BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **21**, No 3
Sofia • 2021 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2021-0037

Anomaly Detection Using XGBoost Ensemble of Deep Neural Network Models

Sumaiya Thaseen Ikram[1], *Aswani Kumar Cherukuri*[1], *Babu Poorva*[1], *Pamidi Sai Ushasree*[1], *Yishuo Zhang*[2], *Xiao Liu*[2], *Gang Li*[2]

[1]*School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India*
[2]*School of Information Technology, Deakin University, Australia*
E-mails: sumaiyathaseen@gmail.com cherukuri@acm.org poorvababu@gmail.com ushasreepamidi1@gmail.com xiaou.liu@deakin.edu.au gang.li@deakin.edu.au

Abstract: *Intrusion Detection Systems (IDSs) utilise deep learning techniques to identify intrusions with maximum accuracy and reduce false alarm rates. The feature extraction is also automated in these techniques. In this paper, an ensemble of different Deep Neural Network (DNN) models like MultiLayer Perceptron (MLP), BackPropagation Network (BPN) and Long Short Term Memory (LSTM) are stacked to build a robust anomaly detection model. The performance of the ensemble model is analysed on different datasets, namely UNSW-NB15 and a campus generated dataset named VIT_SPARC20. Other types of traffic, namely unencrypted normal traffic, normal encrypted traffic, encrypted and unencrypted malicious traffic, are captured in the VIT_SPARC20 dataset. Encrypted normal and malicious traffic of VIT_SPARC20 is categorised by the deep learning models without decrypting its contents, thus preserving the confidentiality and integrity of the data transmitted. XGBoost integrates the results of each deep learning model to achieve higher accuracy. From experimental analysis, it is inferred that UNSW_ NB results in a maximal accuracy of 99.5%. The performance of VIT_SPARC20 in terms of accuracy, precision and recall are 99.4%. 98% and 97%, respectively.*

Keywords: *Accuracy, Backpropagation network, Intrusion detection, Multilayer perceptron, Long short term memory.*

1. Introduction

A substantial evolution has been observed in the growth of network traffic for different types of technologies such as the Internet of things, smart grids and 5G communications. This has raised serious security concerns over the insecure communication protocols used on the Internet. IDSs can be built to protect against cyber-attacks by adopting additional security processes such as encryption, access control and authentication mechanisms. A network attack is where an attacker gains unauthorised access to the network to perform malicious activities, which may be of two types: Passive and Active. In the former attack, access to the network is obtained and monitored. Valuable information may also be phished, but no changes are made to the data or system. The active attack, despite gaining unauthorised access, also modifies the data.

Deep learning enforces the analysis of a tremendous amount of data, and as it is a self-adaptive algorithm, it improves the study and produces better results. Several classification engines like BPN and DNN [22] handle complex classification tasks, and hence they are successfully applied to intrusion detection.

Deep neural network techniques are implemented in IDS for the following reasons:

- to detect modern attacks in the network [3, 11, 25] excluding those being used in the training model;
- to identify attributes in a set of packets or in a record flow that are significant [24] for an attack identification.
- To enhance the accuracy of the model and categorise the various type of attacks in the network.
- Can learn necessary knowledge for its final result and correct classification by directing the entire network's parameters.

The contributions of this work are:

- Three deep learning approaches, namely MLP, BPN and LSTM, are stacked to form an ensemble. Each of these models is analysed on various test ratios, learning rate, epoch size, and batch rate to identify the best-fit parameters that result in optimal accuracy for both datasets. The XGBoost ensemble integrates the results of the individual models.
- Benchmarking the proposed model with other twinning deep learning models on various performance measures.
- A well-known benchmark dataset is tested along with the VIT_SPARC20 dataset containing unencrypted normal and malicious traffic, encrypted normal and malicious traffic of different MIME types.

This paper is structured as follows: Section 2 discusses an extensive review of IDS using deep learning techniques in recent studies. An anomaly identification model is developed, which is discussed in Section 3. The performance of individual deep learning approaches on various learning parameters is analysed. Also, the optimal accuracy is determined in Section 4 using different test sets on the proposed ensemble. We also compare the performance of various test ratio sizes and machine learning techniques. Lastly, Section 5 presents the conclusions.

2. Related work

Deep learning approaches effectively detect sophisticated associations within raw samples with various stages of abstraction without human interference. Feature learning and classification tasks in IDS have been implemented by many deep learning techniques as discussed in the literature given below in chronological order in the last few years. A CNN data filter is deployed for packet-data anomaly detection [6] by storing incident signatures. An accuracy of 98.7% is obtained in the fully connected layer with 512 neurons. Backpropagation neural networks with sample-query and attribute-query have been proposed to develop IDS [8] which is feasible and results in effective execution by choosing the specific attributes to analyse, model and identify the complex attacks on a network. The benchmark KDD dataset has been used to evaluate and prove that the proposed system classifies effectively with less training cost. A review of 10 DNN papers, 7 RNN papers and 7 CNN papers with different hidden neurons and learning rates is available [11]. In comparison to DNN and RNN, CNN exhibits higher accuracy of 97.376 % for the CIC_IDS2018 dataset and 98.371% for the Bot-IT dataset, respectively, with 100 hidden neurons and a 0.5 learning rate.

Deep neural networks and association analysis [13] are deployed for a two-level anomaly detection system. The raw data is collected and preprocessed, which is then used to train the model to categorise the data. The association rules between the various features of the dataset are framed using the apriori approach. Then the classified data is matched with the rules, and mismatched information is identified as malicious traffic and alarm logs are generated. This system results in reduced false-positive rates when tested against the KDDCup 1999 dataset. The deep neural network with four hidden layers and using the ReLu activation function reveals the highest accuracy of 82.74% when compared to networks with different hidden layers and other models. However, the precision is 0.88, which is less in comparison to other models.

Similarly, the DNN-4 model is trained and tested with other networks to prove that the accuracy of intrusion detection is higher with four hidden layers. An ensemble of machine learning techniques, namely, decision tree, random forest, KNN and DNN, is deployed for intrusion detection [14]. NSL-KDD dataset is used to test the proposed model, resulting in an accuracy of 85.2%, which is effective compared to other models.

A network intrusion detection based system using LSTM [16] is developed, which acts as a multi-class classifier to detect the anomalies and classify the attacks as normal, suspicious, unknown, attacker and victim. It avoids long-term dependency problems by achieving disappearing gradient descent to identify the weights in the network. RMSprop optimiser is used to efficiently calculate large datasets with a learning rate of 0.01, 6 hidden layers, and 200 epochs, which results in higher accuracy, precision, and recall than SVM, MLP and Naive Bayes. Multi-channel IDS by LSTM neural network is developed [17] and reported a detection rate, accuracy and FAR of 99.23%, 98.94% and 9.86%, respectively, on the NSL-KDD dataset. The authors analysed several neural network models with different activation functions and learning algorithms [20]. The results obtained are diversified to greater extents depending on the group of data used to analyse the best activation functions. There is no much difference in improvements on using recurrent networks over multilayer networks; however, characterisation of the features and selecting the most appropriate activation function is essential to get the best performance in terms of accuracy.

A cyber-physical IDS is developed [21] to identify cyber-attacks against vehicles. The system uses both RNN and deep MLP to achieve higher accuracy with maximum consistency than machine learning approaches such as SVM and k-Means clustering. The proposed method is tested for vulnerabilities such as command injection, DoS and malware attacks. Deep learning techniques [25] enhance the functionality of intrusion detection systems to deal with real-time threats in a reactive manner. The raw data collected is preprocessed, transformed, allied to the association rules, and time-based hold out validation is performed. The models perform better in terms of minimal prediction error and better steadiness. LSTM derivatives are the best models for performance. The performances of the models are worse if very few data samples are present. Different machine learning techniques have been performed on the VIT_SPARC20 dataset and have determined that Random Forest performs superior to other approaches with 98% accuracy [26].

Deep learning is used for detecting anomalies in real-time by Restricted Boltzmann Machine (RBM) and Deep Belief Networks (DBN) [29]. RBM and AE are self-learning algorithms that extract features from unlabeled data and stacking them with undirected connections. A deep belief network is created by passing an RBM with one hidden layer to another RBM. The dataset is fine-tuned by Logistic Regression (LR) approach using multi-class softmax. Contrastive divergence is used for the efficient training of the network. The network is fine-tuned by adding the hidden units set corresponding to the labels and implementing a wake-sleep algorithm, LR with softmax classifier, which outperforms other algorithms in fine-tuning to classify more than one type of attack. DBN improves the detection rate to 97.9% with a minimal FNR of 2.47%. An IDS [31] using deep neural networks is developed to automatically identify and categorise cyber-attacks at the network and host levels. The unforeseen and unpredictable cyberattacks are dangerous to locate because of the continually changing nature of the malicious attacks and occurrence in large volumes causing tremendous effects. Various publicly available datasets have been experimented with by selecting the optimal network parameters. It is observed that the performance of deep neural networks is higher than the classical machine learning classifiers. The results outperformed SVM and MLP. An intrusion detection model is built [38] using chi-square feature selection and integration of classifiers like SVM, Modified Naive Bayes (MNB) and LPBoost. The class label is predicted by majority voting of SVM, MNB and LPBoost which is an optimal solution in comparison to a single classifier.

A network intrusion model using a convolutional autoencoder is built and tested on the CTU-UNB and the Cotnagio-CTU-UNB datasets [33]. The neural network model is made by the Theano tool. The pre-training and fine-tuning process have used a learning rate of 0.001 and 0.1, respectively. A 0.99 value of the ROC curve is obtained for both 6-class and 8-class classification. The model has achieved an accuracy of 99.59%. Anomaly traffic detection is developed using a neural network with two layers [34]. An improved LetNet-5 CNN is used in the first layer to extract the spatial attributes. The temporal features of the flow are extracted using LSTM in the second layer. CICIDS2017 dataset has been used for analysis, and the performance has been surpassed by 94%. The proposed system achieves higher accuracy, recall, precision, and F1-score than other supervised machine learning and ensemble approaches. A model is built using CNN for an online transaction in a viable bank [35]. One month data has been divided into train and test sets. A precision of 91% and a recall of 94% have been obtained. A Deep-Full Range (DFR) is developed [36], a lightweight framework for identifying novel attacks in encrypted traffic.

Table 1 summarises the most cited IDS research in the recent few years using deep learning approaches. Besides, the structure of the model, datasets used and performance results are also discussed. Literature on the domain of deep learning for anomaly detection deploy older datasets (generally KDD99 and NSL-KDD) for model training and evaluation. The drawback of these datasets is that they do not contain contemporary network traffic, encrypted traffic and incidents. Therefore, these datasets cannot be used for validating novel techniques. It is highly recommended recent intrusion detection datasets [28] to be used by researchers to build models. All models in the literature use flow records, and no analysis at the packet level is performed [12]. Different varieties of ensemble neural networks can also be used in building IDS to analyse the performance.

Table 1. IDS using Deep learning techniques

Model type	Model structure	Dataset	Results
LSTM with 20 hidden nodes [29]	20 hidden nodes	KDD99	Acc=93.82%
DBN [4]	4 layers in DBN with 150,122,90 and 50 nodes with a softmax layer	NSL-KDD (40% subset)	Acc=97.45%

LSTM [19]	One hidden LSTM layer of 80 nodes	KDD99	Acc=96.3%,
			Rec=98.88%,
			FAR=10.04%

DBN+LR [5]	4 layers of DBN with 72,52,40 and 5 nodes integrated with logistic regression	KDD99 (10% subset)	Acc=97.9%,
			Recall=97.5%,
			FNR=2.47%

DBN+DNN [18]	DNN is initialized by pre trained weights	Custom CAN dataset	Acc=97.8%,
			FAR=1.6%,
			FNR=2.8%

Autoencoder + Softmax [1]	3 layers with 150,120 and 50 nodes in stacked AE with a softmax	NSL-KDD	Acc=99.2%;
			Recall=99.27%;
			FAR=0.85%

DNN [10]	245 nodes in four hidden layers. Activation function is ReLu	NSL-KDD	Acc=98.27%,
			Recall=96.5%

Multiple LSTM [17]	Training different LSTM nets with one hidden layer and integrated by majority voting	NSL-KDD	Acc=98.94%,
			Recall=99.23%,
			FAR=9.86%

DNN [31]	5 hidden layers containing 1024,768,512,256 and 128 nodes. Activation function is ReLu	NSL-KDD	Acc=78.5%,
			Prec=81.0%,
			Rec=78.5%,
			F1=76.5%

3. Proposed model

3.1. Motivation

An ensemble of LSTM, BPN and MLP is chosen for the proposed method as a two-level deep machine learning model is well suited for anomaly detection [23]. Diverse deep learning solutions can analyse and estimate their suitability and performance for different network traffic loads. This is due to the following reasons [2]:

- LSTM is well-suited to process and predict time series lags of unspecified length.
- Insensitivity to gap length is a benefit of LSTM over different RNNs, hidden Markov models and other sequence learning approaches.
- Backpropagation is a fast, simple, standard and flexible method that does not require prior knowledge about the network.
- There is no assumption in MLP about essential probability density functions. Besides, no assumption about the pattern classes probability in comparison to other probability-based approaches.

3.2. Proposed model

Fig. 1 shows the proposed ensemble deep learning model for intrusion detection. Two different datasets are used for analysis, namely benchmark UNSW-NB dataset and VIT_SPARC20. Preprocessing is performed initially to remove any anomalies like missing values or outlier values in the dataset. Overfitting or under-fitting problems are also checked and are prepared to fit the model well. A best-first search classifier for feature reduction in the dataset is highlighted in appendix table B to enhance the model accuracy. The dataset is split into train and test sets. Three network models, namely MLP, BPN and LSTM, obtain the training data as input. Each model performs 10-fold cross-validation with various hyperparameters. These parameters include the number of epochs, learning rates, activation function and optimisers. The validation sets which yield the best cross-validation accuracy and low MSE are detailed in the experimental section. Each model output is sent to the XGBoost in the model evaluation stage. In the model prediction stage, XGBoost is fed with new data samples to determine the class labels. Different test sets are given as input to the XGBoost, and the average metrics of all test sets are considered the final result.

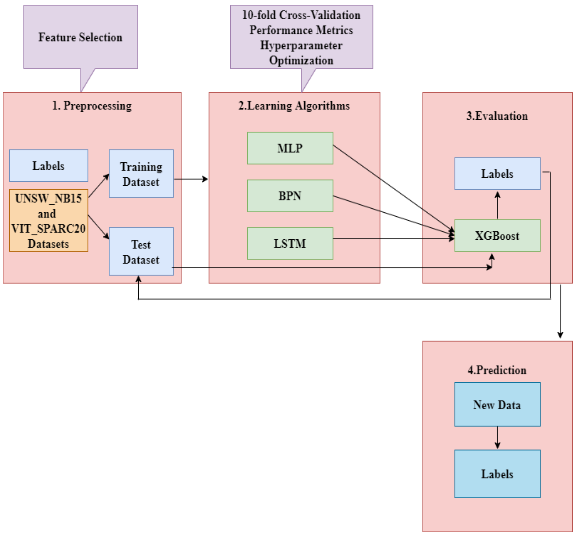

Fig. 1. Proposed High Level Intrusion Detection Model

3.3. Parameter analysis of Deep learning models

Fig. 2 shows the parameter tuning approach of the MLP neural network model. MLP neural network is trained with various hidden nodes, epochs and learning rate. Each of the deep learning models such as BPN and LSTM are analyzed in a similar manner comprising of parameter setting, model training and model verification. The results are discussed in experimental analysis section. In BPN, the parameter settings can be tuned for hidden layer, learning rate, momentum term, number of hidden neurons, and learning cycle. In LSTM, the parameters which can be tuned are hidden nodes, timesteps, input dimensions and dropout value. Parameter settings play a significant role in the increase in accuracy and reduction in error rate in the learning models.

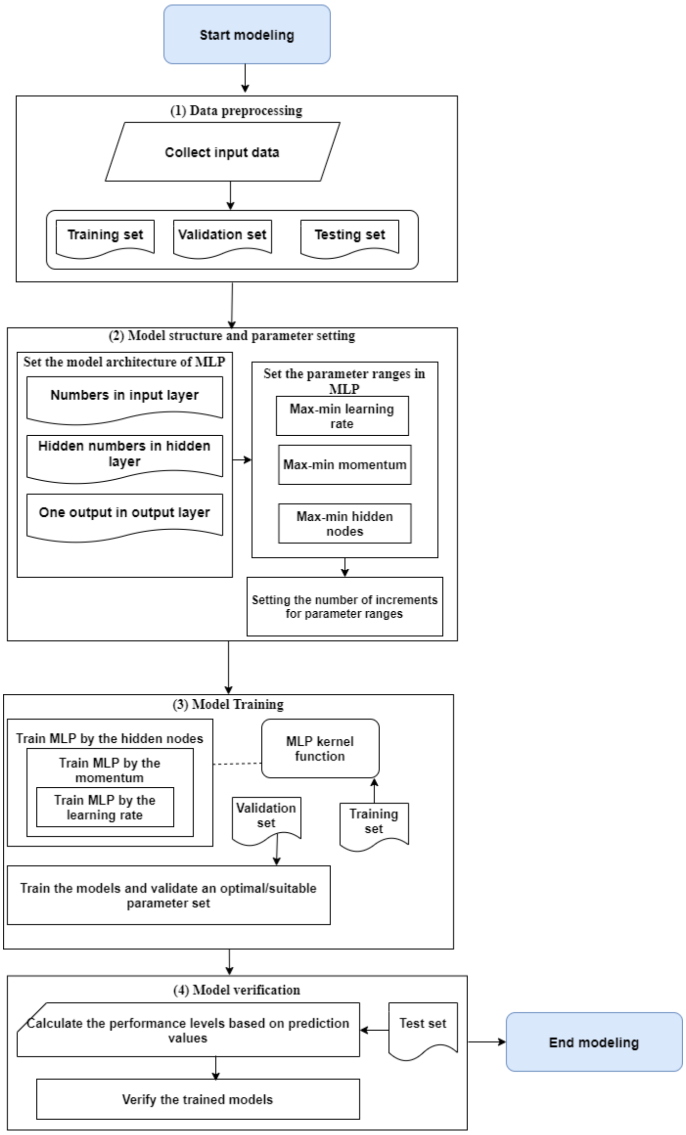
Start modeling
(1) Data preprocessing
Collect input data
Training set
Validation set
Testing set
(2) Model structure and parameter setting
Set the model architecture of MLP
Numbers in input layer
Hidden numbers in hidden layer
One output in output layer
Set the parameter ranges in MLP
Max-min learning rate
Max-min momentum
Max-min hidden nodes
Setting the number of increments for parameter ranges
(3) Model Training
Train MLP by the hidden nodes
Train MLP by the momentum
Train MLP by the learning rate
MLP kernel function
Validation set
Training set
Train the models and validate an optimal/suitable parameter set
(4) Model verification
Calculate the performance levels based on prediction values
Test set
Verify the trained models
End modeling

Fig. 2. Parameter Tuning for the MLP neural network

4. Experimental results

4.1. UNSW_NB15

This dataset contains real recent normal events and synthetic current attack activities generated in the ACCS Cyber Range lab using the IXIA PerfectStorm tool. The tcpdump tool has been utilised to capture raw traffic approximately around 100GB. There are 49 features in the dataset, including the class label generated using the Argus and Bro-IDS tools.

4.2. VIT_SPARC20

This dataset is generated in our institute, VIT, and is captured using network monitoring tools. The entire dataset of VIT_SPARC20 is described in table B of the appendix section. Different packet types are generated using protocols like HTTP and HTTPS. Different mime types are sent from source to destination in the testbed network like a text document, audio, video, and image. Malicious viruses embedded in one of the MIME types are also sent over HTTP and HTTPS. Thus, different types of data that are generated and captured are normally unencrypted traffic, normal encrypted traffic, malicious unencrypted traffic and malicious encrypted traffic. Certain features in the traffic that play a minimal role in classification are removed and extracted using a best-fit search. Table 2 shows a sample of different types of packets generated by sending normally unencrypted traffic, malicious unencrypted traffic, normal encrypted traffic and malicious encrypted traffic. These are the four broad categories of traffic. Also, major traffic categories are split based on the nature of messages communicated between two ends of the network [30]. For example, unencrypted traffic over a network is categorised into HTTP, HTTP get and HTTP post; unencrypted malicious traffic contains specific HTTP ok packets. In total, there are thirty-two attack categories of packets. This is described in the GitHub link (**https://github.com/sparc2020/VIT_SPARC20)**.

4.3. Analysis

Various deep learning models, namely MLP, BPN and LSTM, are trained and tested against different epochs, learning rates, and other optimisers to identify the superior model. Each model performs 10-fold cross-validation. Different epoch values of 10, 20, 50 and 100 are cross-validated with various learning rates of 0.01, 0.1 and 0.3 with varying validation sets. Table 3 displays the cross-validation accuracy results for the sample sets on the chosen parameter values.

The experiments on the first MLP model in the ensemble are performed with four hidden layers. The weights of the neurons initialised and bias values are calculated based on the attributes and weights of the initial neuron. Hidden layers use the sigmoid activation function, and the output layer uses the ReLu function. The final result obtained is optimised using softmax cross-entropy classifier and Gradient Descent optimiser.

Table 2. Sample Dataset Attribute Values of VIT_SPARC20

Attributes	Normal traffic packet	Encrypted traffic packet	Malicious traffic packet	Encrpyted malicious traffic packet
Frame size	342	92	349	1454
Protocols in frame	HTTP	TLS	HTTP	TLS
Colouring rule name	HTTP	TCP	HTTP	TCP
Sequence	155	1777	1	1401
Ack	1	5623	314	518
Duration	288	38	295	1400
Data in File	27	0	0	0
Total size	328	78	335	1440
Time to live	51	128	58	88
TCP payload	288	38	295	1400
TCP segment data	288	0	0	1277

The model exhibits higher accuracy for 100 epochs and a learning rate of 0.3 for both datasets when the test size is 0.2. The use of linear and softmax activation functions reduces the error rate but exhibits the same or lower accuracy values than sigmoid and ReLu functions. The MLP model results in an accuracy of 93.23% and 97.79% for the UNSW-NB15 and VIT_SPARC20 datasets, respectively, for 100 epochs, 0.3 learning rate and 80 % training.

Table 3. Comparative Analysis of Various MLP Attributes

Dataset	Validation ratio	Epochs	Learning rate	Cross-Validation Accuracy	MSE	Execution time (in min)
UNSW-NB15	0.8	100	0.3	93.23	5.2967	86.1
	0.7	100	0.3	91.53	9.1545	60.1
	0.6	100	0.3	92.86	6.3105	100.2
VIT_SPARC20	0.8	100	0.3	97.79	6.9053	88.1
	0.7	100	0.3	96.34	1.9505	91.3
	0.6	100	0.3	96.32	19.5900	87.9

The comparative measures of the MLP parameter values are specified in Table 3. Table 4 shows the second model of the ensemble, the BPN and its parameters tested on various test sizes of 20, 30 and 40%, respectively. BPN is experimented with two hidden layers, the number of repetitions set to 2, and the network training algorithm applied is backpropagation. The start weights of the neurons are initialised to random values. The model is tested against learning rates 0.01, 0.1 and 0.3 and threshold values of 0.1, 0.2 and 0.5. UNSW-NB15 dataset produces higher accuracy at 0.3 learning rate. Maximum accuracy of 96.99% is obtained for the 0.2 train ratio. Higher accuracy of 94.31% is obtained for the self-generated dataset when the learning rate is 0.01, and the threshold is 0.1. The highest results obtained for the various model parameters are listed in Table 4.

The final model in the ensemble is the LSTM model, which displays different results for different optimisers, like Adam, Anagrad, SGD, Nadam and RMSprop, respectively. These optimisers are tested against three batch sizes of 1, 32, and 128 with various test ratios of 20%, 30% and 40% for the two datasets implemented in the study with ReLu activation function and 100 epochs. However, the results obtained both for train and test accuracies are undesirable. This model has exhibited diverse accuracy values for any test ratio, batch size, and optimiser used. The highest cross-validation accuracy of 99.5% is obtained for the UNSW-NB15 dataset when the validation ratio is 80 %; the batch size is 32, and Adam optimiser is used. The execution time to obtain the best parameters is 72 minutes. However, a cross-validation accuracy of 99.4% is obtained for the self-generated dataset when the the validation ratio is 30% with a batch size of 1, and Adam and Nadam optimisers are deployed. The execution time of the model is 27 min.

Table 4. Comparative Analysis of Various BPN Attributes

Dataset	Validation ratio	Threshold	Learning rate	Cross-Validation Accuracy	MSE	Execution time (in min)
UNSW-NB15	0.8	0.1	0.3	96.99	8.693438×10^{-14}	90.4
	0.7	0.2	0.3	94	0	92.6
	0.6	0.5	0.3	93	0	70.7
VIT_SPARC20	0.8	0.1	0.01	94.31	0.0152	20.2
	0.7	0.2	0.01	92.38	0.06143	19.1
	0.6	0.5	0.01	90.09	0.0471	18.9

The corresponding results with higher values for the different attributes and optimisers are depicted in Table 5.

Table 5. Comparative analysis of various LSTM attributes

Dataset	Validation size	Batch size	Optimiser	Cross-Validation accuracy (%)	Execution time (in min)
UNSW-NB15	0.8	1	SGD	96.5	15.3
		32	Adam	99.5	27.8
		128	Adagrad	95.12	20.2
	0.7	1	SGD	92.76	50.4
		32	SGD	91.1	56.5
		128	RMSprop	90.1	60.1
	0.6	1	Adam	92.9	28.1
		32	Nadam	93.7	22.5
		128	Adagrad	90.1	10.1
VIT_SPARC20	0.8	1	Adagrad, SGD	95.5	28.4
		32	SGD	96.3	27.2
		128	SGD	95.6	27.1
	0.7	1	Adam, Adagrad, Nadam	99.4	20.0
		32	RMSprop	95.4	30.1
		128	SGD	95.5	27.1
	0.6	1	SGD	95.5	72.8
		32	SGD	97.7	60.2
		128	SGD	96.4	60.1

4.4. Discussion

All the deep learning models, namely; MLP, LSTM, and BPN, are deployed and evaluated on UNSW-NB15 and VIT_SPARC20 datasets. XGBoost integrates all three deep learning models to result in a maximum accuracy of 99%. Table 6 shows the performance metrics of the proposed model on different test sets on the

UNSW-NB15 dataset. The reason for selecting ten different subsets is 10-fold cross-validation which is deployed in the datasets. However, the VIT_SPARC20 dataset has fewer samples. The additional performance metrics are derived on the complete dataset in a single iteration with an accuracy of 99.4%, precision of 98%, recall and an F1-Score of 97%. Table 7 shows that the proposed deep learning ensemble performs superior to other deep learning-based IDS developed in recent years. However, there is an increase in training time compared to the existing DBN+DNN approach because of the three learning approaches deployed in the proposed model ensemble. The training time can be reduced by minimising the number of nodes in the final hidden layer of deep learning models.

LSTM has been tuned with different hyper-parameters, namely batch size and optimisers, to produce an optimal result in the model training phase. The optimal hyper-parameters for the maximum cross-validation accuracy are the batch size of 32 tuned with Adam optimiser. Nearly 60000 iterations are executed, and fusion by XGBoost is performed to obtain optimal accuracy of 99.5% on the UNSW-NB15 dataset. Also, a cross-validation ratio of 80% also resulted in superior results.

Table 6. Performance Metrics obtained on Proposed Ensemble on different Test Sets on UNSW-NB15 dataset

Subset size	Accuracy	Precision	Recall	F1-score
10%	99	96	94	95
20%	98	95	93	92
30%	97	97	95	96
40%	97	96	98	97
50%	98	97	96	95
60%	99	99	97	96
70%	97	95	96	97
80%	99	98	97	95
90%	99	96	98	97
100%	99	97	96	96

Table 7. Accuracy Comparison on the various machine and deep learning techniques used for IDS on UNSW-NB dataset

Technique	Accuracy (in %)	Precision (in %)	Recall (in %)	F-score (in %)	Time to train (in min)
Stacking ensemble [27]	94	96	93	95	80
SVM	89.63	88.99	90.27	88.12	49
CNN+BiLSTM [15]	77.16	75.80	77.26	78.78	52.7
Pelican [32]	86.64	80.30	86	88	66
TSDL [37]	89.13	81.06	88.89	82	93
RF [9]	92.5	88.9	88.7	92.93	60.92
Random Tree [9]	92.16	90.6	90.3	89.9	65
Neural Network with minimal feature set [2]	95.85	98.07	97.19	98.36	105.5
Proposed XGBoost deep learning ensemble	99.5	99.45	99.42	99.52	70.5

5. Conclusion

Intrusion detection plays a primary role in identifying novel incidents in network traffic. Neural network-based anomaly detection proves beneficial for professionals and researchers as they improve accuracy and minimise FAR. Industry solution providers such as CISCO use machine learning and deep learning models extensively in various security solutions such as advanced threat solutions, intrusion detections, etc. In this paper, we have built an anomaly identification model integrating approaches like MultiLayer Perceptron (MLP), BackPropagation Network (BPN) and Long Short Term Memory (LSTM) by XGBoost. Each of the individual approaches is executed on various cross-validation sizes and various learning parameters. Multiple iterations have been run with various learning parameters to obtain the best cross-validation accuracy. XGBoost integrates the trained results of all three models to produce the best prediction result. It is observed that during training, the Adam optimiser deployed on LSTM resulted in an increase in cross-validation accuracy on UNSW-NB15 and VIT_SPARC20 dataset. LSTM is highly sensitive to different attribute modifications and random weights of the neurons, which is the reason for maximum accuracy. However, it also takes a longer time compared to other models to train the network. LSTM can be exceptionally functional to classify the packets into various types without decrypting their contents. Also, it can predict new types of attacks for which the model is not trained by learning from complex relations between the attributes and does not impose any restrictions on the input variables. The proposed model is compared with the existing deep learning ensembles to show the accuracy and other derived metrics. Further, it will be possible to deploy such an ensemble for online neural network-based anomaly detection as an enhancement.

***Acknowledgements*:** This research was undertaken with the support of the Scheme for Promotion of Academic and Research Collaboration (SPARC) Grant SPARC/2018-2019/P616/SL "Intelligent Anomaly Detection System for Encrypted Network Traffic".

References

1. Abeshu, A., N. Chilamkurti. Deep Learning: The Frontier for Distributed Attack Detection in Fog-to-Things Computing. – IEEE Communications Magazine, Vol. **56**, 2018, No 2, pp. 169-175.
2. Ahmad, H., A. Arif, A. M. Khattak, A. Habib, M. Z. Asghar, B. Shah. Applying Deep Neural Networks for Predicting Dark Triad Personality Trait of Online Users. – In: Proc. of 2020 International Conference on Information Networking (ICOIN'20), IEEE, 2020, pp. 102-105.
3. Aldweesh, A., A. Derhab, A. Z. Emam. Deep Learning Approaches for Anomaly-Based Intrusion Detection Systems: A Survey, Taxonomy, and Open Issues. – Knowledge-Based Systems, Vol. **189**, 2020, pp. 105-124.
4. Alom, M. Z., V. R. Bontupalli, T. M. Taha. Intrusion Detection Using Deep Belief Networks. – In: Proc. of 2015 National Aerospace and Electronics Conference (NAECON'15), IEEE, 2015, pp. 339-344.
5. Alrawashdeh, K., C. Purdy. Toward an Online Anomaly Intrusion Detection System Based on Deep Learning. – In: Proc. of 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA'16), IEEE, 2016, pp. 195-200.
6. Basumallik, S., R. Ma, S. Eftekharnejad. Packet-Data Anomaly Detection in PMU-Based State Estimator Using Convolutional Neural Network. – International Journal of Electrical Power & Energy Systems, Vol. **107**, 2019, pp. 690-702.
7. Berman, D. S., A. L. Buczak, J. S. Chavis, C. L. Corbett. A Survey of Deep Learning Methods for Cyber Security. – Information, Vol. **10**, 2019, No 4. 122.

8. Chang, R.-I., L.-B. Lai, W.-D. Su, J.-C. Wang, J.-S. Kouh. Intrusion Detection by Backpropagation Neural Networks with Sample-Query and Attribute-Query. – International Journal of Computational Intelligence Research, Vol. **3**, 2007, No 1, pp. 6-10.
9. Dahiya, P., D. K. Srivastava. Network Intrusion Detection in Big Dataset Using Spark. – Procedia Computer Science, Vol. **132**, 2018, pp. 253-262.
10. Diro, A., N. Chilamkurti. Leveraging LSTM Networks for Attack Detection in Fog-to-Things Communications. – IEEE Communications Magazine, Vol. **56**, 2018, No 9, pp. 124-130.
11. Ferrag, M. A., L. Maglaras, S. Moschoyiannis, H. Janicke. Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. – Journal of Information Security and Applications, Vol. **50**, 2020. 102419.
12. Gamage, S., J. Samarabandu. Deep Learning Methods in Network Intrusion Detection: A Survey and an Objective Comparison. – Journal of Network and Computer Applications, Vol. **169**, 2020. 102767.
13. Gao, M., L. Ma, H. Liu, Z. Zhang, Z. Ning, J. Xu. Malicious Network Traffic Detection Based on Deep Neural Networks and Association Analysis. – Sensors, Vol. **20**, 2020, No 5. 1452.
14. Gao, X., C. Shan, C. Hu, Z. Niu, Z. Liu. An Adaptive Ensemble Machine Learning Model for Intrusion Detection. – IEEE Access, Vol. **7**, 2019, pp. 82512-82521.
15. Guo, K., S. Han, S. Yao, Y. Wang, Y. Xie, H. Yang. Software-Hardware Codesign for Efficient Neural Network Acceleration. – IEEE Micro, Vol. **37**, 2017, No 2, pp. 18-25.
16. Gwon, H., C. Lee, R. Keum, H. Choi. Network Intrusion Detection Based on LSTM and Feature Embedding. – arXiv. Preprint arXiv:1911.11552, 2019.
17. Jiang, F., Y. Fu, B. B. Gupta, F. Lou, S. Rho, F. Meng, Z. Tian. Deep Learning Based Multi-Channel Intelligent Attack Detection for Data Security. – IEEE Transactions on Sustainable Computing, 2018.
18. Kang, M.-J., J.-W. Kang. Intrusion Detection System Using Deep Neural Network for In-Vehicle Network Security. – PloS One, Vol. **11**, 2016, No 6. e0155781.
19. Kim, J., J. Kim, H. L. T. Thu, H. Kim. Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection. – In: Proc. of 2016 International Conference on Platform Technology and Service (PlatCon'16), IEEE, 2016, pp. 1-5.
20. Larriva-Novo, X. A., M. Vega-Barbas, V. A. Villagrá, M. S. Rodrigo. Evaluation of Cybersecurity Data Set Characteristics for Their Applicability to Neural Networks Algorithms Detecting Cybersecurity Anomalies. – IEEE Access, Vol. **8**, 2020, pp. 9005-9014.
21. Loukas, G., T. Vuong, R. Heartfield, G. Sakellari, Y. Yoon, D. Gan. Cloud-Based Cyber-Physical Intrusion Detection for Vehicles Using Deep Learning. – IEEE Access, Vol. **6**, 2017, pp. 3491-3508.
22. Ma, T., F. Wang, J. Cheng, Y. Yu, X. Chen. A Hybrid Spectral Clustering and Deep Neural Network Ensemble Algorithm for Intrusion Detection in Sensor Networks. – Sensors, Vol. **16**, 2016, No 10, 1701.
23. Maimó, L. F., Á. L. P. Gómez, F. J. G. Clemente, M. G. Pérez, G. M. Pérez. A Self-Adaptive Deep Learning-Based System for Anomaly Detection in 5G Networks. – IEEE Access, Vol. **6**, 2018, pp. 7700-7712.
24. Montavon, G., W. Samek, K.-R. Müller. Methods for Interpreting and Understanding Deep Neural Networks. – Digital Signal Processing, Vol. **73**, 2018, pp. 1-15.
25. Nguyen, G., S. Dlugolinsky, V. Tran, Á. L. García. Deep Learning for Proactive Network Monitoring and Security Protection. – IEEE Access, Vol. **8**, 2020, pp. 19696-19716.
26. Pandey, A., S. Thaseen, C. A. Kumar, G. Li. Identification of Botnet Attacks Using Hybrid Machine Learning Models. – In: Proc. of International Conference on Hybrid Intelligent Systems, Cham, Springer, 2019, pp. 249-257.
27. Rajagopal, S., P. P. Kundapur, K. S. Hareesha. A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets. – Security and Communication Networks, Vol. **2020**, 2020.
28. Ring, M., S. Wunderlich, D. Scheuring, D. Landes, A. Hotho. A Survey of Network-Based Intrusion Detection Data Sets. – Computers & Security, Vol. **86**, 2019, pp. 147-167.

29. Staudemeyer, R. C. Applying Long Short-Term Memory Recurrent Neural Networks to Intrusion Detection. – South African Computer Journal, Vol. **56**, 2015, No 1, pp. 136-154.

30. Thaseen, I. S., B. Poorva, P. S. Ushasree. Network Intrusion Detection Using Machine Learning Techniques. – In: Proc. of 2020 International Conference on Emerging Trends in Information Technology and Engineering (IC-ETITE'20), IEEE, 2020, pp. 1-7.

31. Vinayakumar, R., M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, S. Venkatraman. Deep Learning Approach for Intelligent Intrusion Detection System. – IEEE Access, Vol. **7**, 2019, pp. 41525-41550.

32. Wu, P., H. Guo, N. Moustafa. Pelican: A Deep Residual Network for Network Intrusion Detection. – In: Proc. of 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W'20), IEEE, 2020, pp. 55-62.

33. Yu, Y., J. Long, Z. Cai. Network Intrusion Detection through Stacking Dilated Convolutional Autoencoders. – Security and Communication Networks, Vol. **2017**, 2017.

34. Zhang, Y., X. Chen, L. Jin, X. Wang, D. Guo. Network Intrusion Detection: Based on Deep Hierarchical Network and Original Flow Data. – IEEE Access, Vol. **7**, 2019, pp. 37004-37016.

35. Zhang, Z., X. Zhou, X. Zhang, L. Wang, P. Wang. A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection. – Security and Communication Networks, Vol. **2018**, 2018.

36. Zeng, Y., H. Gu, W. Wei, Y. Guo. "Deep-Full-Range": A Deep Learning Based Network Encrypted Traffic Classification and Intrusion Detection Framework. – IEEE Access, Vol. **7**, 2019, pp. 45182-45190.

37. Khan, F. A., A. Gumaei, A. Derhab, A. Hussain. A Novel Two-Stage Deep Learning Model for Efficient Network Intrusion Detection. – IEEE Access, Vol. **7**, 2019, pp. 30373-30385.

38. Sumaiya, T. I., C. A. Kumar, A. Ahmad. Integrated Intrusion Detection Model Using Chi-Square Feature Selection and Ensemble of Classifiers. – Arabian Journal for Science and Engineering, Vol. **44**, 2019, No 4, pp. 3357-3368.

Received: 29.07.2021; Accepted: 03.09.2021