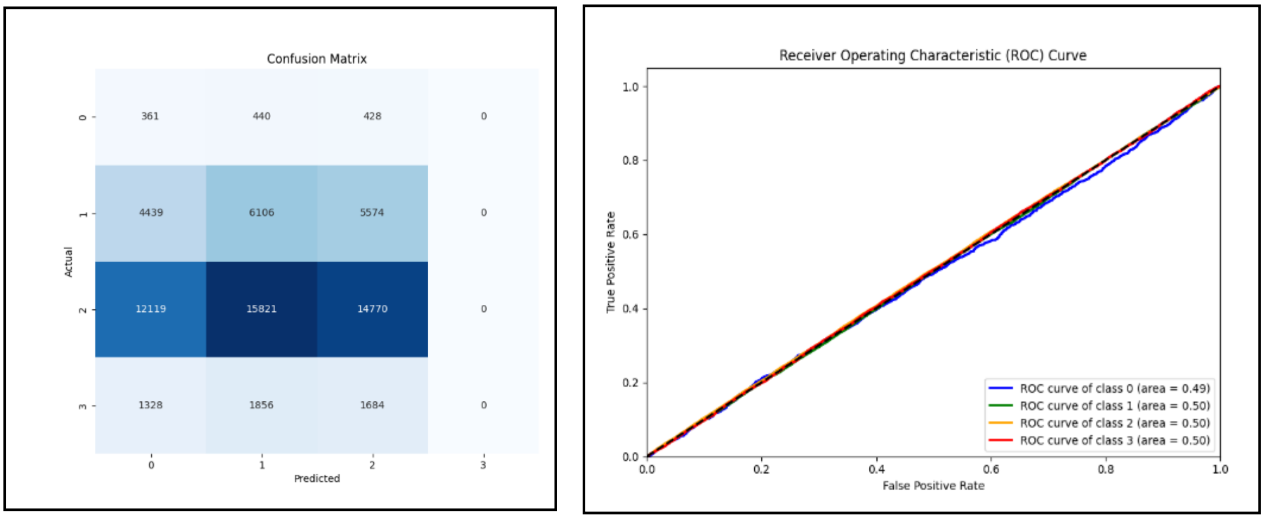


Fig. 4 *Bar chart of specificity, sensitivity, and accuracy comparison for CNN, LSTM, and CNN+LSTM models*



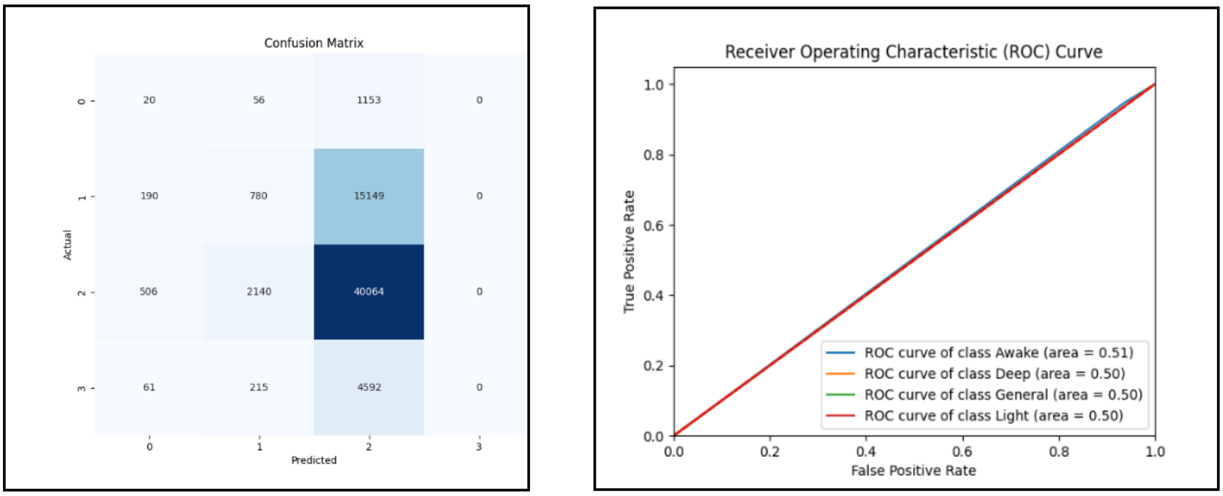
(a) **(b)**

Fig. 5 *(a) Confusion matrix of CNN model; (b) ROC curve of CNN model*

For the CNN model, the confusion matrix in Figure 5 (a) reveals misclassification across classes. Class 0 (awake) shows poor performance, with most instances misclassified into other classes, while Class 1 (Deep anesthesia) demonstrates moderate performance but is frequently confused with Class 2 (General anesthesia).

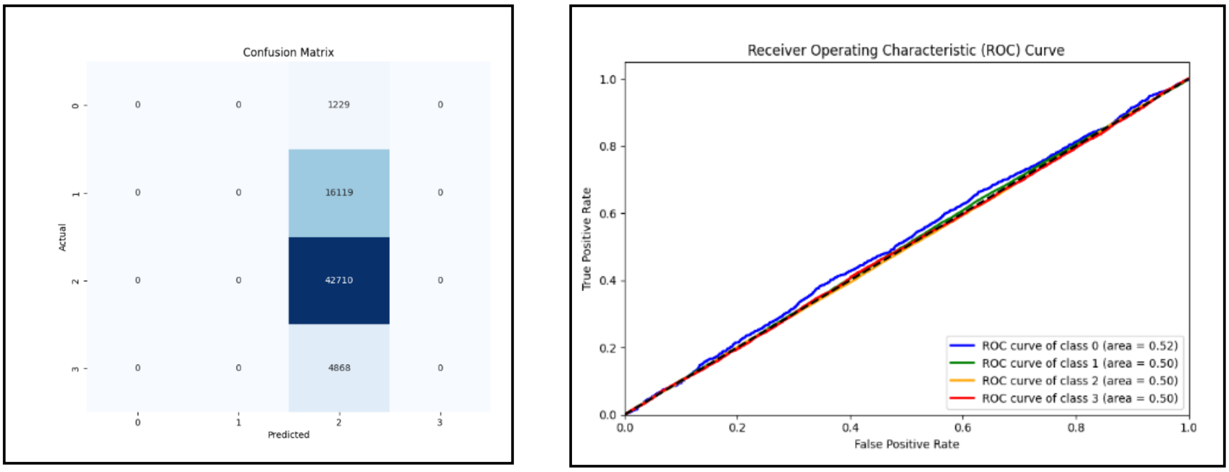
Class 3 (Light sedation) also exhibits poor predictions, with many samples being classified incorrectly into other categories. These results suggest that the CNN model has difficulty distinguishing between the different anesthesia states. This limitation is further supported by the ROC curves in Figure 5(b), where the curves for all classes lie close to the diagonal line and the AUC values are approximately 0.50, indicating that the model performs close to random classification.

The confusion matrix for the LSTM model in Figure 6 (a) highlights significant misclassification across various classes. Class 0 (awake) shows poor performance, with most instances being misclassified into other classes. Class 1 (deep anesthesia) demonstrates moderate performance but is frequently confused with Class 2 (general anesthesia). Despite having the largest representation, Class 2 suffers from widespread misclassifications, indicating that the model struggles with feature separation for this class. Most of the samples fall into other categories than the actual category. The ROC curves in Figure 6 (b) for this LSTM model highlight its limitations, as the curves for all classes remain near the diagonal, with AUC values close to 0.50. This demonstration that the model struggles to differentiate between the classes, reflecting weak classification performance.



(a) (b)

Fig. 6 *(a) Confusion matrix of LSTM model; (b) ROC curve of LSTM model*



(a) (b)

Fig. 7 *(a) Confusion matrix of CNN+LSTM model; (b) ROC curve of CNN+LSTM model*

The confusion matrix for the CNN+LSTM model in Figure 7 (a) reveals a strong tendency toward predicting a single class. All instances of Class 0 (awake), Class 1 (deep anesthesia), and Class 3 (light sedation) are inaccurately classified as Class 2 (general anesthesia). Although Class 2 records the highest number of correct predictions, some misclassifications into other classes are still observed. This pattern suggests that the CNN+LSTM model is heavily biased toward predicting Class 2, which may be influenced by issues such as class imbalance or ineffective feature learning during training. The ROC curves in Figure 7(b) further demonstrate the limitations of this model, as the curves for all classes remain close to the diagonal line and the AUC values are approximately 0.50. These results indicate that the model has unable to distinguish between the different classes.

4. Conclusion & Recommendations

The performance evaluation of CNN, LSTM, and CNN+LSTM models reveals notable strengths and weaknesses. The CNN model shows strong spatial feature extraction capabilities, with its best performance in the General class (precision: 65.95, recall: 55.99, F1-score: 60.56), but struggles significantly with temporal aspects of EEG signals, leading to poor performance in the Awake and Deep classes and complete failure in the Light class. The LSTM model outperforms CNN by capturing temporal dependencies, particularly excelling in the General class (precision: 93.8, recall: 77.29, F1-score: 85.13). However, it similarly struggles with the Awake and Deep classes and fails to classify the Light class. The CNN+LSTM hybrid model, which integrates spatial and temporal feature learning, achieves the best overall performance, with the highest accuracy (65.78%) and specificity (75.00%). It also performs exceptionally well in the General class, achieving a perfect recall of 100 and a high F1-score of 79.36. Despite these strengths, all models face significant challenges with the Light class and moderate difficulties in the Awake and Deep classes. Among the three models, the CNN+LSTM hybrid model demonstrates the best overall performance. By combining the strengths of CNN's spatial feature extraction and LSTM's temporal feature learning, it achieves the highest accuracy (65.78%), competitive specificity (75.00%), and robust performance in the General class. This makes it the most effective model for classifying the depth of anesthesia (DoA) across four classes.

Acknowledgement

The authors acknowledge the Universiti Kuala Lumpur British Malaysian Institute for sponsoring this research through the UERGS Grant (UER24002).

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Nurul Izzati Che Abdul Patah, Julie Roslita Rusli, Mohd Zubir Suboh;* ***data collection:*** *Nurul Izzati Che Abdul Patah, Izanoordina Ahmad, Balqis Budiman;* ***analysis and interpretation of results:*** *Nurul Izzati Che Abdul Patah, Julie Roslita Rusli, Mohd Zubir Suboh, ;* ***draft manuscript preparation:*** *Nurul Izzati Che Abdul Patah, Husna Hamza, Wan Maziyah Ab. Halim, Sairul Safie, Rohana Sapawi, Sohiful Anuar Zainol Murad, Nor Hidayah Kahar. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] D. Fattahi and R. Boostani, "An iterative spatio-spectral discriminant scheme for EEG classification," Iranian Journal of Science and Technology Transaction B-Engineering, vol. 36, no. E2, pp. 147–161, 2012.

[2] A. M. Brambrink and J. R. Kirsch, Essentials of neurosurgical anesthesia & critical care: strategies for prevention, early detection, and successful management of perioperative complications. Springer Nature, 2019.

[3] A. Nimmo, A. Absalom, O. Bagshaw, A. Biswas, T. Cook, A. Costello, S. Grimes, D. Mulvey, S. Shinde, T. Whitehouse et al., "Guidelines for the safe practice of total intravenous anaesthesia (TIVA) joint guidelines from the association of anaesthetists and the society for intravenous anaesthesia," Anaesthesia, vol. 74, no. 2, pp. 211–224, 2019.

[4] Y. Punjasawadwong, A. Phongchiewboon, and N. Bunchungmongkol,"Bispectral index for improving anaesthetic delivery and postoperative recovery," Cochrane Database of Systematic Reviews, no. 6, 2014.

[5] S. Afshar, R. Boostani and S. Sanei, "A Combinatorial Deep Learning Structure for Precise Depth of Anesthesia Estimation From EEG Signals," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3408-3415, Sept. 2021, doi: 10.1109/JBHI.2021.3068481.

[6] Anand, R. V., Abbod, M. F., Fan, S.-Z., & Shieh, J.-S. (2023). Depth Analysis of Anesthesia Using EEG Signals via Time Series Feature Extraction and Machine Learning. Sci, 5(2), 19. https://doi.org/10.3390/sci5020019

[7] F. H. Khan, U. Ashraf, M. A. Bin Altaf and W. Saadeh, "A Patient-Specific Machine Learning based EEG Processor for Accurate Estimation of Depth of Anesthesia," 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), Cleveland, OH, USA, 2018, pp. 1-4, doi: 10.1109/BIOCAS.2018.8584828.

[8] Li, T., Huang, Y., Wen, P. et al. Accurate depth of anesthesia monitoring based on EEG signal complexity and frequency features. Brain Inf. 11, 28 (2024). https://doi.org/10.1186/s40708-024-00241-y

[9] AlMeer, M.H., Abbod, M.F. (2019). Deep Learning in Classifying Depth of Anesthesia (DoA). In: Arai, K., Kapoor, S., Bhatia, R. (eds) Intelligent Systems and Applications. IntelliSys 2018. Advances in Intelligent Systems and Computing, vol 868. Springer, Cham. https://doi.org/10.1007/978-3-030-01054-6_11

[10] Lee, H.-C., & Jung, C.-W.(2018). Vital recorder—a free research tool for automatic recording of high-resolution time-synchronized physiological data from multiple anesthesia devices. Scientific Reports, 8 (1), 1–8.

[11] RUSLI J R, Syahirin Danial Noor Shahrin M, Binti Che Abdu Patah N I, Ahmad I, Mohamad Maharum S M, Izwan Safie S. Development of a Digital Stethoscope for Enhancing Real-Time Respiratory Diagnostics. IJEEE 2025; 21 (2) :3646-3646 URL: http://ijeee.iust.ac.ir/article-1-3646-en.html

[12] D. Jarchi, R. Boostani, M. Taheri, and S. Sanei, "Seizure source localization using a hybrid second order blind identification and extended rival penalized competitive learning algorithm

[13] S. Afshar and R. Boostani, "A Two-Stage Deep Learning Scheme to Estimate Depth of Anesthesia from EEG Signals," 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME), Tehran, Iran, 2020, pp. 7-12, doi: 10.1109/ICBME51989.2020.9319416.

[14] Hight, D. F., Kaiser, H. A., Sleigh, J. W., & Avidan, M. S. (2020). An updated introduction to electroencephalogram-based brain monitoring during intended general anesthesia. Canadian Journal of Anesthesia/Journal Canadien D Anesthésie, 67(12), 1858–1878. https://doi.org/10.1007/s12630-020-01820-3.

[15] Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. Journal of neural engineering, 16(3), 031001. https://doi.org/10.1088/1741-2552/ab0ab5

[16] Kaur, H., Pannu, H. S., & Malhi, A. K. (2020). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys, 52 (4), 1–36. https://doi.org/10.1145/3343440

[17] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[18] Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning (pp. 161–168). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143865

[19] Gatfan, S. (2024). A review on deep learning for electroencephalogram signal classification. Journal of Al-Qadisiyah for Computer Science and Mathematics, 16, 137–151. https://doi.org/10.29304/jqcsm.2024.16.11453

[20] Gupta, S. S., & Manthalkar, R. R. (2020). Classification of visual cognitive workload using analytic wavelet transform. Biomedical Signal Processing and Control, 61, 101961. https://doi.org/10.1016/j.bspc.2020.101961

[21] S. Shanmugapriya, P. Nagaraj, K. Ajay Kumar Reddy, S. Akshay, G. Bhanuprakash and C. Venkat, "Classification of Brain States using CNN under EEG Anesthesia," 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2023, pp. 1108-1113, doi: 10.1109/ICESC57686.2023.10192939.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.

[23] Y. Pamungkas, A. D. Wibawa and Y. Rais, "Classification of Emotions (Positive-Negative) Based on EEG Statistical Features using RNN, LSTM, and Bi-LSTM Algorithms," 2022 2nd International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), Jakarta, Indonesia, 2022, pp. 275-280, doi: 10.1109/ISMODE56940.2022.10180969

[24] M. Islam and T. Lee, "Wavelet based Emotion Detection from Multi-channel EEG using a Hybrid CNN-LSTM Model," TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON), Hong Kong, Hong Kong, 2022, pp. 1-6, doi: 10.1109/TENCON55691.2022.9978122.

[25] Ravichandra Madanu, Farhan Rahman, Maysam F. Abbod, Shou-Zen Fan, Jiann-Shing Shieh. Depth of anesthesia prediction via EEG signals using convolutional neural network and ensemble empirical mode decomposition[J]. Mathematical Biosciences and Engineering, 2021, 18(5): 5047-5068. doi: 10.3934/mbe.2021257

[26] I. Latreche, S. Slatnia and O. Kazar, "CNN-LSTM To Identify The Most Informative EEG-Based Driver Drowsiness Detection Brain Region," 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2022, pp. 725-730, doi: 10.1109/ISMSIT56059.2022.9932696

[27] Mossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5, 1–11. https://doi.org/10.5121/ijdkp.2015.5201

[28] Mendoza, J., Mycroft, J., Milbury, L., Kahani, N., & Jaskolka, J. (2022). On the effectiveness of data balancing techniques in the context of ML-based test case prioritization. Proceedings of the 18th International Conference on Predictive Models and Data Analytics in Software Engineering (pp. 72–81). Association for Computing Machinery. https://doi.org/10.1145/3558489.3559073

[29] Orozco-Arias, S., Piña, J. S., Tabares-Soto, R., Castillo-Ossa, L. F., Guyot, R., & Isaza, G. (2020). Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements. Processes, 8(6), 638. https://doi.org/10.3390/pr8060638