Intelligent Intrusion Detection System for Apache Web Server Empowered with Machine Learning Approaches

Muhammad Ubaid Ullah[1] ,Arfa Hassan[2] ,Muhammad Asif[3] ,Muhammad Sajid Farooq[4] ,Muhammad Saleem[5]
[1]Department of Computer science, Minhaj University Lahore,Pakistan.
[2]Department of Computer science, Lahore Garrison University, Pakistan
[3]Department of Computer science, NCBA & E, Pakistan
[4]Department of Computer science Lahore Garrison University, Pakistan
[5]Department of Computer science, NCBA & E, Pakistan

Abstract- Nowadays, the online communication between vendors and customer are most familiar ways due to covid-19 pandemic. The to make this communication more effective and secure, the system requires more accurate and efficient algorithms. So, in this research work an intrusion detection system for Apache web servers is proposed. The proposed method uses the Naive Bayes machine learning algorithm for training. The data set for training is taken from IEEE. The cross-validation accuracy of proposed system is 98.6%.

Keywords: Intrusion detection, Naïve Bayes, Machine Learning, Intrusion prediction.

1 Introduction

Intrusion Detection Systems (IDS) are algorithms that search a network device for malicious attacks or operation. The misuse-based IDS, on the other hand, is unable to keep up with the exponentially rising number of threats. Exploits and vulnerabilities Anomaly-based IDSs are intended to detect anomalies.to detect any deviation from standard behavior profiles. As a result, they are better suited to detecting unknown or novel attacks without any prior information than misuse-based detection systems. [1] As a result, cybersecurity has become more important in defending networks from various cyber-attacks such as intrusions, denial of service (DoS), overhearing, and rush assaults, among others. In today's security, a standard Intrusion Detection System (IDS) coupled with the grouping approach is critical. It still has limitations in terms of intelligently analyzing massive amounts of data to detect anomalies. The suggested reliably processes large data sets on existing hardware. Numerous network bases are used in real time for intrusion detection in this suggested research work. Identifies intrusions by forecasting strange test situations and storing the data in a database to reduce upcoming discrepancies, according to the suggested research. [2] The research investigates just how RFID technology may be used to detect traffic crowding at every intersection of the roadway by using RFID bookworms and labels as sensors. The goal of this study is to make traffic signal performance that is fixed and programmed dynamic. The research proposes a different strategy for assembly signal judgement proportionate to traffic flow congestion by any agreed time. The suggested intelligent system can maintain the active timings of traffic lights by sensing the density of traffic to reduce congestion using IoT enabled sensors, which supply current and strong communication technologies for people[3]

To detect attacks and anomalies in networks, an Intrusion Detection System (IDS) is used. Anderson produced IDS for the first time in 1980. Based on various data types and analysis techniques, IDS are divided into two modes. Two types of hosts: network-based and host-based.[4] Host IDS dependent on the keeps track of a on its own host's behavior and detects any malicious activity. HIDS primarily monitors method activities and ensures that files, logs, and registry keys are safe. Individual systems run host-based intrusion detection systems, which provide techniques for collecting and analyzing data on that device. To network-based secure a system from attacks, a network-based intrusion detection system (NIDS) monitors and analyses data from network traffic. The data from a system's log files is monitored and analyzed by a host-based intrusion detection system (HIDS). a basic framework Intrusion detection system can also be categorized according to how they detect intrusions. Misuse detection, specification-based detection, the identification of and anomaly based detection techniques are divided into three groups.[5]

2 Literature review

Signature based Intrusion Detection Systems (SBIDS) notice basic patterns, documented malicious instruction sequences, byte sequences in network traffic, and known software vulnerabilities in a database of previous attack signatures. Each intrusion leaves a unique collection of malicious signatures, such as failed logins, failed programme attempts, data packets, as well as failed file and folder access. Similar attacks are recognized and prevented by SBIDS. [6] Intrusion Detection Systems (IDS) that use anomalies to recognize active intrusion attempts use a baseline or studied pattern of regular system operation. Deviations from this pattern or baseline result in anto be set off an alarm Some activities that come under the radar of an anomaly detection engine trigger incidents. Outside of the generally accepted or predefined behavioral model [6] Anomaly detection is a method for detecting irregular patterns that do not follow the standard. Anomaly detection has a wide variety of uses, ranging from intrusion detection to machine health monitoring, and from credit card fraud detection to fault detection in operating systems. [6] The statistical anomaly-based intrusion detection method utilizes statistical analysis to examine user or system behavior by comparing values. Login session variables, resource overflow flags, and different timers are examples of variables.at regular intervals in this method, it's critical to find specific threshold values and reduce the false alarm rate. In The malicious activity is differentiated using the Statistical Anomaly dependent intrusion detection framework (SABIDS). Statistical properties such as the mean and variance of normal activities, as well as statistical measures that assess the deviation of activities from normal activity, are used to differentiate irregular behavior from normal behavior.[7]

Distance and rapidity estimate, time awareness, visual and auditory awareness, attentiveness, the capacity to determination safely, and action reaction time are all examples of external skills. Cognitive intelligence is an internal process that controls and maintains the entire intelligent system of the driver. These cognitive skills form the boundaries for creating adaptive conduct for dynamic settings. Knowledge, reasoning, decision-making, habit, and cognitive competence are the parameters for comprehending intelligent behavior. Various of these characteristics function concurrently to allow drivers to adapt to current events, according to intelligent behavior modelling. Changes in the environment cause the limitation values to alter, a process that continues till all operations are finished. This research uses a 'driver behavior model to get realistic intelligent driving behavior patterns to simulate intelligent behavior. This paradigm is based on layering patterns that preserve hierarchy and coherence in order to accurately transmit data from one segment to another. These patterns are the result of various modules working together to create acceptable values. Correct patterns were collected using an ANN static and go-ahead nonlinear autoregressive technique, and time series dynamic backpropagation ANN, random sub-space on real world data stayed also employed for further accuracy validation.[8]

With each passing day, the number of multimedia programmers and their users grows. To meet the demands of the next generation of network systems, several multi-carrier systems and space-time coding techniques have been created. For combined channel and multi-user identification method is suggested in this study (CMD). There are two phases to FLeABPNN. The first step calculates the channel parameters, while the second stage detects many users. The proposed solution is based on a neuro-fuzzy hybrid system that combines fuzzy logic with the capabilities of neural networks.[9] Anomaly-based systems build a baseline model by modelling the system's normal behavior and comparing it to the monitored behavior. Signature-based systems analyses the data for attack patterns known as probes and sweeps, which are predetermined and preconfigured. When these elements are present, an alarm is raised. There is an intervention action.[6]

It's a security service that controls and analyses system operations and problems in order to gain access to system resources and detect unauthorized activity. An intrusion in a network is described as a series of events that occur in a network. Attempts to jeopardize the confidentiality, honesty, and protection of the system resource availability. The information systems are effectively protected by preventive security techniques. The aim of the HIDS is to monitor the computer system's state and dynamic behavior. This monitoring system tracks all activities of inspected packets on a regular basis.HIDS recognizes which resources are being used on the network. What software has access to certain tools, etc. If there is a network, if any changes or adjustments are made, the system will be notified. Get a few network warnings HIDS is becoming more and more common. It’s important to keep the host computer's architectures and its software up to date. Activities in a network. The attribute feature of the target device is NIDS, and function modules are visible throughout the network. NIDS investigations can be done manually or automatically. Anti-thread programme is mounted on the servers by NIDS to monitor the incoming and outgoing threads. It is critical to ensure protection in a variety of areas, including application in the government, economy, industry, and education [7]

Misuse detection systems and anomaly detection systems are two types of intrusion detection systems. These two techniques can be merged to create a combination. method of identification. When signatures are uncertain or the attack pattern changes from the real signature pattern, misuse identification fails or provides less successful outcomes. Furthermore, this kind of protection method has the same issue as antivirus software in that it needs regular updates.Updates are made on a regular basis to track new types of threats. By evaluating and observing, the Anomaly Detection system produces a regular profile. The normal or baseline profile describes the normal behaviour of a network system.[8]

Using NIDS to identify the attack phase early from the network is one of the focused areas for rapidly overcoming cyber-attacks. NIDS (network intrusion detection systems) are intended to detect malicious behaviours such as viruses, worms, and other malware.DDoS attacks are a form of cyber-attack. The speed, accuracy, and reliability of abnormality detection are important success factors for NIDS. SDN is a modern architecture that separates network control and forwarding functions such that network control can be configured directly.[9] An IDS's core job is to keep an eye on information sources like computers and networks for unauthorised access attempts. IDSs gather knowledge from a number of applications and networks. Sources and review the details for future risks [9] Some of the evaluation metrics used to compare the performance of algorithms in NIDS include accuracy, false negative rate (FNR), false positive rate (FPR), time used, memory used, and kappa statistics. NIDS evaluation criteria such as accuracy, FNR, and FPR are often used. Based on different efficiency parameters, a study of three NIDS recognition methods. [9].

The IDS wanted to find out whether the tracked user action or network traffic was malicious. A warning will be activated if a malicious attack was detected. IDSs may use a variety of strategies to differentiate between attacks.Anomaly identification and threat signatures, for example IDS's success is dependent on these factors, according to the source.strategies One of the most important determinants of the efficiency of the feature construction determines the IDS's efficacy.as well as the function discovery algorithm.[10]

The attacker trying to hack into the device is referred to as an active attack. The hacker can add data during an active attack.as well as actually modifying data inside the system a framework Active threats include distributed denial of service (DDoS), and session hijacking. Replicate and impersonate. Viruses, worms, and Trojan horses are examples of malicious software. Active attacks are an example. The passive assault wants to find out what's going on. Or make use of data from the device without disrupting it resources accessible to the framework Such examples include tapping, encryption, and screening. Different forms of passive attack An assault may also be carried out by an animal. An outsider or a business insider. An insider attack is a malicious attack on a network or computer device by someone who has been granted access to the system. UBS PaineWebber is a financial services firm. Insider attacks are one form of insider attack. An outsider mounted an offensive. Through making improper use of the device Spoofing, Spam, and Spin are only a few examples. Outsider attacks are one of the most frequent forms. [11]

Threats that occur and are controlled by a computer or devices other than those under threat are known as network-based attacks.DOS attacks and distributed-DOS attacks are examples of distributed-DOS attacks. Attacks on the network Intrusion detection systems and firewalls These forms of threats have countermeasures in place. The internals of a device are tracked and evaluated by a host-based IDS system. This approach operates in real time, resulting in a major improvement in precision. To increase protection, a technique is used. Has suggested a system in which the administrator receives an OTP on a registered email address or a cell phone number Photos and video in real time Webcam or snapshots are used to record events.provides a list of attacker behaviours.[11]

In this study number of bugs in operating networks and the ingenuity of threats, unique of the maximum difficult challenges facing network operators nowadays is identifying network assaults. The research presents a deep learning approach for IDS to address this problem. One of the maximum well-known deep learning models, the Deep Auto-Encoder (DAE), is used in our approach. To prevent overfitting and local optima, the proposed DAE model is trained in a greedy layer-wise fashion. In terms of accuracy, identification rate, and false alarm rate, the experimental findings on the KDD-CUP'99 dataset reveal that our method outperforms other deep learning-based approaches.[12] Furthermore, as smart cities begin to use a variety of technologies to provide a variety of high-performance cloud resources, security concerns around connecting organizations that exchange personal requester information persist. To address these issues, the researchers present an integrated stable continuous cloud service availability system for smart connected vehicles that includes a security intrusion detection mechanism and services that satisfy users' quality of service (QoS) and quality of experience requirements (QoE) specifications. TTPs (trusted third-party entities) that serve as mediators between service requesters and providers are chosen as cluster heads for coordination purposes. The most optimal offerings are then provided to the requesters by the chosen service providers. In addition, intrusion prevention is achieved using a three-phase data traffic analysis, elimination, and classification strategy for identifying optimistic, trustworthy sources of information support requests against fake requests that can arise as a result of intrusion attacks. The approach is based on a deep analysis. Machine learning mechanisms such as belief and decision trees are used to reduce and classify data. Correspondingly. The framework is validated through simulations to demonstrate the effectiveness of the solution in terms of intrusion attack detection. [11] In this paper, PSO-FLN is a newly developed learning model for Fast Learning Networks (FLN) based on particle swarm optimization (PSO). The model was tested on the popular dataset KDD99 and applied to the problem of intrusion detection. Our research and development for training the ELM and FLN classifiers, the model was compared to a variety of meta-heuristic algorithms. In terms of learning precision assessment, PSO-FLN has outperformed other learning methods.[10] An intrusion detection system (IDS) is a piece of hardware, software, or a grouping of the two that monitors network or system activity for malicious activity. Scheming a reliable intrusion detection system is one of the most basic and critical problems in computer security. The system's main function is to detect intrusion and issue warnings when a user attempts to do so in a timely manner. When an intrusion is detected using these methods, the IDS will send a warning message to the system administrator. Anomaly detection is a critical problematic that has been studied in a variety of fields and application domains. This survey aims to provide a well-organized and detailed overview of anomaly detection studies. Each method has comparative strengths and limitations among the current anomaly detection techniques. The current state of experimentation in the field of anomaly based intrusion detection is examined, as well as recent studies in this field.[4] The implemented solutions to malicious attacks is the intrusion detection system (IDS). Additionally, attackers are constantly modifying their tools and tactics. However, putting in place a widely agreed IDS scheme is a difficult challenge. Several experiments were conducted and tested in this paper to test various machine learning classifiers using the KDD intrusion dataset. It was able to calculate multiple performance metrics in order to test the classifiers that were chosen. The emphasis was on false negative and false positive performance indicators to improve the intrusion detection system's detection rate.[13]

3 Motivation

As the world become a global village, the people want all their based needs at their doorstep so the exchange of money through online banking become more popular. But in this all process the security of transactions between customer and vendor play an important role. So, Intrusion Detection Systems (IDSs) are used to enhance the protection of systems and applications.

4 Problem statement

The research Intrusion Detection System (IDS) is produced to detect threats by tracking packets transmitted via it. Within and outside intruder’s attacks by using machine learning approaches during real time. It will improve the efficiency and proper detection.

5 Proposed Method

The automated detection process is an essential objective of any intrusion detection system in a network. It is difficult to determine the network traffic either anomalous or legitimate. Automated detection systems to detect anomalous network traffic are mostly based on machine learning methods. The suggested model introduced to the Intelligent Intrusion Detection System for Apache web server Empowered with naïve Bayes as shown in Figure 1

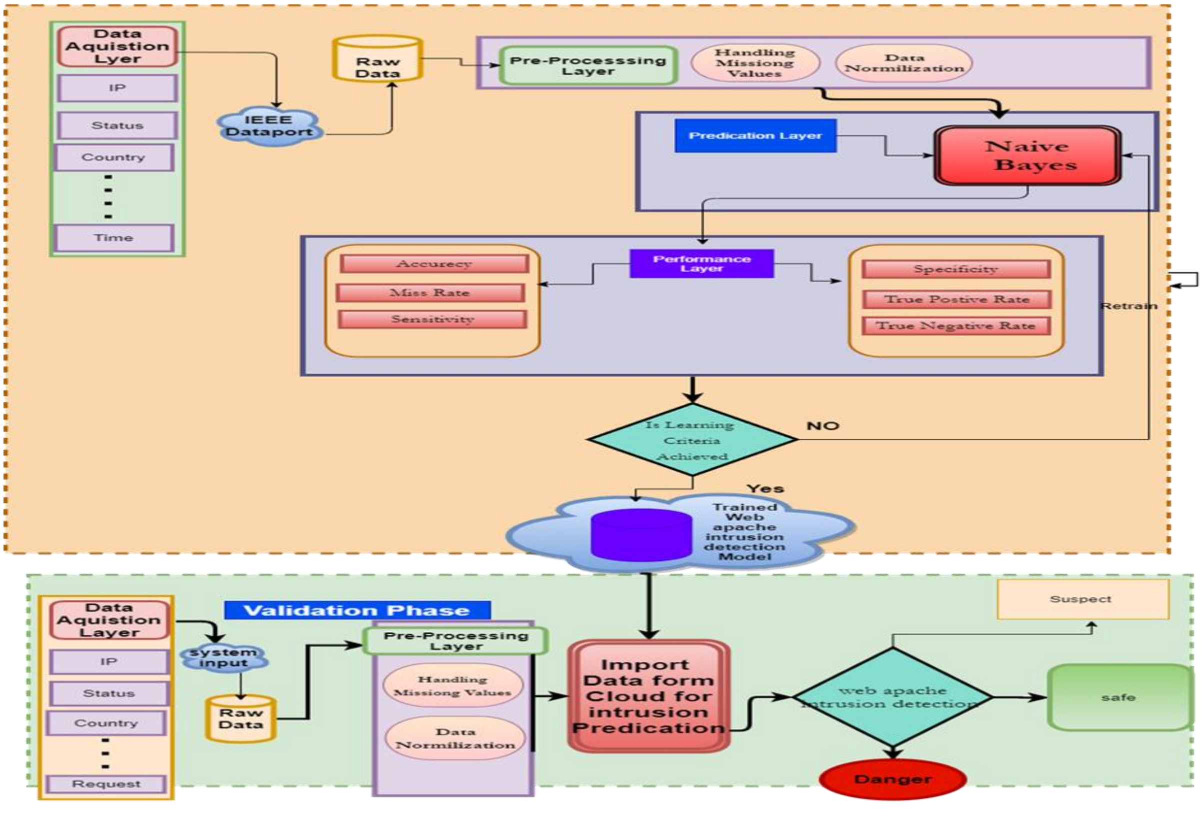

Figure 1: proposed naïve Bayes model for web apache intrusion detection

The proposed model namely training and validation phase. The data acquisition layer was used to store data, After storing data the preprocessing layer activated and handled repetition and missing values using different techniques like mean, mode and moving average method in databases. Support vector Machine algorithm will applied to all three datasets to train the model in application layer. Next, the performance evaluation layer activated and evaluated the model. If the learning criteria did not meet of the proposed model then model need to retrain and if the learning criteria meet then trained data was stored in trained database in cloud for future use.

6 Simulation and Results

For simulation and results MATLAB 2020Ra version is used. And the data set is taken from IEEE dataport .The initial dataset contain total 9 features from which 8 are input features and one is output feature. The dataset is collected from Indonesia so the classes are labeled in Bahasa Indonesia. The details of these variables are shown in table 1.

Table 1: Input variable table

Sr no.	Variable name
1	IP
2	Status
3	Country
4	Referrer
5	Browser
6	Date /time
7	Size
8	Gmt
9	*Detection*

The data to create this dataset is taken from Apache web server log. And assume that the traffic from out the country is strictly prohibited. In data preprocessing and feature extraction phase the propose system handle the missing value by putting 0 and NA and the select only 7 input features. Which are IP, status, Country , Referrer , Browser and size. For training random sample of 300 entities of each class are selected. The proposed system is trained on 5 cross validation in MATLAB classifier learning app and then import it to the work space for independent case testing. The 5 cross validation training accuracy of proposed system is 98.8% as shown in figure 2.

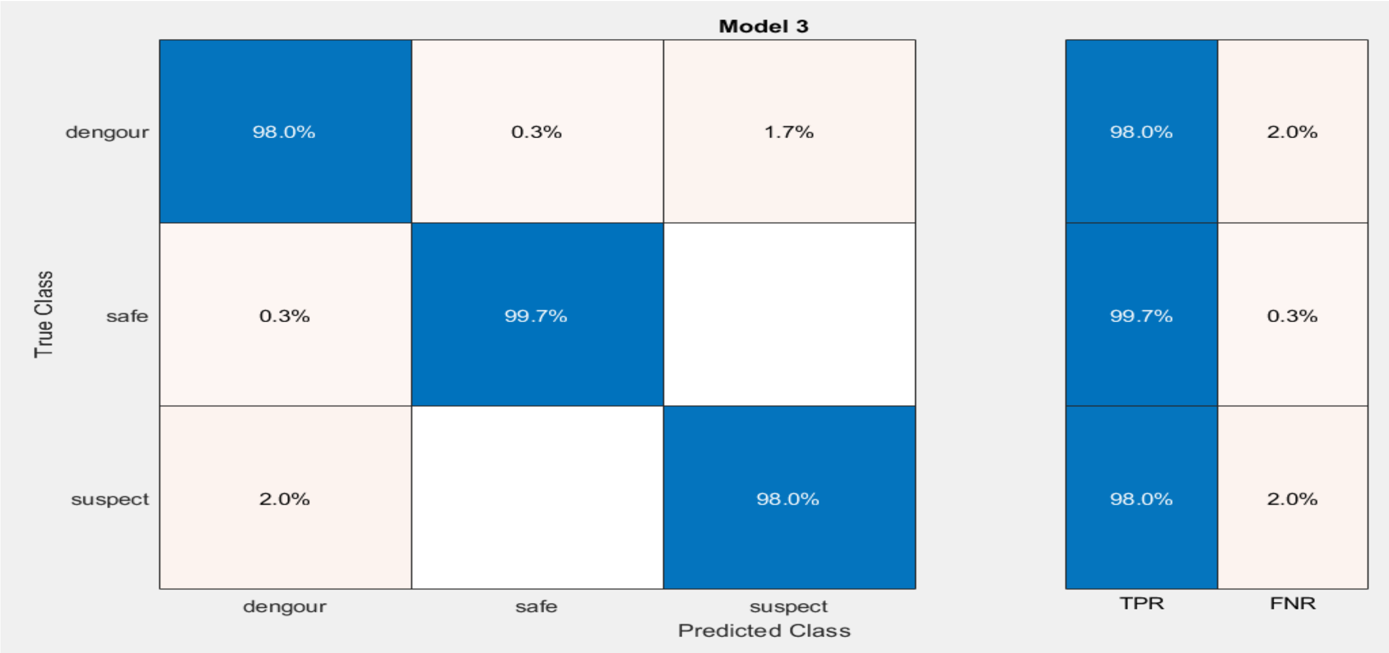

Figure 2: Confusion matrix

Figure 1 shows the True positive and False negative rate of each class. The roc graph of each class is shown in figure 3. Figure 3 shows the roc curve of class 1 ,2 and 3 respectively

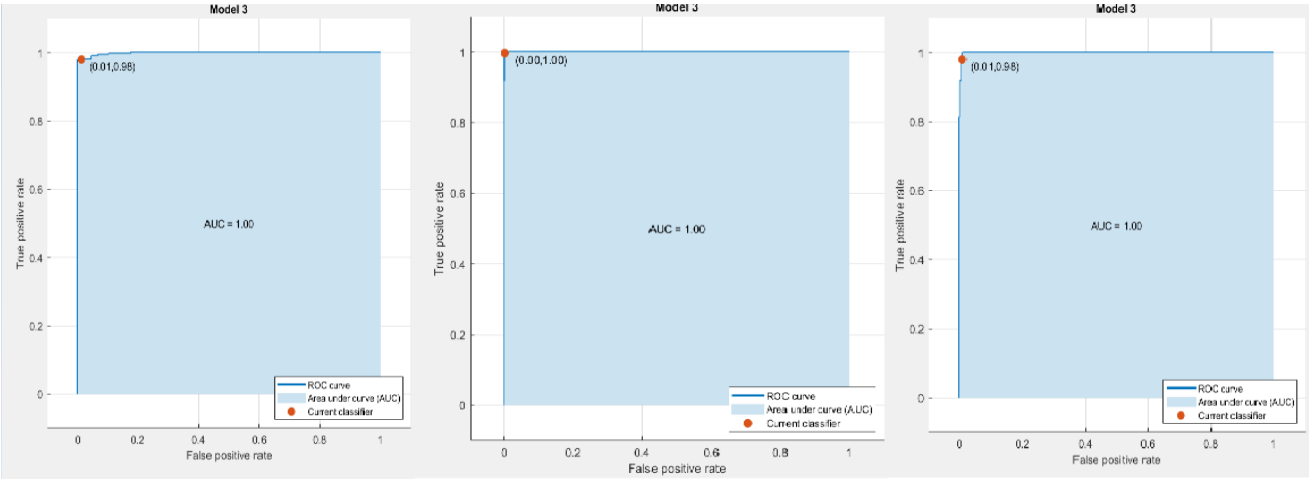

Figure 3 : ROC curve of class danger ,safe and suspect

7 Validation

To evaluate the performance of a model different kind of mathematical test are to be perform. The details of these tests are shown in table 2

Table 2: Model evolution table

Sr no.	Evaluation measure	Class 1	Class 2	Class 3	overall
1	Accuracy	98.0%	99.7%	98.0%	98.57%
2	True positive Rate	98.0%	99.7%	98.0%	98.57%
3	True negative rate	98.8%	98.0%	98.85%	98.55%
4	Sensitivity	97.67%	99.66%	98.33%	98.55%
5	Specificity	98.99%	99.83%	98.99%	99.27%

8 Discussion and Future work

The proposed model shows the 98.57% accuracy rate. The proposed is not suggested previously on this model and design. All the system which are previously proposed are used deep learning base mechanism and their main target for intrusion detection is different. The apache web base intrusion detection with using web server log is only done in this proposed work.

At this stage this work is only worked on for specific countries so in future we can enhance it to all over the world

9 References

[1] C. H. R. Madhuri, G. Anuradha, and M. V. Pujitha, “House Price Prediction Using Regression Techniques: Comparative study,” 6th IEEE Int. Conf. &; amp;quot; Smart Struct. Syst. ICSSS 2019, pp. 1–5, 2019, doi: 10.1109/ICSSS.2019.8882834.

[2] Eshghi and M. Kargari, “Introducing a Method for Combining Supervised and Semi-supervised Methods in Fraud Detection,” Proc. 2019 15th Iran Int. Ind. Eng. Conf. III 2019, pp. 23–30, 2019, doi: 10.1109/IIIEC.2019.8720642.

[3] A. Nur, R. Ema, H. Taufiq, and W. Firdaus,"Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study" . Malang, East Java, Indonesia,‖ Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 10, pp. 323–326, 2017, doi: 10.14569/ijacsa.2017.081042.

[4] D. Banerjee and Dutta, "Predicting the housing price direction using machine techniques", in 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 2998–3000, doi: 10.1109/ICPCSI.2017.8392275.

[5] V. Limsombunc, C. Gan, and M. Lee,"House Price Prediction: Hedonic Price Model vs. artificial Neural Network", Am. J. Appl. Sci., vol. 1, no. 3, pp. 193–201, 2004, doi:10.3844/ajassp.2004.193.201.

[6] S. A. Harmon et al., "artificial intelligence for detecting COVID-19 pneumonia on chest C.T. using multinational datasets", Nat. Commun., vol. 11, no. 1, pp. 1–7, 2020, doi: 10.1038/s41467-020-17971-2.

[7] M. Jethanandani, T. Perumal, and A. Sharma, "Random k-Label sets method for human activity recognition with multi-sensor data in smart home", 2019 IEEE 16th India Counc. Int. Conf. INDICON 2019 - Symp. Proc., 2019, doi: 10.1109/INDICON47234.2019.9030296.

[8] L. M. Rojas-Barahona, "Deep learning for sentiment analysis", Lang. Linguist. Compass, vol. 10, no. 12, pp. 701–719, 2016, doi: 10.1111/lnc3.12228.

[9] N. Shinde and K. Gawande, "Valuation of House Price Using Predictive Techniques',' Int. J. Adv. Electron. Comput. Sci., vol. 5, no. 6, pp. 34– 40, 2018.

[10] C.K. Madhuri, G.Anuradha and M.V. Pujitha, Macch,2019. "House price prediction using regression techniques: a comparative study". In 2019 International Conference on Smart Structures and Systems (ICSSS) (pp. 1-5). IEEE.

[11] P. Durganjali and M.V.Pujitha, "House Resale Price Prediction Using Classification algorithms", 6th IEEE Int. Conf. &amp;amp;amp;amp;amp;quot;Smart Struct. Syst. ICSSS 2019, pp. 1–4, 2019, doi: 10.1109/ICSSS.2019.8882842.

[12] M. Mukhlishin, R. Saputra, and A.Wibowo, "Predicting house sale price using fuzzy logic", Artificial Neural Network, and K-Nearest Neighbor. 2017.

[13] N. Bhagat, A. Mohokar, and S. Mane, "House Price Forecasting using Data Mining", Int. J. Comput. Appl., vol. 152, pp. 23–26, Oct. 2016, doi: 10.5120/ijca2016911775.

[14] V. Limsombunc, C. Gan, and M. Lee, "House Price Prediction: Hedonic Price Model vs. artificial Neural Network", Am. J. Appl. Sci., vol. 1, no. 3, pp. 193–201, 2004, doi: 10.3844/ajassp.2004.193.201.

[15] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques", in 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 2998–3000, doi: 10.1109/ICPCSI.2017.8392275.

[16] E. Beracha, B. Gilbert, T. Kjorstad, and K. Womack, "On the Relation between Local Amenities and House Price Dynamics", Real Estate Econ., Jul. 2016, doi: 10.1111/1540- 6229.12170.

[17] S. Law, "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London", Cities, vol. 60, pp. 166– 179, Feb. 2017, doi:10.1016/j.cities.2016.08.008.

[18] N. Bogin and W. M. Doerner, "Property renovations and their impact on house price index construction", J. Real Estate Res., vol. 41, no. 2, pp. 249–283, 2019, doi: 10.5555/0896-5803.41.2.249.

[19] N.Ali,S. Abbas ,M. Shahid, "Proposed Framework of Smart City for Gawadar, Balochistan Pakistan". Int J Econ Manag Sci. 2017;6(436):2.

[20] M.Asif ,M.A Khan ,S. Abbas ,M. Saleem. "Analysis of Space & Time Complexity with PSO Based Synchronous MC-CDMA System". In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) 2019 Jan 30 (pp. 1-5). IEEE.

[21] T .Batool ,S. Abbas,Y.Alhwaiti,M. Saleem , Ahmad and et al. "Intelligent Model Of Ecosystem For Smart Cities Using Artificial Neural Networks",. INTELLIGENT AUTOMATION AND SOFT COMPUTING. 2021 Jan 1;30(2):513-25.

[22] M.Saleem ,M.A Khan,S. Abbas ,M. Asif ,M. Hassan and et al. "Intelligent FSO Link for Communication in Natural Disasters empowered with Fuzzy Inference System",. In2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) 2019 Jul 24 (pp. 1-6). IEEE.

[23] S.Y.Siddiqui,A. Athar,M.A.Khan, and et al, 2020. "Modeling, Simulation, and Optimization of Diagnosis Cardiovascular Disease Using Computational Intelligence Approaches",. Journal of Medical Imaging and Health Informatics, 10(5), pp.1005-1022.

[24] S.Y.Siddiqui,M.A. Khan,S. Abbas, and F.Khan, in 2020. "Smart Occupancy Detection for Road Traffic Parking using Deep Extreme Learning Machine:,. Journal of King Saud University-Computer and Information Sciences.

[25] M.W.Nadeem,M.A.A. Ghamdi,M. Hussain and et al., 2020. "Brain Tumor Analysis Empowered with Deep Learning: A Review, Taxonomy, and Future Challenges. Brain Sciences", 10(2), p.118.

[26] S.Mehrban,M.W.Nadeem,M. Hussain,M.M. Ahmed and et al., 2020. "Towards Secure FinTech: A Survey, Taxonomy, and Open Research Challenges",. IEEE Access.