Ghafar et al. LGURJCSIT 2023

ISSN: 2521-0122 (Online)
ISSN: 2519-7991 (Print)

LGU Research Journal of Computer Science & IT

doi: 10.54692/lgurjcsit.2023.074524

Vol (7): Issue (4), October– December2023

Hybrid Deep Learning Approach to Identify Intrusion Detection with Imbalance Datasets

Abdul Ghafar[1*], Fazeel Abid[1], Muhammad Farooq[1,2], Muhammad Azam[2], Mohsin Ashraf [3], Ammar Aftab Raja[1]

[1]Department of Information Systems, Dr Hasan Murad School of Management, UMT Lahore, Pakistan.
[2]CS Department, Superior University, Lahore, Pakistan.
[3]CS Department, University of Central PunjabLahore, Pakistan.

Email: abdul.ghafar@umt.edu.pk

ABSTRACT:

The intrusion detection system is a computer-based system that constantly identifies all types of malicious activities by monitoring the network traffic. These intrusions and doubtful activities disturb all business activities performed over the public network, such as the Internet and all connected networks. It is an essential system to provide consistent and reliable transfer of information to complete e-commerce and e-business transactions and private communication using social sites. Various deep learning techniques are used to identify security attacks by observing the typical system usage profile and to restrict all of the network traffic if it is outside the scope of the standard profile. Our proposed system is used to combine various deep learning techniques to develop a hybrid deep learning model to identify any security attack in the network. The proposed hybrid deep learning model is trained using an integrated and balanced dataset by merging already available imbalanced benchmark datasets such as NSL-KDD, ISCX, CICIDS2017, and UNSWNB15. Our proposed system is limited to identifying security attacks in benchmark datasets and restricted to available deep-learning techniques and algorithms.

KEYWORDS: Intrusion Detection System, Network Security Attacks, Hybrid Deep Learning Model, Semi-Supervised Machine Learning.

1. INTRODUCTION

The intrusion detection system is computer-based to provide security by constantly monitoring the network traffic for malicious and suspicious behaviour. It was first proposed by Jim Anderson in 1980. Intrusion detection systems are classified as host-based and network-based. Host-based IDS is deployed on a single host and used to monitor all of the activities of the same host, whereas network-based IDS is used to protect and watch all of the network traffic and detect security threats and malicious activities in the network. It can alsobe divided into signature-based or misuse IDS and anomaly detection-based IDS. Signature-based IDS is based on the signature of the attack pattern stored in the signature database. These are highly effective for already-known issues but cannot handle new security threats. Anomaly detection-based IDS is based on defining a profile for normal activity, so anything deviating from the standard shape will be treated as an anomaly or security threat. It can detect any new security threat, but its accuracy is less than that of signature-based IDS.

Public networks such as the Internet are essential elements of IT infrastructure to share and move organizational data among stakeholders such as customers, suppliers, and employees. Most

business activities, such as e-commerce and e-business operations, are now performed electronically. Public networks and all of the LANs and other networks connected with public networks are more vulnerable to attack by intruders. These intruders can perform various malicious and suspicious activities, which can cause numerous financial losses for the organization and individuals. They also disturb all of the organizational activities to conduct business operations smoothly. An intrusion detection system must be required to perform business operations effectively over the network by avoiding all security threats and malicious activities. As a result, these systems give a reliable and consistent view of the organizational information by protecting and securing the network. They can be helpful for all types of businesses, such as banks, e-commerce sites, educational institutes, government, etc., by protecting all of the transactions moving over the network. They can also protect people's private data that is shared through various social sites such as Facebook, Twitter, Instagram, etc.

Anomaly-based intrusion detection systems can be implemented using different artificial intelligence techniques, classified into deep learning and machine learning techniques. Deep learning techniques are more effective in implementing IDS to cope with new security threats and doubtful activities in different networks. Deep learning intelligence techniques and classification algorithms such as CNN, RNN, MLP, AutoEncoder, DBN, and DNN are commonly used to develop intrusion detection systems. These deep learning techniques and algorithms are based on various statistical and mathematical formulas used to classify and predict different network intrusion types. These approaches are already implemented and tested to solve problems and detect multiple security attacks by designing intrusion detection systems. The performance and accuracy of these deep learning approaches are also satisfactory, so they can be used for further research to improve their performance. However, most of these solutions only apply in certain situations and for certain classes of attacks. These IDS are trained with old and limited feature datasets, so they cannot handle several security threats accurately. Their approaches have also trained the system for all types of security attacks by using the same machine-learning model and ignoring the use of different algorithms for different security threats. So, there is a massive chance of improvement in developing efficient IDS using deep learning approaches. As machine learning techniques are based on various statistical and mathematical formulas and algorithms, they can improve their performance by adjusting and tuning those algorithms to get more optimal solutions for intrusion detection. This research leads to finding the deep learning solution with a hybrid approach to identify security attacks in the network and create a new dataset by combining different benchmark datasets to handle a maximum number of classes of security attacks. This research will try to find whether a hybrid deep learning model will perform better in classifying intrusion detection as well as the impact of merging various datasets on the training of IDS to get a more reliable system.

2. RESEARCH AIM

This research will be a valuable addition to designing IDS and will significantly impact the secure usage of the network to share or transfer organizational data among its various stakeholders. A secured and trusted network increases the reliability of the data moved over it and improves the performance of the various business operations using the web. It will encourage the use of e-business, e-commerce, e-banking, or social networking by avoiding delays in performing operations and without compromising the privacy and confidentiality of organizational and personal data.

2.1. Problem Statement

With the use of the Internet and the increased number of nodes to perform various business operations over the public network that is connected to local LANs and IoTs, all of the connected networks are becoming more vulnerable to security attacks or performed malicious activities by different intruders to disturb the valid business operations. It becomes a desirable time to implement an automated system to identify and classify invalid and malicious activities to avoid all security threats while performing any business operation. Several machine learning and deep techniques have been implemented in designing intrusion detection systems, but they are limited to identifying only a few classes of threats and cannot detect any new security threat as trained with old, imbalanced, and small-sized datasets. Our proposed system implements a hybrid approach by combining

CNN and MLP deep learning algorithms with a supervised learning approach using fuzzy-based rules for data engineering and generating integrated datasets from benchmark datasets using various artificial intelligence techniques.

2.2. Significance of the Study

The significance of our research is to develop a new comprehensive and balanced dataset by removing imbalanced effects in the previously available benchmark datasets; in this way, our dataset can train IDS to identify a maximum number of security threats.

This research will also explore the concatenation of CNN and MLP results to develop an optimal hybrid solution to create an intrusion detection system to achieve the highest possible performance by avoiding maximum security attacks. To achieve better performance, fuzzy-based rules are also implemented during preprocessing to tune the data for better classification of security attacks.

Our proposed system integrates different limited and imbalanced benchmark datasets to generate a new balanced dataset with a maximum number of security threats, so it will be a good addition for future researchers to develop a more efficient IDS. Our proposed system also implements CNN and MLP deep learning algorithms by concatenating their outputs to develop a hybrid deep learning model to build an intrusion detection system that is an essential need of all stakeholders such as network engineers, network administrators, network security in-charge, etc. to avoid all type of network attacks to run the business operations smoothly.

3. LITERATURE REVIEW

Various machine learning techniques, such as decision trees, SVM, and ANN, were used by different researchers to develop intrusion detection systems. Deep learning techniques outperform these machine learning techniques, but they need large-sized datasets and high computation speed. Louati implemented an unsupervised deep learning technique using an AutoEncoder with K-NN and MLP to classify multi-class intrusion detection in the NSL-KDD dataset and found 99% exact identification of intrusions in lab experiments [1]. Alom also implements an unsupervised deep learning approach and has designed Deep Belief Networks DBN to classify multi-class intrusions with the NSL-KDD dataset and found 97.5% accuracy with limited types of security attacks [2]. Xu proposed to perform a supervised deep learning approach implementing RNN with GRU memory units using KDD-99 and NSL-KDD datasets and found above 99% accuracy in identifying the security attacks [3]. Tama also implements a supervised deep learning approach by using a deep neural network to design multi-class IDS to identify security attacks and found 92% average accurate results with UNSW-NB15, CIDDS-001, and GPRS datasets [4]. Ludwig suggests a deep neural network to identify multi-class threats using the NSL-KDD dataset and compare it with other deep learning approaches and found above 92% accuracy with its proposed solution [5]. All of the deep learning techniques, either supervised or unsupervised, have merits and demerits. Supervised methods are much better at identifying types of intrusion than unsupervised techniques, but they cannot identify any unseen intrusion in the system. Unsupervised techniques can be preferred in such situations, but they need large-size datasets. Yang follows a hybrid deep learning model by implementing a deep neural network with conditional AutoEncoder to train the system with supervised techniques, followed by an unsupervised approach on NSL-KDD and UNSWNB15 datasets. Zhang implements MLP, CNN and C-LSTM deep learning techniques to design a hybrid solution to identify different security attacks and control adversarial attacks with a semi-supervised training model and found above 99% accuracy in identifying security attacks [7,8]. Most of the datasets are imbalanced records of various classes of security attacks, so not able to predict the exact error type for all types of security attacks Chowdhary implements a system to control the imbalance records in NSL-KDD and KDDCup99 dataset with limited classes of attacks to identify security threats with real-time accuracy [8]. No single dataset can be used to train model for all types of intrusions as different datasets have variable proportions of various security attacks [8,9] .

Most of the research is based on limited and imbalanced datasets to identify small classes of security attacks, so they are not available to identify all of the threats with the same level of accuracy. They also implement one or a few machine learning or deep learning approaches by using various classifications to identify those security attacks. Few researchers follow hybrid deep learning models, but their accuracy is still

low compared to other approaches, so there is a huge gap to implement hybrid deep learning approach to improve the performance of IDS in detecting security attacks. Our research develops a hybrid deep learning approach by concatenating CNN and MLP deep learning approaches to generate more efficient IDS to identify intrusions with an integrated dataset generated using fuzzy-based rules and artificial intelligence techniques with already available benchmark datasets.

4. RESEARCH METHODOLOGY

Our proposed system will perform a hybrid deep learning approach by concatenating CNN and MLP algorithms with an integrated, combined dataset and balanced copy of several benchmark datasets. For this purpose, we have selected NSL-KDD, KDD-Cup99, CIDDS-001, and GPRS datasets for intrusion detection systems. Most of these datasets have imbalanced records related to various security attacks in different proportions, so our proposed system generates random records by using artificial intelligence techniques to get balanced records for each type of security attack. Each dataset typically covers a few sets of security attacks. Therefore, the proposed system merges these files to create a comprehensive dataset with a maximum number of security attacks to train IDS to identify the maximum number of security attacks.

The KDD-Cup99, also known as KDD99, dataset is the most common and widely used for several years to evaluate IDSs. It contains 39 types of attacks for seven weeks for training data and two weeks for test data. These attacks are divided into four categories: DoS, probe, R2L, and U2R, or otherwise considered normal, as shown in Figure 1. Each record contains 41 features; 34 are continuous, and seven are discrete. These features can be categorized into four categories: basic elements of TCP connections, derived content features, same-host features, and same-service features.

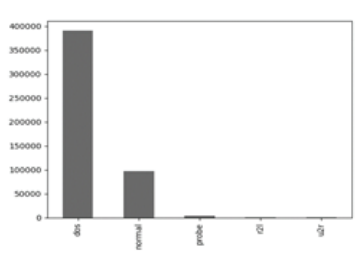

Figure 1: Distribution of types of attack in the dataset

The NSL-KDD is a revised version of KDD99 and is regenerated by removing redundant records. It is more balanced than KDD 99 and contains a reasonable number of records, making it more feasible for training and testing.

Then, a hybrid deep learning model will identify intrusion by concatenating CNN and MLP and training the IDS with a supervised approach. Our proposed system identifies the performance of the proposed model with individual CNN and MLP algorithms to identify the impact of a hybrid deep approach with simple deep learning models in identifying security attacks by IDS. Researchers already use mixed methods and found much better results than other techniques. This research also finds that a better hybrid deep learning model can more effectively classify these security threats. The use of various combinations of deep learning approaches to develop a hybrid model is still a research topic, so our proposed solution will be valid research to get a more accurate and efficient IDS using a hybrid deep learning model by concatenating CNN and MLP algorithms.

For more accurate and efficient results, our proposed system performs various artificial intelligence techniques with fuzzy-based rules in data analysis to remove duplicates, adjust the missing values, and remove irrelevant records. It also analyzes the imbalanced records and regenerates random records to create a balanced dataset to identify each type of security threat equally. A random method is used to split the dataset into training and test data to avoid bias and maintain balanced records in both training and test datasets to train the proposed deep learning model. Performance evaluation matrices are used to analyze the performance of our proposed hybrid model with simple deep-learning models.

4.1. Proposed System

Our proposed system is an IDS based on anomaly detection method and developed with a hybrid deep learning approach that is trained with a balanced dataset. Anomaly detection-based IDS, also known as a behaviour-based intrusion detection system, defines a profile of regular activity. So, any activity that deviates from the standard profile will be treated as abnormal or anomaly behavior. It can detect any unknown and new attack. However, the boundary between normal and abnormal profiles is unclear, increasing the value of FAR. With the application of artificial intelligence, it has become easy to

identify the line between normal and abnormal activity. These are also known as AI-based intrusion detection systems. Various machine learning and deep learning algorithms can be used to develop intelligent and efficient IDS. Supervised machine learning algorithms are essential to classify labelled data, whereas unsupervised machine learning algorithms can extract information from unlabeled data.

4.2. Security Threats

Several datasets are available to develop a model to classify these security attacks using deep learning approaches. These datasets contain various security attacks as an imbalanced set of values against each attack. In this paper, we have developed an integrated dataset by combining NSL-KDD, KDD-Cup99, CIDDS-001, and GPRS datasets to expand the scope of our proposed IDS. For simplicity, we have categorized all these security threats in this combined dataset into the following categories.

A. DoS

DoS, Danial of Service, is a set of different security attacks that prevent legitimate users from using any service available in the system. The system becomes busy due to invalid requests and cannot provide service to any new valid request. The invalid activities by intruders, such as teardrops, mailbombs, smurf, worms, etc., are generally considered DoS attacks.

B. Probe

The probe is another critical type of security attack in which the attacker gains information from the target system without knowing his existence. It is a set of illegal activities such as ipsweep, port sweep, saint, mscan, etc. Intruders can use these to explore and read the system's content.

C. U2R

In the case of a U2R security attack, the attacker tries to gain control of the system without any legal access or privilege. An attacker can easily control or manipulate any information and disturb the execution of any process. It includes activities such as buffer overflow, rootkit, load double, sqlattack, etc.

D. R2L

R2L is the most dangerous security attack, and the attacker gains super user access from his user-level account. In this way, the attacker performs any admin activity to modify or manipulate information in the system. It is conducted by illegal activities such as ftp_write, sendmail, http_tunnel, guess_passwd, snmpguess, etc.

4.3. Hybrid Deep Learning Model

This paper implements a hybrid deep learning model to identify security threats using an intrusion detection system. Using a hybrid deep learning model to build an intrusion detection system can effectively detect and prevent invalid activities. An intrusion detection system (IDS) is a security tool that monitors network traffic to detect and prevent malicious activities. Machine learning techniques have been widely used in IDS to automate detecting network intrusions. We have combined two popular deep learning algorithms, multi-layer perceptron and convolutional neural network, to build a powerful approach for IDS. MLP can identify complex relationships between features, and CNN can extract meaningful features from raw data.

The overall structure of our proposed system can be divided into several components, as shown in Figure 2. The first component is data preprocessing. It is the most critical component to tune and normalize the data in a way suitable for deep learning algorithms. As we have implemented CNN in our hybrid approach, data must be converted into image format. Another task in data preprocessing is to combine and balance different imbalanced datasets into one integrated and balanced dataset. The second principal component is training the model by constantly refining the CNN and MLP model and then concatenating both models' output to the final result. In this step, various parameters are adjusted to improve the model's performance. After the training, testing is performed to confirm the accuracy of the implemented algorithms. If your testing results are undesirable, then it will repeat the last step by backpropagation to refine the results. After successful testing, model performance is evaluated by using various performance metrics. After preprocessing data as in the previous step, deep learning models MLP and CNN are constructed, preprocessed data is given separately as input to both MLP and CNN, and outputs of both models are concatenated to create a hybrid model after compilation of the model, the hybrid model trained by using the training dataset. Finally, the model is evaluated to

measure its performance after testing it.

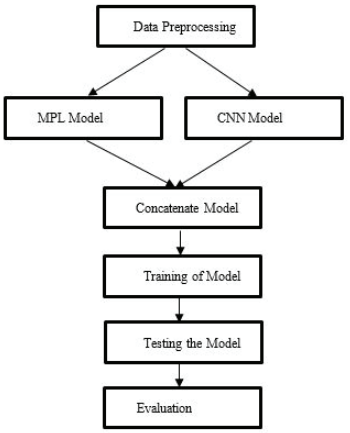

Figure 2: Proposed machine learning model architecture

4.4. Data Preprocessing

The first principal component in our proposed system is to perform preprocessing on the required datasets to convert them into a format suitable for algorithms for deep learning. For this purpose, several cleaning processes and encoding methods are performed to extract the most convenient and reliable data. The cleaning operation removes duplicate rows and missing values, whereas the encoding and normalization process is performed to make it efficient for machine learning. To increase the reliability of data and accurate prediction about any intrusion, all classes with null values or less than 25 iterations are removed. As symbolic or non-numerical data is not practical in analytical applications, all extended data classes are encoded into some suitable format. For example, protocol-type features are mapped using a three-dimensional binary array as (1,0,0), (0,1,0), (0,0,1).

In the same way, flag and service features are converted into an n-dimensional binary array. The output class attack type is also mapped with a one-hot vector. To ensure the accuracy of training data, a normalization process is performed to eliminate differences between dimensional data by setting in the range of 0 and 1. The 1-N encoding is used to convert data between 0 and 1 using mix-max normalization using equation 1. All of the above preprocessing steps are repeated for other datasets as well. After preprocessing all datasets, these are merged to create an integrated dataset.

$$Xn = (x - xmin) / (xmax-xmin) \quad (1)$$

The integrated dataset still has the issue of imbalanced records for different types of security attacks, and will affect the accuracy of predicting the model. To handle this situation, we have applied a Sample-based balancing method. The sampling-based process can be divided into three methods: oversampling, under-sampling, or combining both. The oversampling method increases the number of samples for classes having fewer values, whereas the under-sampling approach reduces the class values. In the hybrid system, class values increase or decrease by calculating the proportion of samples in each class. The sampling-based method is easy, but it may increase the biases in the data. The under-sampling or hybrid approach cuts the actual values that may cause the model's underfitting, so we prefer the oversampling process. For this purpose, we have implemented an artificial intelligence method called SMOTE. It is based on the KNN machine learning approach.

The last step in data preprocessing is to convert input data into image format, as CNN deep learning requires input data in image format. Therefore, the 122-dimensional feature vector input data must be converted into nxn image data. The value of n must accommodate all of the feature vector values. In our case, the optimal value of n is 11 as 121=11x11, so it can hold maximum feature values. We must remove one extra feature with the lowest coefficient variance, such as Equation 2, as our input feature vector has 122 values.

$$V = S.D. / mean \quad (2)$$

After preprocessing, data is randomly divided into training and testing datasets. The training dataset includes 80% of the original and 20% of the testing datasets. The proposed hybrid deep learning algorithm is then trained with a training dataset in the next phase.

4.5. Model Training Phase

The proposed hybrid approach is used to classify intrusion detection by combining various deep learning algorithms in different ways. These deep learning algorithms are typically implemented with artificial neural networks with several hidden layers to perform other mathematical models to extract information from the training dataset to identify regular activity or specific security threats. We have implemented CNN and MLP deep learning algorithms to construct the

proposed hybrid solution.

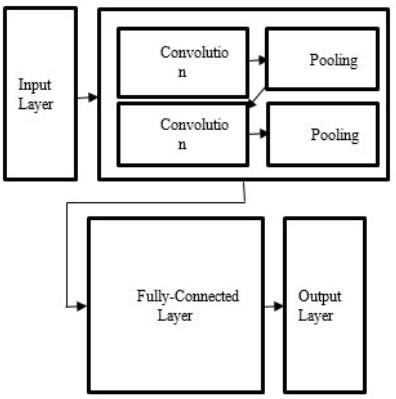

Figure 3: Architecture of CNN model

4.6. Convolutional Neural Network

It consists of a stack of fully connected convolutional and pooling layers to extract more reliable features, combined with a fully connected layer and a softmax classification layer. It performs a supervised learning procedure to remove parts and then use these features for classification. It gets data in image format in its output layer and processes it using convolutional layers to classify it using the softmax layer. In our proposed system, CNN is implemented, as shown in Figure 3, with an input layer with an 11x11 matrix as an image and returns a multi-class of five different security attacks. There are five hidden layers: two are convolutional with 2x2 sized convolution, two pooling layers with 2x2 sized convolution, and one fully connected layer. All neurons in convolutional layers share the same convolutional kernel, which can determine the number of weights w. The value of each neuron is adjusted using a formula with X convolution function with bias variable b with a non-linear activation ReLU activation function. It helps to understand the relationships among the inputs and outputs.

$$Hj = f(hj\text{-}1 \text{ X } wj+bj) \quad (3)$$

The pooling layer reduces the feature image hj to avoid over-fitting using the formula.

$$Hj = pool(hj\text{-}1) \quad (4)$$

After performing two layers of convolutional and pooling, hj reshaped to a vector. The output yi achieved through the fully connected layer. It gets the flattened output and assigns weights to generate output. The error between output yi and expected value is calculated through a loss function. The backpropagation performs the gradient descent to reduce the error in the training phase. Then, finally, the result is returned by the softmax layer as a classification of security attack.

4.7. Multi-Layer Perception

MLP is a famous supervised feedforward artificial neural network. It is usually composed of multiple interconnected layers with several connected neurons. The input layer receives data processed through numerous hidden layers and returns its results by the output layer. The backpropagation approach is used to improve the accuracy in predicting the output. It can quickly learn complex non-linear relationships between data and labels, but it may result in overfitting if data is not handled carefully. In our proposed system, MLP is implemented as shown in Figure 4. It is designed with one input layer, three hidden layers, and one output layer. The input layer, containing neurons equivalent to several features, receives data and passes to the first hidden layer. Hidden layers are fully connected with the previous and next layers. Neurons receive data from the prior layer, and after performing the ReLU activation function and bias value, it transfers it to the neuron in the next layer. The output layer receives data from the hidden layer and returns the class label as output. Backpropagation improves learning by calculating the error between the actual and predicted output using stochastic gradient descent.

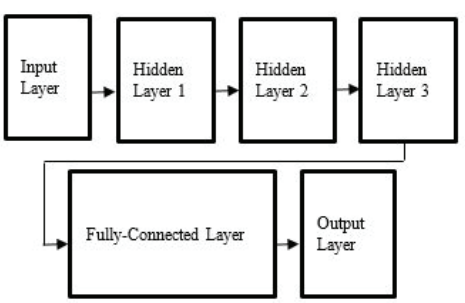

Figure 4: Architecture of MLP model

4.8. Concatenation of both models

The concatenation layer is used to merge the output of both deep learning models with a single softmax layer. Both fully connected layers in these models are associated with the single concatenation layer by combining the production

of both models with a single output of identifying the type of security threat.

4.9. Testing Phase

Test data already generated in the preprocessing phase is used in this phase to test the model. For this purpose, we have tested the model five times to validate the results. The average test results of our proposed model with the same individual deep learning techniques are shown in Figure 5.

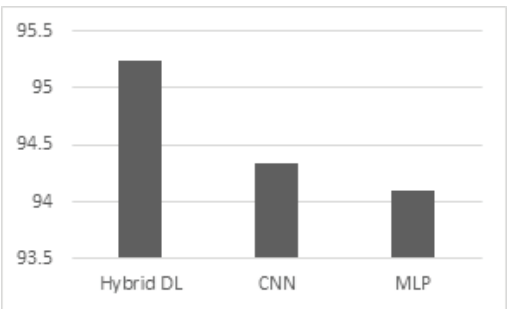

Figure 5: Performance results of hybrid and simple deep learning models

4.10. Evaluation Metrics

After training and testing the proposed deep learning mode, the proposed model's performance and individual deep learning models are evaluated. The most important metric is the Confusion Matrix, a two-dimensional matrix that can further extract TP, FP, FN, and TN evaluation metrics. FP, false positive, is several attack samples as attack class. TP, true positive, is attack sample classes correctly identified. Precision, FAR, Recall, Accuracy, TNR, and F1-score are calculated using these evaluation metrics.

$$\text{Precision} = TP / (TP + FP) \quad (5)$$

$$FAR = FP / (FP + TN) \quad (6)$$

$$\text{Accuracy} = (TP+TN) / (TP+FN+FP+TN) \quad (7)$$

$$\text{F1-score} = 1\ (1/\text{precision} + 1/\text{recall}) \quad (8)$$

5. RESULTS & DISCUSSION

The proposed hybrid model is implemented and trained using Jupiter Notebook powered by Kaggle with TensorFlow support. To evaluate the system's performance, a few hyper-parameters are configured for individual algorithms before execution. A separate test is also performed to train and test the individual CNN and MLP models. To get better, these models are tested five times, and results are stored in an Excel Spreadsheet to find the average of all developments. The TP, TN, FP, and FN values are generated using the confusion matrix generated after testing the model. From these matrices values, other matrices can be calculated using the above equation. The results are shown in Table 1. It shows that the accuracy of the proposed model is 95.24% as compared to individual CNN and MLP models.

Table 1: Results of different executions with deep learning models

Model	Hybrid DL	CNN	MLP
Turn 1	96.12	94.12	94.18
Turn 2	95.36	95.76	93.12
Turn 3	94.62	93.08	94.72
Turn 4	93.82	94.46	94.60
Turn 5	96.28	94.34	93.88
Average	95.24	94.34	94.10

The IDS based on a hybrid deep learning model has achieved high accuracy in detecting network intrusions as compared to other individual deep learning models. Test results shown in Table 1 indicate that the hybrid approach can perform better than unique deep learning algorithms. Average results show that our proposed algorithm leads to 95.24% accuracy in testing compared to CNN and MLP, which achieved 94.34% and 93.10%, respectively. It can also handle large amounts of data to identify unknown and new attacks. However, the actual performance depends on the data's quality and size, the attack type, and many other factors. It would be possible to adjust different hyper-parameters and configurations as per need and type of dataset.

6. CONCLUSIONS

An intrusion detection system based on a hybrid deep learning model has proven highly effective in detecting and preventing security attacks. It uses modern deep learning algorithms to create a hybrid IDS to explore the network traffic data and identify abnormal activity that may be an intrusion. Hybrid deep learning-based IDS has several benefits compared to traditional IDS or developed using simple machine learning or deeplearning algorithms. It can have the ability to handle large amounts of data and recognize any abnormal activity. They can also adapt to new types of security attacks without any significant change in the system. A hybrid deep learning

algorithm that CNN and MLP concatenate is proposed in this paper. It is trained by a balanced, integrated dataset created from multiple imbalanced datasets, showing better performance than individual deep learning algorithms in identifying security attacks by the intrusion detection system.

7. LIMITATIONS AND SCOPE

Our research is based only on the most common available benchmark datasets and covers all the security threats already available in those datasets. However, it performs equally for all known security issues by balancing the datasets with various machine-learning approaches. The proposed solution performs a concatenation of CNN and MLP with supervised training only to develop an optimal hybrid deep learning model to identify security threats over the network. However, the proposed solution does not study the impact of other deep learning approaches.

Our proposed system is limited to correctly identifying all of the known security attacks available in datasets and has limited capability to precisely identify any new security attack. Our research is also not going to design any new deep learning algorithm. Instead it combines CNN and MLP with limited improvements to create a hybrid deep-learning solution for IDS.

REFERENCES

[1] F. Louati and F. B. Ktata, "A deep learning-based multi-agent system for intrusion detection," *SN Applied Sciences*, 2(4), pp. 1-13, 2020.

[2] M. Z. Alom et al., "Intrusion detection using deep belief networks," *In 2015 National Aerospace and Electronics Conference (NAECON),* (pp. 339-344), IEEE, (June, 2015).

[3] C. Xu et al., "An intrusion detection system using a deep neural network with gated recurrent units," *IEEE Access,* 6, pp. 48697-48707, 2018.

[4] B. A.Tama and K. H. Rhee, "Attack classification analysis of IoT network via deep learning approach," *Res. Briefs Inf. Commun. Technol. Evol.(ReBICTE),* 3, pp. 1-9, 2017.

[5] S. A. Ludwig, "Intrusion detection of multiple attack classes using a deep neural net ensemble," In *2017 IEEE symposium series on computational intelligence (SSCI),* (pp. 1-7), IEEE, (November, 2017).

[6] Yang, Y., Zheng, K., Wu, C., & Yang, Y. (2019). Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network. Sensors, 19(11), 2528.

[7] C. Zhang et al., "Tiki-taka: Attacking and defending deep learning-based intrusion detection systems," In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop,* (pp. 27-39), (November, 2020).

[8] J. Lansky et al., "Deep learning-based intrusion detection systems: a systematic review," *IEEE Access*, 9, pp. 101574-101599, 2021.

[9] M. M. U. Chowdhury et al., "A few-shot deep learning approach for improved intrusion detection," In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON),* (pp. 456-462), IEEE, (October, 2017).