Network Intrusion Detection System using Convolution Recurrent Neural Networks and NSL-KDD Dataset

Manjunath H.*, Saravana Kumar
Department of Computer Science and Engineering, CMR University, Bangalore, India
Emails: manjunath.19cphd@cmr.edu.in; sarvana.k@cmr.edu.in

Abstract

Increase in network activity of transferring information online allows network breeches where intruders easily avail the most important information or data. The growth of online functioning and many other governmental data over the internet without security has caused data vulnerability; attackers can easily detect the data and misuse them. Network Intrusion Detection System (NIDS) has allowed this whole process of online data transfer to occur safely and secured transactions. Due to the cloud usage in network the huge amount of traffic is created as well as number of attacks are increased day by day. To prevent the vulnerability and its types are social, environmental, cognitive, military attacks in the network are classified using CRNN model. We used ensemble learning methods in machine learning algorithms are used to detect and prevent the malicious packets in the network. Our model detects the unauthorized users intruding into any network and alerts the organization regarding the same. When a typical firewall is unable to effectively stop certain sorts of attacks on computer system usage and network communications, a network intrusion detection system may be used. First, we are classifying the unauthorized packets using machine learning algorithm. Using our concept, we have used neural networks in this paper to detect any such attack. For the Network Security Laboratory - Knowledge Discovery in Databases data set using CNN and RNN algorithms, we also applied a few well-known techniques as boosting and pasting methods. In this CRNN approach, we demonstrate that neural networks are more effective than other methods at detecting attacks.

Keywords: Neural networks; supervised learning; Network Security Laboratory - Knowledge Discovery in Databases; SVM and Random Forest; CRNN- convolutional and recurrent neural network.

1. Introduction

The usage of the Internet daily drastically increased. Because of this, internet-based information processing methods are susceptible to counters that can result in a wide range of damages and significant losses to the organizations. It is essential that you keep information safe from unauthorized access. When a network is infiltrated, sensitive data is stolen. A network intrusion detection system identifies known and unidentified threats that enable network breaches. We assess our precision in situations with binary and multiclass categorization. Predicting categorical variables can take the form of binary classification, where the output is limited to two major categories. In machine learning, there are various techniques used for binary classification. More than two classes are involved in the classification challenge, for overcoming this problem, a variety of methods are utilized, such as translating the X number of classes to the X number of binary columns that represent each class. Hence, for Multi Classification techniques, a multimodal classifier can be used.

The objective of the NIDS is to prevent the passive attacks in the network and identification malicious packets in the network. After identification of the suspicious or malicious packets are categorized and analyzed using machine learning algorithms. An intrusion detection system's detection mechanism will find counters whose signatures are already identified by the system. Emerging malware counters with unknown signatures can be challenging to find. Antiquated strategy Anomaly-based intrusion detection systems were introduced to

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023

identify the unidentified spyware since new malware is produced quickly.

Machine learning is employed in the anachronism intrusion detection system to create a trustworthy methodological model which is matched with all the approaching and is labeled skeptical if not found in the system. An open-source high-level neural network library called Keras makes it easier to experiment with deep neural networks quickly. It is user-friendly, extendable, and modular. It supports both Convolutional and Recurring Systems separately as well as in combination. In this paper, focus is on developing a Network-IDS based on convolutional neural network and compare our model with various algorithms based on NSL-KDD dataset.

Vulnerability in the network is used to segregate the packet based on the attack type such as malware vulnerability, outdate packets, misconfigured firewalls, and social attacks. Our CRNN approach will give the solution for classification of these vulnerabilities based on binary and multiclass classification.

2. Literature survey

Based on the growth of cloud and IoT based enabled devices connectivity, there will be studies on the work of various algorithmic processes in system for detecting network intrusions. Few well known algorithms such as Random forests, Naive bayes, Ada boost etc. have been used widely in similar models.

In the work carried out in [1], the researchers are used the CNN and bidirectional LSTM to learn deep learning system to learn temporal and spatial data operations in one hot encoding, normalization, and stratified K Fold cross validation methods to compare the bagging of perceptron learning algorithm and logistic regression.

According to [2], the Researchers have used AdaBoost and random forests algorithm along with CICIDS2017 dataset to build an IDS to make sure the system is favorable under high Tolerance systems. In paper [3] the Authors have used Random Forests Naive Bayes, Adaboost and SVM along with network intrusion dataset (UNSW-NB15) and Intrusion Detection Evaluation Dataset (CIC-IDS2017) datasets. In paper [4] Authors have made use of NB, SVM, RF, MLP (multilayer perceptron) and Tree-CNN algorithms for capturing network traffic and analysis filtering of the algorithm using the datasets like CICIDS2017, DARPA98 etc.

The survey in [5] contrasts the most significant earlier deep learning-focused cyber security surveys in addition to looking at them. This research offers an innovative perfectly alright taxonomy that classify the cutting-edge, contemporary learned in the classroom IDSs in accordance with several features, including input data, detection, deployment, and assessment procedures. The algorithm was developed and tested using the Knowledge Discovery in Databases dataset. The neural network was built with five primary layers, a two-dimensional output layer, hidden layer layers, and a central node with six inputs and six cells each. Six characteristics from the selected attributes in the Knowledge Discovery in Databases dataset, which are fundamental and traffic features that may be readily retrieved from the SDN environment, were selected to train the proposed technique. Calculations are made for memory, accuracy, and correctness using suggested approach, yielding an F1-score of 0.763. The model placed sixth out of nine in a further evaluation of eight traditional ML models proposed in [6].

The approach for App Networking outlier detection was expanded by the same author [7], who achieved up to 88.9% accuracy using a rectified linear unit computational model. In order to uplift and accelerate the methodology of learning, process occurs is also done using the Min-Max bootstrapping technique. Also suggested [8] is a basic set of features that are chosen using the Knowledge discovery in databases dataset. This set is predicated on the notions that data pattern critical realists has risen and information accuracy level has decreased. These presumptions allow for the derivation of factual conclusions from incomplete data. Next, a number of RF-using properties are obtained [9].

Researchers in [10-11] also investigated the use of automated learning techniques in various practical uses and attempted to spot various security threats in such applications. Using cluster-based under-sampling and two surface detection and ensemble classifiers, various researchers [12, 13] made effort for enhancing the feasibility of classification for the collection of data.

In [14], which presents the findings of the Logistic Regression, Random Forest, Linear SVM, and Gaussian Bayes, Linear methods Discriminant Analysis for identifying Denial of service attacks, Machine learning algorithms for detecting network intrusions were also investigated. The Intrusion detection and

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023

evaluation dataset [15] was used by the authors as a test and training set for their algorithms. RF, which achieved 95.32% accuracy, was the classification method with the highest efficiency. A method for improving performance and accuracy has been put out in [16]. The best accuracy was raised once the paper found relevant features in the dataset. The classifier's ability to perform was greatly enhanced after the researchers deleted redundant and irrelevant features.

Detection of malicious packets and segregating these packets are challenging task in the real time network, our CRNN approach will give high accuracy in classification of malicious packets and types of vulnerabilities in the firewalls [17].

3. CRNN Approach

Neural Structured Learning-Knowledge discovery in databases Datasets,

- A portion of the Knowledge Discovery Database Test+.txt file called Knowledge discovery in databases Test-21.TXT that excludes records with a difficulty level of 22 out of 22.
- Knowledge discovery in databasesTrain+.TXT: The whole Neural Structured Learning-Knowledge discovery in databases train set in CSV format, which includes attack-type labels and difficulty level.
- Knowledge discovery in databasesTest+.TXT: The complete Neural Structured Learning-Knowledge discovery in databases test set in comma-separated values format that includes attack-type designations and difficulty level.

A. Neural Network

It is one of the artificial intelligence method, Neutral System are a sequence of process that mimics the same operations as in any human brain to recognize the association between large data. They portray the connection of neurons and synapses in the human brain. Neural networks work as deep learning algorithms as they have several processing layers and work very deep [18].

B. Convolutional Neural Network (CNN)

The CNN category is one of the main classifications used in neural network models for image recognition as well as classification. Figure 1 represents the architecture of CNN model, and it contains the different hidden layers, first it detects the image and classifies the images and after classification it applies the convolution method and detects the hidden features and finally classifies the images into

normal and malicious images of two categories. CNN are extensively used in a wide range of applications, including scenario labelling, object recognition, and facial and object tracking. When CNN receives an image, it characterizes and analyses it using terms like "mat," "cat," and other such terms. The image's resolution affects how the computer interprets it as a collection of pixels. Convolutional neural networks use a combination of convolution layer, sharing, layers that are completely linked, and kernels can process every input image. Afterward, an item would be classified using probabilistic values between zero and one using the Soft-max function.

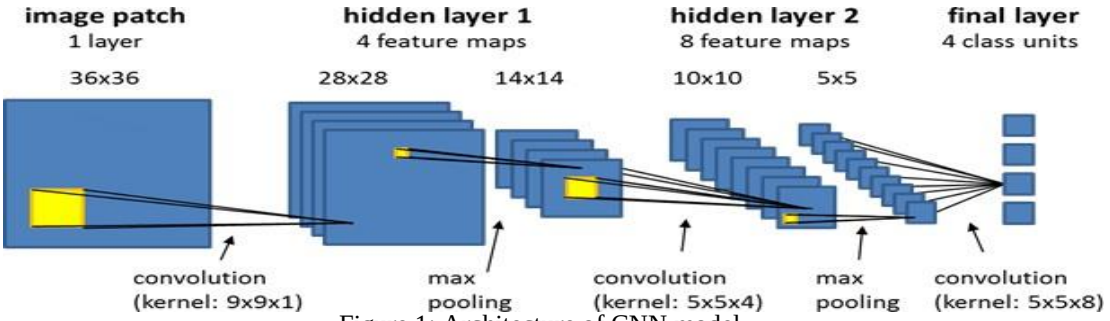

Figure 1: Architecture of CNN model

C. Recurrent Neural Network (RNN)

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023

A type of neural network known as a renewal neural network uses the outputs of one phase as the inputs for a subsequent phase. This issue was fixed by Renewal neural network, which was created with the aid of a Hidden Layer. The main and most significant feature of renewal neural networks is the Integral gain, which retains certain data about a series. The "memory" of renewal neural networks stores all information that relates to calculations. It uses the same settings for each input and performs the same action on each input or hidden layer to produce the output. This makes the parameter set considerably difficult than with other renewal neural networks. Figure 2 shows the architecture of Conventional renewal neural network (RNN) model, here we are introducing the time (t) constraint for the detection process.

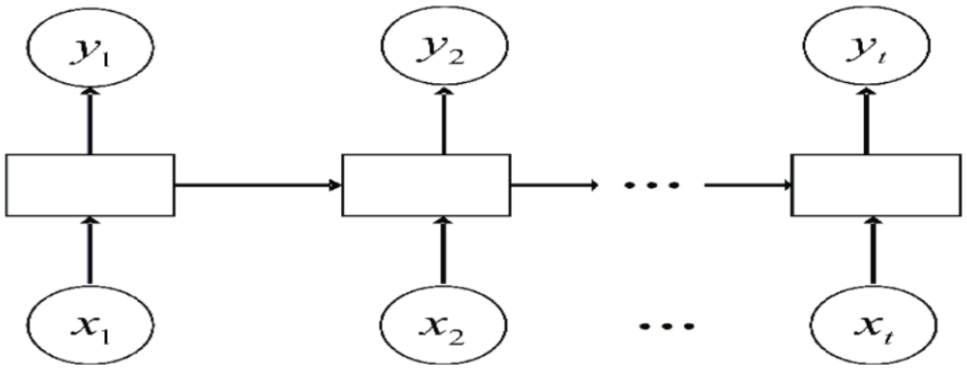

Figure 2: Conventional renewal neural network (RNN) model architecture.

4. **Proposed CRNN Methodology**

Using the NSL-KDD datasets and the proposed Convolutional Recurrent Neural Network (CRNN) approach, packets in firewall may be classified and normalized.

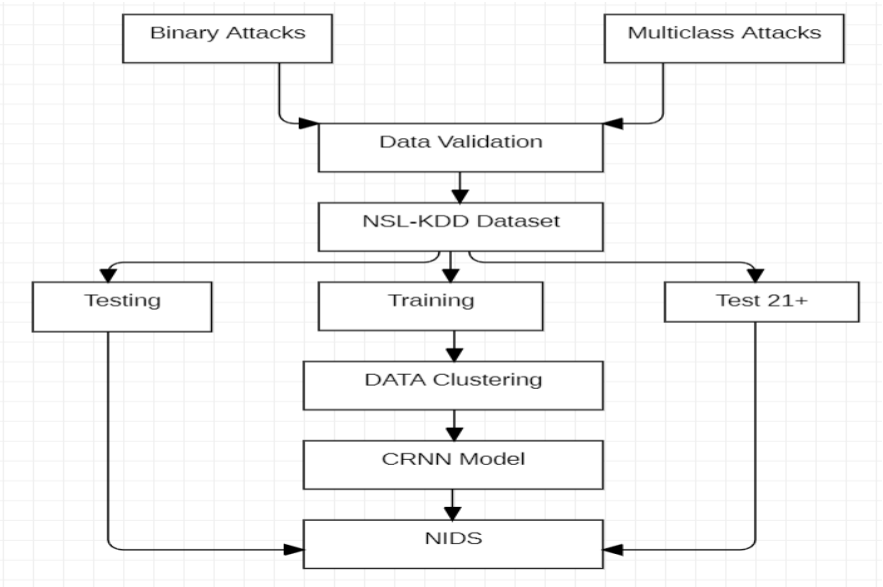

Figure 3: Proposed methodology of CRNN model

The Figure 3 represents the proposed methodology of the Network Intrusion Detection System using Convolution Neural Networks and NSL-KDD dataset. The data is collected from different sources, which will be trained, later segregated in the form of binary attacks, multiclass attacks and remaining is discarded. After the data collection process, data cleaning and data validation process for removing unwanted data will be processed. Data pre-processing is done, and we will train the data using the NSL-KDD algorithm. After training the dataset we need to segregate the data in the form of trained data and test data. Later we will apply KNN algorithm to trained dataset, CNN Model is prepared. Compare the CNN Model to the trained and testing

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023

dataset.

Proposed CRNN Algorithm: Network Intrusion Detection System using Convolution Neural Networks and NSL-KDD.

Input: NSL KDD Data Sets
Output: Detection of malicious attacks and generates alerts
Step1: Data preprocessing
Step2: Visualization and feature extraction
Step3: Data validation process
Step4: Finalization of NSL KDD dataset.
Step5: Apply the clustering method to segregate the malicious packets.
Step6: Apply the AI method to perform the Min Max process in CNN for identification of hidden features.
Step7: Apply the RNN algorithm.
Step8: Finally generates the results of both binary and multiclass classification of Train, Test and Test21+.

The proposed Convolutional Recurrent Neural Network (CRNN) algorithm, the dataset will be cleaned and validated using machine learning techniques, and the classification and normalization of the network packets will be optimized using artificial intelligence techniques.

The normalization and classification of the attacks involved in CRNN approach is as discussed below.

A. Normalization

To recognize each field of the database uniquely, normalization is a mechanism for arranging the data into a relative database to eliminate redundancy of the given huge data. Furthermore, it makes that the data types are appropriate for that field.

It uses the min max algorithm to find the optimality and the to remove the redundancy of the similar packets and it uses DFS methods for pruning the nodes. Our model has the following characteristics: destination bytes (2,2.3107), source bytes (0.1,2.3105), and duration (0,673529). We reduce the differences using the logarithmic scaling technique, and then we use the following formula to translate them to the [0.12,1.24] range:

$$Ai = (Ai - min) / (max - min)$$

B. Classification Of Attacks

- **Binary classification:**

Forty-dimensional features have been translated into eighty two-dimensional features. Consequently, in the binary classification experiment, our mode contains 112 input nodes and two output nodes. We assume that there are 101 epochs total, and that the learning rate is 0.1. Let there be 50, 70, and 113 hidden nodes, respectively, to determine which model is more effective. The batch size is 75, and there are three hidden layers. According to the table 1 shown below, the optimum outcome is produced with a hidden node value of Sixty-Four.

Table 1: Binary Classification

	ANOMALY	**NORMAL**
ANOMALY	8729	982
NORMAL	3408	9406

The studies demonstrate that, when given 101 epochs for the Knowledge discovery in databases Train dataset, the Convolution neural network model operates with a good detection rate (75.512 percentage). According to recent research about artificial neural network methods used when it comes to intruders' identification, the

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023

results of classification algorithms and ANNs both yield a 75.512 percentage. Thankfully, all these findings are based on the same benchmark. The highest accuracy we have managed using conventional approaches is 98.99 percentages. The result of our model is shown in the below table 2 contains Convolutional Neural Network Accuracy for Binary Classification. In comparison to previous classification techniques, the Convolution Neural Network-Intrusion Detection System model performs better in binary classification. Using K-Fold cross validation mechanism the results are simplified and improved compared to the previous approaches.

Table 2: Convolutional Neural Network Accuracy for Binary Classification

TEST	VALIDATION	TRAIN
0.649	63.283	0.870

- **Multiclass classification**

For multiclass classification, we have employed CNN and renewal neural networks (RNN). There are two hidden layers that we have pulled from CNN. There are 10 and 162 epochs, respectively. Nearly all the values used in renewal neural networks are comparable to those in the binary classification experiment. The distinction is that 60 and 150 hidden nodes, respectively, are considered. The table 2 shows that Convolutional Neural Network performs the best when there are 162 epochs, as can be seen.

The table 3 clearly shows the multiclass classification of confusion matrix contains the parameters of common packets, DOS, R2Uand U2R of multiclass classification rate must be calculated. Table 4 CNN Model shows the Accuracy for Multiclass Classification, and it produces the better results compared to the other methods.

Table 3: Multiclass Classification Confusion Matrix

	Common	Denying services	Remote to User (R2U)	User to remote(U2R)	Test
Common	9023	544	63	4	8
Denying services	135	78	0.1	0.1	0.1
Remote to User (R2U)	73	121	6520	0.1	0.1
User to remote(U2R)	31	0.1	0.1	4	3
Test	1720	0.1	0.1	0.1	379

Table 4: CNN Model Accuracy for Multiclass Classification

TEST	VALIDATION	TRAIN
64.85	63.28	86.96

5. **Result and Discussions**

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023

The Free Tier plan from google colab is what we've utilized for all our tests. To examine the effectiveness of the Intrusion Detection System paradigm, we conduct 2 methods. Binary classification comes first. And secondly, classification into five categories like Common, Denial of Service, R2U Remote to user, U2R User to remote, and Test. Using our ensemble learning approaches of machine learning methods, the results we found are better than the other models and comparative results are shown in table 5. Binary classification performance of the Classification algorithm and other classic ML models and it shows that our model has proved 83.34 percentage for Test+ and 80.46 percentage for Test-21.

Table 5: Binary classification performance of the Classification algorithm and other classic ML models.

	Test+	**Test-21**
Simple Bayes	87.40	45.76
Java48	80.00	64.99
NB Tree	81.02	66.10
Random Forest	81.00	64.00
SVM	68.08	42.33
Renewal neural network	93.58	57.54
Our CRNN proposal	83.34	80.46

We compared the performance of several classification and neural network models to our suggested model and discovered that by utilizing the optimizer and epoch number 550, we surpassed other models. Our model outperforms all other neural network models in IDS, according to our comparisons with other models already in use. Nonetheless, classification algorithms perform better for this challenge due to labeled data and the dataset's structure. Table 6 Multiclass classification performances of the Classification algorithm and other classic ML models and it shows that our model has proved 84.30 percentage for Test+ and 64.40 percentage for Test-21. Figure 4, Comparative analysis of Multiclass classification performance of the Classification algorithm and other classic ML models.

Table 6: Multiclass classification performance of the Classification algorithm and other classic ML models.

	Test+	**Test-21**
Simple Bayes	86.65%	45.75%
Java48	75.01%	54.99%
NB Tree	74.28%	66.10%
Random Forest	78.14%	56.30%
SVM	60.16%	52.33%
Renewal neural network	81.26%	68.54%

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023

Our Model CRNN	84.30%	64.40%

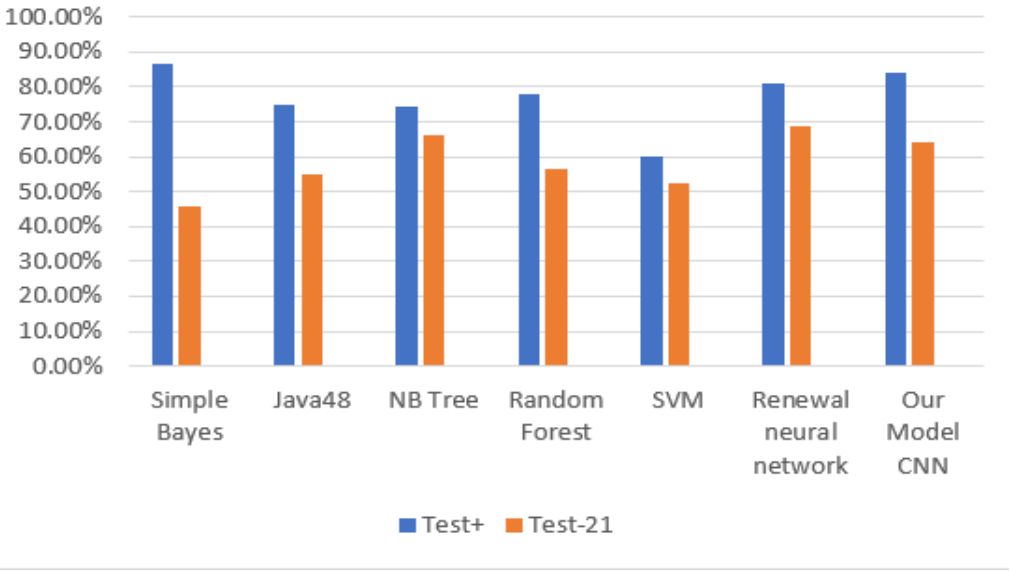

Figure 4: Comparative analysis of Multiclass classification performance of the Classification algorithm and other classic ML models.

If you use binary or multiclass classifications in the CRNN approach, it will give the better performance for detecting attacks in the network, and the CRNN approach is proven as an optimized process of malicious attack detection. Figure 4 shows the Comparative analysis of ensemble learning that demonstrates the various results for different ML methods.

6. **Conclusion**

According to the findings of our CRNN approach, using the same dataset and an intrusion detection model powered by neural networks, classical ML approaches perform less accurately in both binary classification and multiclass classification. Even while our models require more computation time, new hardware can significantly reduce that time. Using CNN and RNN we proved that our model would give the most specific and valid results and our feature aim is to apply this proposed method to the live application in a firewall as well as in cloud and IoT enabled applications. Because hackers are always deciding on new strategies or tactics to steal data or boost network traffic, we must deploy our CRNN model in the real-time system and evaluate the model to solve this issue.

References

[1] Dr. Satya Sandeep Kanumalli, Lavanya K, Rajeswari A, Samyuktha P and Tejaswi M (2023). A Scalable Network Intrusion Detection System using Bi-LSTM and CNN. https://doi.org/10.1109/icais56108.2023.10073719.

[2] Imrana, Y., Xiang, Y., Ali, L., & Abdul-Rauf, Z. (2021). A bidirectional LSTM deep learning approach for intrusion detection. Expert Systems with Applications, 185, 115524.

[3] Yoo, J., Min, B., Kim, S., Shin, D., & Shin, D. (2021). Study on network intrusion detection method using discrete pre-processing method and convolution neural network. IEEE Access, 9, 142348-142361.

[4] Jang, S., & Son, Y. (2019, October). Empirical evaluation of activation functions and kernel initializers on deep reinforcement learning. In 2019 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1140-1142). IEEE.

[5] Mohammadpour, L., Ling, T. C., Liew, C. S., & Aryanfar, A. (2022). A survey of CNN-based network intrusion detection. Applied Sciences, 12(16), 8162.

[6] Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2017, September). Applying convolutional

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023

neural network for network intrusion detection. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1222-1228). IEEE.

[7] Hajisalem, V., & Babaie, S. (2018). A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection. Computer Networks, 136, 37-50.

[8] Hajisalem, V.; Babaie, S. A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection.Comput.Netw. 2018, 136, 37–50.

[9] Mustafa, M., 2021. The Technology of Mobile Banking and Its Impact on the Financial Growth during the Covid-19 Pandemic in the Gulf Region. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), pp.389-398.

[10] Balakrishnan, C.; Ambeth Kumar, V.D. IoT-Enabled Classification of Echocardiogram Images for Cardiovascular Disease Risk Prediction with Pre-Trained Recurrent Convolutional Neural Networks. *Diagnostics* **2023**, *13*, 775. https://doi.org/10.3390/diagnostics13040775

[11] UNB-ISCX: NSL KDD Dataset. http://www.unb.ca/research/iscx/ dataset/iscx-NSL-KDD-dataset.html (2009).

[12] Salama, M.A., Eid, H.F., Ramadan, R.A., Darwish, A., Hassanien, A.E.: Hybrid intelligent intrusion detection scheme. In: GasparCunha, A., Takahashi, R., Schaefer, G., Costa, L. (eds.) Soft Computing in Industrial Applications. Advances in Intelligent and Soft Computing, vol. 96, pp. 293–303. Springer, Berlin (2011).

[13] Y. P. Zhou and J. A. Fang, "Intrusion detection model based on hierarchical fuzzy inference system," in 2009 2nd International Conference on Information and Computing Science, ICIC 2009.

[14] H. Liu, B. Lang, M. Liu, and W. Xu, "CNN and RNN based deep learning methods for wireless intrusion detection system," IEEE Access, vol. 8, pp. 42150–42159, 2020. doi: 10.1109/ACCESS.2020.2976908.

[15] Sowmya G, Navya K, Divya Jyothi G. "Machine learning and mining for social media analytics". Advances in Intelligent Systems and Computing 978- 981(2019).

[16] Gupta, B.B., Tewari, A., Jain, A.K. et al. Fighting against phishing attacks: state of the art and future challenges. Neural Computation & Application 28,3629–3654(2017).https://doi.org/10.1007/s00521-016-2275.

[17] S. Mohammadi, H. Mirvaziri, and M. Ghazizadeh-Ahsaee, ''Multivariate correlation coefficient and mutual information-based feature selection in intrusion detection,'' Inf. Secur. J. Global Perspective, vol. 26, no. 5, pp. 229–239, Sep. 2017.

[18] Madhu, G., Kautish, S., Alnowibet, K. A., Zawbaa, H. M., & Mohamed, A. W. (2023). NIPUNA: A Novel Optimizer Activation Function for Deep Neural Networks. Axioms, 12(3), 246.

Doi: https://doi.org/10.54216/FPA.130109
Received: March 24, 2023 Revised: June 27, 2023 Accepted: September 02, 2023