Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 05 Issue 01
2023

Research Article
https://doi.org/10.56979/501/2023

A Systematic Review of Artificial Intelligence Techniques Used for IDS Analysis

Abdul Majid Soomro[1*], Awad Bin Naeem[1], Muhammad Imran Ghafoor[2], Biswaranjan Senapati[3], and Muhammad Asim Rajwana[1]

[1]Department of Computer Science, National College of Business Administration & Economics, Multan, Pakistan.
[2]Department of Engineering, Pakistan Television Corporation, Lahore, Pakistan.
[3]Department of Computer Science and Data Science, Parker Hannifin Corp, USA.
[*]Corresponding Author: Abdul Majid Soomro. Email: gi180004@siswa.uthm.edu.my

Received: January 18, 2023 **Accepted:** April 28, 2023 **Published:** June 05, 2023

Abstract: Network security is critical for protecting sensitive data, avoiding data breaches, defending against cyber-attacks, ensuring operational continuity, and adhering to regulatory obligations. Security is seen as a danger in today's work environment. When a network begins to behave abnormally, an attack is launched. To get access, attackers use sloppy security processes, code faults such as buffer overflows, and network vulnerabilities. Attackers might be persons with limited access to the system who want greater control over who uses it, or they could be so-called black hat hackers who are just normal internet users attempting to steal crucial information. Intrusion Detection System software monitors network traffic for signals of unauthorised access or suspicious behaviour. It evaluates data from a range of sources, including network traffic logs, system logs, and security events, to detect and inform users of potential security threats. In contrast to an intrusion detection system (IDS), which monitors the whole network, an intrusion detection system (IDS) is a tool that examines network traffic for indicators of odd behaviour and problems, detects, and responds to unauthorised system activity alerts when they occur. In this research, we will employ intrusion detection systems (IDS) to detect suspicious activities.

Keywords: IDS, Network Traffic, Security, Random Forest, KNN.

1. Introduction

This case discusses lives have recently gotten more and more reliant on innovations like computers, the Internet, and other electronics. The majority of these gadgets are connected and essential for humanity in various fields such as health care, business, transportation, and education [1]. For many years, traditional networking technology has attempted to link computers, printers, and large machinery. The complexity and sophistication of modern electronic gadgets are increasing, just as the burgeoning Internet of Things (IoT). As a result, new networking models are emerging for communication that is dependable, trustworthy, and quick. Despite the growing influence of networks on modern life, cyber security has grown into a more vital area of research[2]. The most popular cyber security instruments (IDSs) include firewalls, antivirus software, and intrusion detection systems. An IDS is an example of a detection system that keeps track of the software and hardware configurations running on a network and aids in the defence of cyber security. Automobile accidents claim the lives of millions of people worldwide each year[3]. There is a traffic report that informs how many unintentional deaths and disabilities occur every day, particularly in underdeveloped nations like Ethiopia. High technology firms like Google and Tesla have given transport infrastructure more of their focus[4].

These companies have been developing autonomous vehicles for a long time, and their work has resulted in important advancements. According to 2020 research, between 2018 and 2020, 92% of enterprises supporting the nation's key infrastructures for energy, health, industrial and manufacturing, and transportation suffered at least one hack that resulted in data breaches or severe 3 operational interruptions[5]. According to joint reports from The EU Agency for Cyber Security and Joint Research Centre, autonomous cars are also susceptible to cyber security issues affecting physical sensors, controllers, and their connection methods. Human life will be substantially at risk if hackers gain access to control self-driving cars via network breaches[6].

With minimal emphasis on its security, the majority of research has been devoted to the creation of the SDN-based VANET system. Without first assuring the system's security, it is difficult to attain complete system operation. It is believed that an attacker could exploit the system. If the SDN-based VANET system's security is disregarded, it can hack the system[7]. A strong security mechanism is needed to maintain the system's functionality and defend it from outside attacks in 2013. Intrusion detection systems are therefore a network's secondary line of defence. IDS typically keeps track of the network's behaviour by utilizing various approaches to analyze both its content and statistical data. The most popular areas of research to find anomalies and improper network behaviour are machine learning approaches[8]. For the effective classification of assaults, supervised techniques for machine learning like Random Forest and K-Nearest Neighbor outperform other conventional classifiers when creating IDS for general use, the majority of researchers concentrate on the accuracy of machine learning techniques, but the nature of the networks and the application domain have varied effects on the quality of these IDs[9]. As a result, for a particular network environment known as a Software-defined network based on vehicular ad hoc network, the study is going to construct an IDS employing Random Forest and K-Nearest Neighbors machine learning techniques.

Designing and implementing effective IDS in Software Define Network-based VANETs is essential to defending against the rising danger of assaults launched from inside or external invaders of the networking environment. Most publications did not take the Software Define Network's nature into account when developing IDS[10]. This issue is reflected in and addressed in our suggested work. The system's correctness has a substantial impact on network security. Generally speaking, strengthening the security of a Software Define Network-based VANET entails indirectly protecting human lives by shielding vehicles from attackers and other attackers. By safeguarding communication from vehicle to vehicle and from vehicle to infrastructure, this research will improve the reliability of intelligent transportation systems. A wide number of algorithms and techniques are covered under the broad topic of machine learning, which trains computers to make predictions based on data. In addition to reinforcement learning, which involves applying incentives and penalties to train models, it also comprises supervised and unsupervised learning[11].

While Deep Learning, a subset of Machine Learning, utilizes multi-layered neural networks to model intricate data interactions, is used to model complex data relationships[12]. In particular, in fields like computer vision and natural language processing, it has been responsible for many of the advancements in AI during the past several years. In IDS, networks and devices are connected using a firewall and router. Multiclass categorization, in contrast, involves classifying materials into more than two different categories.

Relating to irrational learning, the algorithm has been getting better since it started adjusting based on past performance. An intrusion detection system (I.D. system) is software that uses different machine-learning learning methods to scan networks for unwanted activities[13]. An ID is a piece of hardware or software that maintains an eye out for future attacks or policy violations on a system and alerts the appropriate parties when they happen. Host-based intrusion detection systems (HIDS) and network-based intrusion detection systems (NIDS) are two categories based on the type of system (NIDS).

The major goal of our these are to develop an IDS for SDN-based VANETs using machine learning algorithms. The following list includes the precise goals of the paper:

- To perform preprocessing activities on categorical data using a one-hot encoder and scale up encoded data using a standard scalar.
- Research various feature selection methods and use recursive elimination of features on transformed datasets.
- NSL-KDD dataset is out of the 41 features, find additional relevant features for a software-defined vehicular network environment.
- To create an IDS model for an SDN-based VANET employing these chosen features.
- To assess the effectiveness of the machine learning algorithms Random Forest and KNearest Neighbors for the chosen features.
- To assess our model's performance against three earlier studies that focused on developing IDS for certain software-defined networks.

As a result, both ML and DL are essential elements of AI, and they work well together. Without ML, AI would not be able to learn from data, and without DL, AI would not be able to tackle increasingly difficult tasks that call for in-depth knowledge of and representation of the data. One of the most significant flaws of signature-based IDS solutions is their inability to detect unknown assaults. Malicious actors can quickly alter the attack sequences they use in malware and other types of attacks to avoid being discovered.

A significant flaw of signature-based systems is that they can only defend against known threats. Purely signature-based intrusion-detection systems have not been effective in past years. Code Red and Nimda, two recent Internet worms, highlighted the necessity for systems that can identify and stop unknown attacks. The research in studies returned some studies based on selected articles cross-reference. A wide number of algorithms and techniques are covered under the broad topic of machine learning, which trains computers to make predictions based on data. In addition to reinforcement learning, which involves applying incentives and penalties to train models, it also comprises supervised and unsupervised learning. This work is organized as follows: Section 2 contains the literature review. Section 3 presents the methodology of the study. Section 4 discusses the results. In last, the conclusion and future work is described in section 5.

2. Literature Review

According to the Study ML uses a statistical modeling technique to learn from earlier data sequences. The system then predicts one of the most likely outcomes using these sequences and their specific properties. The IDS uses a combination of anomaly-based detection, signature-based detection, and a variety of the two to identify malicious activities. The signature of an attack pattern can be utilized to recognize it via signature-based detection. It may acquire information from online I.D. systems, analyze it, and identify previously unidentified dangers to address network security issues[14].

Intrusion detection systems can benefit conventional industrial control systems and the Industrial Internet of Things. A massive and complex system is known as the Industrial Internet of Things. The entire system could quickly suffer severe harm if any component fails or behaves strangely. Therefore, to stop network attacks, you must be able to detect them quickly and accurately. Since it allows for quick detection of network breaches, an intrusion detection system (IDS) is essential to network security defense. In 2019, showed how to use deep learning to help the IICS find unusual things. Information that can be used to train users and monitor their behavior is contained in TCP/IP data packets. Using the new UNSWNB15 data set, developed multilayer deep neural networks. As a result, they were able to predict Industrial IoT risks for 2020[15].

However, the speed, quantity, and complexity of modern multidimensional data are beyond the systems mentioned above' capabilities. An advanced training program is typically required when working with Industrial IoT data. There is a need for more precision as a result. Industrial IoT made it possible to connect production, monitoring, and control systems in previously impossible ways. The control room contains several devices[16]. The handling of many forms of corporate information is made easier through management integration. Because of the Industrial Internet of Things and accessibility, network security challenges are growing. Deep R.L. can maximize reward in a known network environment, and its exploration function can uncover new relevant network environment data automatically. The model immediately complies[17]. Introduced a reinforcement learning-based intrusion detection system and compared it with adaptive machine learning and hybrid cluster intrusion detection systems. In 2020, they intended to analyze and monitor sensor networks. Researchers evaluated the efficiency of several machine learning (ML) models on the NSL-KDD[18]. They achieved this by utilizing multiple ML strategies and attribute extraction techniques. The model does not detect new attacks because negligible attacks have limits that are ignored and because it concentrates on signature-based threats and has a high FPR. In the past, it was common for people to under or over-evaluate their favored model on various data sets. In a label IDS using feature extraction is presented. When the ensemble classifier and feature selection approach are used, the accuracy and speed of intrusion detection are improved. The widely known NSL-KDD dataset, as well as the recently released IDS2017-CIC and AWID databases, were all used in the article[19].

The CFS-BA technique was used to obtain the data. The ensemble-based approach makes it easier to categorize items into various groupings. The model performed best when evaluated against the AWID dataset. Internal intrusion detection systems (IIDS) and intrusion detection systems (IDS), employ forensic and data mining technologies to operate in real-time. A new kind of security technology called intrusion detection monitors systems for hostile activity. For supporting intrusion detection in cyber analytics, data mining techniques are presented[5]. These traits can be employed in a real-time system to identify internal intruders and their malevolent behaviors while developing a new IDS. It will be a reliable IDS that can be utilized by many businesses and MNCs to safeguard their sensitive data since it can accurately identify internal intruders in real time[20]. While the strategies we suggest enhanced accuracy and detection rate up to 95%, the maximum accuracy and detection rate was 90.12%. As a result, both ML and DL are essential elements of AI, and they work well together. Without ML, AI would not be able to learn from data, and without DL, AI would not be able to tackle increasingly difficult tasks that call for in-depth knowledge of and representation of the data[21].

The IDS improve efficiency and dependability as a result. They used the Kyoto2006+ dataset to evaluate their approach, which is more engaging than the most popular yet out-of-date datasets. Despite many false positives, their work is incredibly accurate. A real-time hybrid IDS used anomaly detection for new attacks and signature-based detection for common assaults[22]. The computer's accuracy rapidly increased to an incredible 92.65% by the end, lacking the most recent and relevant attack labels. Current information indicates that the detection rate is low. This is because you are now unable to delete all unnecessary columns. There were way too many false positives. This happens when non-aggressive web traffic is mistakenly classified as such. An IDS will be more challenging to use and less effective if there are a lot of false positives. The performance of anomaly-based IDS should be improved, especially in the FPR, according to a study by which the performance of the extreme-gradient boosting and Ada Boost models was assessed using the NSL-KDD dataset. To improve the performance of the IDS, hybrid or ensemble ML classifiers are needed even with adequate accuracy. Because they lacked a technique to extract features, many prior studies could not address the problems of delayed detection and extended execution times[23].

They achieved this by utilizing multiple ML strategies and attribute extraction techniques. The model does not detect new attacks because negligible attacks have limits that are ignored and because it concentrates on signature-based threats and has a high FPR. In the past, it was common for people to under or over-evaluate their favored model on various data sets[23]. In a label IDS using feature extraction is presented. When the ensemble classifier and feature selection approach are used, the accuracy and speed of intrusion detection are improved[24].

3. Materials and Methods

The intended SLR is carried out by the instructions derived from the Cochrane Manual for the systematic investigation of Interventions, the proposed reporting Items for systematic research, and the Meta-Analysis statement with various criteria. This section covers in full the systematic research process utilized to perform this review.

This study research commenced by identifying primary studies and other relevant studies on IDS detection and analysis in AI on electronic databases. This research process has included all well-reputed journals, a workshop (International Workshop on this topic), and the Inter-National Workshop and conference proceedings. Repositories, which we used, are ACM Digital Library, Web of Science, IEEE Explore, Science Direct, Springer, and Scopus. To achieve the research goal, the following major search keywords were used to formulate the search query. Then, using Boolean operators, an initial pilot search string was generated by connecting primary keywords with alternate terms and synonyms. At last, as a focal gathering for distributing S.L.R., we physically read the titles of the International Conference on EASE papers from 2016 to 2022, which are open on the web.

Snowballing Search, Backward searches (references), Forwarded searches (references), Criteria based on systematic literature review, and Database searches.

We performed a series of operations to achieve work objectives and flow of work as shown in (Figure 1).

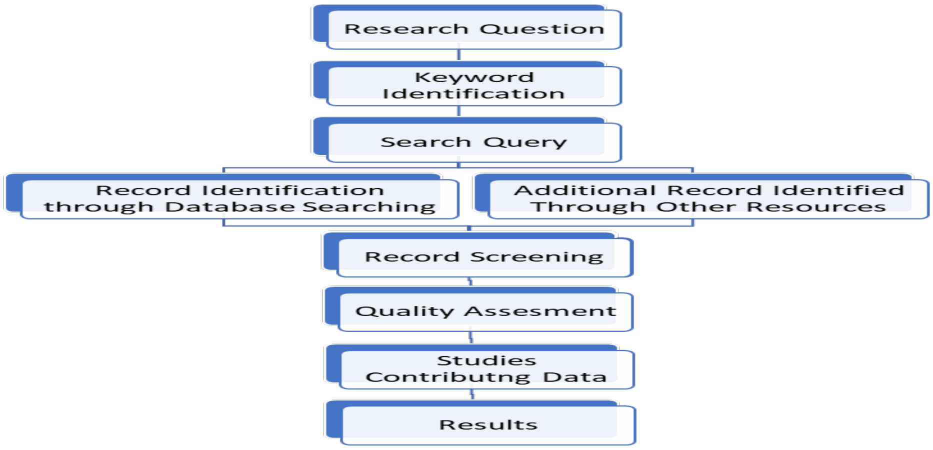


Figure 1. Series of Operations For SLR

3.1. Data Collection

(Table I) show a database which explores 29 articles from IEEE, 12 articles explore from ScienceDirect, 164 articles explore from Semantic Scholar and also search 12 articles from Microsoft Academic. Totals analyse 217 publications, and after eliminating the same duplicate articles that remained at 182, the next step removes articles and chooses 134 based on the title. Finally, in the examination of the abstract and content, 100 articles are selected, and the remaining 50 papers are for a systematic review.

3.2 Data Collection

This study research commenced by identifying primary studies and other relevant studies on IDS detection and analysis in AI on electronic databases. This research process has included all well-reputed journals, a workshop (International Workshop on this topic), and the Inter-National Workshop and conference proceedings. Repositories, which we used, are ACM Digital Library, Web of Science, IEEE Explore, Science Direct, Springer, and Scopus. To achieve the research goal, the following major search keywords were used to formulate the search query, which is given in (Table 2). Then, using Boolean operators, an initial pilot search string was generated by connecting primary keywords with alternate terms and synonyms.

- Criteria based on systematic literature review.
- Database searches.

The following search string in (Table 1) shows to find the research articles by using primary and secondary keywords.

Table 1. Shows the Primary and Secondary Keywords.

Primary keyword	**Secondary keyword**	**Additional keyword**
AI, IDS	Data sets, Methods, Techniques	Design, IOT, HIDS

The following search (table 2) shows the search string used to find the research articles by using a digital library.

Table 2. Shows the Digital Library.

Digital Library	**Search String**
Wiley IEEE Springer, Scopus ACM Semantic Scholar Science Direct	(("IDS " AND "detection" OR "analysis" OR "method" OR technique)) AND ("AI" OR "ML" OR "DL")

3.3 Population of Sampling

Papers selected from 2016-2022 during the past seven years (figure 2) show a prima flow diagram. After that, different steps are performed for the quality assessment of articles. The first title-based filtering is performed. Then abstract and keyword-based filtering of papers is used to select documents. After that, a paper quality assessment is completed. Inclusion and exclusion criteria for papers are given in the following. The related studies are identified using different algorithms and tools based on their titles and abstracts. The selection criteria for the study are carefully followed to raise the quality of the research study.

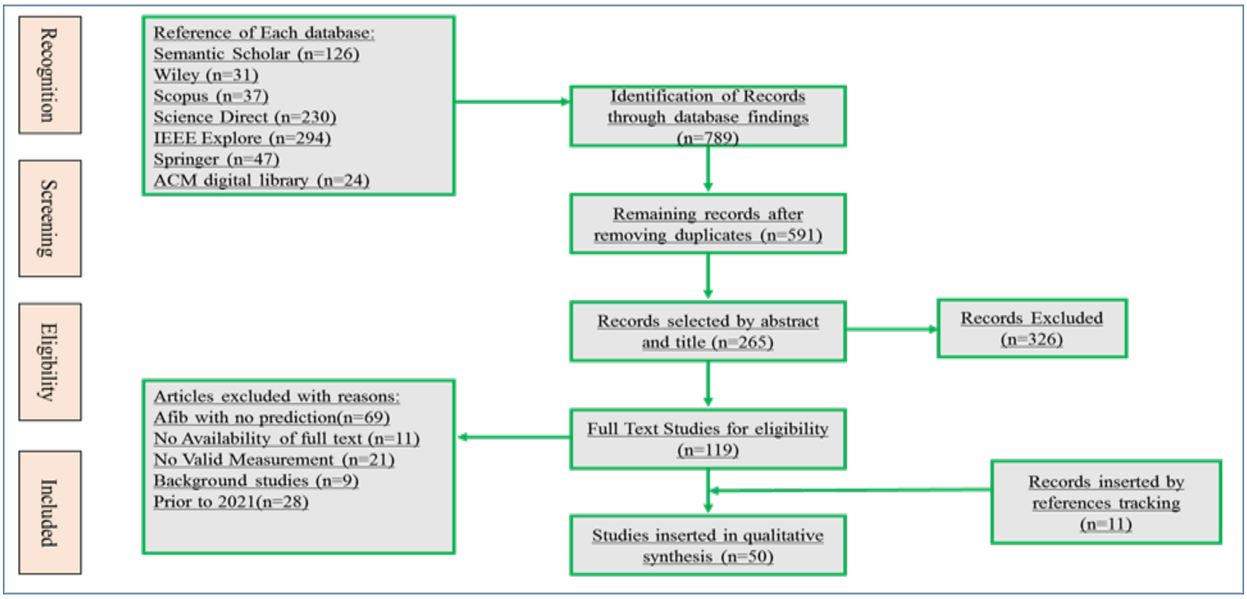

Recognition
Screening
Eligibility
Included
Reference of Each database:
Semantic Scholar (n=126)
Wiley (n=31)
Scopus (n=37)
Science Direct (n=230)
IEEE Explore (n=294)
Springer (n=47)
ACM digital library (n=24)
Identification of Records through database findings (n=789)
Remaining records after removing duplicates (n=591)
Records selected by abstract and title (n=265)
Records Excluded (n=326)
Articles excluded with reasons:
Afib with no prediction(n=69)
No Availability of full text (n=11)
No Valid Measurement (n=21)
Background studies (n=9)
Prior to 2021(n=28)
Full Text Studies for eligibility (n=119)
Records inserted by references tracking (n=11)
Studies inserted in qualitative synthesis (n=50)

Figure 2. Prisma Flow Diagram

3.4 Machine Learning Algorithm

3.4.1 Logistic Regression

Logistic regression (LR) is a common and successful technique for supervised classification. It is an extension of ordinary regression since it can only model a binary variable, which typically signals whether an event will occur or not. LR makes it easy to determine if a new instance belongs to a certain class. Because it is a probability, the result will be between 0 and 1. As a result, when employing the LR as a binary classifier, a threshold must be established to differentiate between two classes. Instances of input are categorized as "class A" if their probability value is more than 0.50; otherwise, they are classified as "class B".

3.4.2 Support Vector Machine

The support vector machine approach can classify both linear and nonlinear data. Each data set is initially mapped into an n-dimensional feature space, where n represents the number of features. The next step is to find the hyperplane that divides the data into two groups while maximizing marginal distances for both classes and minimizing classification errors. The marginal distance for that class is the distance between the decision hyperplane and the nearest instance of that class. Each feature's value is the sum of the 21 values of a certain coordinate, and each data point is first represented as a point in an n-dimensional space. The categorization is then carried out by selecting the hyperplane with the largest margin of separation between the two classes. The topology of multilayer perceptron neural networks and support vector machine models are similar. The concept of a margin, which is the region on each side of a hyperplane dividing two classes of data, lies at the heart of SVMs.

3.4.3 Naïve Bayes

Naive Bayes detectors are well-known for being simple but effective linear classifiers. The Bayes' theorem underpins the probabilistic model employed by naive Bayes classifiers; the word "naive" refers to the naive assumption that characteristics in a dataset are independent of one another. Even though their autonomy assumption is often violated in practice, Naive Bayes classifiers function wonderfully under this erroneous assumption. For limited data sets, naïve Bayes classifiers may outperform more complex alternatives. Bayes classifiers are often less accurate than more complicated learning algorithms "like ANNs". However, it was shown that the Naive Bayes classifier was sometimes superior to the other learning schemes, even on datasets with significant feature dependencies, and outperformed state-of-the-art algorithms for decision tree induction, instance-based learning, and rule induction on popular benchmark datasets. The Bayes classifier's attribute independence issue was handled using averaged one dependency estimators.

3.4.4 K-Nearest Neighbor

One of the oldest and most fundamental classification approaches is the K-nearest neighbor. It is a shortened version of an NB classifier. In contrast to the NB approach, the KNN algorithm does not need the consideration of probability values. The K in the KNN algorithm stands for the number of nearest neighbors considered while casting a "vote." Multiple classification outputs for the same sample item may be achieved by varying the value of "K." K-nearest-neighbor classification is one of the most basic and straightforward categorization approaches, and it is employed when there is little to no prior knowledge about the structure of the data. It should be one of the top possibilities for classification research. Nearest-neighbor classification was created in response to the need for discriminant analysis when trustworthy parametric estimates of probability densities are unavailable or computationally unfeasible.

KNN is a powerful supervised learning strategy that has been explained for a range of challenges, including security approaches. It determines the class classification of a test sample based on the classification of its k neighbors. The K-nearest neighbor algorithm is based on the grouping of components having similar qualities. The value of k in the KNN is affected by the size of the dataset and the kind of classification problem. KNN categorizes the target based on its neighbors, as seen in (Figure 3).

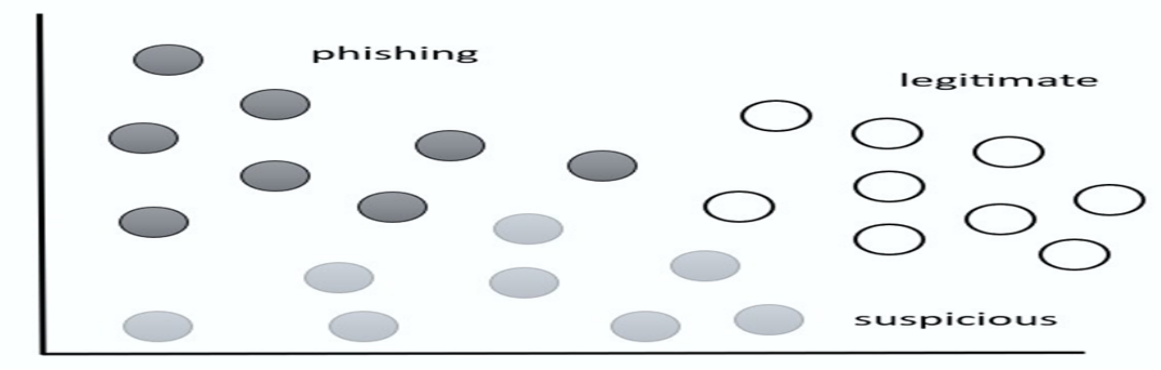


Figure 3. KNN-Classification

3.4.5 Random Forest

As a forest is built up of many trees, so is a random forest (RF), which is an ensemble classifier made up of many DTs. Deep DTs typically result in overfitting of the training data, resulting in a significant variation in classification outputs for little changes in the input data. Because they are so sensitive to their training data, they are very prone to mistakes on the test dataset. (Figure 4) depicts an RF classification, with distinct DTs learned using different sections of the training information. To categorize a new sample, the input vector must be transported with each DT of the forest. After each DT has considered a different input vector segment, the classification result is returned.

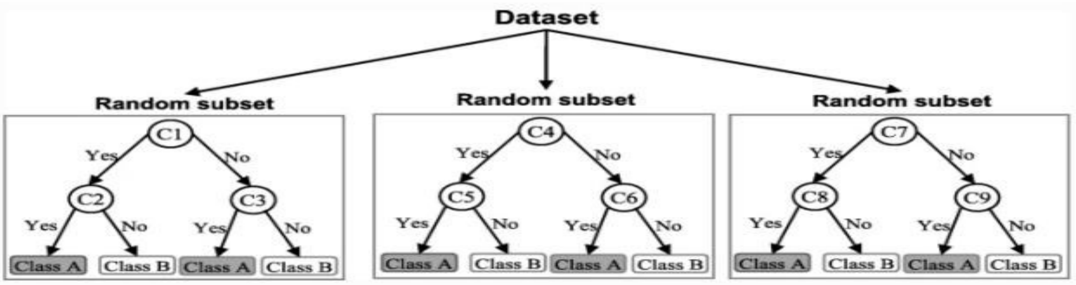


Figure 4. Rf-Classification

3.4.6 Artificial Neural Network

An artificial neural network is a network of interconnected units made up of a large number of neurons as shown in (Figure 5). Each neuron in the network is capable of receiving input signals, processing them, and sending out an output signal. It is made up of a set of weighted synapses, an adder for summing input data weighted by synaptic strength, and an activation function for limiting the amplitude of the neuron's output. Recurrent networks and multilayer feed-forward networks have fundamentally different types of network structures.

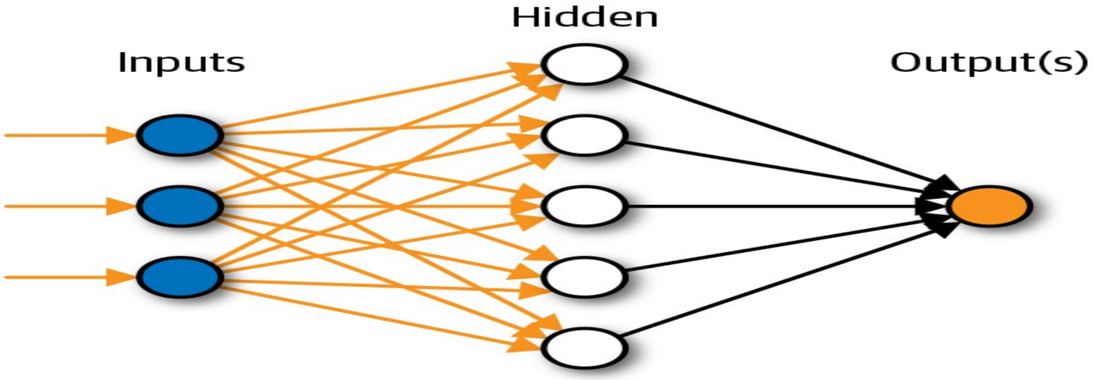


Figure 5. ANN-Classification

3.5 NSL-KDD

The most prominent competition for data mining in the world is the KDD Cup. To solve the problems that the KDD Cup 1999 dataset brought up, the NSL-KDD dataset was developed. Many academics have looked at and analyzed the NIDS problem using the NSL-KDD dataset. The dataset includes attacks of every kind. The dataset consists of 41 features that are grouped into three categories (basic feature, content-based feature, and traffic-based feature), each of which is labelled as either normal or attack. (Table 3) shows NSL- KDD attack type is furthermore clearly identified, along with a brief description of each assault type and its impact. Three groups of qualities have been established. The three kinds of features that are offered are fundamental characteristics, based content characteristics, and traffic-based characteristics.

Table 3. NSL-KDD Attack

Name of Attacks	Type of Attack
Use to root	Root access, password
Denial of service	Resources of memory
Probe	Control of security, network sniffing
Remote to local	Resources of network

For each observation, there are 42 features in the NSL KDD dataset, 34 of which are continuous variables, three of which are nominal, and four of which are binary. In contrast to the test dataset's 30, the training dataset comprises 23 traffic classes. Three groups of qualities are distinguished. The three types of features accessible for the employed NSL-KDD dataset that are shown in (Table 4), which is divided into a training set of 125,653 samples “KDDTrain+” and a testing set of 21,008 samples “KDDTest+” are basic features, based on content characteristics, and traffic-based characteristics.

Table 4: NSL-KDD Training and Testing

Attack	Record	
	Dataset for Training	Dataset for Testing
Local to remote	940	2045
User to root	48	180
Denial of service	45820	7050

Probe	11605	2024
Normal	67240	9709
Total	**125,653**	**21,008**

3.6 Training using 7-characteristics

(Table 5) shows the main 7-characteristics that are used in the training process

Table 5. 7-Characteristics Details

Characteristic's	Details
DST Host Same SRC Port Rate	Percentage of connections from the port services to the destination host.
DST Byte	Number of data bytes from destination to source
Duration Length	"Number of seconds" of the connection
Count	Number of connections to the same destination host as the current connection in the past two
Protocol Type	Types of protocol, such as TCP, UDP, ICMP
SRV Count	Number of connections to the same service as the current connection in the past two seconds
SRC Bytes	Number of data bytes from source to destination

4. Results and Discussions

In our experimentation stage, we use the most well-known classifiers Random Forest and K-nearest neighbour (KNN). Random forests can build powerful ensemble classifiers by combining weak and uncorrelated classification trees. The strength and correlation of the individual classification trees are important factors that contribute to the overall performance of the random forest. K-nearest neighbour is our second classifier for comparison purposes with the Random Forest classifier under similar conditions. Because of its ease of implementation and superior performance, the K Nearest Neighbor method is widely used in data mining and machine learning applications. Both classifiers are going to be trained and tested with similar feature selection methods and several features to identify which classifier scores better performance for our proposed model. During the experimentation, we draft one scenario described as follows in detail. This scenario's presumption is based on several research studies, such as one that manually picks just SDN features from the NSL-KDD dataset. These studies support the idea that while designing an intrusion detection system for SDN, the network's characteristics should be taken into account. NSL-KDD dataset is gathered through wired and wireless network infrastructure; hence it can only be used with IDSs that are not SDN based, according to the argument. There aren't enough SDN-based datasets available online, which is a problem for study. Since the NSL-KDD dataset contains several fundamental and flow-based SDN properties. We provide a model that uses seven features from the NSL-KDD dataset and is based on these ideas. The chapter before goes into great length describing these features. This case uses identical classifiers, and the dataset requires the same preparation steps.

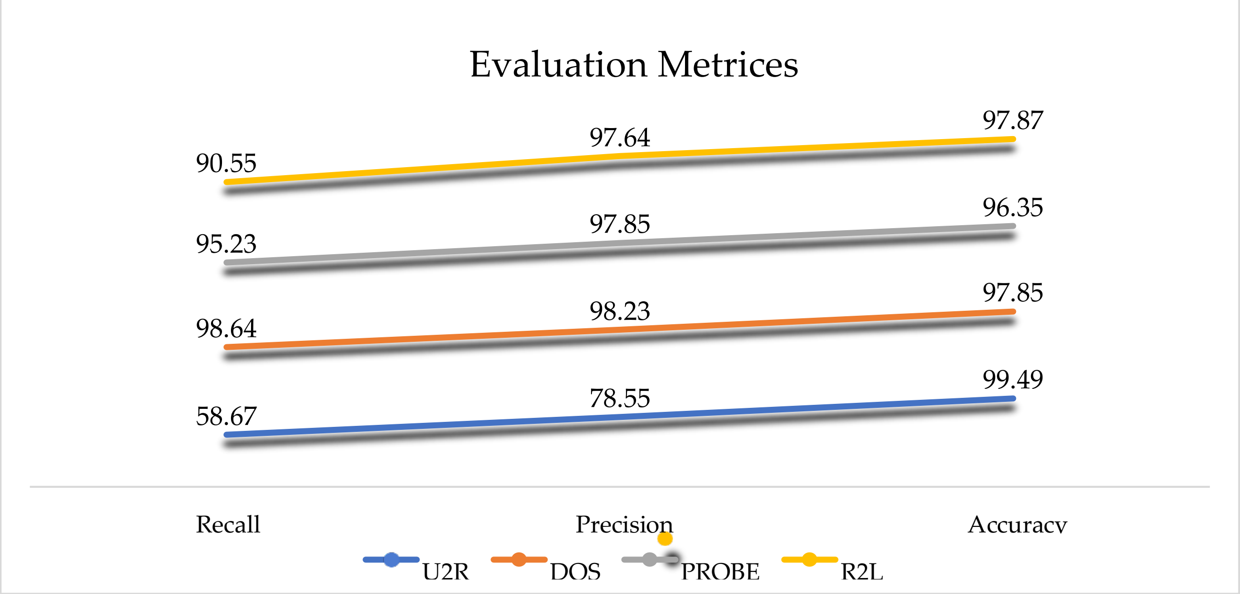
Table 6: KNN and RF On SDN Based Characteristics

Classifier	Selection	Class	Evaluation Metrics

			F1 score	Recall	Precision	Accuracy
KNN	PURPOSELY	U2R	64.54	58.67	78.55	99.49
		DOS	95.64	98.64	98.23	97.85
		PROBE	93.25	95.23	97.85	96.35
		R2L	92.36	90.55	97.64	97.87
		Average	**86.4475**	**85.7725**	**93.0675**	**97.89**
RF	PURPOSELY	U2R	85.78	75.45	90.23	99.84
		DOS	98.65	98.54	98.66	99.92
		PROBE	93.46	96.45	90.11	99.78
		R2L	76.84	81.55	93.25	98.94
		Average	**88.6825**	**87.9975**	**93.0625**	**99.62**

In this case, we tested two classifiers under comparable circumstances, meaning that both Random Forest and KNN were trained and evaluated using manually chosen features from the NSL-KDD dataset. To prepare these seven attributes for the classification stage, we built our model, a useful intrusion detection system for a specific software-defined vehicular network environment, using RF and KNN classifiers applied to these features. (Table 6) KNN was used in experiment two, to create a model with seven features. Except for accuracy, DoS attacks continue to outperform other attacks in this trial in terms of precision, recall, and f1_score.

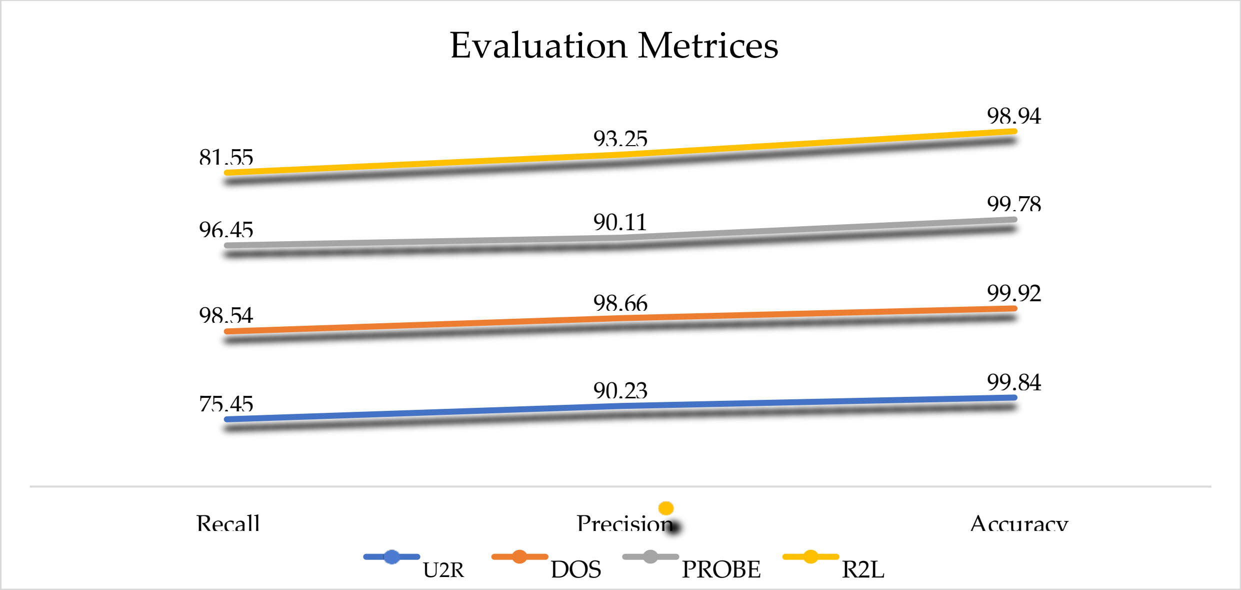
Both classifiers perform on each type of attack. From this table, DoS attack scores the highest accuracy, precision, recall and f1_score among others using seven features with RF. In general, the Random Forest classifier outperformed KNN in this scenario, achieving an average accuracy of 99.62%.



KNN Result

Figure 6. KNN Characteristics

Both classifiers KNN and RF perform on each type of attack. From this (figure 6) and (figure 7), DoS attack scores the highest accuracy, precision, recall and f1_score among others using seven features with RF. In general, the Random Forest classifier outperformed KNN in this scenario, achieving an average accuracy of 99.62%.



RF Result

Figure 7. KNN Characteristics

4.1 Comparison of Related Models

We are going to evaluate the effectiveness of our suggested models to other research in the case of SDN characteristics. Four publications that focus on creating intrusion detection systems for particular software-defined networks were found during our literature review. As a result, we adhered to their ideas and picked basic and flow-based features that are readily discovered in the NSL-KDD dataset. In four of these publications, an IDS for an SDN-specific network was presented, with a focus on the pertinent features from the NSL-KDD dataset. They used various classifier types on the NSL-KDD dataset using constrained and specified features, as shown in (Table 7). They give their model accuracy scores ranging from 75.5% to 97.76%. When we compare our proposed models with these works, both RF and KNN score high accuracy 99.62% and 97.89% respectively.

Table 7. Comparison of Different Models

Proposed Work	Classifier Used	Selected Features from NSL-KDD	Accuracy
01	XGBoost	5	95.5%
02	Deep Learning	6	75.5%
03	Multi-layer (KNN, ELM, HELM)	6	84.29%
04	K-Nearest Neighbor	7	97.76%
Our Work	Random Forest	7	99.62%

Our Work	K-Nearest Neighbor	7	97.89%

After all, comparisons and result analyses show that the suggested model performs well in terms of accuracy. The usage of a well-preprocessed dataset and the thoughtful selection of pertinent features contributed to the effectiveness of the results. Random Forest and KNN model development did not involve the use of hyperparameter tuning because we were successful in producing accurate results using the classifiers' default settings. Throughout the experimental stage, use K-nearest neighbour (KNN) and Random Forest, two of the most popular classifiers. When weak and asynchronous classification trees are combined, random forests can create robust ensemble classifiers. A key element that affects the performance of the random forest as a whole is the strength and correlation of the individual classification trees. For purposes of comparison with the Random Forest classifier under comparable circumstances, the K-nearest neighbour is our second classifier. The K Nearest Neighbor method is popular in data mining and machine learning applications because of how simple it is to use and how well it performs. To determine whether the classifier performs better for our suggested model, both classifiers will be trained and tested using comparable feature selection techniques and numbers of features. This study tested two classifiers under comparable circumstances, meaning that both RF and KNN were trained and evaluated using manually chosen characteristics from the NSL-KDD dataset. Similar to the first case, we preprocess these 7 characteristics to get them ready for classification. To create a useful intrusion detection system for a specific SDVN environment, we then applied RF and KNN classifiers to these features.

Table 7 displays the performance of both classifiers under various assault types. The data in this table reveals that experiment one's DoS assault outperforms all others in terms of accuracy, precision, recall, and f1_score when using seven features with RF. KNN was used to create a model with seven features in the experiment. In this experiment, the DoS assault still achieves higher precision, recall, and f1_score values than other attacks, except for accuracy. In the second situation, the RF classifier fared better than KNN overall, averaging 99.62% accuracy.

5. Conclusion

Vehicle-to-vehicle transmission will soon be feasible, and hackers will be able to monitor people's driving habits. However, the SDVN research field has paid little attention to security. The goal of this research paper was to outline the security issues in the area and provide an IDS based on ML methods. This research paper uses seven features from the NSL-KDD dataset to create an intrusion detection system based on machine learning for SDV networks. With 99.62% and 97.89% accuracy, respectively, the suggested system provides high accuracy. Recursive Feature Elimination with RF and KNN classifiers has been evaluated separately on the same dataset and achieves good results. RF classifier efficiency was very high when compared to previous proposed works, and even with KNN. When weak and asynchronous classification trees are combined, random forests can create robust ensemble classifiers. A key element that affects the performance of the random forest as a whole is the strength and correlation of the individual classification trees. For purposes of comparison with the Random Forest classifier under comparable circumstances, the K-nearest neighbour is our second classifier. The K Nearest Neighbor method is popular in data mining and machine learning applications because of how simple it is to use and how well it performs. To determine whether the classifier performs better for our suggested model, both classifiers will be trained and tested using comparable feature selection techniques and numbers of features. Results are described in the final Systematic Literature Review with Artificial Intelligence techniques used for the analysis of intrusion detection systems. Researchers could use more recent and current data sets in the future to assess implemented algorithms and better prepare for the most common forms of destructive incursions and attacks.

Designing and implementing effective IDS in Software Define Network-based VANETs is essential to defending against the rising danger of assaults launched from inside or external invaders of the networking environment. Most publications did not take the Software Define Network's nature into account when developing IDS. This issue is reflected in and addressed in our suggested work. The system's correctness has a substantial impact on network security. Generally speaking, strengthening the security of a Software Define Network-based VANET entails indirectly protecting human lives by shielding vehicles from attackers and other attackers. By safeguarding communication from vehicle to vehicle and from vehicle to infrastructure, this research will improve the reliability of intelligent transportation systems. A wide number of algorithms and techniques are covered under the broad topic of machine learning, which trains computers to make predictions based on data. In addition to reinforcement learning, which involves applying incentives and penalties to train models, it also comprises supervised and unsupervised learning.

In the field of artificial intelligence, two significant subfields are machine learning (ML) and deep learning (DL) (AI). They are essential for giving computers the ability to carry out tasks like pattern recognition, prediction, and data classification that traditionally require human intellect. A wide number of algorithms and techniques are covered under the broad topic of machine learning, which trains computers to make predictions based on data. In addition to reinforcement learning, which involves applying incentives and penalties to train models, it also comprises supervised and unsupervised learning. In fields like computer vision and natural language processing, it has been responsible for many of the advancements in AI during the past several years. Summarization is both ML and DL are essential elements of AI, and they work well together.

5.1 Future Work

Future work will involve training and testing our model utilizing a variety of benchmarking cyber security datasets, including UNSW-NB15, CIC-IDS2017, and CSE-CIC-IDS-2018. Additionally, SDN-specific datasets will be used to make sure that the suggested approach is never biased. To improve the effectiveness of our model, additional preprocessing and feature selection techniques will be investigated. The effects of each strategy on classification will be compared with each other. Due to the centralized structure of the SDN controller, different assault types affect the brain of the VANET differently. Therefore, we will concentrate on a particular DoS threat that occurs in software-defined vehicular networks. Future testing and comparisons of various classifier types with our model are planned. To find zero-day attacks, hybrid and DL techniques will be put to the test.

Funding: No funding was received.

Data Availability Statement: The authors declare that all data supporting the findings of this study are available within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weber, P., K.V. Carl, and O. Hinz, Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. Management Review Quarterly, 2023.
2. Jan, Z., et al., Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. Expert Systems with Applications, 2023. **216**: p. 119456.
3. Ali, T.E., Y.-W. Chong, and S. Manickam Machine Learning Techniques to Detect a DDoS Attack in SDN: A Systematic Review. Applied Sciences, 2023. **13**, DOI: 10.3390/app13053183.
4. Lewowski, T. and L. Madeyski, Code Smells Detection Using Artificial Intelligence Techniques: A Business-Driven Systematic Review, in Developments in Information & Knowledge Management for Business Applications : Volume 3, N. Kryvinska and A. Poniszewska-Marańda, Editors. 2022, Springer International Publishing: Cham. p. 285-319.
5. Wiafe, I., et al., Artificial Intelligence for Cybersecurity: A Systematic Mapping of Literature. IEEE Access, 2020. **8**: p. 146598-146612.
6. Khan, H.U., et al., Transforming the Capabilities of Artificial Intelligence in GCC Financial Sector: A Systematic Literature Review. Wireless Communications and Mobile Computing, 2022. **2022**: p. 8725767.
7. Gamage, S.H.P.W., J.R. Ayres, and M.B. Behrend, A systematic review on trends in using Moodle for teaching and learning. International Journal of STEM Education, 2022. **9**(1): p. 9.
8. Frizzell, T.O., et al., Artificial intelligence in brain MRI analysis of Alzheimer's disease over the past 12 years: A systematic review. Ageing Research Reviews, 2022. **77**: p. 101614.
9. Chaki, J., et al., Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University - Computer and Information Sciences, 2022. **34**(6, Part B): p. 3204-3225.
10. Yeng, P.K., et al. Data-Driven and Artificial Intelligence (AI) Approach for Modelling and Analyzing Healthcare Security Practice: A Systematic Review. in Intelligent Systems and Applications. 2021. Cham: Springer International Publishing.
11. Sharma, V., et al., Video Processing Using Deep Learning Techniques: A Systematic Literature Review. IEEE Access, 2021. **9**: p. 139489-139507.
12. Abdullahi, M., et al. Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: A Systematic Literature Review. Electronics, 2022. **11**, DOI: 10.3390/electronics11020198.
13. Alghawazi, M., D. Alghazzawi, and S. Alarifi Detection of SQL Injection Attack Using Machine Learning Techniques: A Systematic Literature Review. Journal of Cybersecurity and Privacy, 2022. **2**, 764-777 DOI: 10.3390/jcp2040039.
14. Nassif, A.B., et al., Machine Learning for Cloud Security: A Systematic Review. IEEE Access, 2021. **9**: p. 20717-20735.
15. Gary, S.C., et al., Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open, 2021. **11**(7): p. e048008.
16. Friedrich, S., et al., Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations. European Heart Journal - Digital Health, 2021. **2**(3): p. 424-436.
17. Faruk, M.J.H., et al. Malware Detection and Prevention using Artificial Intelligence Techniques. in 2021 IEEE International Conference on Big Data (Big Data). 2021.
18. Davidson, L. and M.R. Boland, Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. Briefings in Bioinformatics, 2021. **22**(5): p. bbaa369.
19. Ahmad, Z., et al., Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 2021. **32**(1): p. e4150.
20. Amarudin, R. Ferdiana, and Widyawan. A Systematic Literature Review of Intrusion Detection System for Network Security: Research Trends, Datasets and Methods. in 2020 4th International Conference on Informatics and Computational Sciences (ICICoS). 2020.
21. Wang, Y., et al., A systematic review of fuzzing based on machine learning techniques. PLOS ONE, 2020. **15**(8): p. e0237749.

22. Wen, J., et al., Systematic literature review of machine learning based software development effort estimation models. Information and Software Technology, 2012. **54**(1): p. 41-59.

23. Aiyanyo, I.D., H. Samuel, and H. Lim A Systematic Review of Defensive and Offensive Cybersecurity with Machine Learning. Applied Sciences, 2020. **10**, DOI: 10.3390/app10175811.

24. Song, D.Y., et al., The Use of Artificial Intelligence in Screening and Diagnosis of Autism Spectrum Disorder: A Literature Review. Soa Chongsonyon Chongsin Uihak, 2019. **30**(4): p. 145-152.