RESEARCH ARTICLE

Explainable Data Poison Attacks on Human Emotion Evaluation Systems Based on EEG Signals

ZHIBO ZHANG [1,2], **SANI UMAR** [2], **AHMED Y. AL HAMMADI** [1,2,3], **SANGYOUNG YOON** [1,2],
ERNESTO DAMIANI [1,2], **(Senior Member, IEEE)**,
CLAUDIO AGOSTINO ARDAGNA [4], **(Senior Member, IEEE)**,
NICOLA BENA [4], **AND CHAN YEOB YEUN** [1,2], **(Senior Member, IEEE)**

[1] Center for Cyber-Physical Systems, Khalifa University, Abu Dhabi, United Arab Emirates
[2] Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[3] Mohamed Bin Zayed University for Humanities, Abu Dhabi, United Arab Emirates
[4] Dipartimento di Informatica, Università degli Studi di Milano, 20122 Milan, Italy

Corresponding author: Chan Yeob Yeun (chan.yeun@ku.ac.ae)

ABSTRACT The major aim of this paper is to explain the data poisoning attacks using label-flipping during the training stage of the electroencephalogram (EEG) signal-based human emotion evaluation systems deploying Machine Learning models from the attackers' perspective. Human emotion evaluation using EEG signals has consistently attracted a lot of research attention. The identification of human emotional states based on EEG signals is effective to detect potential internal threats caused by insider individuals. Nevertheless, EEG signal-based human emotion evaluation systems have shown several vulnerabilities to data poison attacks. Besides, due to the instability and complexity of the EEG signals, it is challenging to explain and analyze how data poison attacks influence the decision process of EEG signal-based human emotion evaluation systems. In this paper, from the attackers' side, data poison attacks occurring in the training phases of six different Machine Learning models including Random Forest, Adaptive Boosting (AdaBoost), Extra Trees, XGBoost, Multilayer Perceptron (MLP), and K-Nearest Neighbors (KNN) intrude on the EEG signal-based human emotion evaluation systems using these Machine Learning models. This seeks to reduce the performance of the aforementioned Machine Learning models with regard to the classification task of 4 different human emotions using EEG signals. The findings of the experiments demonstrate that the suggested data poison assaults are model-independently successful, although various models exhibit varying levels of resilience to the attacks. In addition, the data poison attacks on the EEG signal-based human emotion evaluation systems are explained with several Explainable Artificial Intelligence (XAI) methods including Shapley Additive Explanation (SHAP) values, Local Interpretable Model-agnostic Explanations (LIME), and Generated Decision Trees. And the codes of this paper are publicly available on GitHub.

INDEX TERMS Cyber resilience, cyber security, data poisoning, EEG signals, explainable artificial intelligence, human emotion evaluation, label-flipping, machine learning.

I. INTRODUCTION

Human daily activities, such as communication, decision-making, and personal development, are significantly

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

influenced by human emotions. Moreover, the unstable emotional states of people who are insiders in an industrial organization, such as current or former employees, can cause industrial insider risk [1]. Therefore, it is important to establish emotional recognition systems to detect the emotional fluctuations of industrial insider humans to avoid unnecessary
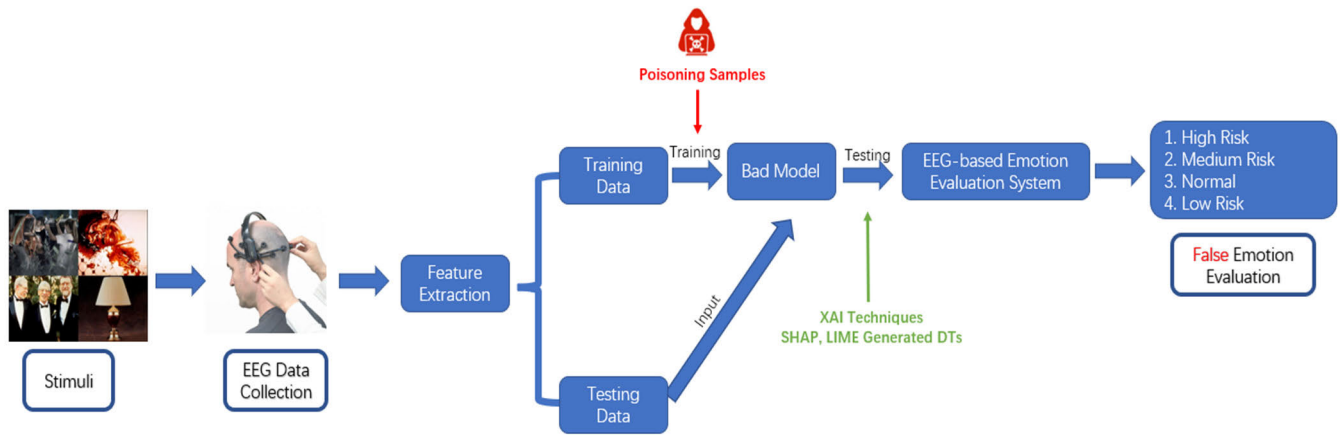
FIGURE 1. The overview of the proposed explainable data poison attacks on EEG-based human emotion evaluation systems.

industrial loss. The classification of human emotions has recently improved thanks to the ongoing development of artificial intelligence (AI) technologies, including deep learning and Machine Learning [2], along with cutting-edge therapeutic therapy.

Traditionally, facial expression [3] and voice [4] information were used to evaluate human emotions and actions of people. However, people are capable of easily masking facial and speech information with the right training [5]. On the other hand, since it is impossible for people to conceal or influence their brainwaves, EEG signals [6] have been used in recent years to evaluate a person's emotional state in order to stop possible industrial insider attacks. Moreover, several EEG signal-based human evaluation systems have employed Machine Learning [7] classifiers in various configurations to analyze human emotions in the context of AI applications.

It has been demonstrated that it is possible to intentionally alter the training data to influence the Machine Learning models' decision-making process, which ultimately results in a complete breakdown during the testing (or inference) phase [8]. On the other hand, attackers could identify and exploit Machine Learning models' weaknesses [9] to reduce the effectiveness of EEG signal-based human emotion assessment systems. To achieve this, attackers employ data poisoning (DP) [10] attack techniques to taint a target Machine Learning model by poisoning it during the training phase. Therefore, the goal of this paper is to design explainable label-flipping-based DP attacks that specifically target the Machine Learning classifiers of EEG signal-based human emotion assessment systems. To quantify the poisoning impacts and vulnerabilities of each ML model, various poisoning thresholds have also been proposed in this research and Explainable Artificial Intelligence (XAI) [11] techniques are used to investigate and clarify the precise impact of DP attacks on EEG signal-based human emotion evaluation systems in terms of features and internal mechanisms.

Therefore, to fill the gap of deploying explainable label-flipping DP attacks against Machine Learning models of

the EEG signal-based human emotion assessment systems from the attackers' perspective, the main contributions of this research paper are mentioned as follows:

1) This study covered diverse types of data manipulation and vulnerabilities of different Machine Learning models in the context of EEG signal-based human emotion evaluation systems.
2) This paper deployed two different DP categories of label-flipping attacks in the training stages of different Machine Learning models to test their resilience against different DP attacks.
3) This study demonstrated the efficiency of label-flipping DP attacks on EEG signal-based human emotion evaluation systems under complicated EEG signal features.
4) A deeper and explainable analysis of the consequences of the DP attacks on various ML models, comparing the results, quantifying the effects of the attack, and identifying each model's vulnerabilities with the help of XAI techniques.
5) This work analyzed different features' impacts on different Machine Learning models' human emotion prediction under DP attacks using XAI techniques.

The rest of this paper is organized as follows: Section II discusses the previous knowledge on various data poisoning attacks and the applications of EEG signals on human emotion evaluation. Section III then introduces the framework of the proposed label-flipping DP attacks to the Machine Learning models of EEG signal-based human emotion assessment systems. Section IV provides experimentation results and analysis in terms of different evaluations of conventional performance metrics and also the explanations of the DP attacks using XAI techniques. Section V concludes this paper and provides prospects and final remarks for future work.

II. RELATED WORKS
There have been studies that concentrate on the topic of emotional reactions when seeing emotional changes. These

studies offer crucial information for predicting people's responses based on changes in their brains.

In [5], the authors collected information from 17 people who were in a range of emotional states. The information was charted and categorized into four risk levels: low, normal, medium, and high. The five electrode Emotiv Insight EEG system is utilized since it is designed to be cost-effective. This study will classify anomalous EEG patterns that could indicate an insider threat and determine whether the employee is fit for duty.

In [12], a computer-centered diagnosis system was developed to diagnose Alzheimer's disease using EEG signals. To extract the features of EEG signals, the filtered signal was then divided into its frequency bands using the Discrete Wavelet Transform (DWT) approach. Many signal features, including logarithmic band power, standard deviation, variance, kurtosis, average energy, root mean square, and Norm, have been integrated into the DWT technique.

Signals from 32 people, including 16 women and 16 men, ranging in age from 19 to 37, are utilized in the most widely used DEAP dataset [13], which incorporates EEG. A total of 40 musical videos were played for the participants as stimuli during the experiment to gather data. Every video was created with the intention of stirring up strong feelings. There are 1,280 different raw data points in the dataset.

In this study [6], the authors proposed a novel method for emotion recognition that combines multichannel EEG analysis with a newly developed entropy called multivariate multiscale modified-distribution entropy (MM-mDistEn) with a model based on an artificial neural network (ANN) to outperform existing approaches. The suggested system outperformed previous approaches in tests using two distinct datasets.

For EEG features learning to perform emotional categorization, Chen et al. [14] presented a hierarchical bidirectional GRU model with an attention mechanism (H-ATT-BGRU). They demonstrated that, as compared to non-hierarchical models like CNN and LSTM, models that investigate hierarchical structures, such as H-AVE-BGRU and H-MAX-BGRU, perform better in classifying EEG features. They discovered that the H-AVE-BGRU model's classification accuracy is 8.4% and 1.9% higher than that of CNN and LSTM, respectively, for categorizing the valence feature and 8.1% and 2.5% for classifying the arousal feature baseline SVM model. The CNN model did not offer a significantly better outcome since its accuracy in categorizing the valence obtained 57.2%, which outperforms the BT model by 1% and its accuracy in classifying the arousal achieved 56.3%; this outperforms the SVM model by 1.8%. In classifying the valence and arousal, the LSTM model performed better than the CNN by 6.5% and 5.6%, respectively.

On the other hand, researchers also suggested DP attacks that would target the ML models used by the EEG analysis systems [15]. To the best of our knowledge, this research offers the first-ever and practically workable
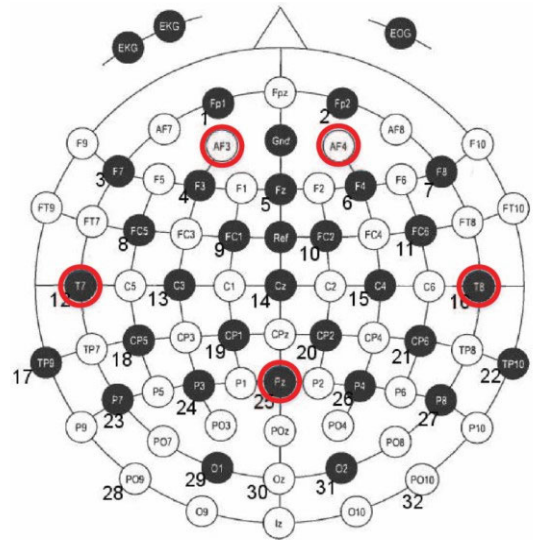

FIGURE 2. Emotiv five channels positions.

recommendation to use a narrow period pulse as a defense against poisoning attacks on EEG-based brain-computer interfaces (BCIs).

The most frequent method of producing this type of poisoning is through willful flipping of the data's labels [16]. Depending on the attacker's objectives, label flipping can be done randomly or deliberately. The former seeks to diminish the overall accuracy of all classes, while the latter does not focus on considerable accuracy reduction but rather on the misclassification of a specific class.

An effective label-flipping poisoning approach that compromises machine learning classifiers is suggested by Paudice et al. in their study [17]. Following an optimization formulation to optimize the target model's loss function, label-flipping operations are carried out. Due to the use of heuristic algorithms that allow label-flipping attacks to downscale the computing cost, this approach is regarded as computationally intractable.

According to Biggio et al. [18], an SVM-based model has been successfully attacked utilizing label-flipping, which maximizes the possibility of classification errors. As a result, the classifier's overall accuracy has been significantly reduced. This method naturally indicates a significant computing overhead as a primary need, which could be a disadvantage.

On the other hand, from the defensors' perspective of label-flipping poisoning attacks, the k-Nearest-Neighbors defensive system [19] seeks to identify harmful material and mitigate its consequences; this defense is also known as label sanitization (LS). The defense of label sanitization (LS) is based on the decision boundary of SVM, which observes the distance of the poisoned samples and recommends that these samples be re-labeled.

As stated in [20], there is yet another method for anticipating outliers. Observe that the main priority in our aims

FIGURE 3. Emotiv Insight device.

TABLE 1. List of all inputs from emotiv electrodes.

Electrode	Input features
AF3	AF3_THETA, AF_ALPHA, AF3_LOW_BETA, AF3_HIGH_BETA, AF3_GAMMA
T7	T7_THETA, T7_ALPHA, T7_LOW_BETA, T7_HIGH_BETA, T7_GAMMA
Pz	Pz_THETA, Pz_ALPHA, Pz_LOW_BETA, Pz_HIGH_BETA, Pz_GAMMA
T8	T8_THETA, T8_ALPHA, T8_LOW_BETA, T8_HIGH_BETA, T8_GAMMA
AF4	AF4_THETA, AF4_ALPHA,
Electrode	**Input features**
	AF4_LOW_BETA, AF4_HIGH_BETA, AF4_GAMMA

is detection and mitigation activities. To prevent an attacker from tampering with the training set, the third sort of protection called data sanitization is intended to be used. As a result, we would pay greater attention to the data's quality before supplying it to the ML model of our choosing.

SVM is particularly prone to label-flipping assaults, which result in complete misclassification due to the computation of incorrect decision boundaries. SVM Resistance Enhancement [21] is designed to prevent these attacks. The suggested method takes a weighted SVM with KLID into account to anticipate the consequences of suspicious data points within the SVM decision border (K-LID-SVM). The K-LID approximation of the Local Intrinsic Dimensionality (LID) metric, which is linked to the outliners in data samples, is introduced in this paper. K-LID computation relies on the kernel distance involved in the LID calculation, allowing LID to be computed in high dimensional transformed spaces. obtaining the LID values in this manner, and as a result, identifying three distinct

label dependant K-LID variations that can mitigate the effects of label flipping.

According to [22], poor generalization performance in Brain-Computer Interface (BCI) classification systems used in separate sessions can be caused by the non-stationarity of EEG signals. This paper concentrated on an experimental investigation of explanations generated by several XAI techniques using a machine learning system trained on a typical EEG dataset for emotion recognition. Results demonstrate that many pertinent elements discovered by XAI approaches are shared between sessions and can be used to create a system that can generalize more effectively.

In [23], the authors proposed an EEG signal analysis-based BCI system that can automatically recognize and decode voluntary eye blinks using Deep Learning. The primary goal of this study was to examine the explainability of the proposed CNN with the ultimate goal of determining which EEG signal segments are most crucial to the process of distinguishing between intentional and involuntary blinks. XAI methods were used to achieve this. In particular, the Local Interpretable Model Agnostic Explanation (LIME) and Gradient-weighted Class Activation Mapping (Grad-CAM) methods were applied. We were able to visually identify the most important EEG regions, particularly for the detection of voluntary and involuntary blinks, thanks to XAI.

In [24], Al Hammadi et al. gathered one dataset to look into the potential applications of brainwave signals to industrial insider threat identification. The Emotiv Insight 5 channels device was used to connect the dataset. Data from 17 people who agreed to participate in the data collection are included in the dataset. The five Emotiv Insight electrodes, each with five power bands, will provide an EEG signal dataset. The total amount of data columns now equals 25 columns. There is also a timestamp column displayed. And this work will also deploy this dataset for investigation. The authors also implemented and compared Deep Learning techniques [25] and Machine Learning methods [1]. Furthermore, XAI techniques including Permutation Importance and SHAP values were explored in [5].

Although these previous works investigated the applications of using Artificial Intelligence techniques in EEG signal processing for human emotion evaluation, few researchers are concerned about the scenarios of DP attacks on EEG signal-based human emotion evaluation systems. These earlier results motivate further research into the utilization of human brainwave patterns to collect valuable data for identifying future DP attacks and accurately analyzing and explaining them using Machine Learning algorithms and XAI techniques. Therefore, this paper proposed a framework of explainable DP attacks on human emotion evaluation systems based on EEG signals using XAI methods in Section III.

III. RESEARCH METHODOLOGY
In this section, the methods utilized to build the framework of DP attacks on EEG signal-based human emotion assessment systems from the attackers' point of view are introduced,

TABLE 2. Metrics comparison among ML models for the second category of the DP attack.

ML model	Poison rate [%]	Accuracy [%]	Recall [%]	Precision [%]	F1-Score [%]	Log loss
Ada Boost	0	99.68	99.67	99.67	99.66	0.017
	5	96.77	96.85	96.96	96.76	0.097
	25	73.87	74.67	61.68	65.99	7.523
	50	50.32	50.00	33.19	37.34	15.282
	75	24.19	25.00	6.05	9.74	
Random Forest	0	90.97	91.03	91.94	91.04	0.627
	5	86.77	86.94	88.49	86.87	0.697
	25	60.97	61.93	55.66	55.33	3.930
	50	34.51	34.88	31.74	24.78	5.223
	75	24.19	25.00	6.05	9.74	
MLP	0	76.13	75.97	76.24	75.67	0.561
	5	43.55	43.66	47.10	43.67	1.928
	25	42.26	43.38	39.99	35.99	7.314
	50	39.68	39.62	26.73	29.09	10.238
	75	24.19	25.00	6.05	9.74	
KNN	0	80.65	80.57	80.48	80.43	2.138
	5	78.71	78.69	78.70	78.63	2.409
	25	57.42	58.11	48.35	51.31	10.560
	50	41.61	41.52	27.61	30.48	17.719
	75	24.19	25.00	6.05	9.74	
ExtraTree	0	99.35	99.33	99.34	99.33	0.217
	5	95.48	95.58	95.81	95.45	0.286
	25	73.55	74.33	61.27	65.68	8.591
	50	49.68	49.38	33.12	36.94	16.553
	75	24.19	25.00	6.05	9.74	
XG Boost	0	85.48	85.54	86.31	85.46	0.498
	5	84.19	84.39	85.12	84.24	0.577
	25	60.65	61.53	52.53	54.54	2.074
	50	37.10	37.32	31.31	27.45	3.554
	75	24.19	25.00	6.05	9.74	

TABLE 3. Metrics comparison among ML models for the second category of the DP attack.

ML model	Poison rate [%]	Accuracy [%]	Recall [%]	Precision [%]	F1-Score [%]	Log loss
Ada Boost	0	99.68	99.67	99.67	99.66	0.017
	5	96.77	96.78	96.79	96.74	0.118
	25	76.45	76.46	76.67	76.33	0.722
	50	57.74	58.25	63.59	56.61	1.799
	75	21.61	21.76	21.53	21.53	4.136
Random Forest	0	90.97	91.03	91.94	91.04	0.627
	5	88.71	88.85	89.87	88.79	0.659
	25	80.00	80.17	81.19	80.05	0.858
	50	50.00	50.56	57.41	48.02	0.995
	75	14.52	14.52	14.25	14.32	1.490
MLP	0	76.13	75.97	76.24	75.67	0.561
	5	43.55	43.92	46.16	43.42	1.250
	25	39.68	39.41	39.42	39.15	1.772
	50	34.84	35.41	39.80	30.73	2.678
	75	16.77	16.75	16.43	15.99	3.444
KNN	0	80.65	80.57	80.48	80.43	2.138
	5	79.35	79.31	79.39	79.18	2.411
	25	68.39	68.37	69.47	68.20	3.577
	50	50.00	50.38	49.92	48.56	6.625
	75	18.06	18.02	17.76	17.75	16.101
Extra Tree	0	99.35	99.33	99.34	99.33	0.217
	5	95.48	95.49	95.47	95.44	0.287
	25	78.39	78.25	78.73	78.03	0.606
	50	60.32	60.64	64.19	59.62	1.107
	75	20.97	21.07	20.95	20.93	2.342
XG Boost	0	85.48	85.54	86.31	85.46	0.498
	5	84.10	84.09	84.55	84.02	0.534
	25	75.16	75.23	76.34	75.21	0.803
	50	54.19	54.71	60.89	53.52	0.977
	75	14.52	14.53	13.45	13.94	1.591

including the data processing of EEG signals, various ML models used to evaluate human emotions based on EEG signals, DP attacks based on label-flipping, and the XAI techniques deployed to explain the features of EEG signals. Figure. 1 provides a high-level view of the suggested architecture, while the subsections each provide a description of a particular component of the suggested attacking framework respectively.

A. EEG SIGNAL DATA COLLECTION AND PROCESSING
In accordance with the ethical guidelines established by the Khalifa University Compliance Committee, the EEG signal dataset was collected in a special facility at Khalifa university [24]. The dataset was gathered to investigate the possible uses of brainwave signals for identifying insider threats in the industrial setting. The dataset was gathered using a device called the Emotiv Insight 5 channels. This dataset includes
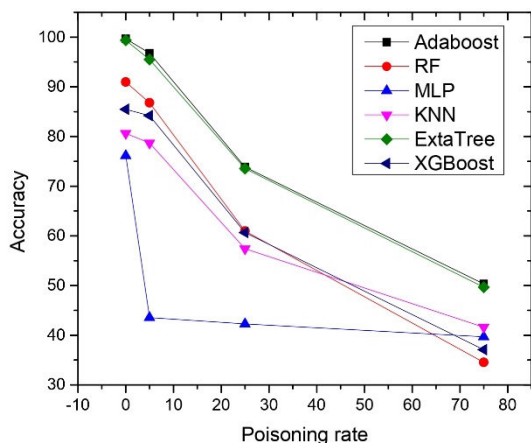
FIGURE 4. Accuracy comparison among ML models under the first category of DP attack.
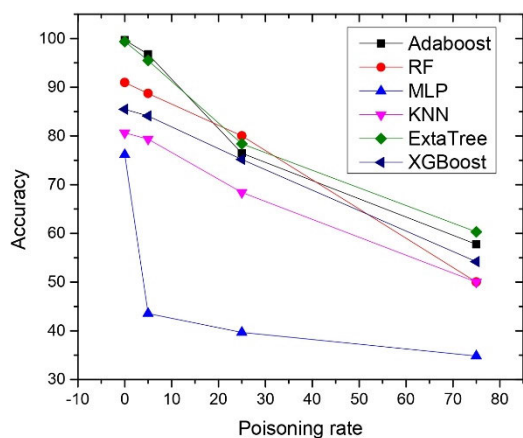

FIGURE 5. Accuracy comparison among ML models under the second category of DP attack.

information from 17 individuals who consented to participate in the data collection.

According to [26], most of n the observed EEG signals fall in the range of $1-100$ Hz. And Delta, Theta, Alpha, Beta, and Gamma brainwave band frequency range from 0 Hz to 4 Hz, 4 Hz to 7 Hz, 8 Hz to 13 Hz, 14 Hz to about 30 Hz, and 31 Hz to 100 Hz respectively. Delta wave is seen normally in adults in slow wave sleep. It may manifest focally with diffuse lesions, diffuse subcortical lesions, metabolic encephalopathy hydrocephalus, or deep midline lesions, or it may manifest generally. Theta wave is seen normally in young children. In older children and adults, it may be observed in drowsiness or arousal, as well as in meditation. Theta wave excess for the age indicates aberrant activity. It can manifest as a diffuse disorder, metabolic encephalopathy, deep midline problems, focal disruption in focal subcortical lesions, or focal disturbance in some cases of hydrocephalus. Alpha wave was the "posterior basic rhythm", which was more pronounced on the dominant side and visible in the posterior regions of the skull on both sides. It appears when the eyes are closed and when one is relaxed, and it recedes when the eyes are opened

or when one is working hard. Beta wave is most noticeable frontally and is typically distributed symmetrically on both sides. Beta activity is associated with motor behavior and is often reduced during vigorous motion.

The four risk categories—High-Risk, Medium-Risk, Low-Risk, and Normal—found in the risk matrix were used to classify each signal for a captured image, and each signal was then given the appropriate label. The data files' 26 features are made up of 25 inputs and 1 output. Each of the five electrodes on the EEG gadget records one of the five brainwave bands: Theta, Alpha, Low Beta, High Beta, and Gamma.

Figure 2 shows the electrodes in the Emotiv Insight device that record brainwaves related to cognitive and deceptive actions. And Figure 3 shows the Emotiv Insight device that was utilized to generate EEG signals. AF3 is connected to making decisions based on emotional cues, assuming others' intentions, and inferential reasoning. AF4 is responsible for decision-making involving incentives and conflicts, planning, and judgment. Pz deals with cognitive processes, whereas T7 and T8 deal with intentions. As a result, Table 1 displays the input data.

B. DP ATTACKS ON MACHINE LEARNING MODELS

Six different ML models—Random Forest, Extra Trees, AdaBoost, MLP, XGBoost, and KNN—are intruded on by DP attacks. These six ML classifiers were applied for the classification tasks of risk assessment based on the obtained EEG signal data without DP attacks initially. The dataset has been split into 80% for training the ML models, and 20% for testing.

Label-flipping technique has been offered as the DP attack on the training sets of EEG signal data, as we covered in the preceding part. Given the attacker's capacity to inject poison samples in vast amounts, it stands to reason that the label-flipping attacks with the fewest harmful samples are the most effective. A key factor to consider while evaluating the training set is the ratio of poisoned samples to all samples. Additionally, from the attackers' point of view, it is beneficial to assess the various resilience capacities of various ML models used in the classification based on the EEG signal data. The training set was therefore subjected to various poisoning rates, including 5%, 25%, 50%, and 75%, and the results were compared to the models' initial performance (0% poisoning).

In addition, two distinct Label Flipping scenarios are described in this paper because there are four labels in the risk assessment system task based on the EEG signal data. Other categories like Medium-Risk, Normal, and Low-Risk would change to High-Risk in the first scenario. This is done to see how changing other risk categories to a High-Risk level may affect the human emotion assessment system. The second scenario, on the other hand, would see Low-Risk turn into Normal, Normal turn into Medium-Risk, Medium-Risk turn into High-Risk, and High-Risk turn into Low Risk. In a similar manner, the second scenario is used to observe the impact on the human emotion evaluation system based on
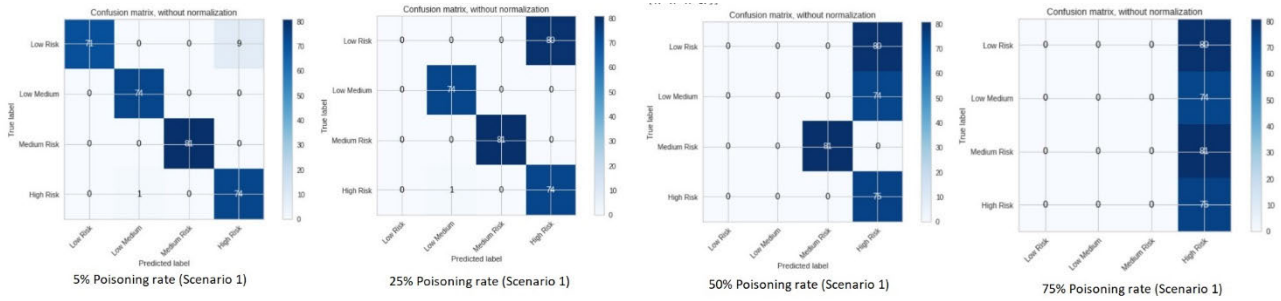
FIGURE 6. Confusion matrix for random forest ML model under the first category of DP attack.
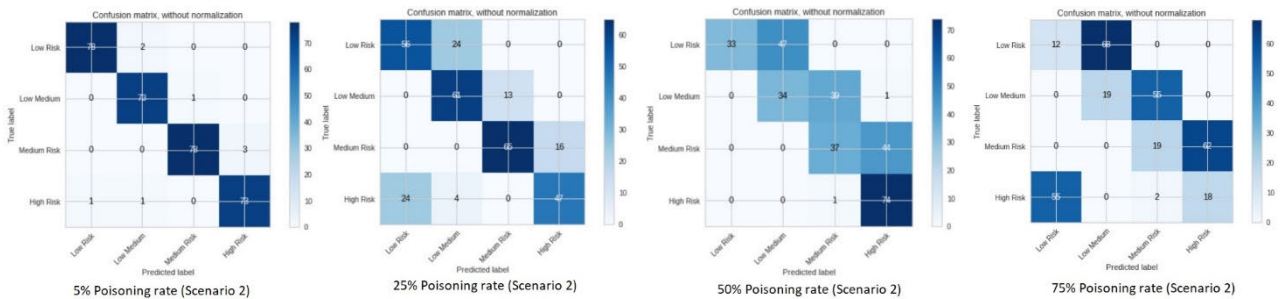

FIGURE 7. Confusion matrix for random forest ML model under the second category of DP attack.

EEG signals after altering danger levels to the other immediate risk levels.

C. EXPLAINABILITY OF THE MODEL

The creation of an explainable framework that will look into the essential features of the suggested DP attacks is one of the objectives of this research. Explainability assures that the obtained DP attacks emerge from explainable conditions rather than a black-box operation and not just improve understanding of the machine learning models.

This paper will use XAI techniques including Shapley Additive Explanation (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Generated Decision Trees to describe the proposed DP attacks framework.

SHAP was introduced as a model-neutral approach to discussing machine learning models in 2017 [27]. Shapley values are determined by comparing the team's performance with and without each individual player in a team game to determine their contribution. By evaluating the difference between the model performance with and without the feature, this method in machine learning determines the influence of each feature. This clarifies the extent to which each feature contributes positively or negatively to the forecast. As stated in [28], SHAP values are thought to be a superior technique of explanation to feature importance. The process of assigning a score to each input characteristic for a certain model is known as feature significance. The relevance of each characteristic is indicated by its score. Only machine learning algorithms can calculate feature importance, which is determined by Gini importance using node impurity.

Local Interpretable Model-agnostic Explanations (LIME) were suggested by Marco et al. in [29]. The primary goal of the LIME approach is to find an interpretable model over the interpretable representation that is both locally true to the classifier and understandable to humans. For a classifier (complex model), you want to use an interpretable model (simple model such as linear programming) with interpretable features for adaptation, and this interpretable model is then locally close to the effect of the complex model in terms of performance.

By deploying the generated decision tree models, this paper also explored the XAI idea to improve trust management. This study divided the choice into numerous little subchoices for human emotion evaluation based on EEG data, using straightforward decision tree algorithms that are simple to read and even mimic a human approach to decision-making. To test this strategy, we extracted rules from a dataset that was being used.

IV. EXPERIMENTAL RESULTS AND ANALYSIS
A. PERFORMANCE METRICS

A Machine Learning-based classifier's four primary performance indicators are as follows:

1) The proportion of test instances with true and expected values of 1 divided by the proportion of test instances having a true value of 1 is known as the True Positive (TP) rate.

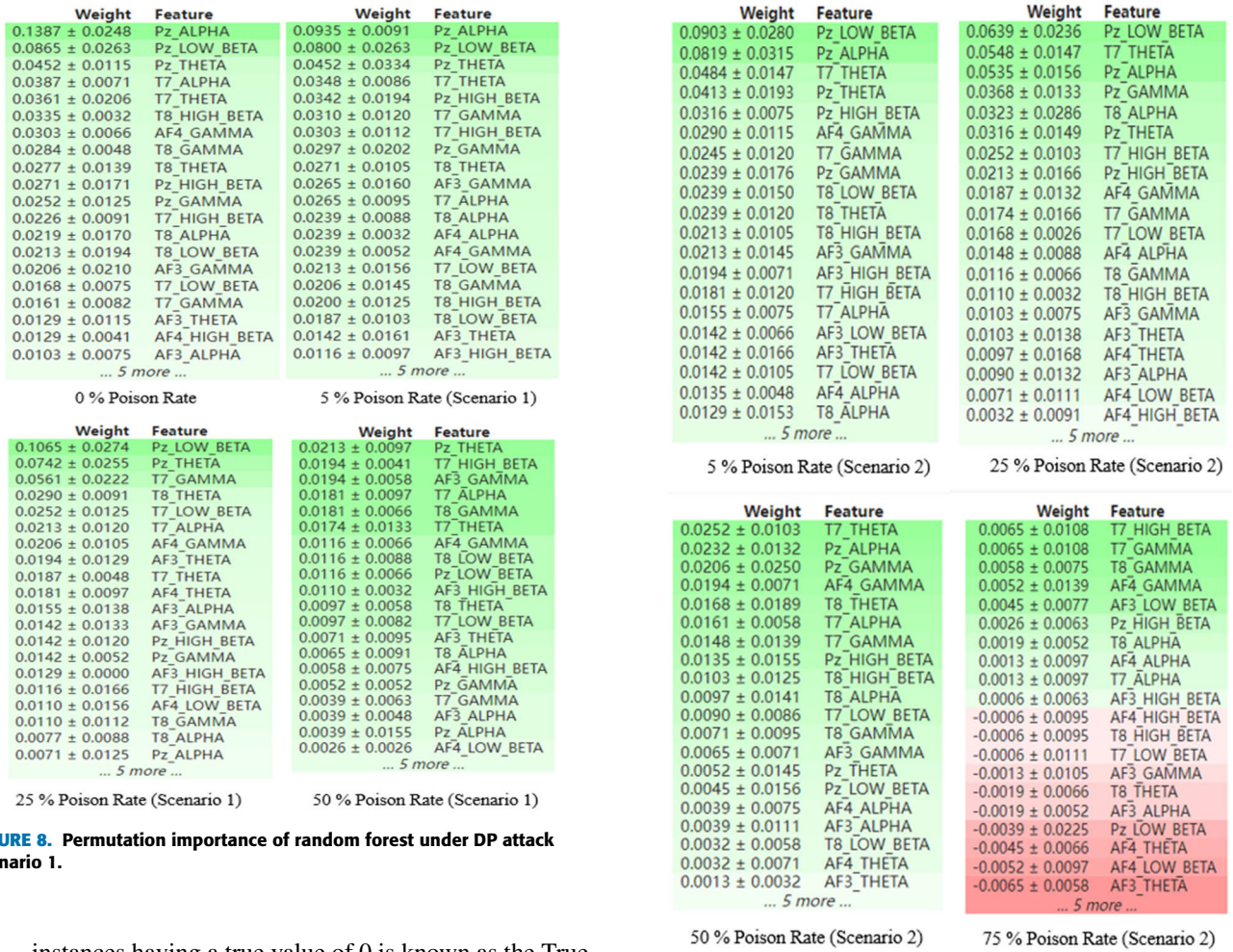2) The number of test instances with true and anticipated values of 0 divided by the total number of test

0 % Poison Rate

Weight	Feature
0.1387 ± 0.0248	Pz_ALPHA
0.0865 ± 0.0263	Pz_LOW_BETA
0.0452 ± 0.0115	Pz_THETA
0.0387 ± 0.0071	T7_ALPHA
0.0361 ± 0.0206	T7_THETA
0.0335 ± 0.0032	T8_HIGH_BETA
0.0303 ± 0.0066	AF4_GAMMA
0.0284 ± 0.0048	T8_GAMMA
0.0277 ± 0.0139	T8_THETA
0.0271 ± 0.0171	Pz_HIGH_BETA
0.0252 ± 0.0125	Pz_GAMMA
0.0226 ± 0.0091	T7_THETA
0.0219 ± 0.0170	T8_ALPHA
0.0213 ± 0.0194	T8_LOW_BETA
0.0206 ± 0.0210	AF3_GAMMA
0.0168 ± 0.0075	T7_LOW_BETA
0.0161 ± 0.0082	T7_GAMMA
0.0129 ± 0.0115	AF3_THETA
0.0129 ± 0.0041	AF4_HIGH_BETA
0.0103 ± 0.0075	AF3_ALPHA
... 5 more ...	

5 % Poison Rate (Scenario 1)

Weight	Feature
0.0935 ± 0.0091	Pz_ALPHA
0.0800 ± 0.0263	Pz_LOW_BETA
0.0452 ± 0.0334	Pz_THETA
0.0348 ± 0.0086	T7_THETA
0.0342 ± 0.0194	Pz_HIGH_BETA
0.0310 ± 0.0120	T7_GAMMA
0.0303 ± 0.0112	T7_HIGH_BETA
0.0297 ± 0.0202	Pz_GAMMA
0.0271 ± 0.0105	T8_THETA
0.0265 ± 0.0160	AF3_GAMMA
0.0265 ± 0.0095	T7_ALPHA
0.0239 ± 0.0088	T8_ALPHA
0.0239 ± 0.0032	AF4_ALPHA
0.0239 ± 0.0052	AF4_GAMMA
0.0213 ± 0.0156	T7_LOW_BETA
0.0206 ± 0.0145	T8_GAMMA
0.0200 ± 0.0125	T8_HIGH_BETA
0.0187 ± 0.0103	T8_LOW_BETA
0.0142 ± 0.0161	AF3_THETA
0.0116 ± 0.0097	AF3_HIGH_BETA
... 5 more ...	

25 % Poison Rate (Scenario 1)

Weight	Feature
0.1065 ± 0.0274	Pz_LOW_BETA
0.0742 ± 0.0255	Pz_THETA
0.0561 ± 0.0222	T7_GAMMA
0.0290 ± 0.0091	T8_THETA
0.0252 ± 0.0125	T7_LOW_BETA
0.0213 ± 0.0120	T7_ALPHA
0.0206 ± 0.0105	AF4_GAMMA
0.0194 ± 0.0129	AF3_THETA
0.0187 ± 0.0048	T7_THETA
0.0181 ± 0.0097	AF4_THETA
0.0155 ± 0.0138	AF3_ALPHA
0.0142 ± 0.0133	AF3_GAMMA
0.0142 ± 0.0120	Pz_HIGH_BETA
0.0142 ± 0.0052	Pz_GAMMA
0.0129 ± 0.0000	AF3_HIGH_BETA
0.0116 ± 0.0166	T7_HIGH_BETA
0.0110 ± 0.0156	AF4_LOW_BETA
0.0110 ± 0.0112	T8_GAMMA
0.0077 ± 0.0088	T8_ALPHA
0.0071 ± 0.0125	Pz_ALPHA
... 5 more ...	

50 % Poison Rate (Scenario 1)

Weight	Feature
0.0213 ± 0.0097	Pz_THETA
0.0194 ± 0.0041	T7_HIGH_BETA
0.0194 ± 0.0058	AF3_GAMMA
0.0181 ± 0.0097	T7_ALPHA
0.0181 ± 0.0066	T8_GAMMA
0.0174 ± 0.0133	T7_THETA
0.0116 ± 0.0066	AF4_GAMMA
0.0116 ± 0.0088	T8_LOW_BETA
0.0116 ± 0.0066	Pz_LOW_BETA
0.0110 ± 0.0032	AF3_HIGH_BETA
0.0097 ± 0.0058	T8_THETA
0.0097 ± 0.0082	T7_LOW_BETA
0.0071 ± 0.0095	AF3_THETA
0.0065 ± 0.0091	T8_ALPHA
0.0058 ± 0.0075	AF4_HIGH_BETA
0.0052 ± 0.0052	Pz_GAMMA
0.0039 ± 0.0063	T7_GAMMA
0.0039 ± 0.0048	AF3_ALPHA
0.0039 ± 0.0155	Pz_ALPHA
0.0026 ± 0.0026	AF4_LOW_BETA
... 5 more ...	

FIGURE 8. Permutation importance of random forest under DP attack Scenario 1.

5 % Poison Rate (Scenario 2)

Weight	Feature
0.0903 ± 0.0280	Pz_LOW_BETA
0.0819 ± 0.0315	Pz_ALPHA
0.0484 ± 0.0147	T7_THETA
0.0413 ± 0.0193	Pz_THETA
0.0316 ± 0.0075	Pz_HIGH_BETA
0.0290 ± 0.0115	AF4_GAMMA
0.0245 ± 0.0120	T7_GAMMA
0.0239 ± 0.0176	Pz_GAMMA
0.0239 ± 0.0150	T8_LOW_BETA
0.0239 ± 0.0120	T8_THETA
0.0213 ± 0.0105	T8_HIGH_BETA
0.0213 ± 0.0145	AF3_GAMMA
0.0194 ± 0.0071	AF3_HIGH_BETA
0.0181 ± 0.0120	T7_HIGH_BETA
0.0155 ± 0.0075	T7_ALPHA
0.0142 ± 0.0066	AF3_LOW_BETA
0.0142 ± 0.0166	AF3_THETA
0.0142 ± 0.0105	T7_LOW_BETA
0.0135 ± 0.0048	AF4_ALPHA
0.0129 ± 0.0153	T8_ALPHA
... 5 more ...	

25 % Poison Rate (Scenario 2)

Weight	Feature
0.0639 ± 0.0236	Pz_LOW_BETA
0.0548 ± 0.0147	T7_THETA
0.0535 ± 0.0156	Pz_ALPHA
0.0368 ± 0.0133	Pz_GAMMA
0.0323 ± 0.0286	T8_ALPHA
0.0316 ± 0.0149	Pz_THETA
0.0252 ± 0.0103	T7_HIGH_BETA
0.0213 ± 0.0166	Pz_HIGH_BETA
0.0187 ± 0.0132	AF4_GAMMA
0.0174 ± 0.0166	T7_LOW_BETA
0.0168 ± 0.0026	AF4_ALPHA
0.0148 ± 0.0088	T8_GAMMA
0.0116 ± 0.0066	T8_HIGH_BETA
0.0110 ± 0.0032	AF3_GAMMA
0.0103 ± 0.0075	AF3_THETA
0.0103 ± 0.0138	AF4_THETA
0.0097 ± 0.0168	AF3_ALPHA
0.0090 ± 0.0132	AF4_LOW_BETA
0.0071 ± 0.0111	AF4_HIGH_BETA
0.0032 ± 0.0091	AF4_HIGH_BETA
... 5 more ...	

50 % Poison Rate (Scenario 2)

Weight	Feature
0.0252 ± 0.0103	T7_THETA
0.0232 ± 0.0132	Pz_ALPHA
0.0206 ± 0.0250	Pz_GAMMA
0.0194 ± 0.0071	AF4_GAMMA
0.0168 ± 0.0189	T8_THETA
0.0161 ± 0.0058	T7_ALPHA
0.0148 ± 0.0139	T7_GAMMA
0.0135 ± 0.0155	Pz_HIGH_BETA
0.0103 ± 0.0125	T8_HIGH_BETA
0.0097 ± 0.0141	T8_ALPHA
0.0090 ± 0.0086	T7_LOW_BETA
0.0071 ± 0.0095	T8_GAMMA
0.0065 ± 0.0071	AF3_GAMMA
0.0052 ± 0.0145	Pz_THETA
0.0045 ± 0.0156	Pz_LOW_BETA
0.0039 ± 0.0075	AF4_ALPHA
0.0039 ± 0.0111	AF3_ALPHA
0.0032 ± 0.0058	T8_LOW_BETA
0.0032 ± 0.0071	AF4_THETA
0.0013 ± 0.0032	AF3_THETA
... 5 more ...	

75 % Poison Rate (Scenario 2)

Weight	Feature
0.0065 ± 0.0108	T7_HIGH_BETA
0.0065 ± 0.0108	T7_GAMMA
0.0058 ± 0.0075	T8_GAMMA
0.0052 ± 0.0139	AF4_GAMMA
0.0045 ± 0.0077	AF3_LOW_BETA
0.0026 ± 0.0063	Pz_HIGH_BETA
0.0019 ± 0.0052	T8_ALPHA
0.0013 ± 0.0097	AF4_ALPHA
0.0013 ± 0.0097	T7_ALPHA
0.0006 ± 0.0063	AF3_HIGH_BETA
-0.0006 ± 0.0095	AF4_HIGH_BETA
-0.0006 ± 0.0095	T8_HIGH_BETA
-0.0006 ± 0.0111	T7_LOW_BETA
-0.0013 ± 0.0105	AF3_GAMMA
-0.0019 ± 0.0066	T8_THETA
-0.0019 ± 0.0052	AF3_ALPHA
-0.0039 ± 0.0225	Pz_LOW_BETA
-0.0045 ± 0.0066	AF4_THETA
-0.0052 ± 0.0097	AF4_LOW_BETA
-0.0065 ± 0.0058	AF3_THETA
... 5 more ...	

FIGURE 9. Permutation importance of random forest under DP attack Scenario 2.

instances having a true value of 0 is known as the True Negative (TN).

3) The ratio of the number of test instances with a true value of 0 and a predicted value of 1 to the number of instances with a true value of 0 is known as the False Positive (FP).

4) The number of test instances with a true value of 1 and a predicted value of 0 divided by the number of test instances with a true value of 1 is known as the False Negative (FN).

These four measurements work together to create the confusion matrix. The confusion matrix is used to assess the effectiveness of the filtering models, and FP, FN, TP, and TN are defined as follows. The following equations represent the statistical measures, which include Accuracy, Recall, Precision, and F1-score:

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

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F1 - Score = \frac{2 \times \text{Á}(Precison \times \text{Á}Recall)}{Precison + Recall} \quad (4)$$

Other than the performance metrics illustrated, log loss is also utilized in this work to measure the loss of classification abilities of the Machine Learning classifiers under the different levels of DP attacks. Log loss indicates how closely the forecast probability matches the associated real or true value. The higher the log-loss number, the more the predicted probability deviates from the actual value. And the log loss function $\mathbf{F}$ in terms of the logarithmic loss function per label $\mathbf{F_i}$ is defined as:

$$F = -\frac{1}{N} \sum_i^N \sum_j^M \left(y_{ij} \cdot Ln \left(p_{ij} \right) \right)$$

$$= \sum_j^M \left(-\frac{1}{N} \sum_i^N y_{ij} \cdot Ln \left(p_{ij} \right) \right) = \sum_j^M F_i \quad (5)$$

where N is the number of instances, M is the number of different labels, y_{ij} is the binary variable with the expected labels and p_{ij} is the classification probability output by the classifier for the i-instance and the j-label.
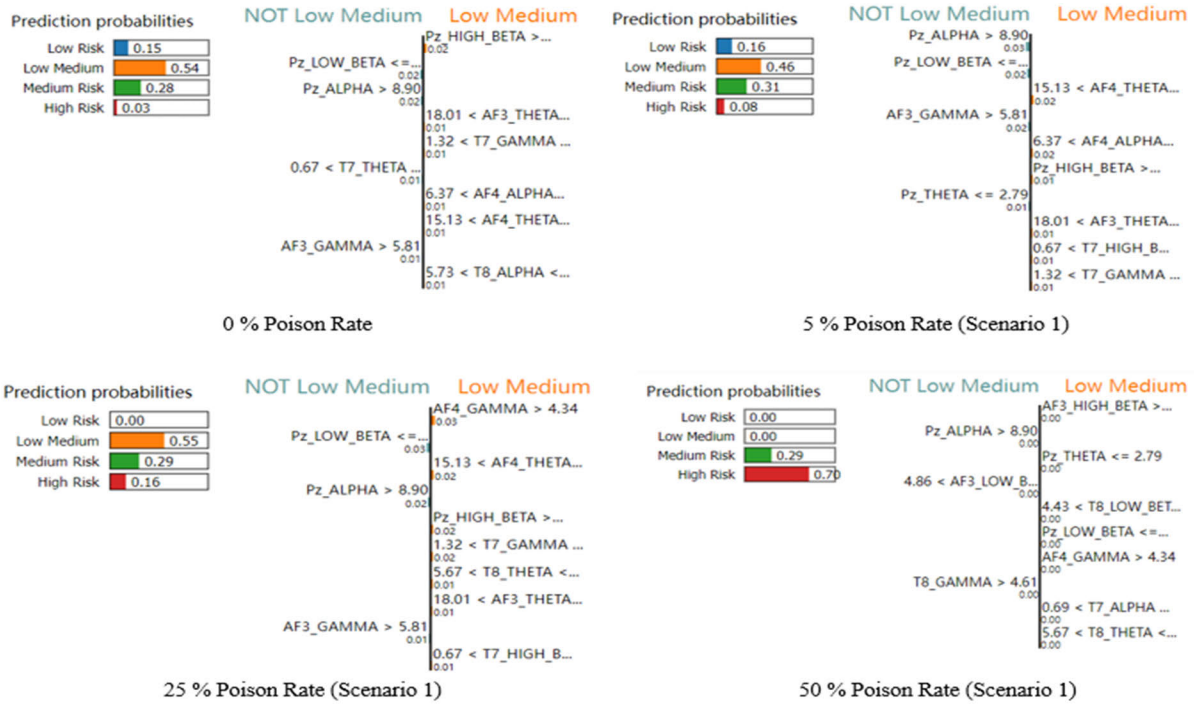
FIGURE 10. Prediction probabilities and feature contribution based on LIME of random forest under DP attack Scenario 1.
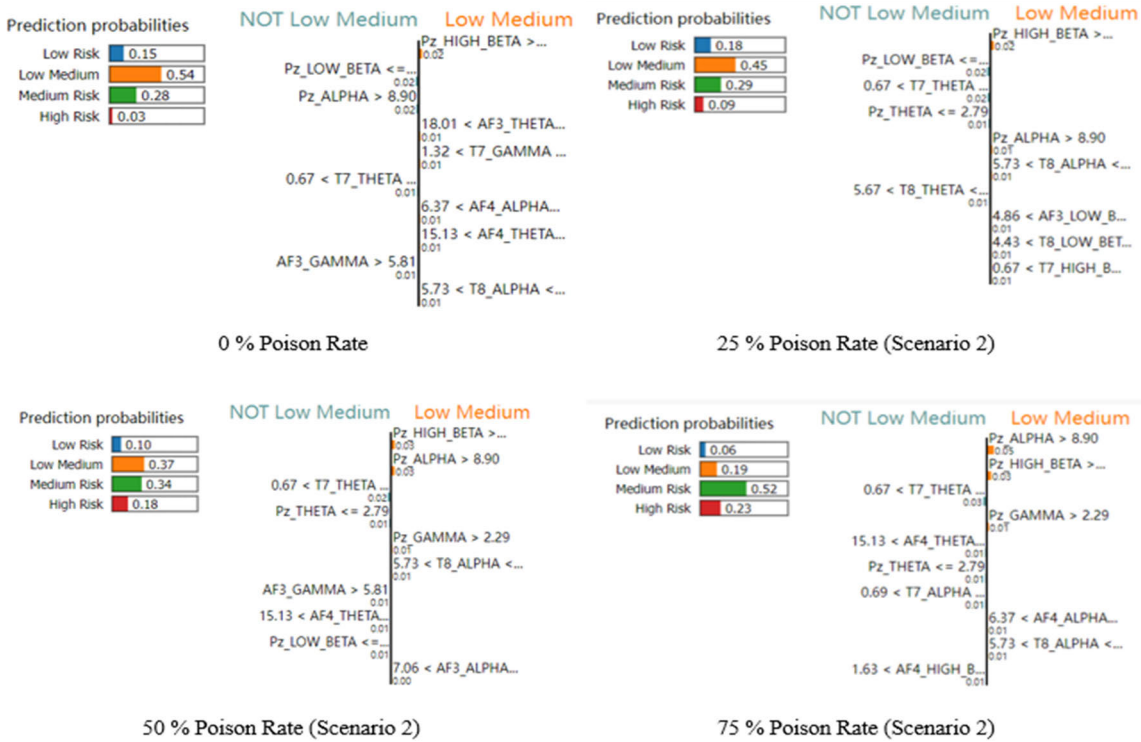

FIGURE 11. Prediction probabilities and feature contribution based on LIME of random forest under DP attack Scenario 2.

B. DP ATTACK RESULTS

This section presents the results of the proposed DP attack scenarios against the EEG-based risk assessment system.

Table 2 and Table 3 highlight the metrics comparison for the proposed DP attack scenarios under different poisoning rates among six different ML models—Random Forest, Extra
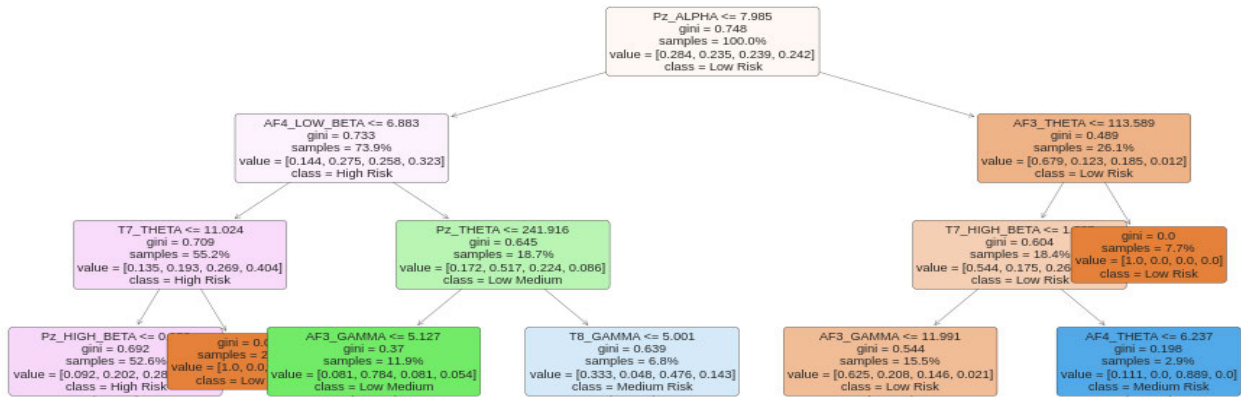
FIGURE 12. Explainable results of random forest classifier under DP attack Scenario 1 using generated decision trees.

FIGURE 13. Explainable results of random forest classifier under DP attack Scenario 1 using SHAP force plot.

Trees, AdaBoost, MLP, XGBoost, and KNN. These metrics include the classification Accuracy of the ML models, Recall, Precision, F1 score, and Log loss. As presented in the tables, although, the six ML models show different resilience to the DP attacks, the attacks were able to degrade the performance of the ML models. For example, in terms of classification accuracy, AdaBoost archived an accuracy of 99.68% with a 0% poisoning rate. Consequently, the accuracy of AdaBoost degraded to 24.19% with a 75% poisoning rate. A similar trend is observed among all other ML models.

Figures 4 and 5 present the accuracy metric comparison for the proposed DP attacks scenarios under different poisoning rates among six different ML models. According to the accuracy metric shown in the figures, AdaBoost and ExtraTree models archived relatively higher accuracy compared to the other ML models. Figures 6 and 7 present the confusion matrix for the Random Forest ML model under different poisoning rates for the proposed DP attack scenarios. The

confusion matrix illustrates an increase in misclassification as the poisoning rates increase from five percent (5%) to seventy-five percent (75%). In the next section, we will present the explanation results of the proposed two DP attack scenarios.

C. EXPLANATION RESULTS

This section presents the explanation results of the proposed two DP attack scenarios against the human emotion evaluation systems based on EEG signals. To explain the employed Machine Learning models' behaviors and the key features that ultimately determined the classification results of the algorithm, this paper deployed several XAI techniques including SHAP, LIME, Generated Trees, and the Permutation Importance to explain the effects of different features under the proposed two DP attacks.

According to the Permutation Importance of Random Forest under DP attack Scenario 1 illustrated in Figure 8,

FIGURE 14. Explainable results of random forest classifier under DP attack Scenario 2 using SHAP force plot.

the permutation importance shows that Pz_ALPHA, Pz_LOW_BETA, and Pz_THETA are the most significant features for the classification of the Random Forest algorithm whereas other features have lower but also positive effects on the decisions of the Random Forest classifier. Under a 5% scenario 1 poison rate, the feature importance ranking has few changes whereas under a 25% scenario 1 poison rate and a 50% scenario 1 poison rate, Pz_LOW_BETA, Pz_THETA, T7_GAMMA and Pz_THETA, T7_HIGH_BETA, AF3_GAMMA are the most significant features respectively. However, other features still contribute positively to the classification decision under all circumstances. For the 75% scenario 1 poison rate, Permutation Importance could not work as all samples are classified into one class in this situation.

According to the Permutation Importance of Random Forest under DP attack Scenario 2 illustrated in Figure 9, for the classification of the Random Forest classifier, under a 5% scenario 2 poison rate, a 25% scenario 2 poison rate, and a 50% scenario 2 poison rate, Pz_LOW_BETA, Pz_ALPHA, T7_ THETA, Pz_LOW_BETA, T7_ THETA, Pz_ALPHA, and T7_THETA, Pz_ALPHA, Pz_GAMMA are the most significant 3 features respectively. Moreover, other features still contribute positively to the classification decision under all circumstances. However, for the 75% scenario 22 poison rate, Permutation Importance shows that T7_HIGH_BETA, T7_GAMMA, and T8_GAMMA contribute most to the classification results. On the other hand, half of the features including AF3_THETA, AF4_LOW_BETA, and AF4_THETA make negative contributions to the final classification results.

More explainable results explained by the Permutation Importance of other Machine Learning algorithms are shown in the appendix.

As shown in Figure 10 and Figure 11, the prediction probabilities and feature contribution based on LIME of Random Forest under DP attack Scenario 1 and 2 illustrate slight differences. In scenario 1, the Random Forest classifier made false predictions in the 50% poison rate whereas the Random Forest classifier made false predictions in the 75% poison rate in scenario 2. More explainable results explained by the LIME of other Machine Learning algorithms are shown in the appendix.

Moreover, this paper used Explainable AI to create Decision Trees (Figure 12) using the IBM SPSS Modeler tool and determine the features that should be prioritized for the classifier. This allowed us to simulate the steps that a possible attacker would take. More explainable results explained by the Generated Decision Trees of other Machine Learning algorithms are shown in the appendix.

This paper also created the SHAP value to help us achieve the goal of integrating Explainable AI. The SHAP value calculates the strength and polarity (positive or negative) of a feature's influence on a prediction. Model semantics can be clearly expressed by SHAP values. SHAP power sites offer a thorough perspective of the model's internal operation, i.e., they demonstrate the model's decision-making process. Figure 13 and Figure 14 depict the impact of a specific record. The model's lower decision in the chosen record is the forecast from the first phase in this situation. The framework outlines the elements that support a given choice. By using this data, the decision-making process of

the model can be contrasted with that in the real world and ensured that the model bases its decisions on pertinent information.

V. CONCLUSION

This study proposed two label-flipping attacks on human emotion evaluation systems based on EEG brainwave signals. Two scenarios of the label-flipping attack were analyzed and evaluated to drastically decrease the overall accuracy and misclassification rate of six different multi-label Machine Learning classifiers including MLP, AdaBoost, Random Forest, KNN, XGBoost, and Extra Trees. On the other hand, the influence of training data poisoning attacks, such as the proposed label-flipping attacks varies on different Machine Learning classifiers and different EEG signal features. The use of AdaBoost, Random Forest, MLP, KNN, Extra Trees, and XGBoost classifiers achieved accuracies of 96.77 %, 86.77%, 43.55%, 78.71%, 95.48%, and 84.19% respectively under 5% poisoning rate. The accuracies of these classifiers decreased to 73.87%, 60.97%, 42.26%, 57.42%, 73.55%, and 60.65% under 25% poisoning rate and 50.32%, 34.51%, 39.68%, 41.61%, 49.68%, and 37.10% under 50% poisoning rate.

To better analyze the impacts of the proposed two label-flipping attacks on different features of EEG brainwave signals and the emotion classification results, this study utilized several XAI techniques including SHAP, LIME, Generated Decision Trees, and Permutation Importance. The results show that the AdaBoost classifier performs best in terms of classification and poisoning resilience. Besides, features of the Pz electrode contribute more to the classification results of emotion assessment according to the XAI technique, Permutation Importance. Moreover, the scenario 1 DP attack can damage the classifier with a lower poisoning rate compared with the scenario 2 DP attack according to the XAI technique, LIME. More details about the explainable results of the label-flipping attacks on human emotion evaluation systems based on EEG brainwave signals using XAI techniques are available in the appendix.

Furthermore, the main future work goal of this research is to increase the robustness of Machine Learning models against tampered training data to be used during re-training with the method of XAI technique. Besides, DP attacks based on label-flipping during the testing phase will be investigated in future work. The resilience of Deep Learning methods on the DP attacks will be evaluated and compared in future work as well. Because this paper concluded that not all features contribute to the human emotion classification prediction, EEG signal generating with fewer features but similar accuracy will be explored in future work.

APPENDIX

Codes and detailed explainable results with XAI are publicly available on GitHub:

https://github.com/qiuyuezhibo/XAIEEG

REFERENCES

[1] A. Y. A. Hammadi, D. Lee, C. Y. Yeun, E. Damiani, S.-K. Kim, P. D. Yoo, and H.-J. Choi, "Novel EEG sensor-based risk framework for the detection of insider threats in safety critical industrial infrastructure," *IEEE Access*, vol. 8, pp. 206222–206234, 2020, doi: 10.1109/ACCESS.2020.3037979.

[2] B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K.-K.-R. Choo, and M. Jamshidi, "Toward artificial emotional intelligence for cooperative social human–machine interaction," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 1, pp. 234–246, Feb. 2020, doi: 10.1109/TCSS.2019.2922593.

[3] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021, doi: 10.1109/TIM.2020.3031835.

[4] M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, "LieNet: A deep convolution neural network framework for detecting deception," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 971–984, Sep. 2022, doi: 10.1109/TCDS.2021.3086011.

[5] A. Y. Al Hammadi, C. Y. Yeun, E. Damiani, P. D. Yoo, J. Hu, H. K. Yeun, and M.-S. Yim, "Explainable artificial intelligence to evaluate industrial internal security using EEG signals in IoT framework," *Ad Hoc Netw.*, vol. 123, Dec. 2021, Art. no. 102641, doi: 10.1016/j.adhoc.2021.102641.

[6] S. T. Aung, M. Hassan, M. Brady, Z. I. Mannan, S. Azam, A. Karim, S. Zaman, and Y. Wongsawat, "Entropy-based emotion recognition from multichannel EEG signals using artificial neural network," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–13, Oct. 2022, doi: 10.1155/2022/6000989.

[7] M.-P. Hosseini, A. Hosseini, and K. Ahi, "A review on machine learning for EEG signal processing in bioengineering," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 204–218, 2021, doi: 10.1109/RBME.2020.2969915.

[8] M. A. Ramirez, S. Yoon, E. Damiani, H. A. Hamadi, C. A. Ardagna, N. Bena, Y.-J. Byon, T.-Y. Kim, C.-S. Cho, and C. Y. Yeun, "New data poison attacks on machine learning classifiers for mobile exfiltration," 2022, *arXiv:2210.11592*.

[9] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, "A taxonomy and survey of attacks against machine learning," *Comput. Sci. Rev.*, vol. 34, Nov. 2019, Art. no. 100199, doi: 10.1016/j.cosrev.2019.100199.

[10] A. Paudice, L. Muñoz-González, A. Gyorgy, and E. C. Lupu, "Detection of adversarial training examples in poisoning attacks through anomaly detection," 2018, *arXiv:1802.03041*.

[11] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable artificial intelligence applications in cyber security: State-of-the-art in research," *IEEE Access*, vol. 10, pp. 93104–93139, 2022, doi: 10.1109/ACCESS.2022.3204051.

[12] K. AlSharabi, Y. Bin Salamah, A. M. Abdurraqeeb, M. Aljalal, and F. A. Alturki, "EEG signal processing for Alzheimer's disorders using discrete wavelet transform and machine learning approaches," *IEEE Access*, vol. 10, pp. 89781–89797, 2022, doi: 10.1109/ACCESS.2022.3198988.

[13] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.

[14] J. X. Chen, D. M. Jiang, and Y. N. Zhang, "A hierarchical bidirectional GRU model with attention for EEG-based emotion classification," *IEEE Access*, vol. 7, pp. 118530–118540, 2019, doi: 10.1109/ACCESS.2019.2936817.

[15] L. Meng, J. Huang, Z. Zeng, X. Jiang, S. Yu, T.-P. Jung, C.-T. Lin, R. Chavarriaga, and D. Wu, "EEG-based brain–computer interfaces are vulnerable to backdoor attacks," 2020, *arXiv:2011.00101*.

[16] X. Liu, H. Li, G. Xu, Z. Chen, X. Huang, and R. Lu, "Privacy-enhanced federated learning against poisoning adversaries," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4574–4588, 2021, doi: 10.1109/TIFS.2021.3108434.

[17] A. Paudice, L. Muñoz-González, and E. C. Lupu, "Label sanitization against label flipping poisoning attacks," in *Proc. ECML PKDD Workshops*, Cham, Switzerland, 2019, pp. 5–15, doi: 10.1007/978-3-030-13453-2_1.

[18] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," 2012, *arXiv:1206.6389*.

[19] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *Proc. Asian Conf. Mach. Learn.*, Nov. 2011, pp. 97–112. Accessed: Dec. 16, 2022. [Online]. Available: https://proceedings.mlr.press/v20/biggio11.html

[20] N. Baracaldo, B. Chen, H. Ludwig, and J. A. Safavi, "Mitigating poisoning attacks on machine learning models: A data provenance based approach," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, 2017, pp. 103–110, doi: 10.1145/3128572.3140450.

[21] S. Chen, S. Chen, M. Xue, L. Fan, S. Hao, L. Xu, H. Zhu, and B. Li, "Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach," *Comput. Secur.*, vol. 73, pp. 326–344, Mar. 2018, doi: 10.1016/j.cose.2017.11.007.

[22] A. Apicella, F. Isgrò, A. Pollastro, and R. Prevete, "Toward the application of XAI methods in EEG-based systems," 2022, arXiv:2210.06554.

[23] M. L. Giudice, N. Mammone, C. Ieracitano, M. Campolo, A. R. Bruna, V. Tomaselli, and F. C. Morabito, "Visual explanations of deep convolutional neural network for eye blinks detection in EEG-based BCI applications," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 01–08, doi: 10.1109/IJCNN55064.2022.9892567.

[24] A. A. A. Hammadi. (Feb. 25, 2021). *EEG Brainwave Dataset*. Accessed: Nov. 18, 2022. [Online]. Available: https://ieee-dataport.org/documents/eeg-brainwave-dataset

[25] A. Y. A. Hammadi, C. Y. Yeun, and E. Damiani, "Novel EEG risk framework to identify insider threats in national critical infrastructure using deep learning techniques," in *Proc. IEEE Int. Conf. Services Comput. (SCC)*, Nov. 2020, pp. 469–471, doi: 10.1109/SCC49832.2020.00071.

[26] B. R. Cahn and J. Polich, "Meditation states and traits: EEG, ERP, and neuroimaging studies," *Psychol. Bull.*, vol. 132, no. 2, pp. 180–211, Mar. 2006, doi: 10.1037/0033-2909.132.2.180.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15. Accessed: Jul. 9, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[28] W. E. Marcilio and D. M. Eler, "From explanations to feature selection: Assessing SHAP values as feature selection mechanism," in *Proc. 33rd SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Nov. 2020, pp. 340–347, doi: 10.1109/SIBGRAPI51738.2020.00053.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

AHMED Y. AL HAMMADI received the B.S. degree in computer systems engineering from United Arab Emirates University (UAEU), United Arab Emirates, and the M.Sc. degree in cryptology, security and coding of information systems from the Grenoble Institute of Technology, France. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology (KU), United Arab Emirates. From 2007 to 2011, he worked with Global Communication and Software Systems as a Security Systems Expert and the acting Department Manager responsible for leading the public key infrastructure (PKI)-based security solution deployments in local governmental authorities and conducting continuous improvement research. From 2011 to 2020, he worked with the Department of Education and Knowledge as the Program Manager in the Special Projects Division of the Strategic Affairs Department to lead mainly different information technology initiatives that impact the entire education sector in Abu Dhabi. Over the last four years, he has specialized in developing artificial intelligence algorithms in access control applications. His current research interest includes developing an electroencephalogram (EEG) risk assessment framework to detect insider threats in national critical infrastructures (NCI) using deep-learning algorithms. He was awarded the Best Paper Award in the Third International Conference on Electric Vehicle, Smart Grid, and Information Technology, South Korea, in 2018. He also participated in the Abu Dhabi Government's efforts to digitalize the public services in the emirates and has been honored to participate in establishing a new governmental social entity in Abu Dhabi.

SANGYOUNG YOON received the M.Sc. degree in mechanical system engineering from Chonbuk National University, Republic of Korea, in 2019. After his M.Sc. degree, he is currently a Research Associate of electrical and computer engineering specializing in artificial intelligence with the Khalifa University of Science and Technology, United Arab Emirates. Previously, he worked as a Research Associate with the Department of Civil Infrastructure and Environmental Engineering, Khalifa University. He is having involved in several research projects, such as cybersecurity, deep learning, optimization algorithms, evaluation of porous media using solitary wave, and impact damage analysis of carbon fiber reinforced composite. His current research interests include non-destructive testing methods to detect composite defects using deep learning algorithms and defenses against data.

ZHIBO ZHANG received the Bachelor of Science degree in mechatronics engineering from Northwestern Polytechnical University, China, in 2021. He is currently pursuing the master's degree in electrical and computer engineering with Khalifa University, United Arab Emirates. His research interests include biometrics, cyber-physical systems, cyber security, explainable artificial intelligence, and federated learning. He received the Award for Outstanding Young Researcher by Khalifa University, in 2022.

ERNESTO DAMIANI (Senior Member, IEEE) is currently a Full Professor with the Università degli Studi di Milano, Italy; the Senior Director of the Robotics and Intelligent Systems Institute; and the Director of the Center for Cyber Physical Systems (C2PS), Khalifa University, United Arab Emirates. He is also the Leader of the Big Data Area, Etisalat British Telecom Innovation Center (EBTIC), and the President of the Consortium of Italian Computer Science Universities (CINI). He is also a part of the ENISA Ad-Hoc Working Group on Artificial Intelligence Cybersecurity. He has pioneered model-driven data analytics. He has authored more than 650 Scopus-indexed publications and several patents. His research interests include cyber-physical systems, big data analytics, edge/cloud security and performance, artificial intelligence, and machine learning. He was a recipient of the Research and Innovation Award from the IEEE Technical Committee on Homeland Security, the Stephen Yau Award from the Service Society, the Outstanding Contributions Award from IFIP TC2, the Chester-Sall Award from IEEE IES, the IEEE TCHS Research and Innovation Award, and the Doctorate Honoris Causa from INSA-Lyon, France, for his contribution to big data teaching and research.

SANI UMAR received the Bachelor of Science (B.Sc.) degree (Hons.) in information systems engineering from Eastern Mediterranean University, Turkish Republic of Northern Cyprus, and the M.Sc. degree in computer networks from the King Fahd University of Petroleum and Minerals, Saudi Arabia. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. His research interests include artificial intelligence for cybersecurity (AI4SEC), machine learning (ML), federated learning (FL), and mobile crowd sensing (MCS). During his B.Sc. degree, he received the Best Graduating Student Award.

CLAUDIO AGOSTINO ARDAGNA (Senior Member, IEEE) has been a Visiting Researcher with BUTP, Khalifa University, and GMU. He is currently a Full Professor with the Università degli Studi di Milano, the Director of the CINI National Laboratory on Big Data, and the Co-Founder of Moon Cloud srl. He has published more than 140 contributions in international journals, conference/workshop proceedings, and chapters in international books. His research interests include the area of cloud-edge security and assurance, and data science.
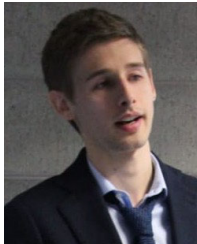
NICOLA BENA is currently pursuing the Ph.D. degree with the Università degli Studi di Milano. His research interests include the area of security of modern distributed systems with particular reference to certification, assurance, and risk management techniques.

CHAN YEOB YEUN (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in information security from the Royal Holloway, University of London, in 1996 and 2000, respectively. After his Ph.D. degree, he joined Toshiba TRL, Bristol, U.K., and later became the Vice President at the Mobile Handset Research and Development Center, LG Electronics, Seoul, South Korea, in 2005. He was responsible for developing mobile TV technologies and related security. He left LG Electronics, in 2007, and joined ICU (merged with KAIST), South Korea, until August 2008, and then the Khalifa University of Science and Technology, in September 2008. He is currently a researcher of cybersecurity, including the IoT/USN security, cyber-physical system security, cloud/fog security, and cryptographic techniques; an Associate Professor with the Department of Electrical Engineering and Computer Science; and the Cybersecurity Leader of the Center for Cyber-Physical Systems (C2PS). He also enjoys lecturing for M.Sc. cyber security and Ph.D. engineering courses at Khalifa University. He has published more than 140 journal articles and conference papers, nine book chapters, and ten international patent applications. He also works on the editorial board of multiple international journals and on the steering committee of international conferences.

• • •