
Article:
Veena, K., Meena, K., Teekaraman, Y. et al. (2 more authors) (2022) C SVM classification and KNN techniques for cyber crime detection. Wireless Communications and Mobile Computing, 2022. 3640017. ISSN 1530-8669

https://doi.org/10.1155/2022/3640017

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3640017, 9 pages
https://doi.org/10.1155/2022/3640017

Research Article

C SVM Classification and KNN Techniques for Cyber Crime Detection

K. Veena,[1] K. Meena,[2] Yuvaraja Teekaraman,[3] Ramya Kuppusamy,[4] and Arun Radhakrishnan[5]

[1]*Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, 600119, Chennai, India*
[2]*Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, India 500043*
[3]*Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S1 3JD, UK*
[4]*Department of Electrical and Electronics Engineering, Sri Sairam College of Engineering, 562106, Bangalore City, India*
[5]*Faculty of Electrical & Computer Engineering, Jimma Institute of Technology, Jimma University, Ethiopia*

Correspondence should be addressed to Yuvaraja Teekaraman; yuvarajastr@ieee.org
and Arun Radhakrishnan; arun.radhakriahnan@ju.edu.et

Received 25 September 2021; Revised 24 November 2021; Accepted 3 December 2021; Published 17 January 2022

Academic Editor: Deepak Kumar Jain

In the digital age, cybercrime is spreading its root widely. Internet evolution has turned out to a boon as well as curse for those confronting the issues of privacy, national security, social decency, IP rights, child protection, fighting, detecting, and prosecuting cybercrime. Hence, there arises a need to detect the cybercriminal. Cybercrime identification utilizes dataset that is taken from CBS open dataset. For identifying the cybercriminal, support vector machine (SVM) in the *C* SVM classification and *K*-nearest neighbor (KNN) models is utilized for determining the cybercrime information. The evaluation of the performance is done taking the following metrics into consideration: true positive, false positive, true negative and false negative, false alarm rate, detection rate, accuracy, recall, precision, specificity, sensitivity, classification rate, and Fowlkes-Mallows Scores. Expectation maximization (EM) calculation is utilized for evaluating the presentation of the Gaussian mixture model. The performance of classifier's presentation is also done. Accuracy is accomplished in the event of grouping by means of SVM classifier as 89% in the supervised method.

1. Introduction

Cybercrime involves attempting a criminal offense via computer and a network wherein the computer can act as a tool, goal, or both of them. Many unauthorized computer enabled activities take place via global electronic networks. The research categorizes cybercrime into the following:

(i) A criminal activity involving computer for the execution of the crime

(ii) A criminal activity involving computer for crime related information storage and not necessarily for the execution of the crime

1.1. Necessity of Detection of Cybercrime. Cybercrime has spread its roots far on a global scale and portrays a great danger towards occurrence of either a criminal or a terrorist activity. These threats can affect both internal and external security (military) without reacting to single authority policing methodologies. There is loss of both personal and financial data if the cybercrime goes undetected. There have already been attacks on information infrastructure and Internet services. Online fraud and hacker attacks are only two instances of computer-related crimes that occur on a daily basis. Cybercrime is said to have caused huge financial damage. Malicious software costs up to USD 17 billion in damages in 2003 alone. According to some estimates,

cybercrime income surpassed USD 100 billion in 2007, surpassing the illegal drug trade for the first time. In the United States, about 60% of firms say that cybercrime is more costly than physical crime. These figures show how critical it is to safeguard information infrastructures.

1.2. Application Methods. Cybercrime detection aids in research, analysis, and investigation of cybercrime and other development areas. It is very well helpful in the cybercrime forensic laboratory as well as most importantly in the network monitoring center. Big enterprises employ cybercrime detection for identifying intruder that gains access to sensitive information. The recommended system in the research can be adopted if user profile info is at hand with no means to negatively that affect the emotions of the victim.

1.3. Issues and Challenges in Cybercrime Detection. Today, the entire world is confronting some or the other crime, and the increasing population has aggravated the issue all the more. The launch and implementation of ICTs in every work culture have reformed the society into a modernized and tech savvy one. This information society is a pool of huge opportunities. Democracy can be supported when there is unrestricted accessing of information since the information flow is merely not in the state authorities control (this could be witnessed in Eastern Europe and North Africa). There is immense advancement and reformation in daily lives of the individual's with the ongoing technical expansion. The launch and integration of ICTs in every domain such as online banking, shopping, mobile data services, and VoIP telephony have made the living easier and comfortable, but development of such information society is simultaneously followed by grave threats and malicious acts. Day to day services like the water and electricity supply, mobiles, traffic control cars, elevators, and air conditioning rely upon ICTs for their smooth conduct. Frequently, from time to time, Internet Services and Information Infrastructures are confronted a lot with several malicious attacks which badly affect the society. Online fraud and hacking attacks are rare computer-related crimes taking place recurrently amongst the extensive cases in cybercrime. In 2003, there was a huge financial loss of USD 17 Billion due to these kinds of cybercrimes as malicious software mentioned in the article.

It is estimated from the revenue that for the first time in the year 2007, there was a surge of around USD 100 billion in the cybercrime, 21 exceeding the illegal trade in drugs. Around 60% of businesses in the USA have admitted that more than the physical crime, cybercrime has incurred huge cost to them. These observations elucidate the urgency of safeguarding the information infrastructures. Maximum of these cybercrime attacks by the hackers targets on more computer infrastructure rather than the critical infrastructure. Though in 2010, "Stuxnet" malicious software caused threats and attacks targeting on the critical infrastructure. It had 4000 plus functions emphasizing on software responsible for computer systems execution for controlling critical infrastructure. Cybercrime has posed innumerable challenges and drawbacks that need to be confronted with urgency.

In this investigation, the clients are classified as genuine data or crime data utilizing the AI procedures. It additionally prescribes a technique to recognize the cybercriminal by assessing the input obtained from various clients. Through grouping, authentic information (cluster 0) is taken out keeping just the unauthentic information (cluster 1). Using the genuine information, false positive is considered as the crime information and utilizing the crime information, and genuine negative is considered as the authenticated information which is then added to cluster 0.Thereafter, there is examination of criminal information by utilizing numerous classes, for example, none, soft, and hard for distinguishing the crook. Different clustering techniques are utilized for different clients with fluctuated characters/credits. At that point, there is investigation of information for different limit esteems. The client profile of the offender relating to the cybercrime is evaluated for different assessments. In supervised learning method, a model is formed for predicting on the basis of evidence under uncertainty. Through adaptive algorithms calculations, information examples can be resolved, causing the PC to take in or fathom from the perceptions. With more number of perceptions, the prescient execution of the PC likewise improves. A regulated learning calculation thinks about the accessible gathering of info information and yield in type of known reactions to the information. The calculation at that point prepares a model for delivering sensible expectations in light of new information.

According to "Social media misuse and Indian cyber law," misusing Twitter is deemed cybercrime. [3 Pavan Duggal, New Delhi [1]] is the author. "Under Indian cyber law, even users can be prosecuted for posting content, thus it is best to take caution before uploading anything,". The relationship between such policies and free expression is discussed from an Indian perspective. "All people shall enjoy the right to freedom of speech and expression," according to Article 19 (1)(a) of the Indian Constitution. The people of India declared their solemn resolve to secure the liberty of thought and expression to all of its citizens in the Preamble to the Constitution of India. "Freedom of speech" is a principle that protects an individual's or a group's right to express their thoughts and beliefs without fear of retaliation, censorship, or legal consequences. The word "freedom of speech" is sometimes used interchangeably; although, it refers to any act of seeking, receiving, or transmitting information or ideas, independent of media.

Cybercrime could be identified using artificial intelligence, such as machine learning or deep learning, according to studies. In this study, a user who is not a criminal is referred to as Genuine. This research is for the greater good of society. Machine learning can be used to detect troll farms, according to the authors [2]. According to the author [Anqi Liu et al.], if the troll factory is not found, it may be a real comment.

2. Literature Survey

This research work includes the background information regarding the identification of Cybercriminal, IDS, and the

appropriate protocols and tools incorporated in IDS. In addition, it illustrates the information regarding the types, impact of network attacks, and the technique of Honeypot.

Akshay et al. [3] discussed that IDS is responsible for monitoring the overall network and the system activities against any fraudulent activities thereby intimating the same to the management station. The rate of false alarms in the existing IDS is quiet high which can be controlled through the honeypots for making the network more reliable and secure. Honeypots are quite simple in usage and can acquire the desired information. Most of the corporate organizations incorporate honeypots for network security. There was also a description regarding building and implementation of a handset detector and score normalization for enhancing the verification performance. Eventually, there was a discussion regarding the representative performance standards and system behavior experiments on NIST SRE corpora. Roshni et.al [4] have inspected that there was medical and legal implementations of deception detection. Two classes of feature have been fetched by employing the classifier yielding an efficiency of 83%.

Alsafi et.al [5] recommended a robust model of integrated IDPS, that merges IDS as well as IPS into one model. Besides, it merges the AD technique and the SD technique which aids in detection of numerous attacks and halting them using the technique of IPS. The research put forwards several IDS techniques employed for combating malicious attacks in Cloud computing environment.

Baykara et.al [6] discussed and recommended a honeypot-based approach for ID/PS. The honeypot approach was integrated with IDS for evaluating real-time data and for carrying out effective operations. In addition, by linking the merits of high and low-interaction honeypots, a hitech hybrid honeypot system was produced. Implementation of honeypots was observed on corporate networks for minimizing the configuration cost, maintenance, and management, thereby making use of virtualization technologies. The recommended and implemented honeypot-based IDPS was capable of visually depicting network traffic on servers with real-time animation.

Data mining resembled an essential field of computer science significant for identifying patterns from voluminous data. Essentially, the approach of data mining implies fetching knowledgeable data from databases and identifying important relationships between voluminous data thus enabling finding of anomalous behavior. Data mining techniques such as clustering, classification, association rule, and mining, hold the utmost significant in IDS and are employed for evaluating and observing the network data and provide information regarding the intrusion. For implementing any of the data mining techniques for intrusion detection, the assembled network data must first undergo preprocessing and thereafter must be transformed into an appropriate form for the process of mining. The reformatted data is then employed for developing a classification.

Khraisat et al. [7] proposed SIDS or knowledge-based detection or misuse detection that relies upon pattern matching approaches for determining a known attack. It made use of matching methods for determining any prior intrusion. That is, on the match of an intrusion signature with a previous existing intrusion signature existing in the signature database, an alarm signal is raised. In SIDS, there was inspection of host's logs for determining sequences of actions/commands that have been detected as malware. Modi et.al [8] represented SIDS as knowledge-based detection or misuse detection. Deka et.al [9] proposed that firewall restricted access amidst networks for intrusion prevention. There was classification of IDS on the basis of network/host detection and the employed detection method.

NIDS helps in monitoring network traffic located at points within the network. NIDS analyses the overall passing traffic in the subnet and matches it with the one forwarded on the subnets to the assembly of known attacks. If there is sensing of any abnormal behavior or an attack is noticed, then the alert is reported to the administrator.

There are two phases of AIDS development: first is the training phase and second is the testing phase. The training phase involves normal traffic profile for comprehending a model of normal behavior, whereas the testing phase involves a new dataset for building the system's capacity and generalizing prior hidden intrusions. Butun et.al [10] proposed that depending on the training methods, AIDS can be classified into knowledge-based, statistical-based, and ML-based. Moore [11] recommended the two algorithms to improvise the system, namely, HRI and HRO. This offered increased flexibility and usability. Customization of each module could be performed on a need basis.

In the paper, Veena and Meena [12] presented a method to analyze the various identities of a user and thus determine if any synthetic identity theft has been committed using three type of data, i.e., input dataset (X), normal dataset (Y), and target dataset (Z). The author Meena K. and Veena [13] worked on cybercrime detection by plotting the noise and the original artifacted EEG signals. The primary motivation behind the research was to explore different methods that could be used for lie or deception analysis. Later, the influence of statistical features for the discrimination of thinking patterns from the normal signals was elucidated. In the paper, Veena and Meena [14], the author, used four different techniques in the determination of cybercrime. Firstly, the detection of synthetic identity theft was checked. Secondly, the intrusion detection was checked using the honeypot security mechanism. Thirdly, the detection was further strengthened using the lie detection technique where the false speech of a person was determined. Finally, by analyzing the user profile, the detection of cyber crime was done using the clustering techniques. The authors Veena and Meena [15] researched on the cyber warfare which was the current greatest challenge to Network security. The greatest challenge in network was addressing the security issues on the server side. The paper recommended to improve the security performance to protect the network from intruders.

The study [Deylami et al. [16]] proposes a cybercrime detection model that uses support vector machines to classify social network (Facebook) datasets (SVMs). The three types of classification algorithms SVMs, AdaBoostM1, and NaiveBayes were utilized to find a high percentage of classification accuracy. Finally, because SVMs apply numerous

kernel functions to increase classification accuracy on the Facebook dataset, they were considered to be the most successful classification algorithm. The classifiers were also tested using the Weka 3.7.4 software on the Facebook dataset. In general, ethical and moral requirements are growing more important. As a result, there is a greater need to detect cybercrime. The emergence of technology has led in unsustainable societal development [17–21].

3. Research Methodology

C SVM: *C* is a regularization parameter that controls the trade-off between the achieving a low training error and a low testing error that is the ability to generalize your classifier to unseen data. Support vector machine: this exploration segment utilizes the cybercrime identification model through SVM for ordering the informational index recovered from CBS information detail line: https://www.cbs.nl/en-gb/our-administrations/open-information. In the SVM classifier training data, SVM classifier utilizes cybercrime discovery datasets through ML devices. In the research work, various clients are ordered into crime user or genuine user.

3.1. SVM Algorithm. Input: train dataset-train, test dataset–test.

Output: cyber dangers classification.
Stage 1: read train dataset.
Stage 2: apply SVM calculation.
Stage 3: generate SVM model for portion work.
Stage 4: read test dataset.
Stage 5: for every characteristic in test data.
Stage 6: extract all the highlights.
Stage 7: apply SVM calculation.
Stage 8: return result of test data.
Stage 9: end.

Cybercrime is identified using data from the CBS open dataset, http://dataworld.com/, and http://kaggle.com/. True positive, false positive, true negative and false negative, false alarm rate, detection rate, accuracy, recall, precision, specificity, sensitivity, classification rate, and Fowlkes-Mallows Scores are all used to determine cybercrime information for identifying cybercriminals in the *C* SVM classification and *K*-nearest neighbor (KNN) models. The presentation of the Gaussian mixture model is assessed using the expectation maximization (EM) approach. There is also a presentation of the classifier. The supervised technique achieves 89 percent accuracy when employing an SVM classifier to group data.

3.2. KNN Classifier

3.2.1. Performance of Different KNN Classification Learners. The KNN algorithm depicts a nonparametric technique employed for performing classification and regression. In either situation, the input comprises of the *k*-closest training exemplar in the characteristic space. KNN algorithm in KNN regression is employed to estimate continuous variables. There is another algorithm that utilizes a weighted average of the k nearest the labeled examples are ordered by increasing distance. Neighbors go through the inverse of their distance. Following is the working of the algorithm. Calculate the Euclidean distance from the query example to the labeled examples. Table 1 displays the KNN classifier with the several classification learners. The numbers of attributes used were 25 implemented for 1000 users. The accuracy for medium KNN and cubic KNN with 88.7% and for optimizable KNN shown in Figure 1 was 88.6%.

3.3. Comparison of SVM and KNN Classifier. Data classification in KNN is on the basis of distance metric but in case of SVM, right training phase is required. SVM is of optimal type; hence, it is ensured that the separated data would be segregated optimally shown in Figure 2. Usually, KNN is employed as multiclass classifiers; on the other hand, the SVM segregates binary data that being a part of either one class. The approach of one-vs.-one and one-vs.-all is adopted for a multiclass SVM. In one-vs.-one concept, $n \times (n-1)/2$ SVMs are trained, that is one SVM for each pair of classes. The pattern is fed that is not known to the entity, and the final decision on the data type is made by majority output from the Overall SVM output. This approach is mainly employed for multiclass classification. Table 2 presents the accuracy of classifiers.

Apparently, SVMs appear to be computationally tough, and once the data is trained, the model can be employed for predicting classes even though new unlabeled data are encountered. But in case of KNN, every time a new unlabeled data is encountered, the distance metric is computed.

So, its required to fix only the *K* parameter and choose distance metric that is apt for classification in KNN where else in SVMs its required to choose the *R* parameter (regularization term) along with the kernel parameters in case the classes are linearly inseparable. When comparing the accuracy of both the classifiers, SVMs depict improvised accuracy in contrast to KNN. Table 3 shows the Accuracy for the different classification learners.

Cyber security datasets: many of the researchers form a repository of various kinds of data from their study point of view as well as providing this collected data for community repositories. The present section describes the prevailing security based datasets through the techniques of ML and artificial intelligent research.

Ecml-Pkdd 2007 dataset: The ECML-PKDD 2007 dataset was built for the European Conference on ML and Knowledge Discovery in 2007. A data mining competition referred to as ECML/PKDD Discovery Challenge took place in conjunction with the 18th European Conference on Machine Learning (ECML). Table 4 depicts the attributes of ECML/PKDD 2007. Description of dataset is using extensible markup language (XML). All samples are expressed using a unique id, and they comprise of the three primary parts namely: context, class, and query.

Collection of cyber crime dataset: for prediction of cybercrime in banking sector, a variety of cybercrime data is gathered by analyzing the crime pattern. The data is gathered from different online sources such as the news feeds, blogs, articles, and police department sites. The cybercrime data that is collected is then stored in the crime database for data handling.

Table 1: Performance of different KNN classification learners.

Classification learner	Total features = 25, predictors = 1, PCA is on						
	Fine KNN	Medium KNN	Coarse KNN	Cosine KNN	Cubic KNN	Weighted KNN	Optimizable KNN
Accuracy	86.2%	88.7%	87.5%	51.8%	88.7%	87.5%	87.3%
Total misclassifications	138	113	125	482	113	125	127
Prediction speed (obs/sec)	~16000	~11000	~14000	~16000	~14000	~11000	~9500
Training time (sec)	0.80359	0.58836	0.5122	0.46068	0.43864	0.5543	85.099
Number of neighbors	1	10	100	10	10	10	3
Distance metrics		Euclidean		Cosine	Min Kowski (cubic)	Euclidean	Mahalanobis
Distance weight	Equal	Equal	Equal	Equal	Squared inverse	Squared inverse	Squared inverse
Standardize data	True	True	True	True	True	True	True
Optimizer options	Hyperparameter options disabled						Hyperparameter options enabled
Feature selection	All features used in the model before PCA						
Misclassification costs	Cost matrix: Default						
For PCA	For PCA training, 2 of 1000 observations were ignored because they contain Infs or NaNs. PCA is keeping enough components to explain 95% variance. After training, 1 components were kept. Explained variance per component (in order): 99.9%						

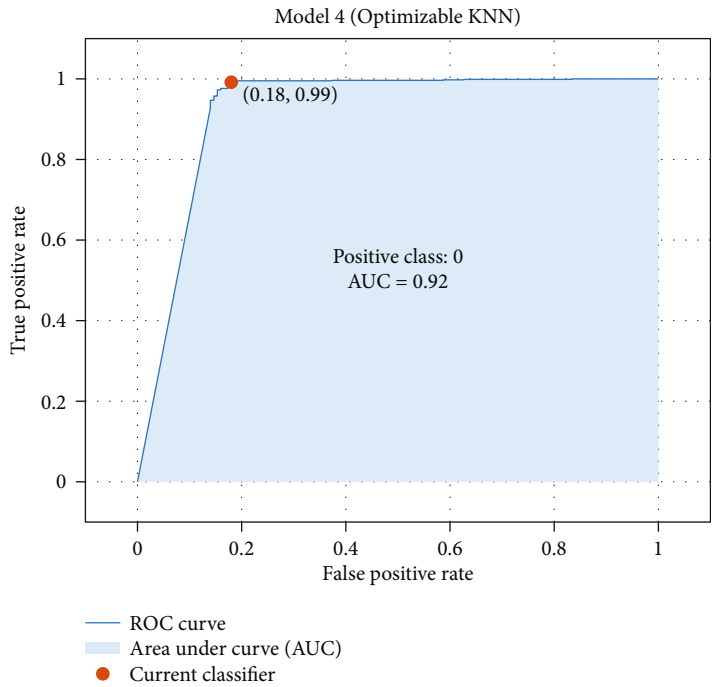

Figure 1: ROC curve for the optimizable KNN classifier.

Preprocessing of cyber crime dataset: the cybercrime dataset stored that is stored in the crime database must undergo preprocessing before the data mining techniques are applied over it. By performing preprocessing missing values, noisy data, etc. can be worked upon.

Data mining techniques: data mining techniques and algorithms are employed overpreprocessed data for identifying any fraud using knowledge innovation from abrupt patterns thereby fighting against cyber credit card fraud. Data mining is an effective tool which helps in resolving the issues

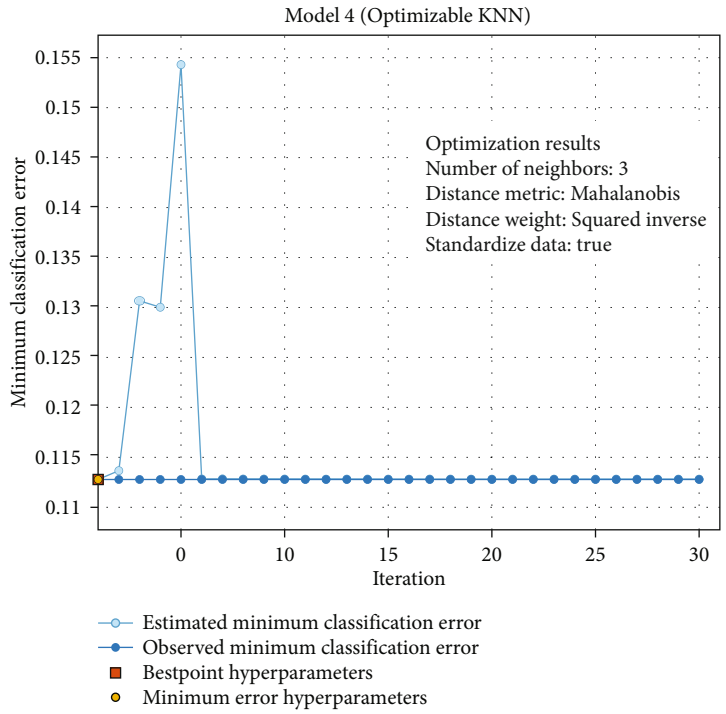

Figure 2: The minimum classification error plot.

Table 2: Accuracy of classifiers.

Classifier	Training set	Test set	Accuracy rate (%)
SVM	8000	2000	98.8
KNN	8000	2000	96.47

confronted in banking sector by identifying patterns, links, and relationships that remain hidden in the business information gathered from the crime databases.

4. Evaluation of Performance of Classifier

Proposed approach: application of honeypot and preventing malfunctioning in wireless networks is carried out in the following manner.

1st phase: it involves general functionality and gathering of information related to simulator, basic honeypot functions, IDS, etc. Moreover, a comparison layout is carried out in this phase.

2nd phase: this phase involves building a network with IDS environment in NS2 simulator. It draws differences pertaining to the wireless network's performance.

3rd phase: the recommended scheme is employed for honeypots for preventing malfunctioning and accomplishing better monitoring methods. In addition, a robust honeypot technique is incorporated in coordination with IDS for attaining high security in the wireless network. Just after the firewall, honeypot is placed, and the intrusion detection system intensely synchronizes with honeypot and the snooping agents too. The snooping agents tend to be an integral part of maximum network activities as it aids in the process of filtering and tagging. These agents are built using interrupt type including traffic pattern information available for communication. Snooping agent are implemented in a style of instruction code that gets synchronized with honeypot. The agents are placed in such a manner that the filtering process becomes robust with overall monitoring of suspicious and abrupt traffic. Monitoring is carried out at packet and pattern level of the traffic. Simulation helps in monitoring and filtering of traffic for identifying any intrusion in the network.

4th phase: the last phase involves testing of the recommended system with standardized factors such as throughput, PDR, delay, and jitter in various wireless networks classes like the constant bit rate and variable bit rate.

For evaluating the efficiency of IDS, various kind of metrics have been formulated are classified in three classes, and these are threshold, ranking, and probability metrics [22]. Threshold metrics comprise of CR, *F*-measure, CPE, cost per example, etc. The prediction is either above or below the threshold and need not necessarily be close to a threshold, and the threshold metric value ranges between 0 and 1. Ranking metrics comprise of FPR, PR-precision, DR-detection rate, CID-intrusion detection capability, and AUC-area under ROC curve, and the value of ranking metrics range between 0 and 1. These metrics rely upon the ordering of the cases rather on the actual predicted values. It does not make any difference until the ordering is preserved. These metrics evaluate the proper ordering of the attack instances prior to the normal instances and areobserved as an outline of model performance pertaining to overall thresholds. Probability metrics comprise of RMSE-root mean square error whose values lie between 0 and 1. The metric decreases when for each attack class, it is

Table 3: Accuracy for the different classification learners.

Classification learner	Accuracy	Total misclassifications	Prediction speed	Training time
Tree	97.5%	25	~14000obs/sec	8.0344 sec
SVM: linear SVM	98.6%	14	~5600 obs/sec	21.179 sec
KNN: fine tree	97.2%	28	~20000obs/sec	59.723 sec
Ensemble: boosted tree	97.3%	27	~5900 obs/sec	48.145 sec

Table 4: Features of ECML/PKDD dataset.

Features	Training set	Test set
Total request	50.116	70, 143
Valid request	35,006 (70%)	42,006 (60%)
Attacks	15,110 (30%)	28,137 (40%)
Cross-site scripting	12%	11%
SQL injection	17%	18%
LDAP injection	15%	16%
XPATH injection	15%	16%

predicted value matches with the true conditional probability of that class being a normal class.

There is a comparison of different IDS with well-known metric like AUC. The CID value ranges from 0 to 1. Performance of IDS is directly proportional to the CID value. That is, high CID value implies high IDS performance. Usually, the confusion matrix helps in the computation of these metrics. The confusion matrix greatly aids in the representation of classification output of the IDS.

Metrics from confusion matrix: though the confusion matrix is highly helpful in representing the classification, it is not adequate and significant enough for comparing the IDSs. In order to combat this issue, several performance metrics are described with respect to the confusion matrix variables. These metrics generate numeric values which are simply comparable and are described as given below.

CR: it signifies the ratio of rightly classified instances and the total no. of instances.

$$\mathrm{CR} = \frac{\text{Correctly classified instances}}{\text{Total number of instances}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \quad (1)$$

DR: DR is measured as the ratio between the no. of accurately detected attacks and the total no. of attacks.

$$\mathrm{DR} = \frac{\text{Correctly detected attacks}}{\text{Total number of attacks}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \quad (2)$$

FPR: FPR is measured as the ratio between the total no. of normal instances which are identified as an attack and the total no. of normal instances.

$$\mathrm{FPR} = \frac{\text{Number of normal instances detected attacks}}{\text{Total n umber of attacks}}. \quad (3)$$

PR: PR is the fraction of positively predicted data instances that are positive in actuality.

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \quad (4)$$

Recall: the recall metric computes the percentage (a missing part from the precision) from the real attack instances included by the classifier. Apparently, it is very well required that the classifier has a high recall value. The recall metric is similar to the DR metric.

There is a ceaseless increase in the level of network-based attacks since the intruders present a fake identity and thus escape from getting trapped or caught. The issue of detecting such cyber criminals has become extremely challenging and complex since the intruders carries a series of attacks via so called the intermediary hosts or referred to as the stepping stones. This implies that the attack from a particular source may traverse through a series of intermediary hosts before actually attacking the victim. Thus, the victim remains clueless about the origin source of the attack rather considers the last intermediary hosts (or stepping stone) as the source from where the attack must have launched. Hence, it becomes highly crucial that an effective and reliable technique is built for the detection of the stepping stone attack to capture the attackers.

Various techniques have been proposed towards resolving the issue of stepping stone attacks. If the information flowing in the network connection is in form of plain text (i.e., not encrypted), then it can be useful for tracing out the source of the attack. By comparing the thumbprints, it can be analyzed if the two connections hold similar text which then can be used to conclude that they belong to the same connection chain. Wang et al. induced identifiable watermarks into the unencrypted traffic, echoing back to the attacker in order to trace back the attack. The content-based approaches are unsuccessful in case of stepping stone connections that are encrypted. Also, if all the packets are padded to the same size, the packet-size-based approaches are ineffective.

For avoiding any detection, an attacker may use an encrypted link via multiple stepping stones for attacking the victim. The encrypted packet that induces an attack may depict varying content between the stepping stones while padding to the same size. In that sense, the packet contents/size cannot be used for the identification or tracing of the attacker. Various approaches are recommended that incorporates just the timing information; although, there are chances that the attacker may avoid being identified by disturbing the timing information. That is before the packet leaves from a stepping stone the attacker may initiate a random delay or induce superfluous packets in form of chaff into the original attack on an intermediary host. The

Table 5: Previous schemes' assumptions.

Scheme	1	2	3	4
On/off	Yes	No	—	Yes
Deviation	No	No	—	Yes
IPD	No	No	—	Yes
Watermark	No	Yes	—	Yes
State-space	Yes	No	—	Yes
Multiscale	Yes	Yes	Yes	Yes
Detect-attacks and detect-attacks-Chaf	Yes	Yes	Yes	Yes

research emphasizes on two situations concerning the issue of stepping stone attribution:

Scenario 1. Producing just delay perturbation without any chaff perturbation.

Scenario 2. Both delay and chaff perturbations are present at the same time.

It is not just the attacker who can produce delay and chaff perturbations rather such kind of perturbations can be introduced by the network too. There are chances that the packets face propagation delay when they traverse through the network. In an unknown network, it is quite possible that the attack connection is caught at some place along with several other connections, unable to distinguish it from the attack flow. Resultant: these normal connections can cause chaff to the attack flow. Table 5 illustrates assumptions of previous scheme. There may be two sources for the delay and chaff perturbations produced in the attack flow: first being the attacker and second being the network itself.

5. Conclusion

In the business world, artificial intelligence (AI) and machine learning (ML) are frequently employed. In fact, by 2020, 37% of businesses and organizations will have incorporated AI into their systems and operations in some form. Using solutions offered by these technologies, businesses can better predict their customers' purchase behavior, resulting in increased income. With the use of ML and AI-powered tools, some organizations, such as Amazon, which was valued at a trillion dollars in 2018, were able to establish highly profitable operations. In this research work, the user profile is used to detect the cybercrime. Here, the fingerprints, gestures, and other objects are not used.

This research work elaborates crime analysis, supervised learning methods incorporating SVM, and KNN classifier and their comparison. The rise in usage of internet has led to the rise in the performance of criminal activities. It is a boon as well as a curse to the mankind involving issues of privacy, national security, social decency, IP rights, child protection, fighting, detecting, and prosecuting cybercrime. The evaluation of the performance was performed taking the following metrics into consideration: true positive, false positive, true negative and false negative, false alarm rate, detection rate, accuracy, recall, precision, specificity, sensitivity, classification rate, and Fowlkes Mallows Scores. The performance analysis of SVM classifier as well as performance analysis of identification of cybercriminal by using cluster computing techniques was done. It discussed the purpose for choosing SVM, methodology employed in different techniques, evaluation of algorithm, time complexity, and performance analysis.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. Duggal, *Under Indian cyber law, even users can be prosecuted for posting content so it is best to exercise caution before uploading anything*, Rand Corporation, New Delhi, 2015.

[2] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo, "A holistic system for troll detection on Twitter," *Computers in Human Behavior*, vol. 89, pp. 258–268, 2018.

[3] A. A. Somwanshi and S. A. Joshi, "Implementation of honeypots for server security," *International Research Journal of Engineering and Technology (IRJET)*, vol. 3, pp. 285–288, 2016.

[4] R. D. Tale and B. P. Harne, "Deception detection method using independent component analysis of EEG signals," *International Journal of Advanced Research in Electronics and Communication Engineering*, vol. 4, pp. 1293–1298, 2015.

[5] H. M. Alsafi, W. M. Abduallah, and A.-S. K. Pathan, *IDPS: an integrated intrusion handling model for cloud*, Networking and International Architecture, 2019.

[6] M. Baykara and R. Das, "A novel honeypot based security approach for real-time intrusion detection and prevention systems," *Journal of Information Security and Applications*, vol. 41, pp. 103–116, 2018.

[7] A. Khraisat, I. Gondal, and P. Vamplew, "An anomaly intrusion detection system using C5 decision tree classifier," in *Trends and applications in knowledge discovery and data mining*, pp. 149–155, Springer, 2019.

[8] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in Cloud," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 42–57, 2013.

[9] R. K. Deka, K. P. Kalita, D. K. Bhattacharya, and J. K. Kalita, "Network defense: approaches, methods and techniques," *Journal of Network and Computer Applications*, vol. 57, pp. 71–84, 2015.

[10] I. Butun, S. D. Morgera, and R. Sankar, "A survey of intrusion detection systems in wireless sensor networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 266–282, 2014.

[11] C. Moore, "Detecting ransom ware with honeypot techniques," in *2016 Cybersecurity and Cyberforensics Conference (CCC)*, pp. 77–81, Amman, Jordan, 2016.

[12] K. Meena and K. Veena, "Lie detection system using input voice signal," *International Journal of Pure and Applied Mathematics*, vol. 117, no. 8, pp. 121–125, 2017.

[13] K. Veena and K. Meena, "Determination of performance to verify the synthetic identity theft by training the neural networks," in *Proceedings of IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, pp. 246–250, Chennai, India, 2017.

[14] K. Veena and K. Meena, "Performance evaluation of cyber criminal detection techniques," *International Journal of Simulation–Systems, Science & Technology*, vol. 19, no. 4, pp. 4.1–4. 11, 2018.

[15] K. Veena and K. Meena, "An intrusion detection system for network security based on an advanced honeypots server," *International Journal of Simulation: Systems, Science and Technology*, vol. 19, no. 4, pp. 2.1–2. 8, 2018.

[16] H. Mohaddes Deylami and Y. Singh, "Cybercrime detection techniques based on support vector machines," *Artificial Intelligence Research*, vol. 2, no. 1, p. 1, 2012.

[17] A. Bharati and R. A. Sarvanaguru, "Crime prediction and analysis using machine learning," *International Research Journal of Engineering and Technology*, vol. 5, no. 9, pp. 1037–1042, 2018.

[18] S. Marsland, *Machine Learning: An Algorithmic Perspective*, CRC press, 2015.

[19] L. I. Song, Q. Zou, and W. Huang, "A new type of intrusion prevention system," in *2014 international conference on information science, electronics and electrical engineering*, pp. 361–364, Sapporo, Japan, 2014.

[20] J. D. A. Silva and E. R. Hruschka, "An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks," *Data and Knowledge Engineering*, vol. 84, pp. 47–58, 2013.

[21] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised Machine learning algorithms: classification and Comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, 2017.

[22] P. Bhuvaneswari and J. Satheesh Kumar, "A note on methods used for deception analysis and influence of thinking stimulus in deception detection," *International Journal of Engineering and Technology*, vol. 7, pp. 109–116, 2015.