Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

AI-DRIVEN INTRUSION DETECTION USING MACHINE LEARNING AN ANOMALY-BASED ANALYSIS OF NETWORK TRAFFIC

Amna Ilyas[1], Fahim Uz Zaman[2], Muqaddas Salahuddin[*3], Faraz Ahmad Zia[4], Muhammad Zohaib Khan[5], Sammia Hira[6], Muhammad Ather Ameen[7]

[1]*Department of computer science and Cybersecurity, Ravensbourne University London, England*
[2]*Digital Technologies, Newcastle College University Centre (NCUC) Rye Hill Campus, Scotswood Rd, Newcastle upon Tyne NE4 7SA, United Kingdom*
[*3]*Faculty of Computer Science and Information Technology, Superior University, Lahore, 54000, Pakistan*
[4]*Government College University Faisalabad*
[5]*Department of Information Technology, Shaheed Mohtarma Benazir Bhutto Institute of Trauma, Karachi, Pakistan*
[6]*Faculty of Computer Science and Information Technology, Superior University, Lahore, 54000, Pakistan*
[7]*Department of Information Technology, Shaheed Mohtarma Benazir Bhutto Institute of Trauma, Karachi, Pakistan*

[1]amna.amu.ilyas1@gmail.com, [2]faheemuzzaman@hotmail.com, [*3]muqaddassalahuddin60@gmail.com, [4]frazzia348007@gmail.com, [5]zohaib_khan2017@yahoo.com, [6]hirac53oort@gmail.com, [7]muhammadatherameen@gmail.com

DOI: https://doi.org/10.5281/zenodo.18045868

Keywords
Network Traffic, Intrusion Detection Systems (IDS), Anomaly-Based, Fuzzy C-Means Clustering, Naïve Bayes (NB), Machine Learning, K-Nearest Neighbor (KNN), Logistic Regression (LR), Feature Selection, Stochastic Gradient Descent (SGD)

Article History
Received: 25 October 2025
Accepted: 09 December 2025
Published: 23 December 2025

Corresponding Author: *
Muqaddas Salahuddin

Abstract
In the modern era, millions of individuals use the internet daily, making cybersecurity a critical concern for protecting users' privacy and network integrity. Ensuring reliable network-based system operation has become increasingly important due to the growing reliance on network technologies. Traditional signature-based intrusion detection systems (IDS) are unable to detect novel attacks, while existing anomaly-based IDS are often limited to specific applications and contexts, leaving them ineffective against all types of new threats. Improving detection rates while reducing false positives remains a major challenge in network intrusion detection systems (NIDS). This study proposes a hybrid IDS model that integrates classification techniques such as Logistic Regression (LR), K-Nearest Neighbor (KNN), Stochastic Gradient Descent (SGD), and Naïve Bayes (NB) with fuzzy C-Means clustering. Advanced feature selection methods are applied to enhance detection accuracy and robustness against evolving cyberattacks. The effectiveness of the proposed approach is evaluated using a network traffic IDS dataset. This study highlights the limitations of conventional intrusion detection systems and demonstrates how machine learning techniques can be leveraged to strengthen network security.

1. INTRODUCTION:

The growth of networked devices and systems has made network security critical in this ever-evolving digital ecosystem. Vulnerabilities are introduced by these networked environments, which could lead to a toxic combination of cyberspace's drawbacks. The data link layer is the most susceptible of the network's levels, and if it is penetrated, network security and integrity could

be completely violated. Intrusion detection, according to [1], is the process of keeping an eye on events taking place in a computer system or network and examining them for indications of invasions. Attempts to jeopardize a computer's or network's availability, secrecy, or security measures are other definitions.

Misuse detection and anomaly detection are the two main categories into which data mining-based intrusion detection approaches often fall. In misuse detection, a learning algorithm is trained over the labeled data after each instance in a data set is classified as "normal" or "intrusion." As long as they have been properly labeled, these methods may automatically retrain intrusion detection models on various input data that contain novel attack types. In contrast to intrusion detection systems that rely on signatures, models of misuse are generated automatically and have the potential to be more complex and accurate than manually generated signatures. The great degree of precision with which misuse detection algorithms identify known attacks and their modifications is one of their main advantages. Their incapacity to identify attacks whose occurrences have not yet been observed is their clear disadvantage. In contrast, anomaly detection creates models of typical behavior and automatically identifies any deviations from them, marking them as suspicious.

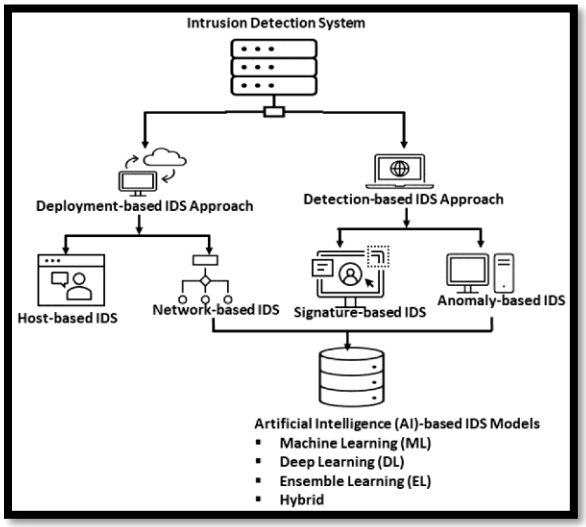

Figure 1: Explainable Artificial Intelligence (XAI) of Intrusion Detection

Thus, new kinds of invasions are detected by anomaly detection systems as departures from typical usage [2]. Although the percentage of false positives (false alarm rate) is a potential disadvantage of these strategies, while being an incredibly potent and innovative instrument. This can occur mainly because system behaviors that were previously undetectable (but legal) may also be identified as anomalies and reported as possible intrusions. IDS has improved as a result of recent ML developments, such as improved feature selection and deep learning techniques that better manage big datasets. ML techniques, such as Machines (SVM) and neural network methods [4]. Beyond basic data analysis, machine learning (ML) has applications in the security of Internet of Things (IoT) devices, including medical IoT systems [5]. Additionally, new technologies like Explainable AI (XAI), blockchain, and federated learning are being used into IDS to provide improved network security interpretability and transparency [6].

The paper provides a comparative analysis of various anomaly detection strategies for detecting new network incursions, after a brief summary of our research on developing predictive models for

learning from unusual classes. This research builds an effective network intrusion detection model using a variety of machine learning algorithms, including LR, NB, SGD, and KNN. The efficacy of selecting a machine learning algorithm for creating such an excellent intrusion detection model has been shown by the dataset's experimental findings. The remainder of the paper is structured as follows. The relevant works are discussed in Section 2, the suggested solution is employed in Section 3, the experimental design is illustrated in Section 4, and our intrusion detection model is evaluated through tests in Section 5. The study concludes with a conclusion and future work.

2. Literature Review:

The core component of cybersecurity tactics are intrusion detection systems (IDS), which combine a number of techniques to identify [7] and stop illegal activity on networks. IDS comes in a variety of forms, such as signature-based IDS, which match system activity patterns with a database of known threats to identify malicious activity [8]. By contrasting network behavior with predetermined norms, behavior-based intrusion detection systems, also known as anomaly-based systems, identify irregularities [9]. Host-based Intrusion Detection Systems (HIDS) are particularly helpful in financial institutions since they monitor individual endpoints or host-based systems with an emphasis on identifying unusual behavior or suspicious behaviors [10]. By examining network flow patterns, network-based intrusion detection systems (NIDS) keep an eye on traffic at critical network nodes and spot questionable activities [11]. While statistical IDS utilize statistical analysis to find departures from expected behavior, protocol-based IDS examine network protocols and headers to find anomalous activity [12].

A number of intrusion detection techniques based on system call trace data have been proposed by Warrender et al. [13]. In order to create a library of typical sequences for testing against test instances, they tried a technique that makes use of sliding windows. They then classified examples based on those in the regular sequence database by comparing the windows in the test instances with the database using a similar technique. For every call made by a process, the function necessitates sequential analysis of a window of system calls. Maintaining a sizable collection of typical system call trace sequences is necessary for this.

The majority of intrusions happen over networks, where targets are attacked utilizing network protocols. Twycross suggested using Danger Theory, a novel immunological paradigm, to create an intrusion detection system [4]. A classification-rule discovery technique that combines fuzzy systems with artificial immune systems (AIS) is presented by Alves et al. [5]. For instance, in order to accomplish his goal during a specific incursion, a hacker must first establish a connection between a source IP address and a target IP address before sending data to attack the target.

In order to minimize processing time, communications overhead, and storage requirements for mining network incursions, Chang et al. have concentrated on combining data reduction and classification with a query-based learning methodology [6]. For this evaluation experiment, the authors created a collection of classifiers using four distinct learning algorithms [7]. According to Mitchell [8], in certain domains, this technique is reported to perform similarly to neural network and decision tree learning. Because of this, it is an intriguing algorithm to employ in experiments involving the assessment of classifiers using a measure function. Panda and Patra [9] concluded that Naïve Bayes is suitable for creating an intrusion detection model after comparing its performance with the Neural Network method.

SVMs provide strong generalization against overfitting and effectively handle both linear and non-linear data by creating hyperplanes in spaces with high dimensions for binary classification [20]. Based on Bayes' theorem, naïve Bayes classifiers provide probabilistic classification; nevertheless, they presume feature independence, which may limit accuracy in some IDS situations [21]. Despite being computationally demanding, the KNN algorithm performs exceptionally well in identifying abnormalities and categorizing network traffic according to similarity criteria [22].

When addressing high-dimensional data, ensemble techniques like Random Forest improve accuracy and robustness by combining predictions from several decision trees. By improving the identification of security risks, the combination of machine-learning techniques with intrusion detection systems represents a major breakthrough in cybersecurity [23]. Researchers can create IDS solutions that improve security in a variety of network situations by tackling these issues and utilizing developments in cybersecurity and machine learning [30].

3. PROPOSED METHODOLOGY:

The first step in the procedure is gathering a cybersecurity dataset that includes records of system activity or network traffic. In order to train and assess the machine learning models, the dataset contains instances of both benign and malevolent activity.

1. The first stage in the procedure is to retrieve information from Google Drive.
2. Pre-processing is performed to improve data quality and ensure compatibility with machine learning algorithms. This step consists of the following sub-steps:
3. Categorical features and class labels are converted into numerical values using label encoding techniques to make the data machine-readable.
4. Feature values are normalized or scaled to ensure uniformity and to prevent attributes with larger ranges from dominating the learning process.
5. Feature importance or selection techniques are applied to identify the most relevant features, reduce dimensionality, and enhance model performance and computational efficiency.
6. The pre-processed dataset is then fed into multiple machine learning classifiers for training and testing. The following algorithms are employed such as LR, SVM, SGD and KNN.
7. In the final step, the trained models classify incoming data instances. Based on the learned patterns, each instance is categorized as either, Normal and Anomalous.

The workflow of the research project is summarized in Figure 2, which shows an overview of the Intrusion Detection Network Security architecture.

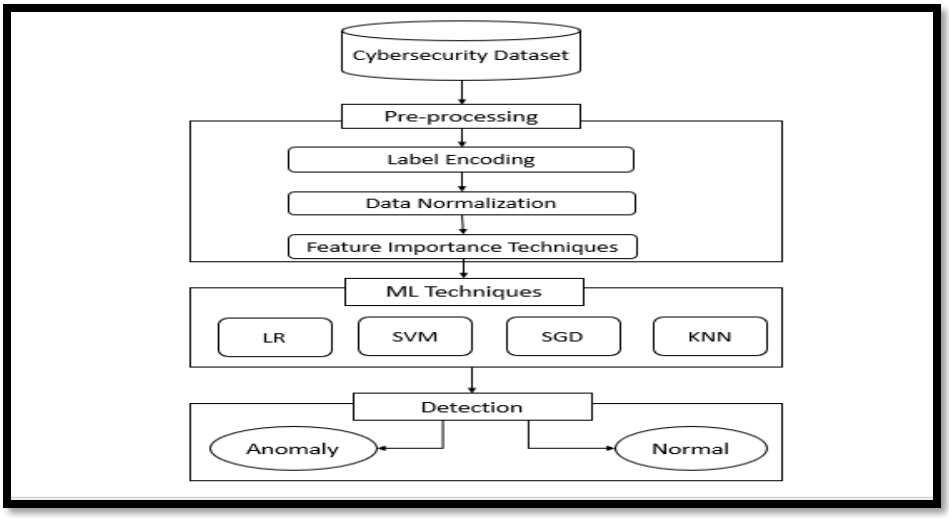

Figure 2: Proposed Methodology

3.1 PREPROCESSING

In order to obtain pertinent and useful information from data, preprocessing is essential for cleaning and grouping the data. By using data cleansing, standardization, and normalizing processes, this step guarantees the quality of the incoming data. Techniques like fuzzy c-means are used in the clustering phase to get the data ready for additional classification.

Data Collection

The Intrusion Detection Systems (IDS) dataset, which was obtained from Kaggle, a well-known platform for machine learning datasets, was utilized in this investigation. The dataset has 25,192 items and 42 attributes associated with Internet of Things (IoT) devices. Of them, 11,743 items belong to abnormalities associated with computer virus-related flaws or other cyber-intrusion occurrences, whereas 13,449 data reflect typical network activity [29]. The frequency and distribution of cyber-intrusion events within the sample are compared in Figure 3. Anomalies indicating cyber invasions are represented by a value of 1, whilst normal activity is represented by a value of 0. The dataset's importance for creating an efficient intrusion detection system is highlighted by this comparative study, which provides insightful information about the trends and frequency of security breaches.

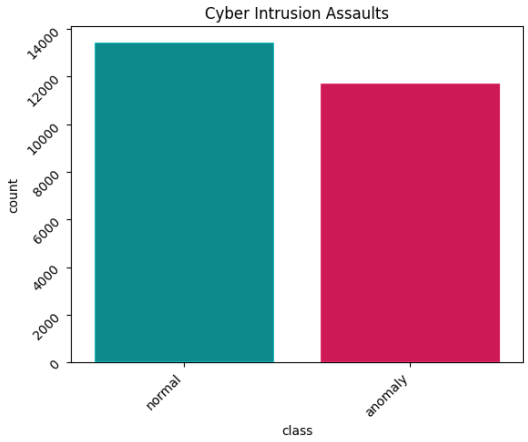

Figure 3: Distribution of Class

FEATURE SELECTION:

In a machine learning model, each input feature is assessed and assigned a score based on its significance, indicating the degree to which it influences the model's predictions. The more significant a trait is to the outcome, the higher its score. By retaining the features with the highest scores and removing those with lower values, which are deemed less significant, this process helps reduce the dimensionality of the model. This decrease simplifies the model and enhances its overall performance and efficiency [32]. These essential traits are depicted in Figure 4, emphasizing their importance in identifying network security threats.

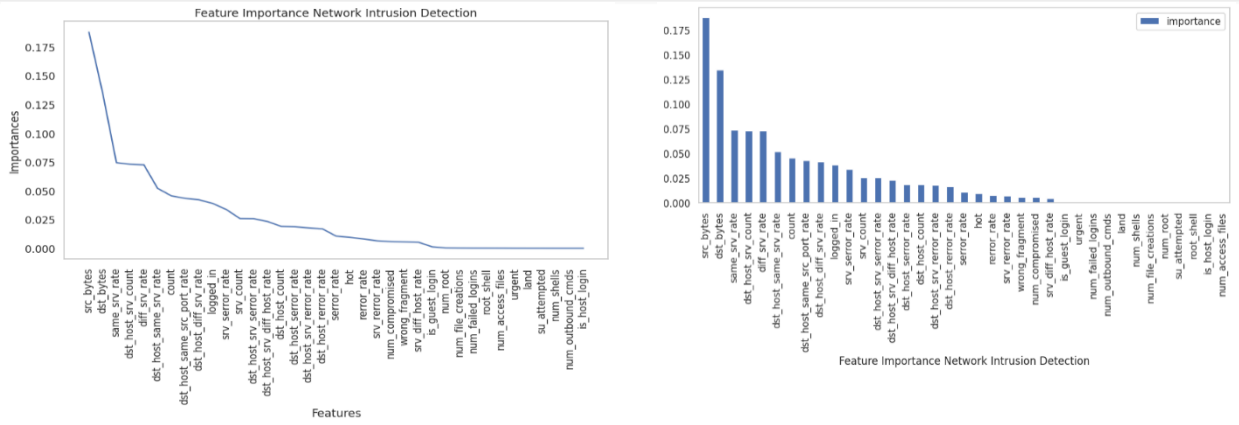

Figure 4: Feature Importance Dataset

Important steps in cyberattack prediction, including feature induction, simulation, preprocessing, and data normalization, are the focus of the current work. These steps are meant to address problems such as data complexity, post-processing efficiency, and system performance. The preparation process begins with data collecting from the Intrusion Detection Systems (IDS) dataset. To ensure uniformity and enhance model correctness, standardization and normalizing are then carried out. The final cleaned and preprocessed dataset is shown in Table 3. Sensitive data patterns are also assessed using fuzzy C-Means groups, which can be visualized by light-yellow circles for the Y values and black circles for the X values.

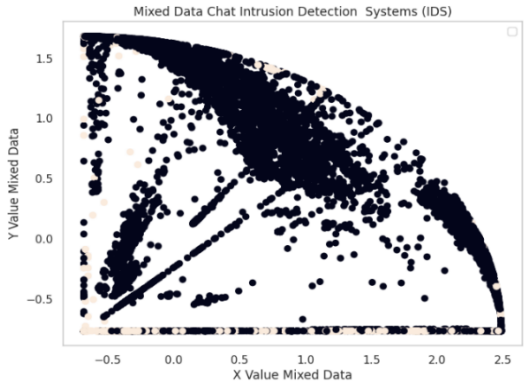

Figure 5: Mixed information Chat Intrusion Detection Systems (IDS)

The study increases the system's effectiveness in managing challenging intrusion detection tasks by utilizing feature selection strategies and incorporating sophisticated clustering techniques like Fuzzy C-Means, which eventually improves the system's performance and predictive capabilities.

Fuzzy C-Means Clustering Method. One popular type of unsupervised learning that is applied in many different domains is clustering. It entails data organization and analysis. The FCM approach assigns a membership value to each data point based on its distance from the centroid of the cluster. The closer a data point is to the centroid, the higher its membership score. Instead of a strict, single membership, our method guarantees that each data point has a degree of association with the cluster [33, 34]. To convert unstructured datasets into more organized

formats, appropriate for clustering, preprocessing approaches are utilized.

By eliminating redundancies, the FCM method is essential for processing and improving big datasets. By dividing the data into three separate clusters, this technique creates a membership matrix that shows how each data point relates to its corresponding cluster. Each cluster's centroids, or core points, show how near data points are to the centroid; closer data points have greater correlations. Eleven dimensions make up the IDS dataset used here, with a target attribute that provides information about cyberattacks and cluster counts. Line graphs and cluster distributions are used in Figures 6 and 7 to depict the clustering results, demonstrating how the conversion of unstructured data into a more ordered structure improves interpretability.

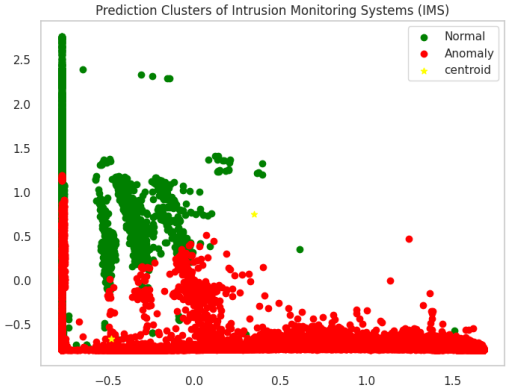

Figure 6: Fuzzy C-Means Three Clusters (IDS)

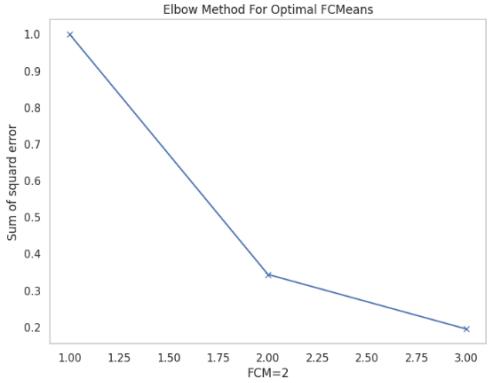

Figure 7: Fuzzy C-Means Sum of Squared Error Line Chart

A range of clustering approaches are used to assess clustering feature detection, and performance is measured by important metrics like F-measure, recall, and precision. Results are classified as genuine, suspect, or illegal based on how much the system deviates from predetermined standards. This method provides a thorough assessment of the system's ability to differentiate and identify various types of outcomes by classifying the data according to predetermined standards. The efficiency of the system in identifying and classifying anomalies is carefully evaluated by utilizing these metrics and classification methods.

3.2 CLASSIFICATION

Data classification is the process of classifying information into discrete groups or classes using training data. Supervised learning methods are used to identify the best classifier for a given dataset. To compare and evaluate the effectiveness of various classification algorithms, a number of tests are carried out. This method guarantees that the best classifier for the dataset is chosen, allowing for precise and trustworthy forecasts. The procedure assists in determining which algorithms yield the best results in terms of classification accuracy and efficiency by analyzing several algorithms.

Stochastic Gradient Descent (SGD) Algorithm. When it comes to optimizing machine learning models, especially for large-scale data, the SGD methodology is a very successful method. It is computationally efficient since it modifies model parameters by examining each individual data point instead of reviewing the entire dataset. Compared to batch processing techniques, this method enables faster convergence and lowers computing overhead. Additionally, SGD adds some randomness to its updates, which improves optimization for complex, non-convex functions by preventing local minima. Because of these characteristics, SGD is essential to attaining the great accuracy exhibited by the hybrid IDS framework **[35].**

Figure 8 displays a confusion matrix for the predictions made by the SGD model. In order to maximize alignment with the real labels, this

method forecasts categories using both previously collected and freshly acquired data after training.

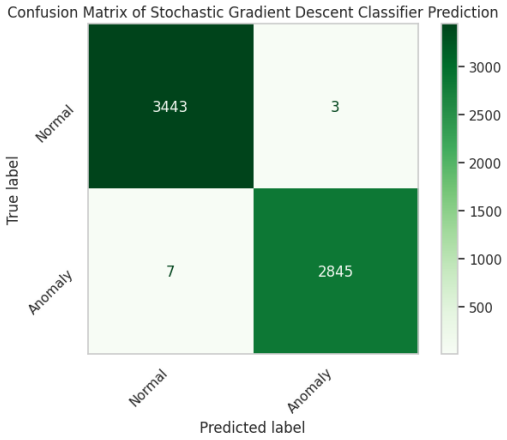

Figure 8: Confusion Matrix of SGD Algorithm

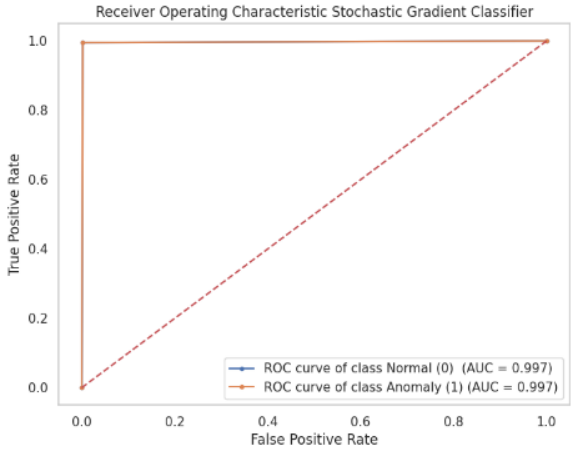

Figure 9: Illustrates the ROC Curve for SGD

The ROC curve provides information about the cost-benefit analysis and aids in understanding the efficacy of the SGD model by assessing a classifier's performance as the grouping threshold varies.

Logistic Regression (LR): A probabilistic hypothesis is used for predictive analysis in logistic regression (LR), a machine learning-based classification technique. In contrast to Linear Regression, which carries out a simple linear transformation, the "Sigmoid function" or "regression model" is a more complex differential equation employed in LR. The LR hypothesis restricts the coefficients to the binary values of 0 and 1. The LR hypothesis cannot be maintained by linear techniques, notwithstanding their attempts to approximate these values **[36, 37].** These processes enable the LR algorithm to iteratively adjust parameters to minimize costs and maximize predictions. LR anticipates the value of new data and labels the values of prior data in order to adjust predictions to match the labels. Figure 10 displays the results of this LR matrix for confusion.

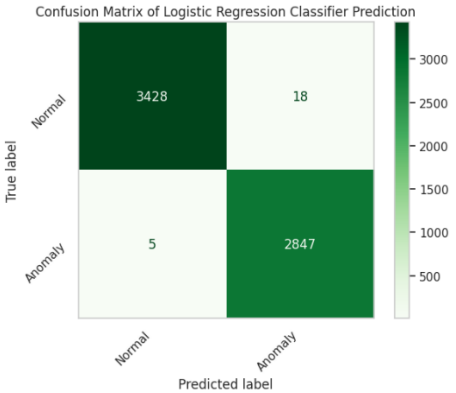

Figure 10: Confusion Matrix for LR Method

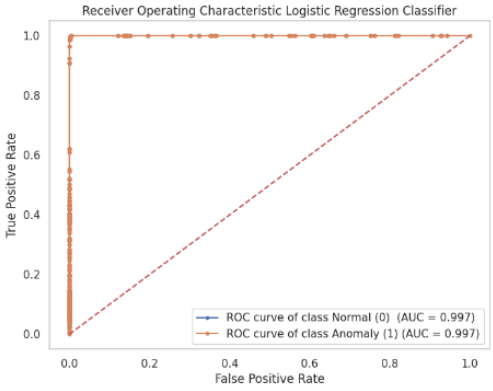

Figure 11: illustrates the LR of ROC Curve.

If the perceptron model specifically, logistic regression is suitable for the given scenario, the confusion matrix is crucial for determining the expected labels for detection and prediction tasks. Figure 11 shows the outcomes of applying the LR model to a given dataset. This Receiver Operating

Characteristic curve is used to assess the LR performance. In this study, the ROC curve helps assess the accuracy of the model's predictions, providing useful information regarding prediction patterns and improving the overall accuracy of the estimating process.

K-Nearest Neighbors (K-NN) Algorithm.
One of the simplest machine learning methods for both classification and regression issues is the KNN. Convergent items are assumed to be identical by the KNN approach. To put it another way, similar objects are next to one another. The KNN technique determines the distance between the item to be classified and every item in the training data in order to classify a new condition. The number of the element to be classified's closest neighbors, or K, is then found to be the optimal value [26,27,28]. To find the ideal value of k, multiple values are typically explored. The outcome of the classification is decided by the neighbors' majority vote **[38].**
During this technique, data is accurately mapped onto multi-dimensional space. After being trained, the algorithm may predict labels for both new and existing data samples with the goal of generating predictions that closely match actual labels. Figure 12 shows the generated confusion matrix, which illustrates the prediction performance of the K-NN method.

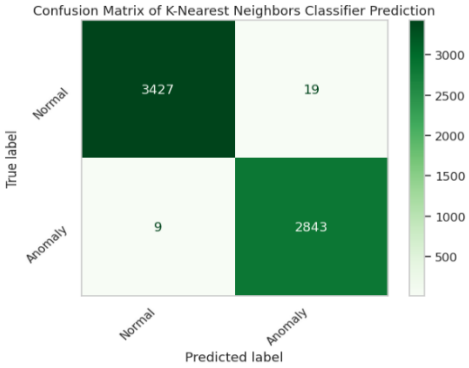

Figure 12: Confusion Matrix of K-NN Method

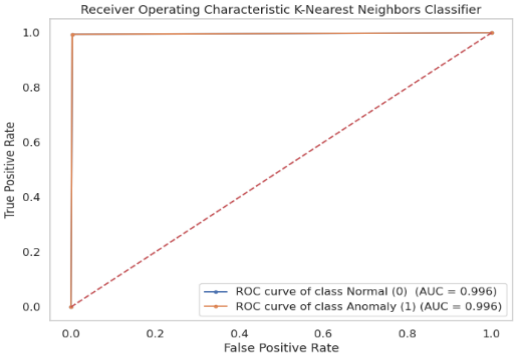

Figure 13: ROC Curve of K-NN

If the K-NN paradigm is suitable for our particular scenario, the confusion matrix is crucial in determining the expected labels for our identification and prediction tasks. Figure 13 shows the outcomes of using the K-NN algorithm on a synthetic dataset. The ROC curve displays the performance of the K-NN model. For this study, the Receiver Operating Characteristic curve is a useful tool for assessing the quality of our model predictions.

Naive Bayes (NB) Algorithm. One extremely simple Bayesian probability model is the Naïve Bayes model [14]. Take into account the likelihood of an outcome given a number of relevant evidence factors in this model. The model encodes both the likelihood that the end result will occur and the likelihood that the evidence variables will occur provided that the final outcome does. It is presumed that the likelihood of an evidence variable given the outcome is independent of the likelihood of other evidence variables given the outcome. This approach is crucial for a number of reasons. It requires no complex iterative parameter estimation algorithms and is relatively simple to develop.
Even people who are unfamiliar with classifier technology may understand the logic behind its categorization because to its simplicity. Finally, it often does surprisingly well: even while it may not

be the best classifier in a particular application, it is usually reliable and does fairly well. The matrix that results from this procedure is shown in Figure 14.

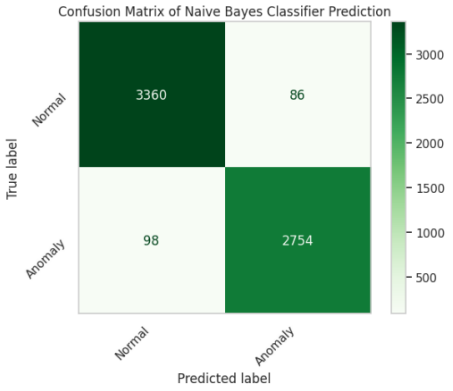

Figure 14: Confusion Matrix of NB Algorithm

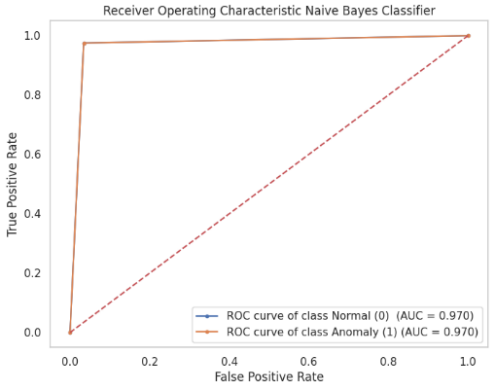

Figure 15: Illustrates the ROC Curve of NB

Because ROC analysis is a crucial technique for assessing classifier performance while adjusting the discrimination threshold, it is used in this study [44]. Making educated decisions is aided by this approach, which is essential for cost-benefit analysis. A graphical depiction of the ROC curve is shown in Figure 15, which aids in understanding the accuracy of the Naive Bayes method in class label prediction and provides a visual explanation of the classifier's performance.

4. RESULTS AND DISCUSSION:

The study assesses how well different hybrid algorithms predict and categorize Intrusion Detection Systems (IDS). The Scikit Learn library, one of the potent libraries used to develop and apply ML techniques and data preparation in Python, is used to build the four ML (RF, DT, KNN, and SVM) techniques. Data that was more than three times the threshold value were examined using the Cook's distance method, and any necessary corrections were performed. Lastly, the T-test statistic and confidence interval calculations were used to assess the data's significance. Our data have been found to fall inside the confidence interval. The accuracy of the tested hybrid algorithms is shown in the following table.

Table 1: Accuracy of Hybrid Algorithms

Hybrid Algorithm	Accuracy of Algorithms
Stochastic Gradient Descent (SGD) **Proposed Method**	99.6506 %
Logistic Regression (LR) **Proposed Method**	99.6348 %
K-Nearest Neighbours (K-NN) **Proposed Method**	99.5554 %
Naive Bayes (NB) **Proposed Method**	97.0784 %

The findings show that the efficacy of hybrid algorithms in anticipating network intrusion threats varies. With an astounding accuracy of 99.6506%, the combination of Fuzzy-C-Means and Stochastic Gradient Descent (SGD) outperformed all other combinations examined. This research shows how these algorithms can be used to create a very accurate model for predicting network invasions. The pairing of Fuzzy-C-Means and K-Nearest Neighbors placed third with an accuracy of 99.5554%, while the combination of Logistic Regression (LR) and Fuzzy-C-Means achieved the second-highest accuracy at 99.6348%. At 97.0784%, the Fuzzy-C-Means and Naive Bayes (NB) hybrid produced the fourth-highest accuracy. These results illustrate the

potential of hybrid algorithms to greatly improve IDS performance and efficiency, underscoring the significance of combining several algorithms to increase predictive proficiency. Several important metrics are used to assess how well integrated algorithms predict cyberattacks.

Table 2: Parameter Scores of Hybrid Algorithms

S/No.	Parameter Score	Stochastic Gradient Descent (SGD)	Logistic Regression (LR)	K-Nearest Neighbours (K-NN)	Naive Bayes (NB)
1	Precision	0.99457257	0.99613041	0.99537098	0.97068911
2	Recall	0.99520583	0.99651169	0.99566534	0.97034083
3	F1-Score	0.99487676	0.99631670	0.99551558	0.97051135
4	Sensitivity	1.0	1.0	1.0	1.0
5	Specificity	1.0	1.0	1.0	1.0

Several important metrics are used to assess how well integrated algorithms predict cyberattacks. The model obtained an accuracy of 0.97068911, perfect specificity of 1.0, F1 score of 0.97051135, recall of 0.97034083, and sensitivity of 1.0 when Naive Bayes (NB) and Fuzzy C-Means were matched. This illustrates how well this combination performs in identifying cyberthreats. With a recall value of 0.99566534, an F1 score of 0.99551558, and a precision of 0.99566534, the combination of Fuzzy C-Means and K-Nearest Neighbors fared better than the others. Additionally, it demonstrated its strong detection capabilities by achieving a specificity of 1.0 and flawless sensitivity. Similarly, recall of 0.99651169, specificity of 1.0, F1 score of 0.99631670, and accuracy of 0.99613041 were obtained when Logistic Regression (LR) and Fuzzy C-Means were combined. This suggests that cyberattack prediction is highly accurate and reliable. Finally, a recall of 0.99520583, specificity of 1.0, precision of 0.99457257, and an F1 score of 0.99487676 were obtained by the Stochastic Gradient Descent (SGD) and Fuzzy C-Means hybrid model, highlighting the efficacy of integrating these methods.

5. Comparative Analysis

Table 3: Algorithms Comparison Across Various Studies

Authors	Algorithms	Results
Debicha et al. (2023)	SVM	91%
Abdulganiyu et al. (2023)	Support Vector Machine	94%
Hassan et al. (2023)	Gradient Boosting Machine	89%
Alkasassbeh and Al-Haj Baddar (2023) [24]	K-Nearest Neighbors	91%
Azam et al. (2023)	Decision Tree	95%
Alotaibi and Rassam (2023)	Neural Network	95%
Heidari and Jamali (2023)	Random Forest	92%
Azar et al. (2023)	Random Forest	93%

The accuracy scores of different algorithm pairings in a hybrid framework for cyberattack detection are shown in Figures 16 and 17. The accuracy of the pairings ranges remarkably from 97.07% to 99.6506%, with fuzzy C-Means and SGD achieving the maximum accuracy of 99.6506%. Fuzzy C-Means and Logistic Regression (LR) achieved 99.6348% accuracy, whilst Fuzzy C-

Means and K-Nearest Neighbors (K-NN) achieved 99.5554%. Finally, the accuracy of the fuzzy C-Means and Naive Bayes (NB) hybrid was 97.0784%. These statistics demonstrate how hybrid algorithms greatly improve cyberattack detection systems' efficacy and provide a potent tool for raising prediction model accuracy.

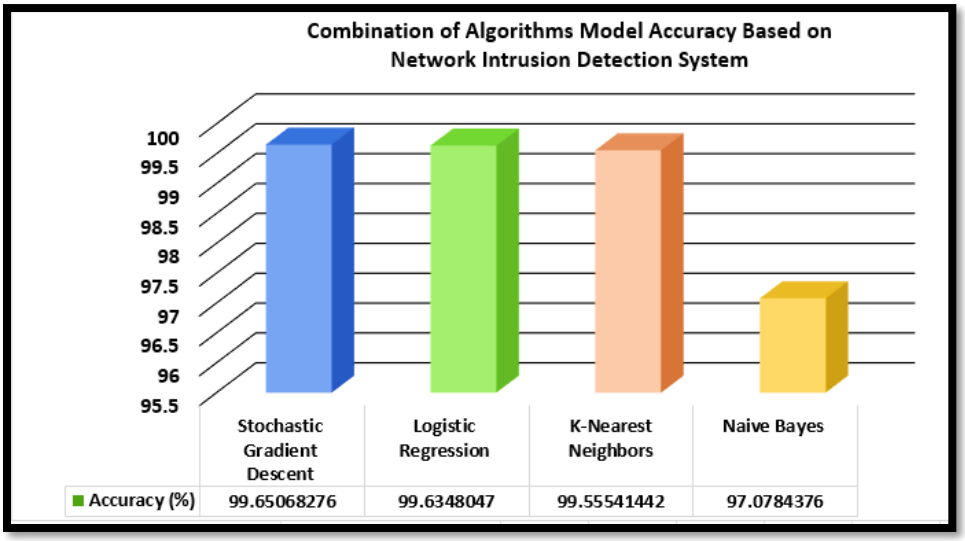

Figure 16: Combination of Algorithms Model Accuracy Based on Network IDS

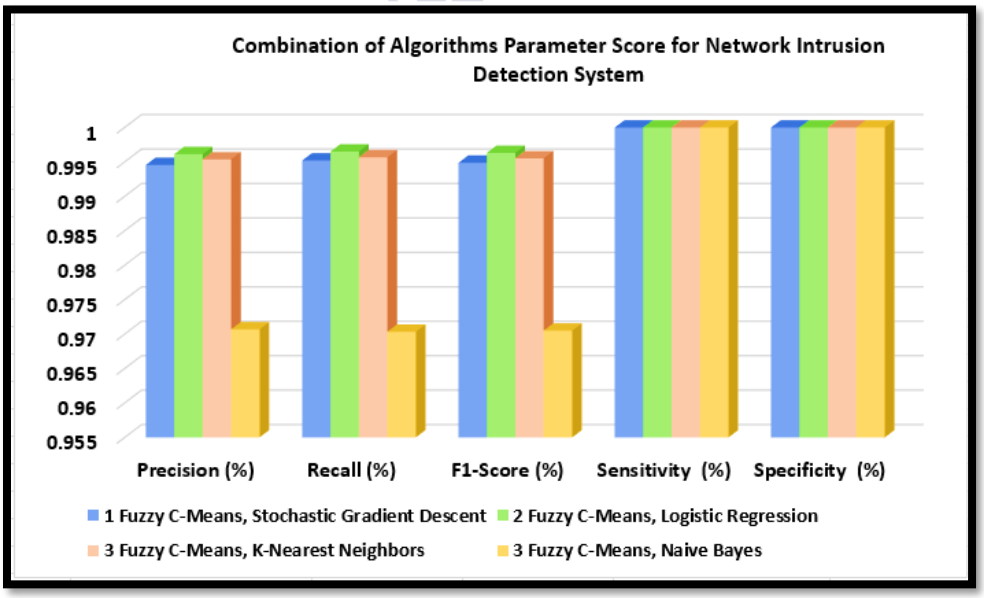

Figure 17: Combination of Algorithms Parameter Score for Network IDS

In this work, different accuracy levels from reference publications were compared to the accuracy of our hybrid model-based intrusion detection systems (IDS) utilizing the NIDS dataset. The hybrid model outperformed all of these benchmarks in terms of accuracy. Further analysis of the experimental design, feature selection strategies, and evaluation procedures is necessary to fully understand each model's performance.

The study's experimental findings show how effective an IDS is at identifying and reducing network intrusions. The IDS's outstanding performance across several evaluation criteria shows that it can significantly improve security measures. However, the study acknowledges its limitations in mimicking real-world scenarios, inherent biases in evaluation procedures, and challenges with scalability and generalizability. Future research should focus on improving machine learning methods like deep learning and reinforcement learning to boost the accuracy and efficacy of IDS. The report also emphasizes the importance of interdisciplinary collaboration among professionals in network technology, cybersecurity, and information science in order to provide innovative ideas and solutions to escalating cyberthreats.

CONCLUSION:

The study uses machine learning methods to create a reliable intrusion detection system (IDS). Cyberattacks that target corporations, institutions, and even individuals have significantly increased. Due to the rapid advancements in technology, attackers have become more skilled, and standard intrusion detection systems are no longer able to identify complex cyberattacks. To identify these costly and damaging attacks, new, sophisticated technologies had to be developed. Numerous studies have used ML approaches to create IDS systems following the enormous successes of ML and DL techniques in a variety of disciplines. An intrusion detection system (IDS) based on feature selection and machine learning approaches is presented in this work. Models such as Logistic Regression (LR), K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), and Naïve Bayes (NB) were investigated, offering insights into their strengths and limitations. A hybrid model achieved accuracy of SGD 99.6506%. The study emphasizes the significance of integrating algorithms to enhance IDS reliability. Future research could be done to categorize the many types of cyberattacks. Using ensemble approaches, which combine multiple classifiers, can help increase classification accuracy.

REFERENCES

1. S. Mahmood, S. M. Mohsin, and S. M. A. Akber, "Network security issues of data link layer: An overview," in 2020 3rd International Conference on Computing Mathematics and Engineering Technologies (iCoMET), Jan. 2020, pp. 1-6.
2. Muqaddas, M., Majeed, S., Hira, S., & Mumtaz, G. (2024). A Systematic Literature Review on Performance Evaluation of SQL and NoSQL Database Architectures. Journal of Compu-ting & Biomedical Informatics, 7(02). https://jcbi.org/index.php/Main/article/view/548/502
3. Salahuddin, M., Zaman, F. U., Mumtaz, G., Khan, M. Z., Kainat, M., Hira, S., Parveen, F., & Mahmood, R. (2025). Integrating network intrusion detection with machine learning techniques for enhanced network security. Spectrum of Engineering Sciences, 3(4), 612–625. https://sesjournal.com/index.php/1/article/view/286
4. Zahra, W. U., Zaman, F. U., Mumtaz, G., Salahuddin, M., Khan, M. Z., Sultan, S. A., Hira, S., & Parveen, F. (2025). Approaches to predict cardiovascular issue using machine learning method. Spectrum of Engineering Sciences, 3(4), 417–429.
5. Parveen, F., Iqbal, S., Mumtaz, G., & Salahuddin, M. (2024). Real-time intrusion detection with deep learning: Analyzing the UNR intrusion detection dataset. Journal of Computing & Biomedical Informatics, 7(02). https://jcbi.org/index.php/Main/article/view/554
6. Afzal, M., Salahuddin, M., Hira, S., Sultan, M. F., Ahmad, S. Z., & Iqbal, M. W. (2024). A systematic literature review of understanding the human-computer interaction collaboration with user experience design. Bulletin of Business and Economics, 13(2), 723–729. https://doi.org/10.61506/01.00386

7. Mahmood, R., Mustafa, S., Asif, H., Raza, A., & Salahuddin, M. (n.d.). Leveraging artificial intelligence to optimize software project management: Enhancing efficiency, risk mitigation, and decision-making. Contemporary Journal. https://doi.org/10.12345/f42z1z57
8. Zaman, F.U., Khan, M.Z., Imroz, A., Khan, A.A., Salahuddin, M. and Kainat, M., 2024, December. Student Performance Analysis in Higher Education Using Integrated Approach of Machine Learning Techniques. In 2024 International Conference on Sustainable Technology and Engineering (i-COSTE) (pp. 1-6). IEEE
9. Abbas, A., Salahuddin, M., Khan, M. Z., & others. (2025). Machine learning-based hybrid technique to enhance cyber-attack perspective. Journal of Cloud Computing, 14, Article 57. https://doi.org/10.1186/s13677-025-00782-5
10. Khan, M. Z., Shaikh, S. A., Khan, A. A., Imroz, A., Salahuddin, M., Bhatti, P., & Bhatti, S. D. (2025). Optimizing heart disease forecasting: Bridging gaps in interpretability, efficiency, and scalability using machine learning. Biomedical Materials & Devices. https://doi.org/10.1007/s44174 025 00504 0
11. Khan, M. Z., Shaikh, S. A., Khan, A. A., Imroz, A., Salahuddin, M., Bhatti, P., & Bhatti, S. D. (2025). Artificial intelligence in dermatology: Current applications and future innovations. [PDF]. ResearchGate.https://www.researchgate.net/publication/396020302_ARTIFICIAL_INTELLIGENCE_IN_DERMATOLOGY_CURRENT_APPLICATIONS_AND_FUTURE_INNOVATIONS
12. H. Ahmad, G. Mumtaz, and M. Salahuddin, "A Hybrid Deep Learning Model for High-Accuracy Brain Tumor," Al-Aasar, vol. 2, no. 3, pp. 1–15, Aug. 2025, doi: 10.63878/aaj737.
13. Moeez, M., Mahmood, R., Asif, H., Iqbal, M. W., Hamid, K., Ali, U., & Khan, N. (2024). Comprehensive Analysis of DevOps: Integration, Automation, Collaboration, and Continuous Delivery. Bulletin of Business and Economics (BBE), 13(1).
14. S. Ali, Q. Li, and A. Yousafzai, "Blockchain and federated learning-based intrusion detection approaches for edgeenabled industrial IoT networks: A survey," Ad Hoc Networks, vol. 152, p. 103320, 2024.
15. R. Kimanzi, P. Kimanga, D. Cherori, and P. K. Gikunda, "Deep Learning Algorithms Used in Intrusion Detection Systems—A Review," arXiv preprint arXiv:2402.17020, 2024.
16. C. EL Asry, I. Benchaji, S. Douzi, and B. EL Ouahidi, "A robust intrusion detection system based on a shallow learning model and feature extraction techniques," Plos One, vol. 19, no. 1, p. e0295801, 2024.
17. A. Heidari and M. A. Jabraeil Jamali, "Internet of Things intrusion detection systems: A comprehensive review and future directions," Cluster Computing, vol. 26, no. 6, pp. 3753-3780, 2023.
18. O. H. Abdulganiyu, T. Ait Tchakoucht, and Y. K. Saheed, "A systematic literature review for network intrusion detection system (IDS)," International Journal of Information Security, pp. 1-38, 2023.
19. H. A. Hassan, E. E. Hemdan, W. El-Shafai, M. Shokair, and F. E. A. El-Samie, "Intrusion Detection Systems for the Internet of Thing: A Survey Study," Wireless Personal Communications, vol. 128, no. 4, pp. 2753-2778, 2023.
20. M. Alkasassbeh and S. Al-Haj Baddar, "Intrusion detection systems: A state-of-the-art taxonomy and survey," Arabian Journal for Science and Engineering, vol. 48, no. 8, pp. 10021-10064, 2023.
21. H. A. Hassan, E. E. Hemdan, W. El-Shafai, M. Shokair, and F. E. A. El-Samie, "Intrusion Detection Systems for the Internet of Thing: A Survey Study," Wireless Personal Communications, vol. 128, no. 4, pp. 2753-2778, 2023.

22. A. Alotaibi and M. A. Rassam, "Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense," Future Internet, vol. 15, no. 2, p. 62, 2023.
23. I. Debicha, R. Bauwens, T. Debatty, et al., "TAD: Transfer learning-based multi-adversarial detection of evasion attacks against network intrusion detection systems," Future Generation Computer Systems, vol. 138, pp. 185-197, 2023.
24. A. T. Azar, E. Shehab, A. M. Mattar, I. A. Hameed, and S. A. Elsaid, "Deep learning-based hybrid intrusion detection systems to protect satellite networks," Journal of Network and Systems Management, vol. 31, no. 4, p. 82, 2023.
25. Network Intrusion Detection Dataset, Kaggle, https://www.kaggle.com/datasets/sampadab17/network-intrusion-detec.
26. A. H. Balla, M. H. Habaebi, E. A. Elsheikh, M. R. Islam, and F. M. Suliman, "The Effect of Dataset Imbalance on the Performance of SCADA Intrusion Detection Systems," Sensors, vol. 23, no. 2, p. 758, 2023.
27. S. Ullah, J. Ahmad, M. A. Khan, et al., "TNN-IDS: Transformer neural network-based intrusion detection system for MQTT-enabled IoT Networks," Computer Networks, vol. 237, p. 110072, 2023.
28. C. Hazman, A. Guezzaz, S. Benkirane, and M. Azrour, "Toward an intrusion detection model for IoT-based smart environments," Multimedia Tools and Applications, pp. 1-2, 2023.
29. Elzaridi, K., & Kurnaz, S. (n.d.). Integration between network intrusion detection and machine learning techniques to optimizing network security. *Research Article*. Information Technologies Department, Altinbas University, İstanbul, Turkey, and Computer Engineering Department, Altinbas University, İstanbul, Turkey.
30. Feature Importance, Built-In (2011), https://builtin.com/data-science/feature-importance
31. Khan, M.Z., Shaikh, S.A., Shaikh, M.A., Khatri, K.K., Mahira Abdul Rauf, Kalhoro, A. and Muhammad Adnan (2023). The Performance Analysis of Machine Learning Algorithms for Credit Card Fraud Detection. International Journal of Online and Biomedical Engineering (iJOE), 19(03), pp.82–98. doi: https://doi.org/10.3991/ijoe.v19i03.35331.
32. Data Clustering Algorithms, Fuzzy C-Means Clustering Algorithm, https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm
33. Stochastic Gradient Descent (SGD), Tutorials Point, https://www.tutorialspoint.com/scikit_learn/scikit_learn_stochastic_gradient_descent.htm
34. Mirza Azam Baig, Sarmad Ahmed Shaikh, Kamlesh Kumar Khatri, Muneer Ahmed Shaikh, Muhammad Zohaib Khan and Rauf, A. (2023). Prediction of Students Performance Level Using Integrated Approach of ML Algorithms. International Journal of Emerging Technologies in Learning (iJET), 18(01), pp.216–234. doi: https://doi.org/10.3991/ijet.v18i01.35339
35. Towards Data Science, Logistic-Regression, https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148
36. Khan, M.Z., Khan, A.A., Laghari, A.A., Shaikh, Z.A., Khani, M.A.K., Morkovkin, D., Gavel, O., Shkodinsky, S., Makar, S. and Taburov, D. (2022). COMPARATIVE CASE STUDY: AN EVALUATION OF PERFORMANCE COMPUTATION BETWEEN SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBORS, K-MEAN, AND PRINCIPAL COMPONENT ANALYSIS. [online] Journal of Tianjin University Science and Technology.
37. K-Nearest Neighbors (K-NN) Algorithm, IBM (1911), https://www.ibm.com/topics/knn

38. Siddiqui, M., Kalwar, H.A., Khan, M.Z., Khan, M.A., Imroz, A., Kalwar, M.A. and Marri, H.B. (2023). Performance Analysis for the Diagnosis of COVID-19 Prediction by Mathematical Modeling & Simulation. [online] International Journal of Artificial Intelligence & Mathematical Sciences (IJAIMS). Available at: http://ijaims.smiu.edu.pk/ijaims/index.php/AIMS/article/view/47 [Accessed 1 Jul. 2023].