Journal of Information Systems and Informatics
Vol. 7, No. 3, September 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i3.1195 Published By DRPM-UBD

# Securing EEG-based Brain-Computer Interface Systems from Data Poisoning Attacks

**Joshua Joshua Tom[1], Frank Edughom Ekpar[2,3], Wilfred Adigwe[4]**

[1]Department of Computer Science and Cyber Security, Elizade University, Ilara-Mokin, Nigeria
[2]Founder, Scholars University Ltd, Port Harcourt, Rivers State, Nigeria
[3]Deprtment of Computer Engineering, Rivers State University, Port Harcourt, Nigeria
[4]Department of Computer Science, Delta State University of Science and Technology, Ozoro, Nigeria
Email: [1]drtomjoshua@gmail.com, [2,3]frank.ekpar@ust.edu.ng, [4]adigwew@dsust.edu.ng

## Abstract

Electroencephalogram (EEG)-based brain computer interface (BCI) is a widely used access technology to aid human-computer interactions. It enables communication between the human brain and external devices directly without the need for actuators such as human hands and legs. The BCI system acquires brain signals from an EEG device and uses machine learning (ML) algorithms to analyze and interpret the signals into actionable commands. However, EEG-based BCI systems are vulnerable to data poisoning attacks, which compromises the accuracy and security of the BCI system, and user safety. The objective of this paper is to protect the BCI systems against backdoor data poisoning attacks for reliable system operations. In this paper, a backdoor detect-and-clean mechanism, code named Bkd-DETCLEAN, to secure EEG-based BCI systems against data poisoning (backdoor) attacks is proposed using the Random Forest Classifier. Two models were designed, trained and validated on both clean and poisoned dataset respectively. The results of experiments on two benchmark EEG datasets shows that our solution achieves a detection accuracy of 98.5%, effectively identifying poisoned samples with a little below 5% false positive rate. Continued data cleaning iterations restored the poisoned training set, resulting in an overall system accuracy improvement from 78.9% to 93%. Based on these results, the proposed model sustained high detection and cleaning efficiency with different poisoning rates, underscoring the effectiveness of the machine learning driven proposed model in ensuring that brain signal integrity is not compromised. The proposed mechanism is also applicable in other areas including healthcare and medical data protection, protecting fraud detection models in financial systems, ensuring the integrity of sensor data in industrial control systems, protecting against user data manipulation in recommender systems, etc.

**Keywords**: Adversarial Perturbation, Backdoors, Brain-Computer Interface, Brain Signals, Electroencephalogram, Machine Learning Models

## 1. INTRODUCTION

There have been significant transformations in human interactions with technology through the deployment of Brain-Computer Interface (BCI) systems [1], which translate neural signals into executable commands. Neural signals are measured using Electroencephalogram (EEG) devices. EEG-based BCIs are famous due to their non-invasive nature, availability, low cost, and ease of deployment [2]. EEG-based BCIs systems are widely deployed in applications areas such as neurorehabilitation, assistive communication, cognitive monitoring, and gaming. BCI determines human functional intent by circumventing the use of actuators (arms and legs through muscles control) instead enabling a direct control of devices in the user's environment [3]. The functionality of EEG-based BCI largely dependent on machine learning algorithms trained to interpret complex neural data patterns [4]. Though these machine learning-based models have exhibited impressive accuracy and usability, their dependence on data-driven methods comes with significant security vulnerabilities [5].

Recent studies have emphasized that machine learning systems are susceptible to data poisoning attacks, where attackers manipulate the training data to harmfully control model behavior. For EEG-based BCI systems, such attacks could cause neural signals to be misclassified, leading to unsolicited system behavior, compromised integrity, and potential harm to users. The technology works by transferring the brain signals measured by the Electroencephalogram (EEG) device to the computer where decoding and interpretation of the signals functional intents is facilitated by machine learning algorithms. The brain signals are generated when neurons communicate with each other (that is, when there is brain activities) and are recorded by the BCI and converted to data (output control signals) for machine learning analytics [6]. This has the potential to assist individuals such as motor paralysis patients in improving communication and control of external devices in the environment through conscious brain activity. EEG-based BCI are vulnerable to a number of attacks [7].

In this paper, we examine intentional adversarial perturbation (backdoor attacks) on the machine learning algorithm [8, 9, 10, 11, 12]. As an application case, [13] developed a comprehensive AI-driven healthcare system characterized by a modular architecture using EEG-based BCI to collect data from the patients by exploiting the signals generated by the patient's brain activity. Despite the increased awareness made through resent studies on the security concerns in machine learning, there has been limited research effort specifically aimed at addressing the security of EEG-based BCI systems against data poisoning. Majority of researches focus generally on adversarial attacks in other areas like image recognition, text processing, etc. with little attention paid to the area of neurotechnology. This gap brings to limelight the importance of investing data poisoning threats against BCI systems and developing possible approaches to mitigate these vulnerabilities.

Machine learning models have received commendable attention in recent times hence it is a well-explored area. This is not the case with studies on defending EEG-based BCI systems against data poisoning attacks. As BCI systems are increasingly utilized in critical applications, designing mechanisms to ensure security against threats is essential. It should be of note that the design of a detection mechanism for detecting threats to a BCI system is not without some challenges. These challenges include EEG data-based complexity as a result of the high dimensionality of EEG signals making it hard to distinguish between malicious backdoor signals and benign brain signals. Again, EEG signals have a lot of variability across individuals and hence require adaptive detection methods that can generalize across diverse users due to the intricate nature of EEG data, the operational constraints placed on EEG-based BCI systems, the need for real-time processing with latency kept as low as possible, the tact with which backdoor triggers are designed, and lastly the need not to trade system usability for accuracy.

These challenges are, however, surmountable when proper attention is paid to deploying suitable machine learning algorithm that can provide the needed optimizations. There are many available machine learning algorithms which have been used in the design of existing threat detection mechanisms. Specifically, algorithms such as SVM, Ada Boost, ANN, XGBoost, KNN, etc. have been recently used by many researchers for this purpose. This paper is aimed at analyzing the susceptibility of EEG-based BCI systems to data poisoning attacks, evaluating the impact of such attacks on system performance and reliability, and proposing an effective defense mechanism to detect and mitigate data poisoning in EEG datasets. This will enable the development of more secure and reliable BCI systems capable of operating securely in adversarial domains.

Despite the multidimensional benefits of the EEG-based BCI technology, its adoption in the diagnosis, detection, and prediction of stroke, autism, epilepsy and other health conditions can be disastrous and counterproductive if adequate attention is not paid to critical security concerns with regards to BCI threats including brain signal data poisoning which could have negative effects on the end user. The key contributions made by this paper include analysis of EEG data poisoning attack particularly data injection attacks, label flipping and backdoor or Trojan attacks on BCI-based system proposed in [13] demonstrating that these attacks can significantly degrade the performance of machine learning models in BCIs. We also contribute to the area of EEG-BCI security by proposing a robust, machine learning-based detection framework designed to detect and remove data poisoning attempts in EEG datasets for BCI system defense against adversarial manipulations. This work demonstrates the effectiveness of the proposed detection mechanism through series of experiments conducted on two publicly available datasets namely the Raw EEG Dataset and the Feedback Error-Related Negativity (ERN) dataset. The paper also contributes by providing practicable recommendations and best practices in the development and deployment of EEG-based BCI systems.

The remainder of this paper is organized as follows: Sections 2 focuses on literature review and review of related works. Section 3: handles the materials and methods used in the research. Section 4 presents the results and discussions. Finally, Section 5 draws conclusions.

## 2. LITERATURE REVIEW

### 2.1. EEG-based BCI Systems

EEG-based Brain-Computer Interfaces (BCIs) allow individuals to control devices or communicate using electroencephalography (EEG) signals, which are generated and recorded from the brain's activity. By using EEG-based BCIs, people with severe motor disorder can communicate with others, home appliances, computers, robots etc. can be con-trolled without the natural motor system (actuators). The system is also an effective method for designing rehabilitation systems [14]. There are different types of EEG-based BCI systems including motor imagery-based BCI, visual imagery-based BCI, auditory imagery-based BCI and P300-based BCI systems [15]. Motor imagery-based BCI is the most common paradigm in BCI systems. It uses EEG to collect brain activity signals associated with imagination of specific motor actions (moving the arms or leg). The data collected is thereafter passed through preprocessing to extract important features such as frequency bands, amplitudes, and latency. Subsequently, the extracted data is decoded and classified by a machine learning model into motor imagery classes like a "move lefthand", "move righthand", "move left foot", "move right foot", etc. or functional intentions like grasping or manipulation of objects. Recent researches have however shown that machine learning-based BCI systems are vulnerable to various adversarial attacks [10]. In these attacks, the attackers inject the EEG brain signals with some noise by either flipping the labels or manipulating the features with the intention of misleading the model into misclassification. This misclassification can have a dire consequence of causing the BCI to misinterpret the user's intentions or make incorrect decisions thereby degrading the performance of the system.

### 1) Machine Learning Models

With the ever-skyrocketing amount of data available today, computers can analyze patterns in data and learn from the data patterns and make efficient and valuable predictions without being programmed using machine learning (ML) mod-el-based algorithms. ML models have a variety of applications including image recognition and classification, natural language processing, speech recognition, predictive modeling, recommender systems, BCI systems, spam filtering, sentiment analysis, and anomaly detection [16]. There are different ML model architectures providing different ways of processing and learning from data. Neural networks, decision trees, random forests, Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs) are some of the most common ML algorithms [17]. Random forests ML algorithm is an ensemble approach which combines multiple decision trees to create a model with improved accuracy and robust-ness. Individual decision tree is trained on a random subset of the dataset. In a decision tree, feature values are used to recursively partition the dataset into subsets by selecting a random subset of features choosing the best feature to split the data. After the data is split, the most common class is selected by voting across individual decision tree outputs.

### 2) Backdoor Attacks Against Machine Learning Models

Today many critical real-world applications in various industries use machine learning models making it ubiquitous [18]. These applications include computer vision, health related apps, recommender systems, brain computer interfaces, etc. The dark side is that these models are vulnerable to adversarial and backdoor attacks, which allows attackers to inject malicious data or tampering with the model training process. Backdoor attacks are of different types including data poisoning, model poisoning, adversarial example, and trigger-based attacks [19]. In the ML scenarios, a backdoor is similar to a covert action of the model, only occurring if the secret trigger is activated causing a misclassification of the input feature vector to the attacker's desired label. Figure 1 shows a simplified architecture of a data poisoning attack during the training phase and its effect on the testing phase of the ML model.

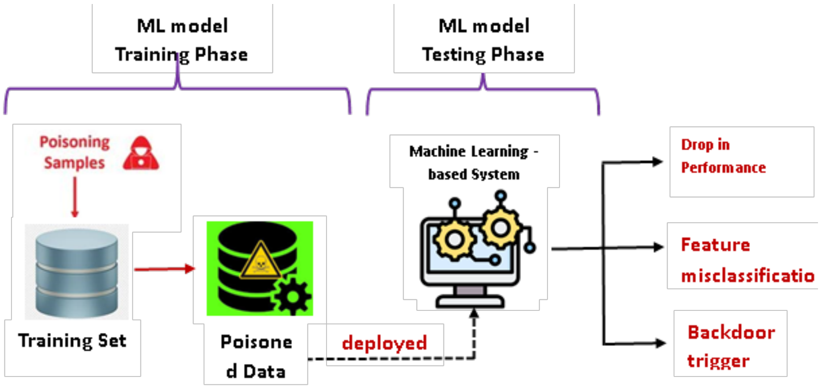

**Figure 1.** A Simplified Architecture of a Data Poisoning Attack During the Training Phase and Its Effect on Model Outputs.

### 2.2. Related Works

#### 1) Adversarial Attacks

Many EEG-based BCIs leverage sophisticated machine learning methods to decode the EEG signals. Studies have also shown that machine learning algorithms are vulnerable to adversarial attacks. [10] proposed a narrow period pulse-based method for poisoning attack of EEG-based BCIs. In this approach, dangerous backdoors are created in the model by injecting poisoning samples into the training set. The model classifies test samples with the backdoor key into the target class which the attacker specifies without the need to synchronize the backdoor key with the EEG trials making adversarial attacks much easier to implement. They showed the effectiveness and robustness of their backdoor attack approach highlighting an unignorable concern for EEG-based BCIs.

[18] reviewed and summarized the research progress in the combined field of Rapid Serial Visual Presentation (RSVP)-based Brain-Computer Interface (BCI) (RSVP-BCI), a system that establishes a direct communication pathway between the brain and external devices, to achieve high-throughput target image retrieval by utilizing the human visual system. RSVP-BCI system has recently made steady and significant progress in research on the paradigm, ElectroEncephaloGram (EEG) decoding, and system applications. Continued studies and research on the RSVP-BCI system application is approaching practical applications but there are challenges such as limited practical applications, difficulties in cross-domain decoding of EEG, and the rapid progress of computer vision. This article reviews and summarizes the re-search progress of RSVP-BCI in recent years and looks forward to the future development direction.

A DNNs classifier can be confused with high confidence leading the model to misclassify. Attackers can achieve this by designing a quasi-imperceptible perturbation to confuse the targeted classifier. [20] proposed a positive–negative detector termed PNDetector based on a positive–negative classifier, PNClassifier. The model was trained with both the positive representations and their negative representations having same structural and semantic features with a very high probability of the feature space of the positive and negative representations of adversarial examples belonging to different categories. Addition of negative example representations into the training set had no impact on performance of the classifier on clean examples. The PNDetector was tested with adversarial examples generated by eight typical attack methods on four typical datasets and the results showed that the proposed detector is efficient in all datasets and under all attack types.

Study by [21] proposed evasion and backdoor attacks on EEG-based BCIs using adversarial filtering to detect and remove adversarial examples. Adversarial filtering technique is particularly useful in detecting backdoor attacks. In their proposal, they designed adversarial filters that can significantly degrade the performance of machine learning models in BCIs. The designed adversarial filter was used as the backdoor key from which the model learns a mapping to the target class. The authors demonstrated the effectiveness of their attack model by conducting experiments on three datasets from different BCI paradigms further highlighting the need for more attention at exploring robust defense mechanisms against adversarial attacks on machine learning based system in general and EEG-based BCIs systems in particular deploying machine learning approach.

### 2) Defense against Adversarial Perturbation

In order to reliably detect attacks in a given set of inputs against machine learning algorithms for averting security risk in applications with real-world consequences, [22] proposed an unsupervised method for detecting adversarial attacks in inner layers of autoencoder (AE) networks by maximizing a non-parametric measure of anomalous node activations. This used subset scanning methods from the anomalous pattern detection domain to enhance detection power unlabeled examples of the noise, retraining or data augmentation methods. Their proposed method assigned a score to the anomality and also returns the subset of nodes within the AE network that contributed to that score. It is believed that this work may transit from detection to visualization and explainability.

Most methods of combating adversarial perturbation rely on the output of DNNs or require training a separate network to detect adversarial examples obviously leading to high computational costs and low efficiency. [23] proposed a simple and effective approach called the entropy-based detector (EBD) to protect DNNs from various adversarial attacks. EBD detects adversarial examples by comparing the difference in entropy between the input sample before and after bit depth reduction. The paper showed that EBD can detect over 98% of the adversarial examples generated by attacks using fast-gradient sign method, basic iterative method, momentum iterative method, DeepFool and CW attacks when the false positive rate is 2.5% for CIFAR-10 and ImageNet datasets.

Though deep neural networks help greatly in environmental perception tasks, they are highly susceptible to adversarial perturbations limiting their use in practical applications. [24] proposed a novel adversarial perturbation detection scheme based on multi-task perception of complex vision tasks including depth estimation and semantic segmentation. The model is designed to detect adversarial perturbations by inconsistencies between extracted edges of the input image, the depth output, and the segmentation output. They developed a novel technique to determine edge consistency loss between all three modalities. This improved the initial consistency and in turn support the detection scheme. To test their model, the authors developed a multi-task adversarial attack aimed at fooling detection scheme.

Informed by the observation that humans are able to recognize objects that appear out of place in a scene or along with other unlikely objects and that most of the perturbation-based attacks target object classifiers, [25] added a system that learns context consistency rules during training and checks for the violations of the same during testing to the DNN classifier. This approach builds a set of auto-encoders, one for each object class, appropriately trained so as to output a discrepancy between the input and output if an added adversarial perturbation violates context consistency rules. They conducted experiments and showed that their method effectively detects various adversarial attacks and achieves high reliability over a state-of-the-art context-agnostic method.

[26] observed that models trained on the dataset will passively implant the backdoor, and triggers on the input can mislead the models during testing. Based on this, the authors proposed a general training pipeline to defend against backdoor attacks actively. To achieve this, Benign models were trained from the unreliable dataset. The learning process was decoupled into three stages, supervised learning, active unlearning, and active semi-supervised fine-tuning. They showed the effectiveness of their approach in numerous experiments across various backdoor attacks and datasets.

Backdoor attacks are a kind of emergent training-time threat to deep neural networks (DNNs) that can manipulate the output of DNNs and possess high insidiousness. [27] proposed a simple and effective textual backdoor defense named ONION. The proposed model is based on outlier word detection to handle all the textual backdoor attack situations. Their experiments demonstrated the effectiveness of their proposed model in defending bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT) systems against five different backdoor attacks.

Backdoor attacks have become a major security threat for deploying machine learning models in security-critical ap-plications. [28] proposed TextGuard, a provable defense against backdoor attacks on text classification. TextGuard divides the (backdoored) training data into sub-training sets, achieved by splitting each training sentence into sub-sentences. This partitioning ensures that a majority of the sub-training sets do not contain the backdoor trigger. Subsequently, a base classifier is trained from each sub-training set, and their ensemble provides the final prediction. Theoretically, they proved that when the length of the backdoor trigger falls within a certain threshold, TextGuard guarantees that its pre-diction will remain unaffected by the presence of the triggers in training and testing inputs. Their evaluation demonstrated the effectiveness of TextGuard on three benchmark text classification tasks, surpassing the certification accuracy of existing certified defenses against backdoor attacks.

Study by [29] proposed a novel backdoor defense method to mark and purify the infected neurons in the backdoored neural net-works. Specifically, we first define a new metric, called benign salience. By combining the first-order gradient to retain the connections between neurons, benign salience can identify the infected neurons with higher accuracy than the commonly used metric in backdoor defense. Then, a new Adaptive Regularization (AR) mechanism is proposed to assist in purifying these identified infected neurons via fine-tuning. Due to the ability to adapt to different magnitudes of parameters, AR can provide faster and more stable convergence than the common regularization mechanism in neuron purifying. Extensive experimental results demonstrate that our method can erase the backdoor in neural networks with negligible performance degradation.

With the emergence of COVID-19 disease in 2019, deep learning networks (DNNs), played a key role in diagnosing the disease in the medical industry due to their superior performance. Because the computational cost of deep learning networks (DNNs) can be quite high outsourcing the training process to third-party providers becomes inevitable. In the light of this, carefulness is key in achieving robustness in DNN-based systems against cyber-security attacks. [30] proposed the dropout-bagging (DB-COVIDNet) algorithm against poisoning backdoor attacks. In this model, the trig-ger-related features will be removed by the modified dropout algorithm, and a new voting method in the bagging algorithm is used to achieve the final results. An attention-guided contrastive CNN (AC-COVIDNet) is employed as the main inducer of the bagging algorithm to evaluate the performance of the proposed method with the malicious COVIDx dataset. The results demonstrated that DB-COVIDNet has strong robustness and can significantly reduce the effect of the backdoor attack.

Study by [31] proposed a textual backdoor defense method based on deep feature classification composed of deep feature ex-traction and classifier construction. The method exploits the distinguishability of deep features of poisoned data and benign data. Backdoor defense is implemented in both offline and online scenarios. defense experiments on two datasets and two models for a variety of backdoor attacks showed the effectiveness of this defense approach and outperform the baseline defense method.

Sensing intrusion is a new threat to information security of many systems, which employs adversarial sample to show sensors fake information, aiming to mislead decision making and eventually achieve the hacker's illegal intention. This is agreeably a novel risk which formulating a countermeasure for it is non-trivial in information security. [32] proposed a method based on semantic similarity check to address this issue. In the model, Scene semantic centroid is introduced where each component of the centroid is used to depict the standard sensing semantics for each sensor in this kind of scene. The algorithm simultaneously detects both abnormal driving scene and the sensors that may be hacked. To handle the high dimensionality of raw sensing semantic space and the high sparsity efficiently, a scene semantic autoencoder was developed to extract scene semantic centroid by semantic embedding. Their experiments on *nuScenes* dataset show that their proposed method is not only feasible to identify the suspicious intruded sensors, but also more effective and accurate than the traditional abnormal sensing data detection.

## 3. METHODS

This section describes the approach and techniques adopted in developing our Bkd-DETCLEAN model to demonstrate that ML-based systems are vulnerable to backdoor data poisoning. Figure 2 shows a flow outline detailing the methodology adopted in this paper. We give detailed explanation of these processes in the experimental setup subsection.

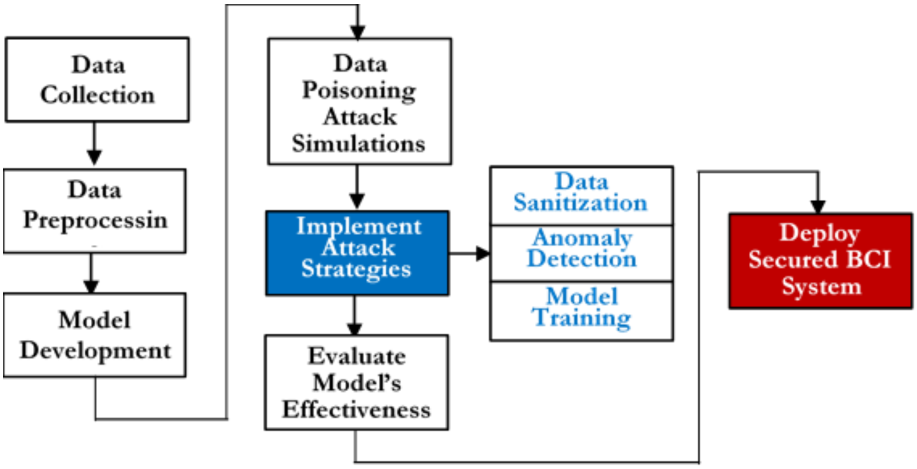
Data Collection
Data Preprocessin
Model Development
Data Poisoning Attack Simulations
Implement Attack Strategies
Evaluate Model's Effectiveness
Data Sanitization
Anomaly Detection
Model Training
Deploy Secured BCI System

**Figure 2.** A Flowchart Outline for the System Design Process.

In this section, a method to detect data poisoning threats such as data injection attacks, label flipping and backdoor or Trojan attacks may be present in BCI data and automatically clean the infected dataset is presented. This provides a defense mechanism against such threats to ML model. Data injection attacks manipulate EEG samples maliciously to insert noise into the training set to cause the model to be bias in interpreting the data leading to the model's degraded performance. In label flipping attack, the attacker mislabels the EEG data connected to particular mental states. This causes the model to learn malicious interpretations of the signals. The modus operandi of backdoors is hiding triggers in training datasets, which are designed to produce unsolicited outputs when specific patterns emerge. A practical implication of these attacks is a possible BCI system takeover, false or malicious recognitions, misclassifying mental states, evidently exposing the user to safety or privacy risk.

## 3.1 Ethics

### 1) Participant Recruitment and Data Collection

Adult participants who gave informed consent were recruited for EEG data gathering. The results presented here were obtained from data captured from an adult male participant. EEG hardware was used to record participant's brain signals, ensuring proper electrode placement, sampling rates, and data quality. This procedure was conducted under controlled conditions to minimize noise as far as possible.

### 2) Ethical Clearance and Statement

Ethical clearance for the studies was obtained from the Research Ethics Committee at Topfaith University, Mkpatak, Nigeria. All studies complied with applicable rules and regulations and proceeded only after obtaining informed consent from participants and informing them of the benefits and circumstances of the studies. We do not intend any harm to anybody or group of persons. The backdoored data poisoning mechanism developed in this paper is only for research purposes.

## 3.2 Experimental Setup

In this section, we illustrate the performance of our Bkd-DETCLEAN defense mechanism against backdoor-based data poisoning attack on EEG BCI systems and justify the effectiveness of our model by evaluating the model experimentally on Windows 10 running on Intel® Core i7 6820 HQ CPU @2.7 GHz. The implementation was done using Python 3.11.3. The designed model was trained as binary classification model to classify the data as either clean data point or infected data point. The Bkd-DETECTCLEAN model utilized the Random Forest classifier as the machine learning algorithm. The justification for using Random Forest classifiers compared to other algorithms in data poisoning detection mechanisms is due to its several offerings within the context of EEG data analytics and BCI systems. Random Forest algorithm is capable of modelling complex and nonlinear relationships in EEG-BCI data, an essential property to identify intricate pattern which could indicate data poisoning attack. The algorithm is existentially an ensemble model suitable to handle the inherently noisy and variable nature of EEG signals, aggregating predictions from multiple Decision Trees lending robustness and resiliency to the host model. The aggregation over multiple Decision Trees also lends credence to the reduction in the risk of overfitting resulting in reliability in detection performance for new data. Other offerings by Random Forest which are not offered by other machine learning algorithms include computational efficiency, easy implementation, and minimal data preprocessing.

### 3.2.1 Datasets

We deployed two publicly available datasets in our experiment.

1) Raw EEG Dataset [13], the Emotiv EPOC Wireless Headset was harnessed for the acquisition of the EEG data analyzed in the studies Validated for the acquisition of research-grade EEG data [33, 34, 35, 36], the Emotiv EPOC system offers an affordable pathway for EEG-based BCI systems and studies. The dataset consists of 24,000 samples with 13 features representing the BCI electrodes.
2) Feedback Error-Related Negativity (ERN) dataset [37]. This dataset is made up of a training set from 16 subjects and a test set from 10 subjects. We only used the training set in this paper. Each subject had 340 trials with two classes (good-feedback and bad-feedback).

### 3.2.2 Data Preprocessing

The EEG datasets used in this paper were loaded into Python code for preprocessing operations. The features of the Raw EEG Data were classified into clean and poisoned data points with “1” representing clean and “0” representing infected or poisoned data. For the Raw EEG dataset, slow drifts and high frequency noise were eliminated using a bandpass filter (typically between 0.5 Hz and 40 Hz). For the ERN dataset, we down-sampled the 32-channel EEG signals to 128Hz and filtered using a [1, 40]Hz band-pass filter. We designed a bandpass finite impulse response (FIR) filter and applied zero-filtering to each channel to prevent phase distortion. The EEG trials between [0, 1]s after each image on set were extracted and standardized by z-score normalization.

### 3.2.3 Proposed Model Architecture

Here, we present the details of our approach depicted by the proposed architecture presented in Figure 3.

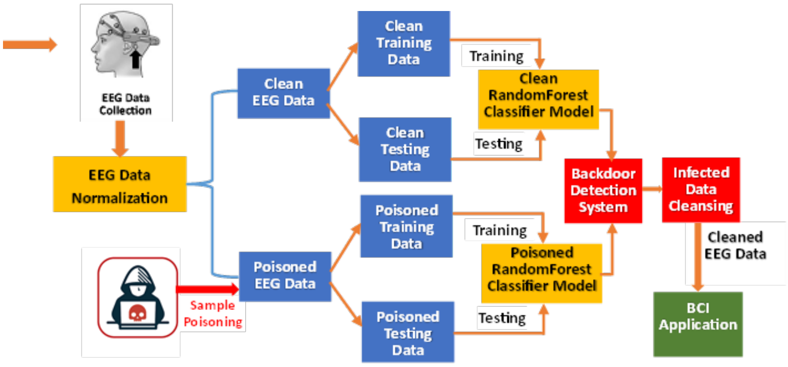

**Figure 3.** Architecture of the Bkd-DETCLEAN Mechanism

### 3.2.4 Model Development

In the experiment, we loaded the clean dataset and divided the dataset into training and testing subsets. The data was normalized using the StandardScaler from the sklearn.preprocessing module. A poisoned dataset was created from the normalized version by defining a backdoor pattern characterized by a 10 Hz frequency component. To do this, a 10 Hz sinusoidal pattern was generated using Equation 1.

$$bp= \sin(2*pi*f*t) \tag{1}$$

where $bp$ represents backdoor pattern, $f$ is the frequency of the backdoor pattern in Hz and $t$ is a time vector range from zero to $1/f$ in specified duration. This pattern was then introduced into the sampled subset of the clean dataset at specific points or channels to simulate a backdoor label flipping and random noise injection, after which the poisoned sample and the clean data were concatenated to obtain a poisoned dataset. Both clean and poisoned version of the dataset were split into training set and test set. Next, we implement a Python system that uses a Random Forest classifier to detect data poisoning attacks and remove malicious data points from the dataset by training the ML algorithm on both the clean data and the poison data obtaining a detection accuracy of 98.5% on the benign dataset with less than 5% false positive rate and an accuracy of 78.9% accuracy on the poisoned dataset. The system flags a data point as malicious if the confidence threshold is low and consequently saves the index of the flagged data in an array. The suspicious data indices array is subsequently used in the threat removal process resulting in a cleaned dataset. Next, we retrained the RF classifier-based Bkd-DETCLEAN system on the cleaned dataset and found that the performance of the system improved from 78.9% to 93%. We executed this manipulation at various poisoning proportions (0 – 100% poisoning rates) to determine the impact of the poisoning on the data integrity and subsequent model performance. We conducted three additional experiments using Ada Boost, Multilayer Perceptron (MLP) and K-Nearest Neighbors (KNN) machine learning models on each of the two datasets. The procedure described above was repeated for all the four models for each dataset instance. We designed two versions of each model, with one trained and evaluated on the clean data and the other on the poisoned data. The performance comparisons of the different models at different poisoning rates are presented in the results and discussions section. We set different poisoning rate of 0%, 5%, 15%, 25%, 40%, 50%, 75%, and 100% for each model trained on each dataset. Tables 1 to 8 show the model performance accuracies obtained using the respective classifiers on the different datasets

## 4. RESULTS AND DISCUSSION

### 4.1. Performance Evaluation

In our experiment, we evaluated the performances of our machine learning model built using Random Forest against Ada Boost, Multilayer Perceptron (MLP) and K-Nearest Neighbors (KNN) machine learning models on both clean and poisoned datasets and found that our model outperformed all others because of the utilization of Random Forest algorithm due to the justification already given in section 3.2. Based on the results obtained from the experiments, Random Forest classifier outperformed other models as shown in tables 1 to 8 and plotted in figures 3 and 4. The high performance of the Random Forest algorithm stems from the fact that the machine learning algorithm is existentially an ensemble model that combines the strength of multiple decision trees. The model has the ability to evaluate features importance in a dataset and is capable of handling data with high dimensionality. This informed our adoption of the Random Forest model as the proposed model's basic backdoor detection engine.

The results with the Random Forest ML-based design showed that the model achieved an accuracy of 100% on the clean dataset, indicating that it was able to accurately classify the data without any errors. However, when the model was tested on the poisoned dataset, its accuracy dropped significantly to 79.8%. as Figure 4 shows the performances in terms of accuracy of all the models in the evaluation. The Random Forest model is represented in green, the Ado Boost model in black, the Multi-Layer Perceptron model in red and the KNN model in yellow. All the models show a decline in performance as the data poisoning rates increase using the Raw EEG dataset.

Figure 5 is the plot of the performances of all the models trained with the ERN dataset. This also show a decline in performances as the data poisoning rates. This behavior of the models given the two datasets is a confirmation of the negative impact of data poisoning attacks on the models' performances.

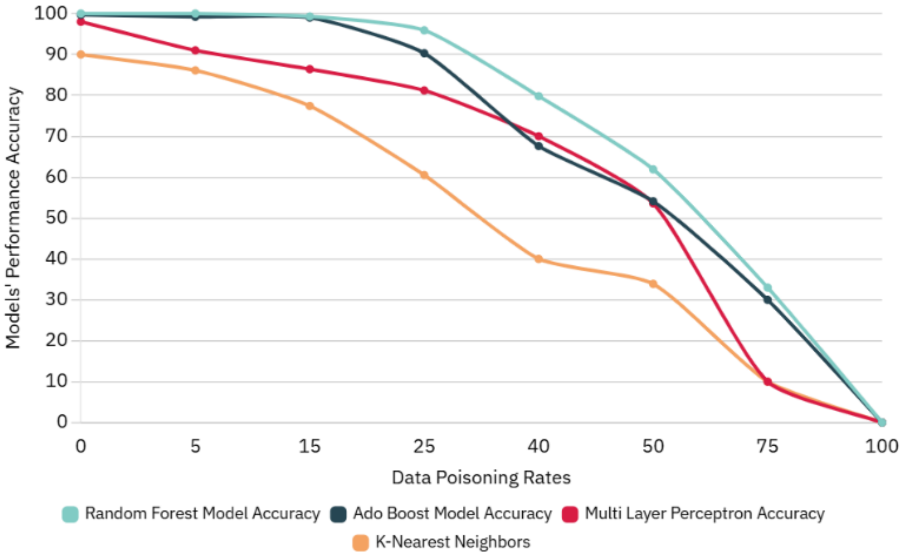

**Figure 4.** Performance of Machine Learning Models Trained on the Raw EEG Dataset at Different Data Poisoning Rates.

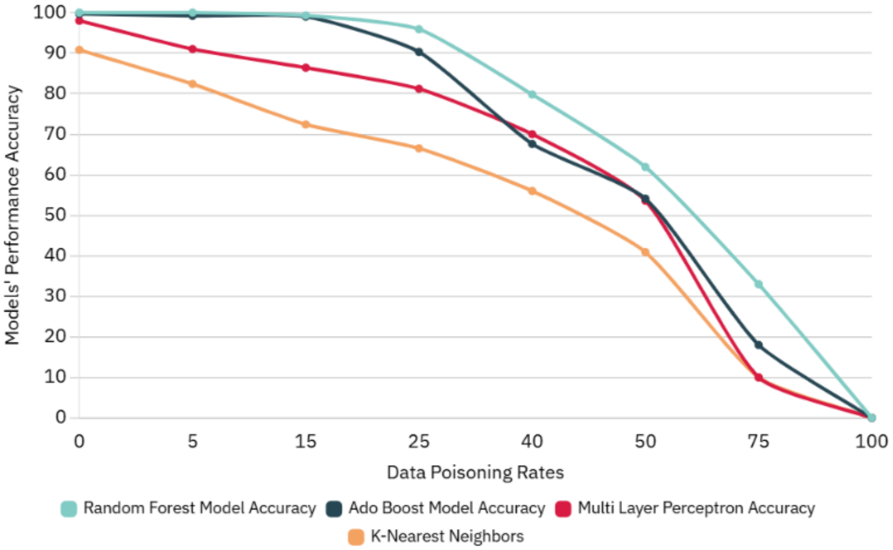

**Figure 5.** Performance of Machine Learning Models Trained on the Feedback Error-Related Negativity (ERN) dataset at Different Data Poisoning Rates

This performance drop of the models irrespective of the dataset suggests that there is a backdoor present in the poisoned data, which negatively affected the model's ability to accurately classify the data. Figure 6 shows the visualizations of the clean dataset and the poisoned data for the Raw EEG Dataset and figure 7 shows the visualizations for the Feedback Error-Related Negativity (ERN) dataset.

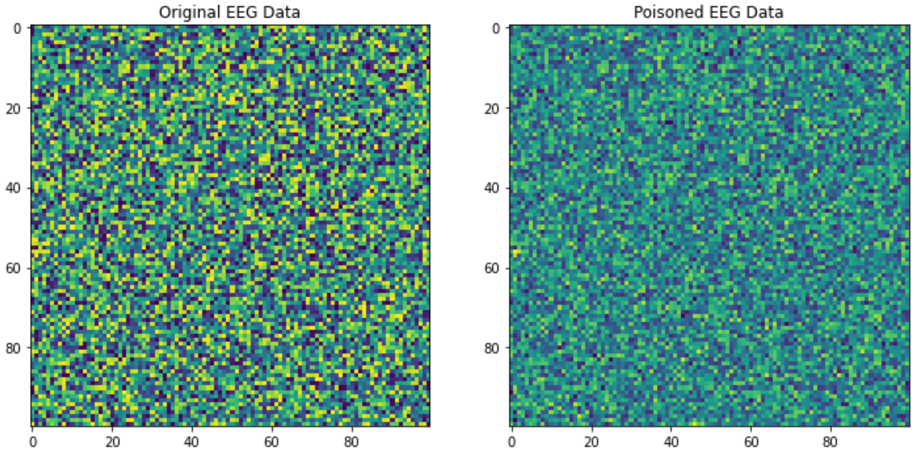

**Figure 6.** Clean versus Poisoned EEG Data Visualization.

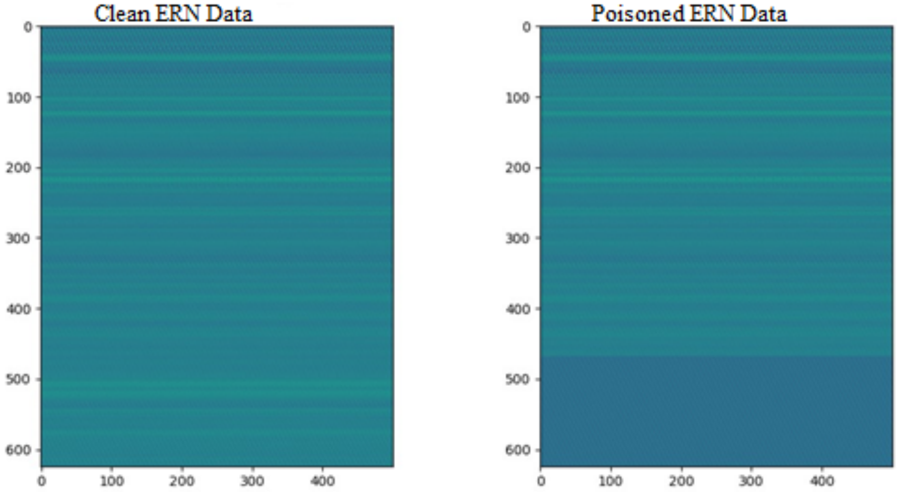

**Figure 7.** Clean versus Poisoned ERN Data Visualization.

The observable differences indicate that there was a backdoor attack. This result reveals that poisoning the datasets by injecting a backdoor into the clean dataset has a degrading effect on the performance of the EEG-based BCI system as the modified the data enabled a successful backdoor attack which led to misleading the random forest classifier machine learning model. It was found that the presence of the backdoor triggered specific misclassifications, making it difficult for the model to accurately classify new data. The results demonstrate the effectiveness of our Bkd-DETCLEAN detection solution in identifying malicious data in machine learning models. The high accuracy of 100% on the clean dataset indicates that our model was able to learn from the clean data and generalize well to unseen data. However, when the model was tested on the poisoned dataset, its accuracy dropped significantly, indicating that there was a backdoor present in the data. We draw unignorable inferences for the security of machine learning models. Based on this, we assert that any well-trained model can be vulnerable to exploitation through backdoor data poisoning attacks. Hence, developing a robust backdoor detection and cleansing technique and incorporating them into ML-based BCI systems and other machine learning pipelines is not optional but an essential component to ensure that models are not compromised. The results in tables 1 to 8 showed that our model using Random Forest had the best detection capability compared to the other machine learning algorithms as it was able to detection data poisoning even 40% poisoning rate compared to other model, which detected data poisoning only at above 40% data poisoning rate.

### 4.2. Discussion

The statistical metrics used to compare the performances of the models in this paper include accuracy, recall, precision, and F1-score. Tables 1 - 8 present the metrics comparison among different ML models at different data poisoning (DP) rates for the Raw EEG Data and the ERN datasets. Our system detects backdoor presence in the dataset by taking the model's predictions on the normal data and the poisoned data. We compared corresponding predictions and take note of the number of times the predictions vary. The model detection strategy is that if this number is greater than the set anomaly threshold, a backdoor attack is detected and then the cleaning mechanism is activated. Our poisoned data cleaning mechanism uses the anomaly indices generated during the detection phase to identify poisoned rows in the poisoned dataset and clean the data by removing or replacing the backdoored rows and the resulting cleaned dataset is fed back to the EEG-based BCI system earlier depicted in Figure 3. A plot of the tables 1 - 8 shows the effects of data poisoning on the Random Forest-based model proposed in this paper, and the other trained models on the two different datasets. As can be seen from the tables, Random Forest consistently performed better at detecting malicious data for all poisoning rates. Tables 1 and 5 show high recall values for RFC validated on both datasets compared to other algorithms. This means that the proposed system has very high sensitivity resulting in low false positives, other models miss more true positives compared to ours.

**Table 1.** Accuracies with different Poisoning Rate for the Proposed RFC-based Model on EEG Dataset.

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| RFC-based Model on EEG Dataset | 0 | 100 | 99.2 | 98.3 | 97.3 | ND* |
| | 5 | 100 | 99.2 | 98.0 | 96.2 | ND* |
| | 15 | 99.3 | 98.7 | 96.2 | 95.0 | ND* |
| | 25 | 95.9 | 94 | 91.4 | 91.5 | ND* |
| | 40 | 79.8 | 79 | 80.0 | 88.2 | detected |
| | 50 | 61.9 | 60 | 58.9 | 76.6 | detected |
| | 75 | 3.3 | 3 | 4 | 3.8 | detected |
| | 100 | 0 | 0.2 | 1.4 | 1.0 | detected |

A skim through Tables 1 to 8 reveals that the highest Recall, Precision and F1-Score values are recorded by the proposed model using Random Forest Classifier. The high Recall indicates the system achieves very high detection accuracy, high true-positives and low true-negatives irrespective of dataset. In terms of precision, the implication is that our proposed system predicts with very high accuracy, reflected in nearly al the predictions beings true. The F1-Score indicates that the proposed system achieves excellent balance between precision and recall. Lower F1-Scores by other models in this evaluation indicates less optimal performance prone with misclassifications or unacceptable balance between precision and recall.

**Table 2.** Accuracies with different Poisoning Rate for Ada Boost Classifier on EEG Dataset.

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| Ada Boost on EEG Dataset | 0 | 99.7 | 81.2 | 93.3 | 90.3 | ND* |
| | 5 | 99.2 | 78.2 | 87.0 | 87.2 | ND* |
| | 15 | 99.0 | 74.5 | 80.2 | 80.0 | ND* |
| | 25 | 90.3 | 74.1 | 73.4 | 71.5 | ND* |
| | 40 | 67.6 | 71.2 | 69.0 | 70.2 | ND* |
| | 50 | 54.1 | 60.4 | 61.6 | 64.6 | detected |
| | 75 | 3.0 | 6.2 | 4.9 | 3.8 | detected |
| | 100 | 0 | 0.2 | 1.4 | 1.0 | detected |

**Table 3.** Accuracies with different Poisoning Rate for MLP on EEG Dataset.

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| MLP on EEG Dataset | 0 | 98 | 92.2 | 92.3 | 94.3 | ND* |
| | 5 | 91 | 90.2 | 91.0 | 92.2 | ND* |
| | 15 | 86.4 | 89.7 | 89.2 | 86.0 | ND* |
| | 25 | 81.2 | 83 | 81.4 | 82.5 | ND* |
| | 40 | 70 | 79 | 80.0 | 81.2 | ND* |
| | 50 | 53.6 | 60 | 58.9 | 56.6 | detected |

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| | 75 | 1 | 3 | 4 | 3.8 | detected |
| | 100 | 0 | 5.2 | 6 | 5 | detected |

**Table 4.** Accuracies with different Poisoning Rate for KNN on EEG Dataset.

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| KNN on EEG Dataset | 0 | 90.0 | 90.2 | 90.3 | 91.4 | ND* |
| | 5 | 86.1 | 86.2 | 87.0 | 87.2 | ND* |
| | 15 | 77.4 | 76.8 | 76.2 | 75.0 | ND* |
| | 25 | 60.5 | 60.0 | 60.4 | 61.5 | ND* |
| | 40 | 40.0 | 43.0 | 40.0 | 42.0 | ND* |
| | 50 | 33.9 | 34.2 | 33.6 | 32.6 | detected |
| | 75 | 10 | 8 | 9.1 | 8 | detected |
| | 100 | 0 | 7.2 | 6.8 | 6 | detected |

**Table 5.** Accuracies with different Poisoning Rate for the Proposed RFC-based Model on ERN Dataset.

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| RFC-based Model on ERN Dataset | 0 | 100 | 98.7 | 95.3 | 97.3 | ND* |
| | 5 | 100 | 97.2 | 91.0 | 95.2 | ND* |
| | 15 | 99.3 | 96.7 | 92.2 | 93.0 | ND* |
| | 25 | 95.9 | 94.8 | 88.4 | 90.5 | ND* |
| | 40 | 79.8 | 80.3 | 80.0 | 84.2 | ND* |
| | 50 | 61.9 | 75.8 | 62.9 | 52.6 | detected |
| | 75 | 33 | 45 | 40 | 31.8 | detected |
| | 100 | 0 | 20 | 15 | 10 | detected |

**Table 6.** Accuracies with different Poisoning Rate for Ada Boost Classifier on ERN Dataset.

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| Ada Boost on ERN Dataset | 0 | 99.7 | 91.8 | 93.3 | 98.3 | ND* |
| | 5 | 99.2 | 89.2 | 87.0 | 87.2 | ND* |
| | 15 | 99.0 | 86.5 | 80.2 | 80.0 | ND* |
| | 25 | 90.3 | 78.0 | 73.4 | 71.5 | ND* |
| | 40 | 67.6 | 74.2 | 69.0 | 70.2 | ND* |
| | 50 | 54.1 | 66.4 | 61.6 | 64.6 | detected |
| | 75 | 18.0 | 38 | 41 | 38 | detected |
| | 100 | 0 | 20 | 14 | 10 | detected |

**Table 7.** Accuracies with different Poisoning Rate for MLP on ERN Dataset.

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| MLP on ERN Dataset | 0 | 98 | 96.2 | 95.3 | 96.3 | ND* |
| | 5 | 91 | 95.1 | 93.8 | 92.2 | ND* |
| | 15 | 86.4 | 88.7 | 90.2 | 86.0 | ND* |
| | 25 | 81.2 | 84 | 87.4 | 80.5 | ND* |
| | 40 | 70 | 71 | 80.0 | 74.2 | ND* |
| | 50 | 53.6 | 62 | 60.9 | 56.6 | detected |
| | 75 | 10 | 30 | 40 | 38 | detected |
| | 100 | 0 | 20 | 10 | 10 | detected |

**Table 8.** Accuracies with different Poisoning Rate for KNN on ERN Dataset.

| ML Model | DP rate (%) | Accuracy (%) | recall | precision | F1-score | DP** Detection |
|---|---|---|---|---|---|---|
| KNN on ERN Dataset | 0 | 90.8 | 90.2 | 92.3 | 91.4 | ND* |
| | 5 | 82.4 | 86.2 | 85.0 | 81.2 | ND* |
| | 15 | 72.4 | 76.8 | 72.2 | 78.0 | ND* |
| | 25 | 66.5 | 60.0 | 60.8 | 64.5 | ND* |
| | 40 | 56.0 | 53.0 | 40.7 | 52.0 | ND* |
| | 50 | 40.9 | 44.2 | 33.6 | 40.6 | detected |
| | 75 | 10 | 34 | 28 | 38 | detected |
| | 100 | 0 | 20 | 14 | 10 | detected |

*ND – not detected. **DP – Data Poisoning

Though the proposed mechanism demonstrates promising results, it is not without some limitations. First and foremost, scalability of the model for large datasets constitutes a serious concern as the proposed system's effectiveness may dwindle given large-scale EEG datasets or in real world BCI applications that utilizes many channels with high sampling rates. We opine that this increase in data volume can pose a challenge with regards to the speed of processing and storage requirements. The second important factor that may limit the implementation of the proposed system is the relatively high computational cost involved. This cost is attributed to the large processing power required by anomaly detection and data sanitization techniques utilized in the security mechanism. Expectedly, these could impede the proposed system’s deployment especially in resource-constrained and portable BCI devices. Also, we assume that clean and well labelled data is always available for training and validating the machine learning models. This might not be realistic in the actual sense. Considering all the highlighted limitations, it is pertinent for future work to be focused on optimizing the machine learning algorithms for efficiency and research on designing lightweight security mechanisms robust against data poisoning attacks with less computational costs.

## 5. CONCLUSION

We designed a system with backdoor data poisoning detection capability, which automatically cleans the poisoned data leaving disinfected EEG BCI signals. By this approach we successfully provided a mechanism for prevention of back-door-based data poisoning on EEG-based BCI systems. This experiment demonstrates the effectiveness of our backdoor detection solution in identifying malicious data in machine learning models especially in EEG-based BCI systems. The results suggest that no matter how well ML models are trained, they are still vulnerable to backdoor poisoning attacks in datasets and highlight the need for more robust evaluation methods and robust backdoor detection techniques. Machine learning-based technique advocated in this paper is just one way of providing robust defense against data poisoning attacks for EEG-based Brain-Computer Interface (BCI) systems. There are several pragmatic approaches to addressing the challenges of detecting data poisoning attacks which can be explored in the future. Adaptive and real-time detection mechanisms for dynamic identification and mitigation of data poisoning attempts as BCI operations is performed had not been fully explored. Incorporating adaptive techniques in data poisoning detection in particular and other threat detection mechanisms in general, will ensure that that emerging, evolving and sophisticated threats in today's threat landscape are aptly checked to ensure system resilience. Explainable AI (XAI) is another promising candidate for integration into a threat detection system architecture. The XAI technology expedites understanding of attack patterns and how the detection system responses to these malicious attempts when detected. This can approach can be very beneficial in a defense system especially where interpretability and trustworthiness of defense strategies is a system requirement. Also, federated learning frameworks stands to provide data privacy in an environment that is privacy-centric to satisfy some stringent data privacy requirements while providing robust security in a collaborative way across distributed BCI systems.

In this paper, we recommend data integrity assurance practices during development of EEG-BCI security mechanisms by implementing rigorous data validation to ensure that corrupted EEG signals are detected before model training and real-time processing. To ensure secure data collection, a resource friendly cryptographic system is recommended to provide encrypted transmission channels for EEG data. This is to prevent interception and tampering attacks between the wireless medium from EEG device to the BCI system. Also, it is important to secure the data storage media by ensuring that stored training data are provided with access controls to detect any threats. Most importantly, users and operators the BCI system should be constantly trained about likely security threats and best practices for safe data handling. Finally, integration of a multilayer approach to building a defense system is a sure way to ensuring BCI system security.

## REFERENCES

[1] U. Asgher, Y. Ayaz and R. Taiar, "Advances in artificial intelligence (AI) in brain computer interface (BCI) and Industry 4.0 for human machine interaction (HMI)," *Frontiers in Human Neuroscience,* vol. 5, no. 17, p. 1320536., Dec 2023.

[2] J. I. Joshiraj, "Brain-Computer Interfaces (BCIs) and AI: The Future of Human-Machine Symbiosis," *Journal of Science, Technology and Engineering Research,* vol. 30, no. 12, pp. 58-65., Dec 2024.

[3] N. Siribunyaphat and Y. Punsawad, "Steady-State Visual Evoked Potential-Based Brain–Computer Interface Using a Novel Visual Stimulus with Quick Response (QR) Code Pattern," *Sensors,* vol. 22, no. 4, p. 1439, 2022.

[4] W. H. Elashmawi, A. Ayman, M. Antoun, H. Mohamed, S. E. Mohamed, H. Amr, Y. Talaat and A. A. Ali., "Comprehensive review on brain–computer interface (BCI)-based machine and deep learning algorithms for stroke rehabilitation," *Applied Sciences,* vol. 14, no. 14, p. 6347, 21 Jul 2024.

[5] O. A. Alimi, ". "Data-Driven Learning Models for Internet of Things Security: Emerging Trends, Applications, Challenges and Future Directions," *Technologies,* vol. 13, no. 5, p. 176, 29 Apr 2025.

[6] S. Ajrawi, R. Rao and M. Sarkar, "Cybersecurity in Brain-Computer Interfaces: RFID-based design-theoretical framework," *Informatics in Medicine Unlocked,* p. 100489, 2021.

[7] L. Miao, W. Yang, R. Hu, L. Li and L. Huang, "Against Backdoor Attacks in Federated Learning with Differential Privacy," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022.

[8] B. Xue, L. Wu, A. Liu, X. Zhang and X. Chen, "Detecting the universal adversarial perturbations on high-density sEMG signals," *Comput Biol Med,* 2022.

[9] H. Wang, J. Hong, A. Zhang, J. Zhou and Z. Wang, "Trap and Replace: Defending Backdoor Attacks by Trapping Them into an Easy-to-Replace Subnetwork," *Adv Neural Inf Process Syst,* 2022.

[10] L. Meng, X. Jiang, J. Huang, Z. Zeng, S. Yu, T. P. Jung, C. T. Lin, R. Chavarriaga and D. Wu, "EEG-based brain–computer interfaces are vulnerable to backdoor attacks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 5, no. 31, pp. 2224-2234, May 2023.

[11] X. Sun, J. Li, X. Li, Z. Wang, T. Zhang, H. Qiu, F. Wu and C. Fan, "A General Framework for Defending Against Backdoor Attacks via Influence Graph," *arXiv preprint,* vol. 21, no. 11, p. 14309, 2021.

[12] J. H. Metzen, T. Genewein, V. Fischer and B. Bischoff, "On Detecting Adversarial Perturbations," *ArXiv,* vol. 17, no. 2, pp. 42-67, 2017.

[13] F. E. Ekpar, "A Comprehensive Artificial Intelligence-Driven Healthcare System," *European Journal of Electrical Engineering and Computer Science,* vol. 8, no. 3, May 2024.

[14] M. A. Khan, R. Das, H. K. Iversen and S. Puthusserypady, "Review on motor imagery based BCI systems for upper limb post-stroke neurorehabilitation: From designing to application," *Computers in biology and medicine,* vol. 123, p. 103843, 2020.

[15] D. Wen, B. Liang, Y. Zhou, H. Chen and T. P. Jung, "The current research of combining multi-modal brain-computer interfaces with virtual reality," *IEEE journal of biomedical and health informatics,* vol. 25, no. 9, pp. 3278-3287, 29 Dec 2020.

[16] S. Tufail, H. Riggs, M. Tariq and A. I. Sarwat, "Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms," *Electronics,* vol. 12, no. 8, p. 1789, 10 Apr 2023.

[17] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of big Data,* vol. 8, no. 1, p. 53, 31 Mar 2021.

[18] A. Salem, R. Wen, M. Backes, S. Ma and a. Y. Zhang, "Dynamic backdoor attacks against machine learning models," in *IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, 2022.

[19] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," *arXiv preprint,* p. 10760, 21 Jul 2020.

[20] W. Luo, C. Wu, L. Ni, N. Zhou and Z. Zhang, "Detecting adversarial examples by positive and negative representations," *Applied Soft Computing,* vol. 117, p. 108383, 2022.

[21] L. Meng, X. Jiang, X. Chen, W. Liu, H. Luo and D. Wu, "Adversarial filtering-based evasion and backdoor attacks to EEG-based brain-computer interfaces," *Information Fusion,* vol. 107, p. 102316, 2024.

[22] C. Cintas, S. Speakman, V. Akinwande, W. Ogallo, K. Weldemariam, S. Sridharan and E. McFowland, "Detecting Adversarial Attacks via Subset Scanning of Autoencoder Activations and Reconstruction Error," in *International Joint Conference on Artificial Intelligence (IJCAI-20)*, 2020.

[23] G. Ryu and D. Choi, "Detection of adversarial attacks based on differences in image entropy," *International Journal of Information Security,* vol. 23, p. 299–314, 2024.

[24] M. Klingner, V. R. Kumar, S. Yogaman, A. B. and T. Fingscheidt, "Detecting Adversarial Perturbations in Multi-Task Perception," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[25] S. Li, S. Zhu, S. Paul, A. Roy-Chowdhury, C. Song, S. Krishnamurthy, A. Swami and K. S. Chan, "Connecting the Dots: Detecting Adversarial Perturbations Using Context Inconsistency," in *European Conference on Computer Vision*, 2020.

[26] Z. Ying and B. Wu, "DLP: towards active defense against backdoor attacks with decoupled learning process," *Cybersecurity,* vol. 6, no. 9, 2023.

[27] F. Qi, Y. Chen, M. Li, Y. Yao, Z. Liu and M. Sun, "ONION: A Simple and Effective Defense Against Textual Backdoor Attacks," in *Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, 2021.

[28] H. Pei, J. Jia, W. Guo, B. Li and D. Song, "TextGuard: Provable Defense against Backdoor Attacks on Text Classification," *ArXiv,* vol. 2311.11225, 2023.

[29] M. Fan, C. Chen, X. Liu and W. Guo, "Defense Against Backdoor Attacks Via Identifying and Purifying Bad Neurons," *ArXiv,* 2022.

[30] S. Shamshiri, K. J. Han and I. Sohn, "DB-COVIDNet: A Defense Method against Backdoor Attacks," *Mathematics,* vol. 11, no. 20, p. 4236, 2023.

[31] K. Shao, J. Yang, P. Hu and X. Li, "A Textual Backdoor Defense Method Based on Deep Feature Classification," *Entropy (Basel),* vol. 25, no. 2, p. 220, 23 Jan 2023.

[32] Z. Zhao, B. Yang, H. Shu, Q. Liu, K. Zhang and L. Peng, "Sensing Intrusion Detection for Automatic Driving System based on Scene Semantic Centroid," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, Macau, China, 2022.

[33] N. A. Badcock, K. A. Preece, B. De Wit, K. Glenn, N. Fieder, J. Thie and G. McArthur, "Validation of the Emotiv EPOC EEG system for research quality auditory event-related potentials in children," *PeerJ,* vol. 3, p. 907, 2015.

[34] N. S. Williams, G. M. McArthur and N. A. Badcock, "It's all about time: precision and accuracy of Emotiv event-marking for ERP research," *PeerJ,* p. 10700, 2021.

[35] N. A. Badcock, P. Mousikou, Y. Mahajan, P. D. Lissa, J. Thie and G. McArthur, "Validation of the Emotiv EPOC EEG gaming system for measuring research quality auditory ERPs," *PeerJ,* vol. 1, no. e38, 2013.

[36] W. Y. Choong, W. Khairunizam, W. A. Mustafa, M. Murugappan, A. Hamid, S. Z. Bong, R. Yuvaraj, M. I. Omar, A. K. Junoh, H. Ali, Z. M. Razlan and A. B. Shahriman, "Correlation Analysis of Emotional EEG in Alpha, Beta and Gamma Frequency Bands," in *J. Phys.: Conf. Ser.*, 2021.

[37] P. Margaux, M. Emmanuel, D. S. ebastien, B. Olivier and M. J. e. emie, "Objective and subjective evaluation of online error correction during P300-based spelling," *Advances in Human-Computer Interaction,* p. 578295, 2012.