IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Novel EEG Sensor-based Risk Framework for the Detection of Insider Threats in Safety Critical Industrial Infrastructure

AHMED Y. AL HAMMADI[1,2], DONGKUN LEE[2,5], CHAN YEOB YEUN[1,2], ERNESTO DAMIANI[1,2], SONG-KYOO KIM[3], PAUL D. YOO[4] and HO-JIN CHOI[5]

[1]Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi 127788, UAE
[2]Center for Cyber-Physical Systems, Khalifa University of Science and Technology, Abu Dhabi 127788, UAE
[3]School of Applied Sciences, Macao Polytechnic Institute, Macau, Macao
[4]CSIS, Birkbeck College, University of London, Malet Street, London, WC1E 7HX, United Kingdom
[5]School of Computing, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

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

"The authors acknowledge support from the Center for Cyber-Physical Systems, Khalifa University, under Grant Number 8474000137-RC1-C2PS-T3."

ABSTRACT The loss or compromise of any safety critical industrial infrastructure can seriously impact the confidentiality, integrity, or delivery of essential services. Research has shown that such threats often come from malicious insiders. To identify these insiders, survey- and electrocardiogram-based approaches have been proposed; however, these approaches cannot effectively detect or predict any malicious insiders. Recently, electroencephalograms (EEGs) have been suggested as a potential alternative to detect these potential threats. Threat detection using EEG would be highly reliable as it overcomes the limitations of the previous methods. This study proposes a proof of concept for a system wherein a model trained using a deep learning algorithm is employed to evaluate EEG signals to detect insider threats. The algorithm can classify different mental states based on four category risk matrices. In particular, it analyses brainwave signals using long short-term memory (LSTM) designed to remember the previous mental states of each insider and compare them with the current brain state for associated risk-level classification. To evaluate the performance of the proposed system, we performed a comparative analysis using logistic regression (LR)—a predictive analysis technique used to describe the relationship between one dependent binary variable and one or more independent variables—on the same dataset. The experimental results obtained suggest that LSTM can achieve a classification accuracy of more than 80% compared to LR, which yields a classification accuracy of approximately 51%.

INDEX TERMS Deep learning, EEG sensors, fitness evaluation, insider threats, LSTM, safety critical industrial infrastructure.

I. INTRODUCTION

Industrial organization always secure measures to mitigate data security threats. It is a general belief that cyber-attacks are executed by outsiders rather than insiders. Given this belief, a considerable amount of effort has been invested in developing security measures to prevent outsider attacks. However, the reality is considerably different. Carnegie Mellon University [1] conducted a survey wherein they consulted 2,134 global executives regarding the nature of cybersecurity and privacy; the results indicated that the most common causes of breaches in the last 12 months were attributed to inadvertent attacks by insiders (36%) and malicious attacks by insiders (25%). This implies that 61% of all security attacks were organized by insiders. Companies can lose approximately $445,000 dealing with one such attack. The reported average of 3.8 successful attacks per year implies damages of up to $1.7 million, coupled with other long-term effects. Lunt [2] identified via survey that the number of intruders has been growing steadily, and currently, there are technical approaches that can handle this security threat

feasibly and effectively.

Safety critical industrial infrastructures (SCIIs) deploy a periodic psychological survey among their employees to detect internal intrusions [3]. However, this survey-based approach cannot effectively detect or predict any insider threats because surveyees become familiar with the survey questions and can fake their responses. Thus, there have been several attempts to automate such a process using electrocardiogram (ECG) sensors [4–6]; unfortunately, recent studies have indicated that humans can control their ECG signals by practicing special breathing techniques [7].

To overcome the limitations of the previous methods, electroencephalograms (EEGs) have been suggested as a potential alternative, and they are gaining research interest in the security community. EEGs can not only capture cognitive processes in the time frame in which a thought occurs, but also record electrical brainwaves corresponding to the different mental states of a human. Unlike ECG, brainwaves cannot be controlled by humans, which makes EEG suitable for detecting attack intentions. Human brainwaves travel at different frequencies; they are categorized into five bands: Delta, Theta, Alpha, Beta, and Gamma. The delta band is linked with relaxed mental states (e.g., when a subject is in a mentally relaxed mode such as deep sleep). The brainwaves move towards the Gamma band as the subject engages in higher cognitive processes such as complex tasks and high intentions.

Expensive and complex EEG sensors that employ 20 electrodes are routinely used for the diagnosis and treatment of mental and brain diseases and abnormalities [8], whereas economic EEG sensors with fewer electrodes are used in Brain–Computer Interfaces (BCI) employed in the gaming and entertainment industries.

Commercial EEGs have recently been adopted for the implementation of various security systems. For example, they could be used as a trustworthy biometric signal for multimodal authentication systems [9]. Further, EEG signals have been the focus of studies investigating privacy attacks because they can be used to intercept brainwaves to extract confidential information [10] and as a source of information about individuals [11].

A wide range of applications utilizing EEG signals are emerging; these include learning attention detection, assisted systems for disabled people, and brain disease diagnosis. Hadoush et al. identified differences between children with mild and severe ASD using EEG signal analysis with empirical mode decomposition (EMD) and second-order differences plot (SODP); they achieved a model sensitivity and specificity of 100% in EMD, 94.7% in the SODP model, and an overall accuracy of 97.2% [12]. Kim et al. proposed and developed a vision-aided brain-machine interface training system for robotic arm control. Their proposal uses Microsoft Kinect to detect and estimate the 3D positions of possible target objects. The predicted velocity vector for a robot arm input is compensated using artificial potential to follow an intended object among all possible targets. The system trained with two participants with cervical spinal cord injury to explore its possible effects [13]. Ieracitano et al. proposed data-driven machine learning for differentiating subjects with Alzheimer's disease, mild cognitive impairment, and healthy control by analyzing EEG signals obtained via a noninvasive interface.

Another proposal suggested evaluating the power spectral density (PSD) of 19-channel EEG traces and reflecting the relevant spectral profiles into 2D grayscale images (PSD-images). A convolutional neural network with one processing module of convolution, rectified linear units (ReLU), and a pooling layer is used to extract suitable features from PSD-images and perform two- and three-way classification tasks. The CNN achieved an average accuracy of 89.8% in binary classification and 83.3% in three-way classification [14]. However, these applications are based on EEG devices that have more than 15 channels. Such EEG signal analysis using an EEG device with a large number of channels is not practical for use in the context of an industry safety environment. Thus, we plan to demonstrate a high-precision lightweight EEG analysis using an EEG device with five channels only.

Understanding and developing novel signal processing techniques for the analysis of EEG signals specifically targeted to detect the perturbed states of the mind are still open problems in CPS security research. Such a development could help meet the increasing requirements for affordable and effective security monitoring measures [15]; however, privacy-related issues such as preventing the exploitation of EEG signals should be considered for EEG signal processing.

This study proposes a novel EEG sensor-based risk assessment framework using deep-learning techniques as a reliable and cost-effective method that can be employed by SCII to protect its data and valuable assets from insider threats. The proposed framework is expected to detect any individuals (insiders) who have knowledge of and access to SCII facilities and its valuable assets and/or who may attempt unauthorized actions, sabotage, or aid outsiders in such purposes.

The proposed framework can also be viewed as an access control scheme that predicts potential insider threats before any malicious actions can occur. framework not only captures the brainwaves of individuals before they enter their work zones within the SCII, but also decides whether to grant them access based on their current mental state and fitness level to perform their daily responsibilities. The experiments stimulate different emotions and intentions in the user by utilizing a scientific open affective standardized image set (OASIS) [16]. Brainwaves are captured using light Emotiv EEG devices that use five electrodes during the emotional stimulation experiment. This creates a large dataset of EEG signals that can be utilized for learning a model from a long short-term memory (LSTM) network. The final trained model is used to map the captured brainwaves to a risk matrix that classifies the criticality of the potential threats.

The main contributions of this paper are as follows.

- Proof that using a lightweight EEG device with reduced

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access

number of electrodes is sufficiently reliable to build a strong access control system for securing SCIIs
- Taught neural networks to exploit and distinguish between the normal and abnormal EEG signals with a given input difference from random data.
- Tested the strength of different deep learning distinguishers and explored the optimal configuration parameters of a neural network to achieve better overall classification accuracy.

The remainder of this paper is organized as follows. Section II describes the related studies to analyze the gaps in current knowledge. Section III provides the details of the EEG risk framework using the LSTM network. The experimental results are discussed in Section IV. Finally, conclusions are presented in Section VII along with suggestions for future work.

II. RELATED WORK

Studies that focus on the aspect of emotional responses when monitoring emotional changes have been conducted [17, 18]. These studies provide key knowledge for determining the reactions of individuals based on the transformation of their brains.

The techniques for assisting in the measurement of emotional responses include frontal alpha asymmetry, laterality index at rest from near-infrared spectroscopy and comfort vector model [17]. Rahman and Oyama [17] examined the applicability of the aforementioned techniques to validate the extent of positive and negative effects caused by depression-related symptoms such as anxiety. NIRS and EEG techniques have been shown to be useful in detecting changes based on different illnesses [17]. Further, Oyama and Sakatari [19] demonstrated that NIRS and EEG signals are effective in measuring emotional responsiveness.

To detect malicious attacks and data breaches, it is necessary to identify malevolently acting insiders. Current security methods have been unable to assist in handling such issues effectively. Therefore, studies that focus on utilizing human biological data to detect and predict threats caused by individuals within an organization are being conducted [18]. Meanwhile, the "affect-as-information" hypothesis can be used to explain emotional influences [20, 21]. Zadra and Clore [20] argued that the visual perception of an individual is of considerable importance in determining his or her environment. This means that a person can develop positive or negative moods based on their surroundings and actions. Suh and Yim [18] investigated emotional changes in participants; they focused on identifying EEG signals that can be used to detect the desire and behavior of an intruder using an Epoc Emotive wearable device and electrodes. The aforementioned procedures were guided toward measuring how ECG data can supplement the identification of the emotional changes in an individual. In this study, a piece of significant evidence was observed based on the behavior of a person. As a malicious insider would act differently, it is possible to employ EEG indicators to categorize the judgement of an individual.

There is an interesting aspect of human brain functions mandated by various cognitive processes. Suh and Yim [22] listed thinking, judging, and controlling one's emotions to the process. To investigate how a malicious insider can be detected using potential approaches such as EEG analysis, human emotion data play an important role. This school of thought suggests that an emotion can assist in identifying any potential fluctuations in the human brain [22]. The evident changes in human emotional responses could be identified based on various reactions. In particular, behavioral patterns are expressed in the form of feelings and reasoning by perpetrators. This is an important indicator of variations in the brainwaves of an individual [22]. Thus, the analysis of any abrupt changes in one's bodily behaviors can help determine his or her motive as a malicious insider.

There are different theoretical approaches to designing brain networking models. Some studies have focused on information related to the theory of networks [23, 24]. Many of these studies seek to understand different hypotheses formulated to assist in explaining the workings of the human brain. *Stam and Reijneveld* [21] developed the concept of classical graph theory. The success of this approach was determined by a graph that shows a significant transition as a crucial factor in connected networks. Understanding the working nature of the graph therefore assists in interpreting the evident relationship between interaction and action, which means that a network is an ordered interconnection based on graphical representations that need to be appreciated. The basic network types—ordered, small-world, and random [23]—are shown in Fig. 1.

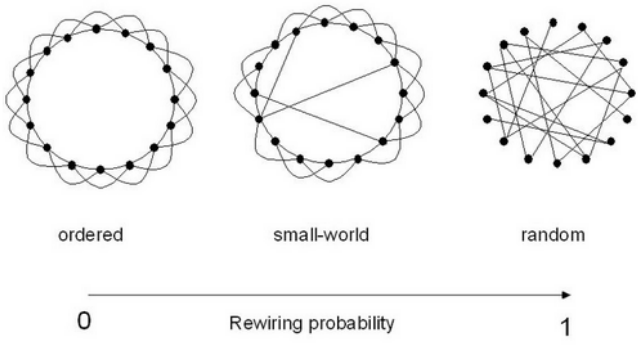


FIGURE 1: Basic network types

The basic network types have unique characteristics. For example, the ordered network is a ring with 16 vertices connected to four neighbors. A graph of such a nature has a high clustering coefficient. In the case of a random network, the procedure involves choosing a random vertex, which results in an increasing rewiring probability. Data mining is a crucial technique in various analyses. The approach to data discovery indicates that there is a way to extract useful information or patterns [24]. Any evident patterns from databases can be recognized, thereby making a unified

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access

conclusion. On the contrary, one needs to understand the source of the data for conducting an EEG analysis. The technique comprises recording operations of the human brain using electrodes [24]. In particular, the neocortex can be considered a major source of such data. As discussed, EEG entails the monitoring of the electrical impulses of the human brain under different scenarios. Interestingly, EEG big data must be comprehended when considering an investigation to detect insider threats to create attacks.

An investigator may need to seize an opportunity in the prediction by employing strategies such as a BCI. As such, the application of the analysis of EEG signals can be effective as it can offer much-needed real-time solutions based on electrographic behavior analysis [25]. The effectiveness of the EEG signals has been appreciated as a working technology to identify that the response of a person to a particular scenario lies in the use of the right algorithm in ensuring effective prediction and processing of real-time data [26].

Jeonghyun et al. [27] proposed a novel methodology for a convolutional neural network (CNN) based on motor imagery (MI) classification using a new form of input. Continuous wavelet transform is applied to the input EEG signal to track the features of MI. They considered the real and imaginary parts of the transformed signal to exploit the magnitude and phase information simultaneously. The features extracted are then fed to a basic CNN with only one convolution layer, one max-pooling layer, and one fully connected layer. Two public BCI datasets were used to test the classification accuracy: BCI competition IV dataset IIb that has data from nine subjects and BCI competition II dataset III that has data from one subject. The proposed method achieved at best 78.3% accuracy using BCI competition IV dataset IIb and an average of 89.6% accuracy rate using BCI competition II dataset III.

Shingchern et al. [28] reported the use of EEG signal recordings to assess if a user likes a video clip he/she is watching. The experiment was performed based on a ready dataset called Database for Emotion Analysis, using physiological signals that provided by Queen Mary University of London in UK. The dataset contains EEG signals that are segmented, and features from each segment are computed prior to classification. They used the k-nearest neighbors, support vector machine (SVM), multilayer neural network, and autoencoder neural networks. The autoencoder achieved the best classification accuracy rate of approximately 80%, whereas other classifiers yielded only approximately 60% accuracy.

Haoyan et al. [29] proposed a new lightweight EEG classification model based on EEG devices with a small number of channels. They used an EEG sensor that uses five channels to perform blind source signal separation to obtain data from several sources, including noise signals such as electro-oculography. Further, they used a combination of EEG and EOG signals and decomposed the signal into different frequency bands, and then, they input them into a lightweight CNN and SVM to perform data classification and obtain five classifiers. The prototype system showed a final accuracy rate of 74.7% using SVM and 80.1% using the CNN.

Chen et al. [30] proposed a hierarchical bidirectional GRU model with an attention mechanism (H-ATT-BGRU) for EEG features learning to perform emotional classification. They showed that the model that explores hierarchical structures such as H-AVE-BGRU and H-MAX-BGRU shows better performance in EEG features classification compared to non-hierarchical models such as CNN and LSTM. They found that the classification accuracy of the H-AVE-BGRU model is 8.4% and 1.9% greater than the accuracy obtained using CNN and LSTM, respectively, for classifying the valence with 8.1% and 2.5% for classifying the arousal feature. The best accuracy was obtained by the H-ATT-BGRU model (66.5%), which is 12% more than the accuracy achieved by the best shallow baseline SVM model. The CNN model did not show a significantly better result because its accuracy in classifying the valence achieved 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%. The LSTM model outperformed the CNN by 6.5% and 5.6% in classifying the valence and arousal, respectively.

Alkalin et al. [31] proposed emotional stimulation using a visual and audio experimental setup and collected physiological signals data from 20 participants using Neurosky EEG device. They used three different classification models: SVM, random forest, and deep learning. Their research showed success in achieving an accuracy rate of 77.04% using SVM, 79.76% using random forest, and 62.86% using deep learning.

These findings provide encouragement for investigations into the use of human brainwave signals to extract useful information to find potential attacks in SCII and analyze them with a high accuracy rate using deep learning algorithms.

III. RESEARCH METHODOLOGY AND ANALYSIS

As depicted in Fig. 2, the first phase of this study focuses on preparing the experimental setup and collecting data from digital marketing campaigns. The second phase includes manipulating data, preprocessing, clearing data, and wrangling and extracting features that ensure the collected data are in the state to be fed into the deep-learning algorithm. The third phase learns a model from the deep-learning algorithm using the preprocessed data obtained during the second phase. The final phase performs a proof of concept considering the full and realistic architecture of SCII.

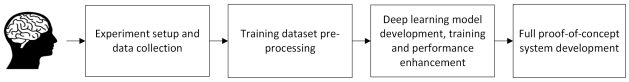


FIGURE 2: Development phases

A. EXPERIMENTAL SETUP AND DATA COLLECTION

This process started with complying to all ethical procedures defined by the Khalifa University Compliance Committee.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access

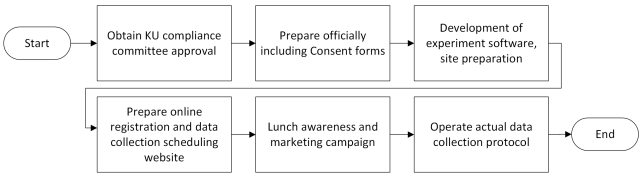


FIGURE 3: Experimental setup and data collection tasks

A dedicated lab for our experiments was setup to facilitate the required conducive atmosphere. A Java-based application that simulates the visual expressions of the participant upon displaying the OASIS was then developed and made available with five main screens, as shown in Fig. 4.

Data collection was conducted in a dark, square room (1.7 m $\times$ 1.7 m) with one study table, the desktop hosting the experimental application, a recording camera, and one chair; this room was specially designed for this research. Figure 5 shows an actual photo of the data collection room.

Finally, marketing campaigns that promote the awareness of this study were launched. An online registration form that allows users to register their preferred date and time to participate in the experiment was created, and confirmations were sent over the phone. These campaigns and registrations were conducted following the guidelines set by the university compliance committee.

B. PRE-PROCESSING

The data collected during this experiment were pre-processed following the processes depicted in Fig. 6

Pre-processing started with using the Emotiv application that removes noise signals (Fig. 7). Emotiv Insight generates a file in the EDF file format, which is converted to a CSV file to ease data manipulation using MATLAB. The signals for each captured image were separated and divided into four categories identified in our risk matrix; the signals were labelled accordingly. The procedure followed to label the data using MATLAB is shown in algorithm 1.

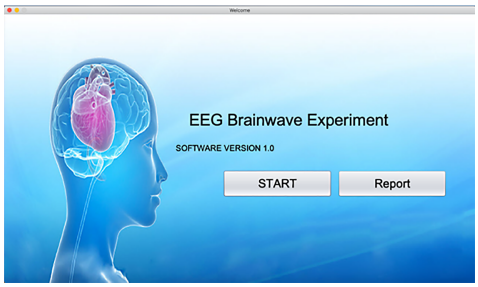
Algorithm 1 MATLAB implementation

procedure DATA LABELING(EEG signal records)

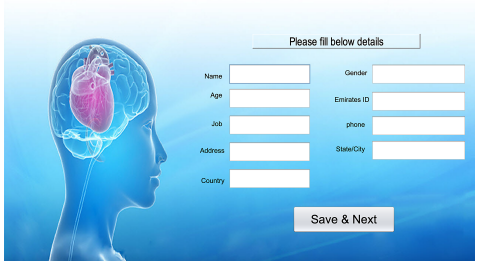
1) Input EDF file format including captured EEG signals from all subjects
2) Convert EDF file format to CSV file format
3) Convert timestamp associated with the EEG signal to local time
4) Segregate the EEG signal for each OASIS image period using the converted timestamp
5) Label each EEG signal corresponding to the OASIS image based on the arousal and variance value of the image

end procedure

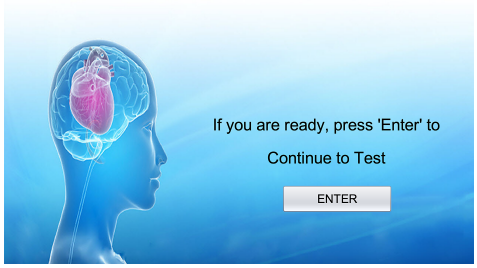
The data files have 26 features including 25 inputs and one output. Each of the five electrodes of the EEG device captures the five brainwave bands: Theta, Alpha, Low_Beta,



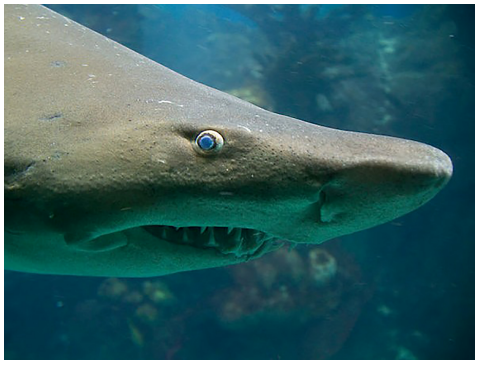
(a) Welcome page



(b) Subject credentials page



(c) Launch experiment page



(d) Experiment images sample

(e) Thank you page

FIGURE 4: Experiment application

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access

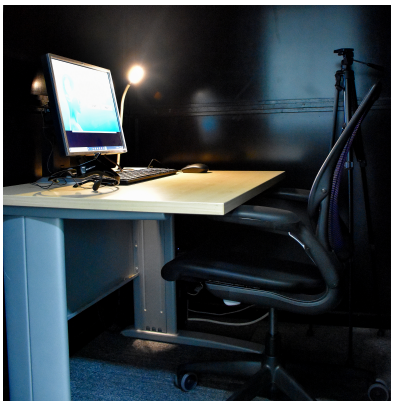


FIGURE 5: Data collection room

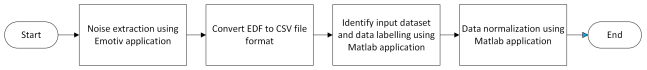


FIGURE 6: Data pre-processing steps

High_Beta, Gamma. We then performed feature extraction by identifying features using the band-to-band power ratio of beta-to-alpha ($\frac{\beta}{\alpha}$) and gamma-to-alpha ($\frac{\gamma}{\alpha}$) waves as they are considered to be dependent variables.

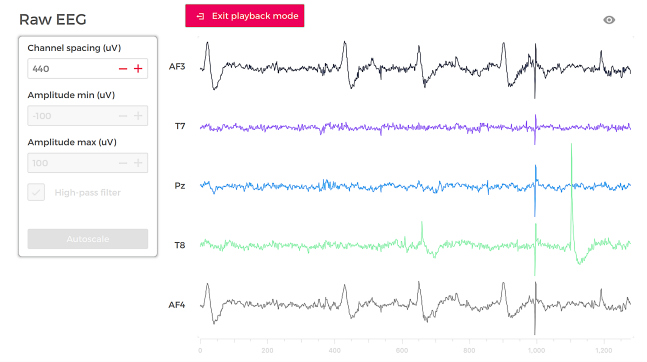


FIGURE 7: Emotiv Insight signal noise sample

This helps to find informative features that can be fed into the deep-learning algorithm to preserve the identities of the subjects, ensuring anonymity. Feature extraction may not be required if it does not improve the deep learning results. In this case, only data normalization needed to be performed. The source code and the dataset are provided in the appendix for validating and reproducing the results.

C. MODEL BUILDING AND OPTIMIZATION

This phase describes the three main activities shown in Fig. 8. To learn a model from the dataset, we preprocessed data from the previous phase, and we utilized a deep learning algorithm. The intermittent results were analyzed and turned by adjusting its parameters and structures.

The dataset was split into training, validation, and testing sets. The training dataset was used to learn the model,

FIGURE 8: Deep learning implementation action

whereas the validation dataset was used for model optimization purposes; finally, the test dataset was used for measuring and evaluating the performance of the model built using an LSTM network scaled up based on the performance analysis to ensure a high (90% - 98%) accuracy rate.

Changes to the accuracy and loss of the model at each epoch were plotted, and the confusion matrix with and without normalization were compared to validate the usefulness of the normalization technique applied to the dataset. The code in algorithm 2 shows the procedure followed in our Keras implementation.

Algorithm 2 LSTM-RNN Keras implementation procedure

1) Input $x = (x_1, x_2, ..., x_t)$, where x_i is a vector of size 25 representing the EEG signal features
2) Outputs $y = (y_1, y_2, ..., y_t)$, where y_i is a vector of size 4 representing hot encoding for four output classifications
3) Shape Input(x) and Output(y) data
4) Define Train data size AND Test data size
5) For $i = 1, 2, \ldots t$ do
6) $i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + b_i)$
7) $o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + b_o)$
8) $f_t = 1 - i_t$
9) $tc_t = g(W_{cx}x_t + W_{cm}m_{t-1} + b_c)$
10) $c_t = f_t \odot c_{t-1} + i_t \odot tc_t$
11) $m_t = o_t \odot hc_t$
12) $y_t = \phi(W_{ym}m_t + b_y)$
13) EndFor
14) Show accuracy rate for training and test dataset
15) Show confusion matrix without nomrmalizatio and with nomalization

In algorithm 2, i denotes the input gate, and o and f are the output gate and the forget gate, respectively. tc represents the input to the memory cell, c denotes the activation vector, and m is the output of the memory cell. Further, "W" denotes the weight matrices from x to the input gate i. b is the bias, and g and h are the activation function of the cell input and cell output, respectively. $\odot$ is the SoftMax activation function.

D. NETWORK ARCHITECTURE AND SPECIFICATION

The sequential model API was used to create the LSTM models, where an instance of the sequential class and model layers were created and added to it. The model was initialized with three LSTM layers of size 128 with parameters batch_input_shape, which indicates that the sequential classification of the LSTM network can accept input data of the defined batch size only, thereby restricting the creation of any variable dimension vectors.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access

The second parameter return_sequence was set true for the first two layers, and it indicates that the recurrent layer of the LSTM network should return its entire output sequence (i.e., a sequence of vectors of specific dimension) to the next layer of the network followed by a dense layer that outputs the prediction.

The model calculates the loss function using the mean_absolute_error and optimizes it using the Adam optimizer function. The model uses the SoftMax function as the activation function to output a vector that represents the probability distributions of a list of potential outcomes and acts as a fully connected layer. The epochs parameter was set to 500 so that the model could parse through the training set 500 times, thereby increasing the overall accuracy of the model.

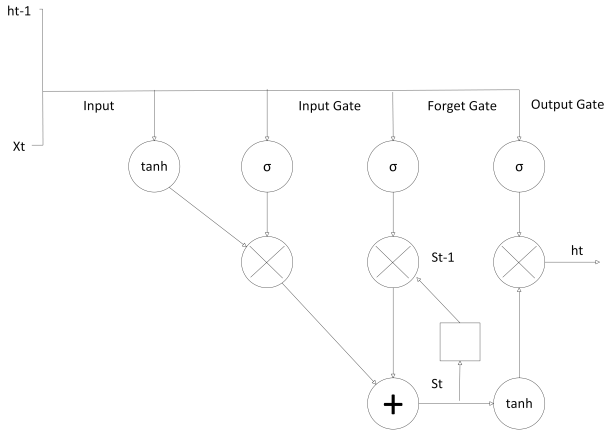


FIGURE 9: LSTM network architecture diagram

E. MATHEMATICAL IMPLEMENTATION

The LSTM is composed of a linear unit surrounded by three logistic gates, They are

- the "Input" gate, which handles the writing of data into the information cell;
- the "Output" gate, which handles the sending of data back onto the recurrent network; and
- the "Keep" gate, which maintains and modifies the data stored in the information cell.

The three gates are the centerpiece of the LSTM unit. The gates, when activated by the network, perform their respective functions. The Input gate will write whatever data are passed into the information cell, the Output gate will return whatever data are in the information cell, and the Keep gate will maintain the data in the information cell; this gate can modify the data based on the signal they are sent.

The Keep gate decides whether to keep or forget data currently stored in the memory. It receives both the input and the state of the recurrent network, and it passes the input through its Sigmoid activation function. If K_t has a value of one, it means that the LSTM unit should keep the data stored perfectly, and if K_t has a value of zero, it means that it should forget it entirely. Consider S_{t-1} as the incoming (previous) state, X_t as the incoming input, and W_k, B_k as the weight and bias for the Keep gate, respectively. In addition, consider Old_{t-1} as the data previously in memory. What happens can be summarized by the following equations.

$$K_t = \sigma(W_k \times [S_{t-1}, x_t] + B_k) \tag{1}$$

$$Old_t = K_t \times Old_{t-1} \tag{2}$$

As Old_{t-1} multiplied by the value was returned by the Keep gate(K_t), this value is written in the memory cell.

Then, the input and state are passed on to the Input gate, in which another Sigmoid activation is applied. Concurrently, the input is processed as normal by whatever processing unit is implemented in the network, and then it is multiplied by the Sigmoid activation's result I_t, shown in

$$I_t = \sigma(W_i \times [S_{t-1}, x_t] + B_i), \tag{3}$$

which is similar to the Keep gate. Consider W_i and B_i as the weight and bias for the Input gate, and the C_t as the result of the processing of the inputs using the recurrent network.

$$New_t = I_t \times C_t \tag{4}$$

New_t shown in equation 4 is the new data to be input into the memory cell. This is then added to whatever value is still stored in the memory, as shown in 5.

$$Cell_t = Old_t + New_t \tag{5}$$

The obtained candidate data are to be kept in the memory cell. The conjunction of the Keep and Input gates work in an analog manner, making it so that it is possible to retain part of the old data and add only part of the new data.

The Output gate functions in a similar manner. To decide what we should output, we take the input data and state and pass it through a Sigmoid function. The contents of our memory cell, however, are pushed onto a Tanh function to bind them between values of -1 to 1. Consider W_o and B_o as the weight and bias for the Output gate shown in

$$O_t = \sigma(W_o \times [S_{t-1}, x_t] + B_o) \tag{6}$$

$$Output_t = O_t \times tanh(Cell_t) \tag{7}$$

Further, $Output_t$ is output into the recurrent network.

F. FULL PROOF-OF-CONCEPT SYSTEM

The proof-of-concept system has two main components: Registration and Access Granting. Registration or pre-registration is the stage where all internal employees of an organization pass through the data collection point. This captures the brainwaves of the employees which are then used to build a model from the LSTM network. The second stage is the access granting stage. If an employee wishes to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access

TABLE 1: Demographics of Participants

#	Gender	Age	Education	Country
1	Female	28	PhD Student	South Korea
2	Male	26	PhD Student	Thailand
3	Male	28	PhD Student	China
4	Male	33	PhD Student	Mongolia
5	Male	21	Student	Vietnam
6	Male	22	Student	Vietnam
7	Male	21	Student	Vietnam

access a critical zone within the SCII, then he or she can be asked to perform a quick brainwave capture using the EEG device. Here, the signals are matched into different risk categories and the access may either be granted or rejected. Fig. 10 shows the architecture of the full proof-of-concept system.

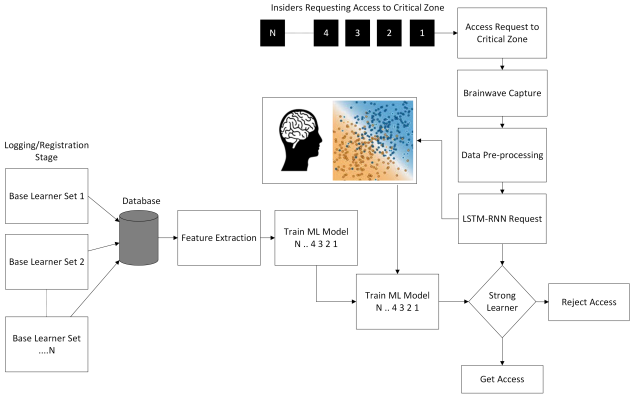


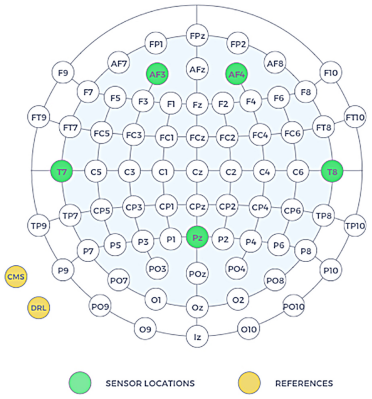
FIGURE 10: Proof-of-concept system architecture

Fig. 11 shows the positions of the electrodes in Emotiv Insight (e.g., green dots) during both data collection and the full proof-of-concept phases. This helps obtain brainwaves related to deceptive and cognitive behaviors. AF3 and AF4 are mapped to Brodmann Areas 9 and 10, respectively. These areas are responsible for gathering information and coordinating intellectual functions and actions. AF3 is related to making decisions about emotional stimuli, attribution of intention to others, and inferential reasoning. AF4 is in charge of decision-making such as conflict and rewards, planning and judgment. T7 and T8 are in the attribution of intentions and Pz is for cognitive process.

IV. RESULTS

Our dataset included seven participants: one female aged 28 years and six males aged between 21 and 33 years. All participants were in good health and did not suffer from any brain illnesses. Table 1 shows the filtered demographics of the subjects collected using the Java application.

More than 10,000 brainwaves were collected. However, after applying several data filtering techniques including the removal of noise signals and margin from the start and end of



(a) Standard 10-20 positions

(b) Emotiv Insight device

FIGURE 11: Emotiv five channels positions

TABLE 2: Data labeling

Category	label_0	label_1	label_2	label_3
0	1	0	0	0
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1

each picture showing time, only 1550 brainwaves are ready to be fed into the LSTM network. The final dataset is split into three different sets: training (60% of entire dataset), testing (40% of entire dataset) and validation set (40% of the training dataset). The target dataset was labelled into four groups that will be used as the final risk matrix where 0 indicating the lowest risk and 3 the highest. Table 2 shows the labelling using the library "Dummies."

Each picture in the OASIS library is attached with two main values: Valence, which shows the degree of positive or negative effect the image evokes, and Arousal, which shows the intensity of the effect the image evokes. Each of these two values has a scale from 1–7 : a Valence value

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access

TABLE 3: 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, AF4_LOW_BETA, AF4_HIGH_BETA, AF4_GAMMA

of 1 reflects "very negative"; 2, "moderately negative"; 3, "somewhat negative"; 4, "neutral"; 5, "somewhat positive"; 6, "moderately positive"; and 7, "very positive." Further, for Arousal, 1 reflects "very low"; 2, "moderately low", 3, "somewhat low"; 4, "neither low nor high"; 5, "somewhat high"; 6, "moderately high," and 7, "very high." We created a four-dimension risk matrix for mapping each image group to one risk category as follows:

- Images with valence value equal to one are labeled as zero: "High Risk"
- Images with valence value equal to two and three are labeled as one: "Medium Risk"
- Images with valence value equal to four and five are labeled as two: "Normal"
- Images with valence value equal to six and seven are labeled as three: "Low Risk"

All images selected as part of this experiment had arousal values more than five to ensure their intense impact on the participants.

The input data to the LSTM network are a collection of five data inputs generated from each of the five Emotiv electrodes. The input data are shown in Table 3

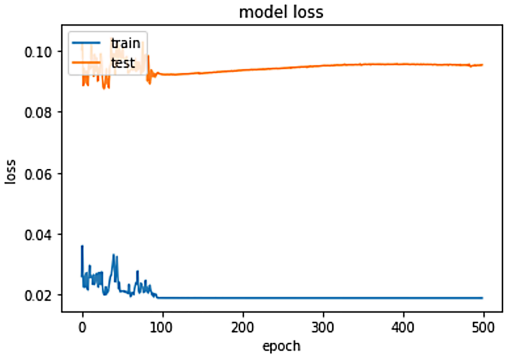
The neural network developed in this experiment had four layers: three LSTM layers each with 128 memories and 1 output *Dense* layer with the same memory size as the LSTM with the *SoftMax* activation function. The network was trained using the *mean absolute error* loss function and the *Adam* optimizer. Table 4 provides the summary of the model.

The model was learned from the training dataset using 500 epochs. We then optimised the model with the validation dataset using 500 steps. The validated model was finally run on the test dataset for evaluation. With the architecture we

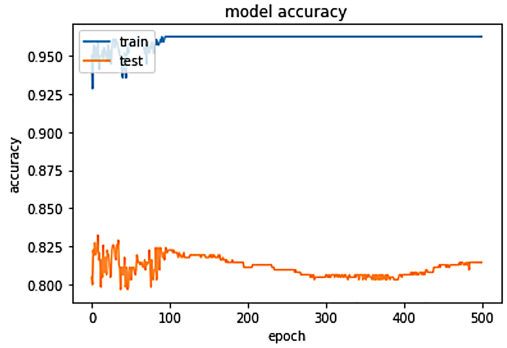
TABLE 4: LSTM model summary

Layer (type)	Output shape	Param #
lstm_1 (LSTM)	(None, 1, 128)	78848
lstm_2 (LSTM)	(None, 1, 128)	131584
lstm_3 (LSTM)	(None, 128)	131584
dense_1 (dense)	(None, 4)	516

Total params: 342,532 Trainable params: 342,532
Non-trainable params: 0



(a) Model loss



(b) Model accuracy

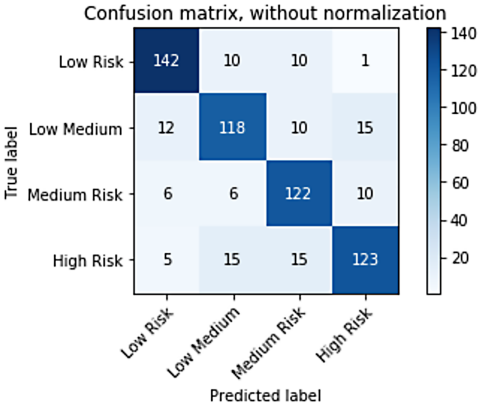
FIGURE 12: Changes in accuracy and loss model

used, we managed to obtain 80.91% accuracy. Fig. 12 shows the model lost and accuracy changes over 500 epochs.

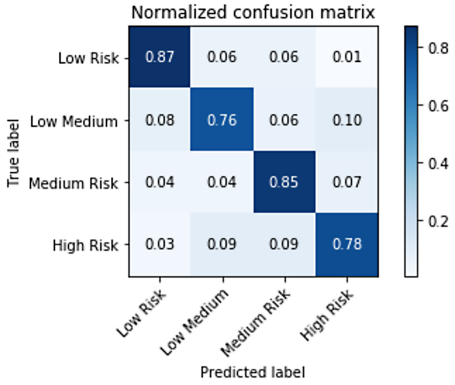
After the first 100 epochs, the training accuracy was stabilized. This phenomenon was also shown with the validation dataset. After the first 100 epochs, the accuracy started decreasing slowly over the next 300 epochs until it increased again over the last 100 epochs. The model loss in both training and testing showed that it aligned with the model accuracy. Over the first 100 epochs, it was unstable; however, it stabilized over the last 400 epochs as shown in Fig. 14.

Fig. 13 shows the confusion matrix with and without normalization. As shown in the figure, it reaches an accuracy of 80.91%. Although we aim to obtain a result that is 90% or above, the accuracy we achieved shows a great potential for further research. True positives and true negatives are within

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access



(a) Confusion matrix without normalization



(b) Confusion matrix with normalization

FIGURE 13: Emotiv five channels positions

an acceptable level for the dataset with some level of false positives and false negatives that are expected because of the noise available in the dataset.

V. EXTENDED EXPERIMENT USING LOGISTIC REGRESSION

A. INTRODUCTION

A Naive Bayes classifier is a probabilistic machine learning model that is used for the classification task. The classifier is based on Bayes theorem, which is used to determine the probability of the hypothesis happening, given that the evidence has occurred. The assumption made here is that the predictors/features are independent. The presence of one particular feature does not affect the other. Hence, it is called naive. The method can calculate conditional probability, which is the probability of an event based on previous knowledge available on the events.

B. ARCHITECTURE

Initially, the model needs to decide what to use as features. It excludes things that may be known but are not useful to the model. The probability is transformed to calculate the required result. For this, some basic properties of probabilities, and the Bayes' theorem are used.

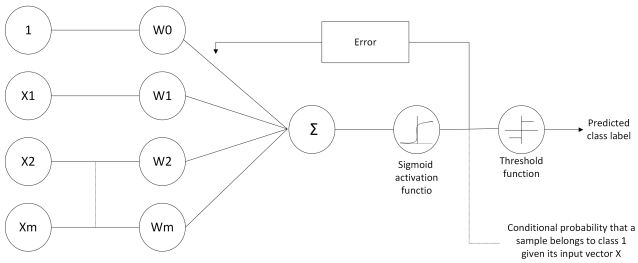


FIGURE 14: Logistic regression architecture

From a statistical point of view, MLE sets the mean and variance as parameters for determining the specific parametric values for a given model. This set of parameters can be used for predicting the data needed in a normal distribution. The MLE assumes a joint probability mass function.

The Sigmoid function, also called the logistic function gives an 'S' shaped curve that can accept any real-valued number and map it to a value between 0 and 1. If the curve goes to positive infinity, y predicted will become 1, and if the curve goes to negative infinity, y predicted will become 0.

C. MATHEMATICAL IMPLEMENTATION

This is a special case of linear regression where the target variable is categorical in the nature; it uses a log of odds as the dependent variable. LR predicts the probability of occurrence of a binary event by utilizing a logit function.

The liner regression equation is given as

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn \tag{8}$$

where y is the dependent variable and x1, x2 ..., and Xn are the explanatory variables.

The Sigmoid function is given by

$$p = \frac{1}{1 + e^{-y}} \tag{9}$$

Applying the Sigmoid function on linear regression,

$$p = \frac{1}{1 + e^{\beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn}} \tag{10}$$

D. MODEL DEVELOPMENT

During the initial stage, the data set is prepossessed into a suitable format for actual processing. The dataset is shuffled using the re-index and the random permutation function, and therefore, the test set contains all classes that achieve better accuracy. Unnecessary features are dropped from the dataset. Feature scaling is performed using the standard scaler function to standardize the independent features present in the data in a fixed range; it is performed during the data preprocessing step to handle highly varying magnitudes or values. Then, the dataset is split into the train and test sets; with a test set size of 20%.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037979, IEEE Access

The LR is imported using LogisticRegression from Sklearn. Gridsearch is used for parameter tuning that will methodically build and evaluate the model for each combination of the algorithm parameters specified in a grid where the parameters are a "penalty" that is used to specify the norm used in the penalization, and 'c' is the inverse of the regularization strength.

E. RESULT

1) Accuracy

The model only achieved a train accuracy of 56% and a test accuracy of 51%. It is the ratio of number of correct predictions to the total number of input samples. The accuracy is generated using the accuracy score metrics.

2) Confusion Metrics

The confusion matrix, also known as an error matrix, is a specific table layout that allows the visualization of the performance of the random forest algorithm. Confusion metrics with and without normalization are created.

The confusion metrics with and without normalization are generated for a wider range of analysis. The model only attained about 51% accuracy, which is not really good. True positives and true negatives are not at an acceptable level for the dataset with several false positives and false negatives.

VI. COMPARISON BETWEEN LSTM AND LOGISTIC REGRESSION

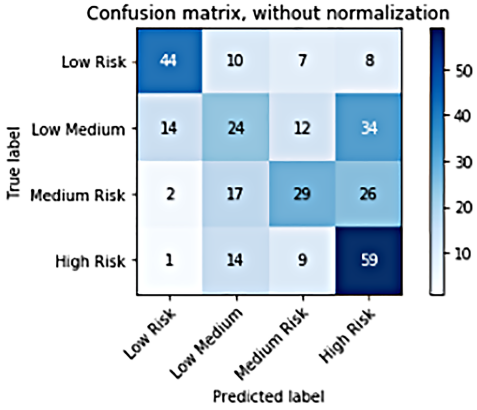
The test accuracy reaches its optimal value in the case of the LSTM model. The LSTM model can utilize the past activities to predict the next activity and achieve the highest accuracy. Since the model is implemented using a deep learning algorithm, it achieved higher complexity. The LSTM model achieved 80.91% accuracy.

Further, the model is compared with the classical LR method. Because the LSTM is more complex than LR, it is understood that the solution achieves a considerably higher accuracy than that of LR. Further, LR is less prone to overfitting but it can overfit in high-dimensional datasets. The LR model achieved an accuracy of 51%.

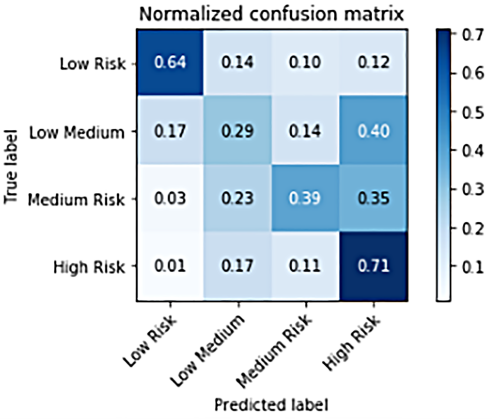
VII. CONCLUDING REMARKS

This paper presented a novel deep-learning-based framework that detects insider threats using human brainwaves. Brainwaves were collected using EEG sensors from seven different subjects and fed into an LSTM network to build a detection model. This framework is particularly designed to protect SCIIs' assets using a comparatively reliable electrophysiological monitoring method that cannot be manipulated or controlled.

The theories of LSTM and its novel architecture were applied to the problem of insider threats using EEG signals. The network included three LSTM layers followed by a single "Dense" layer, each with 128 neutrons. The "Dense" layer uses "SoftMax" as an activation function that has the same number of nodes as the output layer, while both "mean



(a) Confusion matrix without normalization



(b) Confusion matrix with normalization

FIGURE 15: Emotiv five channels positions

absolute error" is used as a loss function and "Adam" as the optimizer. The usefulness and effectiveness of this architecture has been proved by achieving an accuracy of 80.91%, which outperformed the accuracy achieved in a similar work with the H-ATT-BGRU model (66.5%) and random forest (79.76%). The LR model only achieved 51% test accuracy because the model over-fitted on high-dimensional datasets.

We believe that the proposed framework could be refined to achieve a detection accuracy of 90% or more. To achieve this, in the future, the dataset will be increased by collecting additional samples while reducing the number of levels in target classification. The reduction in the noise presented in the dataset will also lead to an improvement in performance significantly.

The future work will include tuning the parameters for optimal performance and experimenting with other potential algorithms to achieve higher accuracy.

APPENDIX

Python deep learning libraries were used for the development of this novel EEG risk framework that identifies insider threats in SCII using deep learning. The corresponding TensorFlow codes and the dataset are available at [32] for users

to reproduce the results obtained in this study.

ACKNOWLEDGEMENT

The authors acknowledge support from the Center for Cyber-Physical Systems, Khalifa University, under Grant Number 8474000137-RC1-C2PS-T3.

REFERENCES

[1] "Analytic approaches to detect insider threats," Software Engineering Institute, Carnegie Mellon University, 4500 Fifth Avenue, Pittsburgh, Tech. Rep., December 2015. [Online]. Available: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=451065

[2] T. F. Lunt, "A survey of intrusion detection techniques," Computers Security, vol. 12, no. 4, pp. 405 – 418, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167404893900295

[3] M. A. M. Frank M. G. and O. M., Human behavior and deception detection. Wiley, 2008.

[4] S. Kim, C. Y. Yeun, E. Damiani, and N. Lo, "A machine learning framework for biometric authentication using electrocardiogram," IEEE Access, vol. 7, pp. 94 858–94 868, 2019.

[5] E. Al Alkeem, S. Kim, C. Y. Yeun, M. J. Zemerly, K. F. Poon, G. Gianini, and P. D. Yoo, "An enhanced electrocardiogram biometric authentication system using machine learning," IEEE Access, vol. 7, pp. 123 069–123 075, 2019.

[6] S.-K. Kim, C. Yeun, and P. Yoo, "An enhanced machine learning-based biometric authentication system using rr-interval framed electrocardiograms," IEEE Access, vol. PP, pp. 1–1, 11 2019.

[7] K. N. . K. I. Perry, S., "Control of heart rate through guided high-rate breathing," Scientific Reports, vol. Sci Rep 9, p. 1545, 2019.

[8] S. Sanei and J. Chamber, EEG signal processing. Wiley, 2007.

[9] A. Azzini, E. Damiani, and S. Marrara, "Ensuring the identity of a user in time: a multi-modal fuzzy approach," in 2007 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, June 2007, pp. 94–99.

[10] I. Martinovic, D. Davies, M. Frank, D. Perito, T. Ros, and D. Song, "On the feasibility of side-channel attacks with brain-computer interfaces," in Proceedings of the 21st USENIX Conference on Security Symposium, ser. Security'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 34–34. [Online]. Available: http://dl.acm.org/citation.cfm?id=2362793.2362827

[11] V. Abootalebi, M. H. Moradi, and M. A. Khalilzadeh, "A new approach for eeg feature extraction in p300-based lie detection," Computer Methods and Programs in Biomedicine, vol. 94, no. 1, pp. 48 – 57, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260708002484

[12] A. E. Hadoush H, Alafeef M, "Automated identification for autism severity level: Eeg analysis using empirical mode decomposition and second order difference plot," Behavioural Brain Research, pp. 362:240–248, 2019.

[13] N. H. L. W. e. a. Kim, Y.J., "Vision-aided brain–machine interface training system for robotic arm control and clinical application on two patients with cervical spinal cord injury," BioMed Eng OnLine, pp. 362:240–248, 2019.

[14] A. B. A. H. C. Ieracitano, N. Mammone and F. Morabito, "A convolutional neural network approach for classification of dementia stages based on 2d-spectral representation of eeg recordings," NEUROCOMPUTING, vol. 323, pp. pp. 96–107, 2019.

[15] P. Lang, M. Bradley, and B. Cuthbert, International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual. NIMH, Center for the Study of Emotion & Attention, 2005. [Online]. Available: https://books.google.ae/books?id=VEW2PgAACAAJ

[16] L. S. . B. M. Kurdi, B., "Introducing the open affective standardized image set (oasis)," IEEE Access, vol. Behav Res 49, p. 457–470, 2017.

[17] L. Rahman and K. Oyama, "Long-term monitoring of nirs and eeg signals for assessment of daily changes in emotional valence," in 2018 IEEE International Conference on Cognitive Computing (ICCC), July 2018, pp. 118–121.

[18] Y. Suh and M. Yim, "An investigation into the applicability of biodata from health wearable devices to insider threat detection in npps," in 57th Annual Meeting for Nuclear Materials Management, 2016.

[19] K. Oyama and K. Sakatani, "Temporal comparison between nirs and eeg signals during a mental arithmetic task evaluated with self-organizing maps," in Oxygen Transport to Tissue XXXVIII, Q. Luo, L. Z. Li, D. K. Harrison, H. Shi, and D. F. Bruley, Eds. Cham: Springer International Publishing, 2016, pp. 223–229.

[20] J. R. Zadra and G. L. Clore, "Emotion and perception: the role of affective information," Wiley Interdisciplinary Reviews: Cognitive Science, vol. 2, no. 6, pp. 676–685, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.147

[21] C. J. Stam and J. C. Reijneveld, "Graph theoretical analysis of complex networks in the brain," Nonlinear Biomedical Physics, vol. 1, no. 1, p. 3, July 2007. [Online]. Available: https://doi.org/10.1186/1753-4631-1-3

[22] Y.-A. Suh and M.-S. Yim, ""high risk non-initiating insider" identification based on eeg analysis for enhancing nuclear security," Annals of Nuclear Energy, vol. 113, pp. 308 – 318, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306454917304218

[23] F. D. V. Fallani, L. d. F. Costa, F. A. Rodriguez, L. Astolfi, G. Vecchiato, J. Toppi, G. Borghini, F. Cincotti, D. Mattia, S. Salinari, R. Isabella,

and F. Babiloni, "A graph-theoretical approach in brain functional networks. possible implications in eeg studies," Nonlinear Biomedical Physics, vol. 4, June 2010. [Online]. Available: https://doi.org/10.1186/1753-4631-4-S1-S8

[24] A. Flexer, "Data mining and eeg," in Statistical Methods in Medical Research, 2000, pp. 395–413.

[25] M. Hosseini, D. Pompili, K. Elisevich, and H. Soltanian-Zadeh, "Optimized deep learning for eeg big data and seizure prediction bci via internet of things," IEEE Transactions on Big Data, vol. 3, no. 4, pp. 392–404, Dec 2017.

[26] X. Zhao, Y. Chu, J. Han, and Z. Zhang, "Ssvep-based brain–computer interface controlled functional electrical stimulation system for upper extremity rehabilitation," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, no. 7, pp. 947–956, July 2016.

[27] J. Kim, Y. Park, and W. Chung, "Transform based feature construction utilizing magnitude and phase for convolutional neural network in eeg signal classification," in 2020 8th International Winter Conference on Brain-Computer Interface (BCI), 2020, pp. 1–4.

[28] S. D. You and C. Liu, "Classification of user preference for music videos based on eeg recordings," in 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech), 2020, pp. 1–2.

[29] H. Xu and X. Xu, "Lightweight eeg classification model based on eeg-sensor with few channels," in 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2019, pp. 464–473.

[30] 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. 118 530–118 540, 2019.

[31] N. Akalin and H. Kose, "Emotion recognition in valence-arousal scale by using physiological signals," in 26th Signal Processing and Communications Applications Conference, SIU 2018, Izmir, Turkey, May 2-5, 2018. IEEE, 2018, pp. 1–4. [Online]. Available: https://doi.org/10.1109/SIU.2018.8404632

[32] A. Alhammadi. (2020) Brainwave EEG Dataset ieee dataport. [Online]. Available: http://ieee-dataport.org/2138