**Journal of Computer Science and Technology Studies**
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# Advanced Cybercrime Detection: A Comprehensive Study on Supervised and Unsupervised Machine Learning Approaches Using Real-world Datasets

**Duc M Cao[1], Md Abu Sayed[2], Md Tanvir Islam[3], Md Tuhin Mia[4], Eftekhar Hossain Ayon[5] ✉ Bishnu Padh Ghosh[6], Rejon Kumar Ray[7], Aqib Raihan[8]**

*[1]Department of Economics, University of Tennessee, Knoxville, TN, USA*
*[2]Department of Professional Security Studies, New Jersey City University, Jersey City, New Jersey, USA*
*[3]Department of Computer Science, Monroe College, New Rochelle, New York, USA*
*[46]School of Business, International American University, Los Angeles, California, USA*
*[5]Department of Computer & Info Science, Gannon University, Erie, Pennsylvania, USA*
*[7]Department of Business Analytics Business Analytics, Gannon University, USA*
*[8]Computer science New Jersey City University Jersey City, New Jersey*
**Corresponding Author:** Eftekhar Hossain Ayon, **E-mail**: ayon001@gannon.edu

**| ABSTRACT**

In the ever-evolving field of cybersecurity, sophisticated methods—which combine supervised and unsupervised approaches—are used to tackle cybercrime. Strong supervised tools include Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), while well-known unsupervised methods include the K-means clustering model. These techniques are used on the publicly available StatLine dataset from CBS, which is a large dataset that includes the individual attributes of one thousand crime victims. Performance analysis shows the remarkable 91% accuracy of SVM in supervised classification by examining the differences between training and testing data. K-Nearest Neighbors (KNN) models are quite good in the unsupervised arena; their accuracy in detecting criminal activity is impressive, at 79.56%. Strong assessment metrics, such as False Positive (FP), True Negative (TN), False Negative (FN), False Positive (TP), and False Alarm Rate (FAR), Detection Rate (DR), Accuracy (ACC), Recall, Precision, Specificity, Sensitivity, and Fowlkes–Mallow's scores, provide a comprehensive assessment.

**| KEYWORDS**

Cybercrime Detection; on Supervised and Unsupervised Machine Learning; Real-world Datasets

**| ARTICLE INFORMATION**

**ACCEPTED:** 15 December 2023 **PUBLISHED02** 01 January 2024 **DOI:** 10.32996/jcsts.2024.6.1.5

## 1. Introduction

The growth of cybercrime in the cyber age demands a targeted approach to cybersecurity research, especially in the area of access control. The study emphasizes how important it is to identify cyber users and notify cybercrime investigators of any illegal activity as soon as possible. This allows investigators to carry out in-depth investigations and file lawsuits against offenders. Because there is no prior data, handling cybercrime situations is a special problem that emphasizes the need for machine learning models. These models are essential for accurately categorizing data using in-depth analysis and for making efficient use of features in class prediction. Enhancing network security and strengthening it against possible attackers is the main goal. Using real-time datasets and cluster computing tools, the study carefully compares the effectiveness of several approaches to cybercrime detection. The assessment extends to the efficacy of classifiers, seeking to enhance overall cybersecurity measures and mitigate the risks posed by malicious actors in the digital landscape.

Innovative and cutting-edge tactics are needed to combat cybercrime in the quickly changing field of cybersecurity. Using state-of-the-art instruments, including Support Vector Machines (SVM), K-nearest neighbors (KNN), K-means clustering, and Gaussian mixture models, this study explores both supervised and unsupervised approaches. The centerpiece is the extensive collection of personal traits of one thousand crime victims found in the CBS open data StatLine. Performance analysis is a rigorous testing process that uses several datasets to demonstrate the effectiveness of SVM, which has an amazing 91% accuracy rate in supervised classification. In the unsupervised realm, on the other hand, Gaussian mixture models excel, exhibiting an astounding 79.56% accuracy rate in crime identification.

True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN), False Alarm Rate (FAR), Detection Rate (DR), Accuracy (ACC), Recall, Precision, Specificity, Sensitivity, and Fowlkes–Mallows scores are among the evaluation metrics used to provide a comprehensive assessment. The Expectation-Maximization (EM) technique is used to carefully examine the Gaussian mixture model's performance to provide a thorough assessment of its effectiveness in cybercrime detection.

The complexity of gathering, classifying, and preparing cybercrime data is further examined in this article, which also highlights the significance of abiding by data privacy laws like the General Data Privacy Regulation (GDPR) of the European Union. Based on a variety of characteristics, the dataset is divided into discrete groups, and patterns are found to forecast cybercrime in particular industries, including banking. The approach includes the use of supervised learning techniques, wherein SVM is the primary tool used to create a cybercrime detection model. Real-time dataset input, classification using clustering approaches, SVM-based classification, cluster classification, and SVM evaluation based on new classes are among the comprehensive procedures. After a thorough analysis of performance metrics, training data properties, and SVM classifier results, an 89% classification accuracy is demonstrated.

The study also looks at using KNN for regression and classification tasks, emphasizing the importance of the K parameter and its weighted average method. A comparison of the computational requirements of SVM and KNN classifiers highlights the former's aptitude for binary classification, while the latter performs better as a multiclass classifier. Self-guided Hebbian learning, a feature of the unsupervised learning methodology, aids in pattern identification in the absence of predetermined labels. Two essential elements of this strategy are principal component analysis and cluster analysis. The performance of the suggested research is assessed using a real-time dataset in the article's conclusion, with a focus on gathering and preparing cybercrime data. An examination of IDS indicators along with performance metrics for the SVM classifier offers a comprehensive perspective on the efficacy of the research.

## 2. Literature Review

Buczak and associates' (2016) extensive survey study describes a targeted investigation of the literature about approaches for data mining (DM) and machine learning (ML) used in cyber analytics to enable intrusion detection. Every ML/DM approach has a brief tutorial description to go along with it, and the papers that best exemplify each method are chosen based on emerging importance and citation frequency. A thorough reading process was followed by summaries to ensure that every recognized approach was conveyed in its entirety. Considering the critical role that data plays in ML/DM techniques, significant cyber datasets that are used here are explained. The article explores the nuances of machine learning and deep learning algorithms, explores the difficulties in using them in cybersecurity, and provides suggestions for the best use of particular techniques.

Khater and associates' (2020) extensive examination looks closely at methods for identifying and stopping cybercrime. First, many types of cybercrimes are examined, together with the risks they represent to computer system security and privacy. The report explores the methods that cybercriminals use to commit these crimes against people, companies, and communities. The effectiveness of current cybercrime detection and prevention strategies is evaluated, with a critical examination of their weaknesses and an objective assessment of their merits. The paper concludes with recommendations for the creation of a sophisticated cybercrime detection model that can detect cybercrimes more successfully than existing methods.

Khan and associates' (2015) research looks at the structures and sources of the spamming botnets that are used to spread a significant amount of email spam. It then goes on to give detailed accounts of spamming botnets, presenting an organized picture of the chronological flow of events and significant advancements in these botnets' history. The aim of this work is to provide a thorough analysis of several email spamming botnet detection techniques that have been presented in previous research. It tries to classify various approaches according to their detection strategies as well as their defensive attributes, and it presents, contrasts, and compares their advantages and disadvantages in great detail. Additionally, a qualitative examination of these methods is provided. Finally, the report describes future directions and challenges in the field of email detection.

The Samarthrao group's (2020) research aims to improve cybersecurity by developing a novel spam detection methodology. The suggested approach consists of several phases, such as the acquisition of datasets, feature extraction, selection of the best features,

and detection. First, a benchmark email dataset is collected that includes text and image datasets. Then, two feature sets—text features and visual features—are used for feature extraction. While visual features include color correlograms and Gray-Level Co-occurrence Matrix (GLCM), text characteristics require extracting Term Frequency-Inverse Document Frequency (TF-IDF). An ideal feature selection procedure is carried out in order to handle the feature vector's increased length. This procedure uses the Fitness Oriented Levy Improvement-based Dragonfly method (FLI-DA), a revolutionary meta-heuristic method. Following the identification of the best features, detection is carried out by a hybrid learning strategy that combines the strengths of two deep learning methodologies: convolutional neural network (CNN) and recurrent neural network (RNN). In order to improve these deep learning methods' performance, FLI-DA is used to optimize the number of hidden neurons in CNN and RNN. In the end, data is divided into spam and ham categories using the optimum hybrid learning method that combines CNN and RNN. Empirical findings highlight the effectiveness of the suggested approach in classifying spam emails via improved deep learning.

Bouyeddou et al. (2021) present a technique that uses the Kullback-Leibler distance (KLD) to identify abnormalities in the form of DOS and DDOS flooding attacks, such as TCP SYN flood, UDP flood, and ICMP-based attacks. The method combines the sensitivity of an exponential smoothing scheme with the advantageous features of KLD, which is well-known for its capacity to quantitatively distinguish between two distributions. The goal of combining data from previous and present samples in the decision rule is what drives the addition of exponentially smoothed KLD measurements (ES–KLD) to increase the sensitivity of the detector to small abnormalities. Additionally, a nonparametric strategy employing kernel density estimation is employed to establish a threshold for the ES-KLD decision statistic, aiding in the identification of attack occurrences. Evaluations conducted on three publicly available datasets demonstrate the enhanced performance of the proposed mechanism in comparison to traditional monitoring techniques for cyber-attack detection.

## 3. Methodology

### *3.1 Data Collection and Classification*

A vast database of crime statistics is kept up to date in police files, with annual case counts recorded across the country. These records are compiled under the direction of the National Crime Bureau of Records. The collected data are typically unprocessed and frequently contain errors or missing values. As a result, in order to address these problems and properly arrange the information, data preparation becomes essential. Both data cleansing and preprocessing methods are included in this procedure. The dataset used is from the CBS open data Stat Line and focuses specifically on cybercrime detection. Open data for Stat Line tables is available to everyone via a Web service (API). This makes it easier to obtain the most recent stable version, which can be requested and downloaded from within Stat Line. Automation of data handling is enabled through the OData API. It's important to note that the protection of personal data adheres to the guidelines set forth by the European Union's General Data Protection Regulation (GDPR); Figure 1 shows the entire flow of our work.

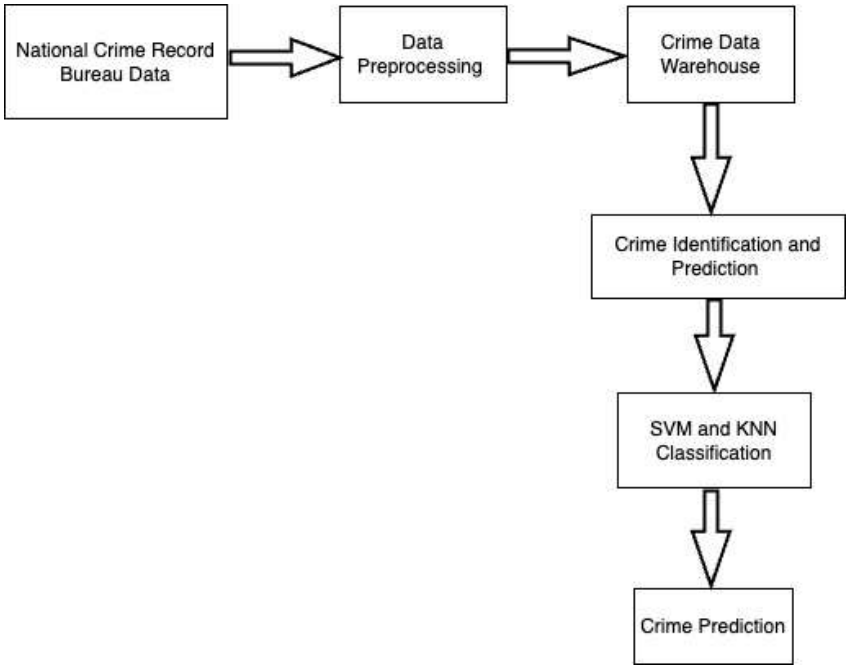


Figure 1: Entire Workflow of our Model

The dataset is partitioned into discrete groups based on qualities that are intrinsic to the data items. The method of categorization involves grouping crimes according to the states and cities in which they occur. This classification encompasses several categories of offences, enabling a thorough comprehension of the dataset. By utilizing the K-means technique, similar features in the data can be efficiently grouped or clustered, improving the information's overall organization and analysis.

### *3.2 Pattern Identification and Prediction*

This complex procedure includes figuring out the patterns and trends that are present in crime data. The key to this pattern identification is identifying the patterns of crime that are connected to places. Relevant location-related factors, such as weather, notable events, local sensitivity, and the existence of criminal gangs, are carefully considered in this context. Law enforcement personnel benefit greatly from the insights gained from these patterns, which help them perform more smoothly and effectively. For every place, a customized model is built, creating a focused method for crime investigation. A prediction software system receives the current date and pertinent attributes to identify crime-prone areas. The data are graphically presented using visualization tools, providing a thorough and understandable picture of the recognized criminal patterns. With the help of this integrated strategy, which combines attribute consideration, pattern detection, and predictive modeling, police officers may make proactive decisions and use resources strategically in their ongoing efforts to uphold public safety.

### *3.3 Supervised Learning Method*

In the context of supervised learning approaches, a model is carefully constructed to generate predictions based on data in an environment that is, by nature, unpredictable. With more observations, the computer's predictive power is significantly increased. Algorithms in the supervised learning paradigm work by taking into account a set of input data and associated known responses. A heterogeneous matrix represents the overall structure of the input dataset, with rows representing instances, observations, or examples and columns representing attributes, predictors, or features. The variables that contain the measurements for every user are represented by rows and columns. The column vector that is made up of the replies contains the output that is associated with every observation in the input dataset. When a supervised learning model is being trained, an appropriate algorithm is selected, and it is then given the input and response data to process. Through the process of iteration, the model is improved and optimized, learning and generalizing patterns from the given data and improving its prediction power. In the face of uncertainty, supervised learning becomes an effective tool for deriving relevant insights and producing precise forecasts thanks to this methodical methodology.

This research segment employs a cybercrime detection model utilizing a Support Vector Machine (SVM) for the classification of a dataset sourced from CBS data StatLine, accessible at https://www.cbs.nl/en-gb/our-services/open-data. SVM plays a pivotal role in the training process, facilitating the prediction of a particular user as either a Genuine or a potential Crime User based on multiple attributes.

The sequential steps in this methodology are outlined as follows:

1. **Input Real-time Dataset:** Commencing with the real-time dataset as input.
2. **Clustering Techniques Classification:** Utilizing clustering techniques for the initial classification.
3. **SVM-Based Classification:** Performing classification through Support Vector Machine (SVM).
4. **Cluster Classification and SVM for New Classes:** Employing cluster classification based on average data and conducting SVM classification for new classes across ten different attributes, predictors, or features.
5. **Performance Evaluation:** Employing various performance metrics, including True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN), False Alarm Rate (FAR), Accuracy (ACC), Detection Rate (DR), Specificity, Sensitivity, Precision, Recall, and Fowlkes–Mallows scores for diverse attributes.
6. **Training Data Metrics Determination:** Assessing mean-squared error for regression via 10-fold cross-validation (cvMSE), misclassification rate via stratified 10-fold cross-validation (cv MCR), and confusion matrix via stratified 10-fold cross-validation (cfMat). Extracting SVM structural components, such as Support Vectors, Alpha, Bias, and Support Vectorization, while also identifying the minimum and maximum values for the training attributes.
7. **SVM Utilization for Classification Accuracy:** Achieving a commendable classification accuracy of 89% through the use of SVM.

This methodological approach integrates advanced techniques, leveraging SVM and clustering, to enhance cybercrime detection accuracy and comprehensively evaluate model performance across diverse attributes.

### *3.4 SVM Classifier Training Data*

The SVM classifier utilizes machine learning tools for processing datasets related to cybercrime detection. The training data for the SVM classifier is illustrated in Table 1.

| Sl. no. | Training dataset (average) | SVM Struct. Support Vectors | SVM Struct. Scale Data Shift | SVM Struct. Scale Data Scale Factor | SVM Struct. Alpha | SVM Struct. Bias | SVM Struct. Support Vectorization | Mean-squared error for regression using 10-fold cross-validation: | Misclassification rate using stratified 10-fold cross-validation: | Confusion matrix using stratified 10-fold cross-validation: |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | cvMSE | cvMCR | cfMat |
| 1 | 7.850 | 0.314 | −7.102 | 0.260 | 0.625 | 1.958 | 5.000 | 0.003 | 0.150 | 0 1 0<br>12 69 2<br>0 0 16 |
| 2 | 8.040 | 0.403 | −7.102 | 0.260 | 0.625 | 1.958 | 9.000 | 0.003 | 0.150 | |
| 3 | 8.000 | 0.379 | −7.102 | 0.260 | 0.625 | 1.958 | 10.000 | 0.003 | 0.150 | |
| 4 | 7.750 | 0.317 | −7.102 | 0.260 | 0.625 | 1.958 | 11.000 | 0.003 | 0.150 | |
| 5 | 7.730 | 0.462 | −7.102 | 0.260 | 0.625 | 1.958 | 12.000 | 0.003 | 0.150 | |
| 6 | 8.310 | 0.384 | −7.102 | 0.260 | 0.625 | 1.958 | 13.000 | 0.003 | 0.150 | |
| 7 | 7.780 | 0.743 | −7.102 | 0.260 | 0.625 | 1.958 | 45.000 | 0.003 | 0.150 | |
| 8 | 7.670 | 0.936 | −7.102 | 0.260 | −1.250 | 1.958 | 51.000 | 0.003 | 0.150 | |
| 9 | 8.220 | 0.749 | −7.102 | 0.260 | 0.625 | 1.958 | 50.000 | 0.003 | 0.150 | |
| 10 | 8.650 | 0.837 | −7.102 | 0.260 | −2.500 | 1.958 | 51.000 | 0.003 | 0.150 | |
| 11 | 8.560 | 0.533 | −7.102 | 0.260 | 0.625 | 1.958 | 52.000 | 0.003 | 0.150 | |
| 12 | 8.170 | 0.689 | −7.102 | 0.260 | 0.625 | 1.958 | 53.000 | 0.003 | 0.150 | |
| 13 | 8.320 | 0.754 | −7.102 | 0.260 | 0.625 | 1.958 | 53.000 | 0.003 | 0.150 | |
| 14 | 8.880 | 0.790 | −7.102 | 0.260 | −2.500 | 1.958 | 54.000 | 0.003 | 0.150 | |

### *3.5 KNN Classifier*

In this scenario, the classification and regression tasks are performed through the application of the K-nearest neighbors (KNN) technique. KNN is employed not only for classification but also for estimating continuous variables in the context of KNN regression. An alternative approach involves calculating a weighted average from the two closest neighbors, where the value of k is set to 2 in this study. The labeled examples are organized based on increasing distance, with neighbors prioritized by the inverse of their distance. The algorithm's functionality entails computing the Euclidean distance between the query and labeled examples to determine the closest neighbors.

### *3.6 SVM and KNN Classifiers*

In KNN, the basis for data categorization relies on the distance metric, while SVM requires a proper training phase to ensure optimal segregation of divided data. SVM is particularly well-suited for binary classification, dividing data into two classes, while KNN is often employed as a multiclass classifier. For multiclass SVMs, both one-vs-one and one-vs-all approaches are utilized. Under the one-vs-one strategy, n*(n − 1)/2 SVMs are trained, with one SVM dedicated to each pair of classes. In this method, when an entity encounters an unknown pattern, the data type is determined by the majority output from the aggregate SVM output,

primarily employed in multiclass categorization. The classification involves distinguishing data as Genuine or Crime data, with specific user ranges assigned to each category's demonstrate computational demand, particularly when predicting classes with additional unlabeled data after the initial training phase. Conversely, in KNN, the distance metric is recalculated each time new unlabeled data is encountered. While KNN only requires fixing the K parameter and selecting an appropriate distance metric for classification, SVMs necessitate the choice of the regularization term along with kernel parameters, especially when dealing with linearly inseparable classes. In terms of accuracy, SVMs outperform KNN, as indicated in Table 2.

Table 2: TP, TN, FN, and FP classification for attribute.

| UID | Group | | Attribute 1 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Cluster classification based on average | GD = 0 CD = 1 | New class SVM classifier: Attribute 1 | GD = 0 CD = 1 | TP 0 | TN 1 | FP 11 | FN 10 |
| **1** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **2** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **3** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **4** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **5** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **6** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **7** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **8** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **9** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **10** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **11** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **12** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **13** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **14** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **15** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **16** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **17** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |
| **18** | Genuine User | 0 | "Crime User" | 1 | FALSE | TRUE | FALSE | FALSE |

### *3.7 Unsupervised Learning Method*

Unsupervised learning represents a self-guided Hebbian learning process, adept at recognizing previously unknown patterns within datasets devoid of predefined labels. This autonomous method, also known as self-organization, enables the creation of probable density distributions for provided inputs. Within the broader spectrum of machine learning, unsupervised learning stands alongside supervised and reinforcement learning as an essential component. The primary techniques employed in unsupervised learning encompass principal component analysis and cluster analysis.

### *3.8 The Proposed Research's Performance Utilizing a Real-Time Dataset.*

In community repositories, numerous researchers curate and distribute a variety of datasets from their work. This section uses research on machine learning and artificial intelligence to describe popular security-related datasets. The goal is to forecast cybercrime in the banking industry by analyzing crime trends and gathering data from a variety of online sources, including blogs, publications, news feeds, and police agency websites. After being gathered, the cybercrime data are kept for later processing in a specific crime database. To solve problems like missing values and noisy data, the stored cybercrime dataset must go through necessary preprocessing before data mining algorithms are applied. The goal is to combat cyber credit card fraud by detecting fraud by harnessing knowledge innovation from abrupt trends using data mining techniques and algorithms on preprocessed data. Using data mining to uncover hidden patterns, connections, and links in business data obtained from crime databases is a useful tactic for resolving issues facing the banking sector.

## 4. Results

### *4.1 Evaluation of Performance of Classifier*

Three groups of metrics—threshold, ranking, and probability metrics—have been developed to evaluate the efficacy of intrusion detection systems (IDS). Threshold metrics function according to whether the prediction is above or below a predetermined threshold, with values ranging from 0 to 1. Examples of threshold metrics are CR, F-measure, and CPE (cost per example). The sequence of examples is the subject of ranking metrics like FPR, precision (PR), intrusion detection capability (CID), detection rate (DR), and area under the ROC curve (AUC), which range from 0 to 1. These metrics assess how well attack instances are arranged in relation to normal instances, giving a general picture of the model's performance with regard to thresholds. In a probability statistic with values ranging from 0 to 1, the root-mean-square error (RMSE) decreases as the anticipated value for each attack class approaches the actual conditional likelihood of that class being normal. We compare different IDSs with well-known metrics like AUC. Higher numbers indicate a better IDS rating. The CID value, which ranges from 0 to 1, is closely correlated with IDS performance. The confusion matrix is a useful tool for representing the results of IDS categorization and is frequently used in the computation of these measurements.

Table 3: Performance metrics using the SVM classifier in cross-validation partition.

| **Cross-validation partition** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Attributes used** | **cvMSE** | **cvMCR** | **cfMat** | | | **Type** | ***N*** | **Num test sets** | **Train size** | **Test size** |
| **1,2,3,8** | 0.0025 | 0.15 | 0 | 1 | 0 | K-fold | 100 | 10 | 80 | 20 |
| **1,2,3,4** | 0.0025 | 0.03 | 12 | 69 | 2 | | | | | |
| | | | 0 | 0 | 16 | | | | | |

This study examines the examination of criminal activity, investigates supervised learning approaches using SVM and KNN classifiers, and offers a comparative evaluation. The research also includes unsupervised learning approaches, such as the use of the FCM algorithm for quasi-random data clustering, the selection of K-means clustering with justification, and clustering using Gaussian mixture models and the EM algorithm. In addition, the study assesses user profiles using a variety of clustering strategies to detect possible cybercriminals. Performance assessments are conducted using multiple datasets to identify machine learning algorithms that exhibit superior results. As indicated by the investigation, the Gaussian clustering technique outperforms other clustering methods in unsupervised scenarios. The results affirm the precise detection capability of cybercrime through these methodologies.

## 5. Conclusion and Discussion

Finally, this thorough research paper addresses both supervised and unsupervised approaches for cybercrime detection by carefully examining and assessing a wide range of machine learning and data mining techniques. A wide range of techniques is covered in the paper, such as Gaussian mixture models, K-nearest neighbors (KNN), K-means clustering, Support Vector Machines (SVM), and more. To get insight into the effectiveness of these algorithms in cybercrime detection, the research applies them to real-world datasets, such as the CBS open data Stat Line. Extensive testing and assessment measures show that supervised learning techniques, especially SVM, have remarkable accuracy rates—up to 91% in the case of SVM. Furthermore, the KNN classifier is presented for both classification and regression tasks, adding to the general comprehension of the advantages and processing requirements of different machine-learning approaches.

The study highlights the superior performance of Gaussian clustering approaches over other clustering methods in the field of unsupervised learning. Expectation-maximization (EM) algorithms are incorporated to further enhance the assessment, offering an extensive examination of cybercrime detection through unsupervised methods. The research proposal assesses current approaches and contributes by presenting new ones, including a hybrid learning technique that combines CNN and RNN for spam detection, an advanced feature extraction-integrated spam detection model, and a meta-heuristic algorithm for optimal feature selection. The outcomes highlight the usefulness of the suggested approach in classifying spam emails and highlight the significance of ongoing innovation in strengthening cybersecurity defenses.

The paper acknowledges and discusses the difficulties in detecting cybercrime, stressing the need for precise data collection, preprocessing, and compliance with data protection laws. The potential of machine learning models to identify crime patterns and improve overall security measures is highlighted in the discussion of its application in forecasting cybercrime, particularly in the

banking industry. The study offers a comprehensive grasp of the advantages, disadvantages, and trends in machine learning and data mining approaches for cybercrime detection, making it a useful resource for the cybersecurity community. The incorporation of authentic datasets and the utilization of varied approaches enhance the comprehensiveness of the results, rendering this piece a thorough manual for scholars, professionals, and decision-makers operating in the domain. From a larger perspective, the research indicates that staying ahead of cybercrime activities requires a holistic approach that incorporates both supervised and unsupervised methodologies, along with ongoing innovation in model building. This is because cyber threats are constantly evolving. This article's approaches and insights set the foundation for future developments in cybersecurity research and the continuous fight against cyber threats.

**Funding:** This research received no external funding.
**Conflicts of Interest:** The authors declare no conflict of interest.
**Publisher's Note**: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

**References:**

[1] Aljarboua E. F., Bte D M. and Bakar A. A., (2022). Cyber-Crime Detection: Experimental Techniques Comparison Analysis, 2022 International Visualization, Informatics and Technology Conference (IVIT), Kuala Lumpur, Malaysia, 2022, 124-129, doi: 10.1109/IVIT55443.2022.10033332.
[2] Ahsan, M., Gomes, R., Chowdhury, M. M., & Nygard, K. E. (2021). Enhancing machine learning prediction in cybersecurity using dynamic feature selector. *Journal of Cybersecurity and Privacy*, *1*(1), 199-218.
[3] Al-Khater W. A., Al-Maadeed S., Ahmed A. A., Sadiq A. S. and Khan M. K. (2020). Comprehensive Review of Cybercrime Detection Techniques, in IEEE Access, 8, 137293-137311, 2020, doi: 10.1109/ACCESS.2020.3011259.
[4] Ahmed, A. H., Ahmad, S., Sayed, M. A., Ayon, E. H., Mia, T., & Koli, T. (2023). Predicting the Possibility of Student Admission into Graduate Admission by Regression Model: A Statistical Analysis. *Journal of Mathematics and Statistics Studies*, *4*(4), 97-105.
[5] Abu S, Tayaba, M., Islam, M. T., Md Eyasin Ul, Islam P, Md Tuhin M, Eftekhar H A, Nur N, & Bishnu P G. (2023). Parkinson's Disease Detection through Vocal Biomarkers and Advanced Machine Learning Algorithms. *Journal of Computer Science and Technology Studies*, *5*(4), 142–149. https://doi.org/10.32996/jcsts.2023.5.4.14
[6] Buczak A. L. and Guven, E. (2015). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection, in IEEE Communications Surveys & Tutorials, 18, 2, 1153-1176, Second quarter 2016, doi: 10.1109/COMST.2015.2494502
[7] Bouyeddou, B., Harrou, F., Kadri, B. (2021). Detecting network cyber-attacks using an integrated statistical approach. *Cluster Comput* **24**, 1435–1453 (2021). https://doi.org/10.1007/s10586-020-03203-1
[8] Berry, H., Abdel-Malek, M. A., & Ibrahim, A. S. (2021, March). A machine learning approach for combating cyber attacks in self-driving vehicles. In *SoutheastCon 2021* (pp. 1-3). IEEE.
[9] Das, R., & Morris, T. H. (2017, December). Machine learning and cyber security. In *2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE)* (pp. 1-7). IEEE.
[10] Dushyant, K., Muskan, G., Annu, Gupta, A., & Pramanik, S. (2022). Utilizing Machine Learning and Deep Learning in Cybesecurity: An Innovative Approach. *Cyber Security and Digital Forensics*, 271-293.
[11] Haque, M. S., Amin, M. S., Miah, J., Cao, D. M., Sayed, M. A., & Ahmed, S. (2023). Retail Demand Forecasting: A Comparative Study for Multivariate Time Series. *Journal of Mathematics and Statistics Studies*, *4*(4), 40-46.
[12] Khan, R. H., & Miah, J. (2022, June). Performance Evaluation of a new one-time password (OTP) scheme using stochastic petri net (SPN). In *2022 IEEE World AI IoT Congress (AIIoT)* (pp. 407-412). IEEE.
[13] Khan R. H., Miah J., Arafat S. M. Y., Syeed M. M. M. and Ca D. M., (2023) Improving Traffic Density Forecasting in Intelligent Transportation Systems Using Gated Graph Neural Networks, 2023 15th International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2023, 104-109, doi: 10.1109/IIT59782.2023.10366426.
[14] Khan W. Z., Khan M. K, Bin M F. T., Aalsalem M. Y. and Chao H. -C., (2015). A Comprehensive Study of Email Spam Botnet Detection, in IEEE Communications Surveys & Tutorials, 17, 4, 2271-2295, Fourthquarter 2015, doi: 10.1109/COMST.2015.2459015.
[15] Khan R. H., Miah J., Rahman M. M. and Tayaba M., (2023). A Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer, 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, 647-652, doi: 10.1109/CCWC57344.2023.10099106.
[16] Kayyum S. et al., (2020). Data Analysis on Myocardial Infarction with the help of Machine Learning Algorithms considering Distinctive or Non-Distinctive Features, 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020. 1-7, doi: 10.1109/ICCCI48352.2020.9104104.
[17] Miah J., Khan R. H., Ahmed S. and Mahmud M. I. (2023). A comparative study of Detecting Covid 19 by Using Chest X-ray Images– A Deep Learning Approach, 2023 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2023, 0311-0316, doi: 10.1109/AIIoT58121.2023.10174382.
[18] Miah, J., Haque, M. S., Cao, D. M., & Sayed, M. A. (2023). Enhancing Traffic Density Detection and Synthesis through Topological Attributes and Generative Methods. *Journal of Computer Science and Technology Studies*, *5*(4), 69–77. https://doi.org/10.32996/jcsts.2023.5.4.8
[19] Miah, J., Haque, M. S., Cao, D. M., & Sayed, M. A. (2023). Enhancing Traffic Density Detection and Synthesis through Topological Attributes and Generative Methods. *Journal of Computer Science and Technology Studies*, *5*(4), 69–77. https://doi.org/10.32996/jcsts.2023.5.4.8
[20] Miah, J., Ca, D. M., Sayed, M. A., Lipu, E. R., Mahmud, F., & Arafat, S. M. (2023). Improving Cardiovascular Disease Prediction Through Comparative Analysis of Machine Learning Models: A Case Study on Myocardial Infarction. *arXiv preprint arXiv:2311.00517*.
[21] Miah, J., Cao, D. M., Sayed, M. A., & Haque, M. S. (2023). Generative AI Model for Artistic Style Transfer Using Convolutional Neural Networks. *Journal of Computer Science and Technology Studies*, *5*(4), 78-85.

[22] Mia, M. T., Ray, R. K., Ghosh, B. P., Chowdhury, M. S., Al-Imran, M., Das, R., Sarkar, M., Sultana, N., Nahian, S. A., & Puja, A. R. (2023). Dominance of External Features in Stock Price Prediction in a Predictable Macroeconomic Environment. *Journal of Business and Management Studies*, *5*(6), 128–133. https://doi.org/10.32996/jbms.2023.5.6.10

[23] Miah J., Ca D. M., Sayed M. A., Lipu E. R., Mahmud F. and Arafat S. M. Y., (2023). Improving Cardiovascular Disease Prediction Through Comparative Analysis of Machine Learning Models: A Case Study on Myocardial Infarction, 2023 15th International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2023, pp. 49-54, doi: 10.1109/IIT59782.2023.10366476.

[24] Omar, M. (2022). Application of Machine Learning (ML) to Address Cybersecurity Threats. In *Machine Learning for Cybersecurity: Innovative Deep Learning Solutions* (pp. 1-11). Cham: Springer International Publishing.

[25] Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., Chen, S., Liu, D., & Li, J. (2020). Performance comparison and current challenges of using machine learning techniques in cybersecurity. *Energies*, *13*(10), 2509.

[26] Salih, A., Zeebaree, S. T., Ameen, S., Alkhyyat, A., & Shukur, H. M. (2021, February). A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection. In *2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC)* (pp. 61-66). IEEE.

[27] Samarthrao, K V and Rohokale, V M. (2022). Enhancement of Email Spam Detection Using Improved Deep Learning Algorithms for Cyber Security'. 1 Jan. 2022: 231 – 264.

[28] Syeed, M. M., Khan, R. H., & Miah, J. (2021). Agile Fitness of Software Companies in Bangladesh: An Empirical Investigation. *International Journal of Advanced Computer Science and Applications*, *12*(2).

[29] Sarkar, M., Ayon, E. H., Mia, M. T., Ray, R. K., Chowdhury, M. S., Ghosh, B. P., Al-Imran, M., Islam, M. T., Tayaba, M., & Puja, A. R. (2023). Optimizing E-Commerce Profits: A Comprehensive Machine Learning Framework for Dynamic Pricing and Predicting Online Purchases. *Journal of Computer Science and Technology Studies*, *5*(4), 186–193. https://doi.org/10.32996/jcsts.2023.5.4.19

[30] Tayaba, M., Ayon, E. H., Mia, M. T., Sarkar, M., Ray, R. K., Chowdhury, M. S., Al-Imran, M., Nobe, N., Ghosh, B. P., Islam, M. T., & Puja, A. R. (2023). Transforming Customer Experience in the Airline Industry: A Comprehensive Analysis of Twitter Sentiments Using Machine Learning and Association Rule Mining. *Journal of Computer Science and Technology Studies*, *5*(4), 194–202. https://doi.org/10.32996/jcsts.2023.5.4.20

[31] Veena, K., Meena, K., Kuppusamy, R., Teekaraman, Y., Angadi, R. V., & Thelkar, A. R. (2022). Cybercrime: Identification and Prediction Using Machine Learning Techniques. *Computational Intelligence and Neuroscience*, *2022*.

[32] Yogesh, K., Karthik, M., Naveen, T., & Saravanan, S. (2019, September). Design and Evaluation of Scalable Intrusion Detection System Using Machine Learning and Apache Spark. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)* (pp. 1-7). IEEE.