International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(14s) 2026 pp. 1285-1295
https://doi.org/10.70917/ijcisim-2026-4341
Article

Impact of Missing Data Imputation on Intrusion Detection Performance: An Empirical Evaluation on NSL-KDD

Jyoti Gupta[1], Hemant Pal[2]

[1]Department of Computer Science, Medicaps University, Indore, Madhya Pradesh, India.
Email: jyotig9414@gmail.com
[2]Department of Computer Science, Medicaps University, Indore, Madhya Pradesh, India.
Email: hemantpal.scs@gmail.com

Abstract: – For reliable detection performance, intrusion detection system uses machine learning algorithm for converting missing value data to superior quality.becauses of daily traffic on network dataset contain some values that are missing due to sensor failure, packet loss or concern about privacy. This makes intrusion detection system less precise. Often some research neglect data missing and some other use basic techniques to deal with it without exploring. In this paper we examine all thevarious missing data mechanism like MCAR, MAR and MNAR mechanism on NSL-KDD dataset. We also use variety of missing value imputation strategies such as stastical and machine learning based approaches. Based on result, Random Forest-based imputation significantly improves accuracy and the F1-scores under all missing situation. The findings highlight the importance mechanism-aware missing data treatment is for effective intrusion detection.
Keywords: - IDS, Missing Data, Data Imputation, Machine Learning, NSL-KDD, Network Security

1. Introduction

The continuous growth of digital communication technologies including cloud computing platforms and IOT environment has increasing the complexity of modern network. If the network expanded threats also increases. To address this challenge intrusion detection system has become an essential component of modern cyber security architecture. These system monitor network traffic and handle suspicious behavior. In current year IDS uses machine learning technique because they are capable to handle complex pattern to large volume of data. And distinguish between normal network behavior and malicious activities.

In practical scenarios, network monitoring data frequently contain incomplete or missing values. Such missing information may arise due to several factors, including packet loss during transmission, malfunctioning monitoring sensors, encryption processes that hide certain attributes, or deliberate manipulation by attackers attempting to evade detection. The presence of missing values can negatively affect the performance of machine learning models by reducing prediction accuracy, introducing bias during model training, and increasing the likelihood of false alarms. Furthermore, missing data do not occur randomly in many situations; instead, they often follow specific statistical patterns. In the world of data analysis, these patterns are commonly described as MCAR, MAR, and MNAR. Each of these mechanisms influences the distribution of dataset features differently, which can subsequently affect how effectively IDS models learn patterns related to network intrusions.

While values are missing at a random location (MCAR) it means they are missing on their own. While values are missing at random (MAR), thenthey havedisappeared due to the values of other. In MNAR, the value of one feature are missing given of the values of the same feature.

Although various studies have proposed advanced machine learning algorithms for intrusion detection, comparatively limited attention has been given to systematic analysis of how missing data patterns impact IDS performance and how different imputation strategies can mitigate these effects. Mean, median, or mode replacements are example of Traditional imputation techniques such as are computationally simple but may distort underlying feature relationships. Conversely, advanced methods—including KNN, MICEand deep learning–based reconstruction approaches—can preserve structural information but introduce computational overhead and model complexity. Determining which imputation strategy is most effective for IDS datasets therefore remains an open research problem.

To solve this issue, this paper investigates the influence of missing data mechanisms and imputation techniques on machine learning–based intrusion detection performance. Using benchmark datasets such as NSL-KDD, the research systematically introduces controlled missingness patterns, applies multiple imputation methods, and for performance evaluation itutilizesnormal metrics. By analyzing the interaction between missing data characteristics and model behavior, this work aims to provide a deeper understanding of how data incompleteness affects IDS reliability and to identify robust imputation strategies that enhance detection accuracy under realistic conditions.

2. Literature review:

Recent studies have demonstrated that missing data significantly influences machine learning performance, particularly when different missingness mechanisms are considered. Missing data is typically categorized into MCAR, MAR and MNAR, each affecting model behavior differently.

“Analyzing the Effect of Imputation on Classification Performance under MCAR and MAR Missing Mechanisms” This paper works onthe different imputation methodsbased on classification performance if there aremissing values along with MAR or MCAR mechanisms. It finds that Random Forest-based imputation methods generally perform well, while the effectiveness of imputation varies depending on the classifier and missingness pattern [1].

“Effect of Missing Data Types and Imputation Methods on Supervised Classifiers: an Evaluation Study:” This paper evaluates how different types of data missingand howimputation methods affect performance of five classification models. It highlights that classifiers are more sensitive to missing data patterns and that using the Matthews correlation coefficient provides a better assessment of their performance compared to accuracy alone [2].

“A simulation study on missing data imputation for dichotomous variables using statistical and machine learning methods” This paper focus mainly on imputation method of missing data in dichotomous variables, comparing eight techniques including traditional stastical and machine learning methods. It uses various machine learning methods , particularly SVM, ANN and Decision Tree, and offer higher accuracy and stability, emphasizing the importance of understanding variable correlations and distributions for effective imputation [3].

“A novel approach for handling missing data to enhance network intrusion detection system” This study highlights how important it is to correctly imputation missing data, particularly when it is related to IDS, where datasets commonly have a high incompleteness ratio. This research provides a novel approach termed DeepLearning_Based_MissingData_Imputation and analyses the way it performs with various models of machine learning [4].

"A survey on missing data in machine learning” ,This paper analyses an issue of missing data, such as types, mechanisms, and different approaches to handling it in various applications. It provides researchers direction regarding how to choose appropriate techniques. Using the Iris and ID fan datasets, a test evaluating the performance of KNN and random forests (RF) imputation revealed that the use of KNN performed better on the Iris dataset while RF greatly outperformed the performance of KNN on the ID fan dataset. The study arrives at a finding that no one method is consistently superior to another one, and that the efficiency of imputation algorithms depends heavily on the type of data [5].

“HANDLING MISSING VALUES IN NUMERIC DATASET USING MACHINE LEARNING TECHNIQUES: A REVIEW” this paper mainly work on large amount of raw data for extracting important information and transform it into a form that can used effectively. This paper mainly works on various classification strategies and compare them with their advantages and disadvantages [6].

"A Comparative Analysis of Machine Learning Imputation Techniques for MAR Missingness." This paper based on mar missingness and missing value imputed using KNN, MICE, and MissForest techniques. And compare all the results. After that miss forest method perform better other than other [7].

"Missing Value Estimation using Clustering and Deep Learning within Multiple Imputation Framework" with the help of this paper by enhancing the result of MICE algorithm through ensemble learning and deep neural network. This finding indicatesthe lead to better improvement and classification performance across various missingness type and percentage [8].

"Analysis of Machine Learning Based Imputation of Missing Data" this paper mainlyworks on three dataset from UCI repository. And compare the result of machine learning based method to traditional method. And result is machine learning method improve its accuracy across traditional methods across various dataset [9].

"Advanced methods for missing values imputation based on similarity learning." In this paper author challenge the missing value by proposing two hybrid imputation methods KI and FCKI .which combines K-nearest neighbor and iterative imputation technique. This method aim to enhance the accuracy against various dataset with different missing data types [10].

3. Methodology:

This research done by various steps from data source finding to then choosing and identifying correct dataset then prepare dataset according to requirement i.e. finding missing values in different percentage and then apply different imputation technique and evaluate them accordingly. There are various phases to do this work. This are -

1. **Dataset Description**: this experiment works on the NSL_KDD dataset. It is an enhancement of KDD Cup 1999 dataset.it is mainly used to address redundancy and class imbalance issues. It contains 41 features including both categorial and numerical attributes along with class labels indicating attack or normal traffic. The KDD train+ and KDD test+ subset employed for training evaluation respectively.

To ensure unbiased evaluation it was divided into two set called as training and testing set using stratified sampling so that class distribution preserved across both the set.

2. **Data preprocessing**:firstly use one hot encoding for transforming categorical attributes to numerical representations prior to a model was trained.Since Random Forest chosen classifier, is insensitive to variations in feature magnitude, feature scaling was not necessary. Following preprocessing, the dataset was divided into70 to 30% ratio of training and testing subsets. The test data remained unchanged to imitate real-world evaluation conditions, while the training data was used for model training and missing value injection.

Let the original dataset be represented as:

$$D = \{(P_i, Q_i)\}_{i=1}^{N}$$

Where:

- $P_i \in R^d$represents the feature vector,
- $Q_i \in \{1,2,\ldots,C\}$denotes the class label,
- N is number of samples,
- d is the number of features,
- C Is the total number of classes?

The dataset is divided into training and testing subsets:

$$D = D_{train} \cup D_{test}$$

Witha 70:30 split ratio.

3. **Missing data Techniques**:it was simulated according to missing mechanism like MCAR, MNAR and MAR.

3.1 MCAR

In this mechanism, missingness is not depending on both the feature like observed and unobserved values:

$$R(M_{ij} = 1 \mid P, Q) = r$$

Where:

- M_{ij} indicates whether feature j of sample i is missing,
- r Is the missing rate.

Random masking is applied uniformly across the feature matrix.

3.2 MAR

In the context of MAR, a missingvalue depends only on the observed variable

$$R(M_{ij} = 1 \mid P_{obs}, P_{mis}) = R(M_{ij} = 1 \mid P_{obs})$$

Missing values are includedconditionally based on features which are selected.

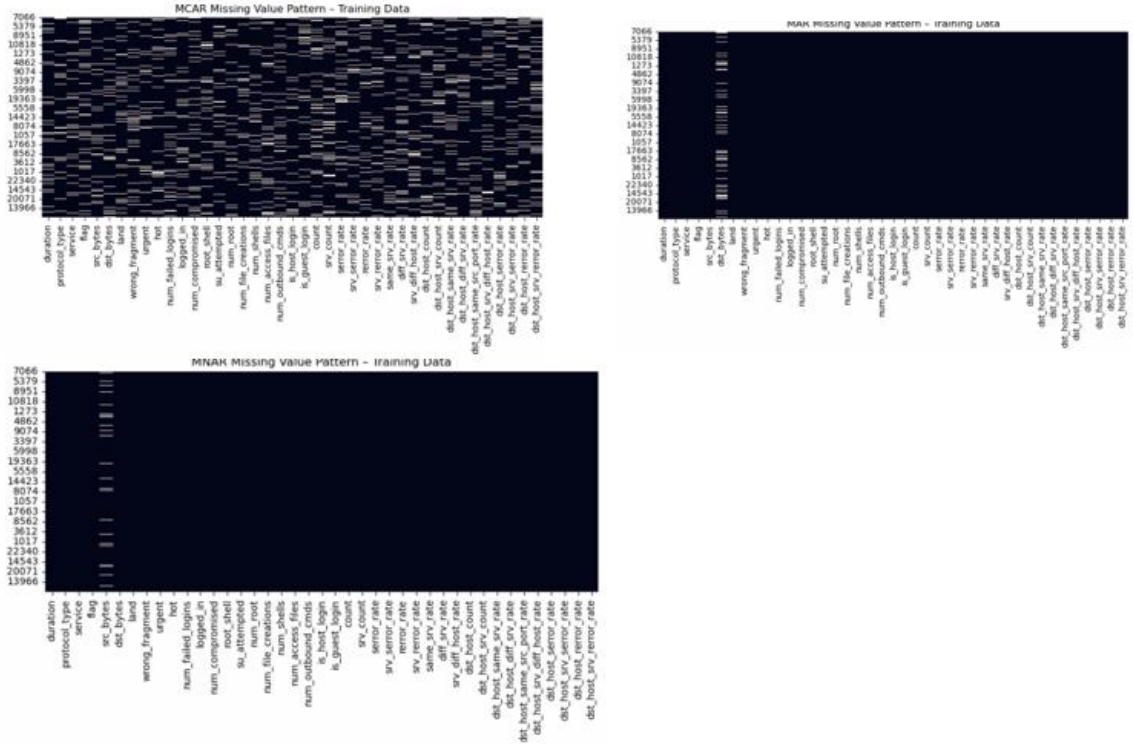
3.3 MNAR

In MNAR, a missing value is depends on the value that was not recognized.

$$R(M_{ij} = 1 \mid P_{mis}) \neq R(M_{ij} = 1)$$

This mechanism simulates systematic information loss by masking values according to feature-dependent thresholds.

4. **Missing value injection strategy:** For analyzing the effect of incomplete data, artificial missing value were injected into the training dataset at predefine missingrate like 5%, 10%, 20% and 30%. The test dataset checks the performanceof missingness pattern based on MAR, MCAR and MNAR On different missingness percentage.



5. **Imputation Strategies**:

Mechanism-specific imputation methods were used since multiple missing data mechanisms impose different statistical assumptions, mechanism specific imputation were employed:

Mechanism	Imputation Method

MCAR	Mean Imputation
MAR	Correlation-based Imputation
MNAR	Median Imputation

Classification Model: Due to its robustness, capability for handling high-dimensional data, resistance to overfitting, and high accuracy in intrusion detection tasks, a Random Forest classifier was chosen as the base learning model. Even with partially incomplete data, Random Forest produces stable and dependable classification performance by building multiple decision trees using bootstrap sampling and aggregating their predictions through majority voting.

Models' selection for prediction: This paper uses statical and machine learning based model for predict the missing value and check the accuracy of each model. This study provides comprehensive practical analysis of missing value in ids using supervised learning algorithms like KNN, MICE, RF, SVM, and Gradient Boosting. Mainly following models are used in this study -

5.1 K nearest neighbors – It is supervised learning algorithm which uses for regression classification tasks. It utilizes distance to arrange points with based on the majority vote of nearest neighbour.KNN is sensitive to data quality.

For a sample with missing feature x_{ij}, the imputed value is computed using the average of its $k$nearest neighbors:

$$\hat{x}_{ij} = \frac{1}{k} \sum_{l \in N_k(i)} x_{lj}$$

Where:

- $N_k(i)$= set of k nearest neighbors of sample i
- Distance metric: Euclidean distance

$$d(x_i, x_l) = \sqrt{\sum_{m=1}^{d} (x_{im} - x_{lm})^2}$$

5.2 Iterative Imputation (MICE) – Iterative imputation also called as MICE. Is a multivariant missing value treatment technique which feature containing missing value is observed with other feature in the dataset.

MICE models each feature with missing values as a function of other features:

$$x_j = f_j(X_{-j}) + \epsilon$$

Where:

- X_{-j}= all features except j
- f_j= regression model (e.g., Random Forest or Linear Regression)
- ϵ= error term

Imputation is performed iteratively until convergence.

5.3 RF Imputation–itis an ensemble learning method. That create multiple decision trees and merge their output.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

Where:

- B= number of trees
- $T_b(x)$= prediction of the b-th tree

For classification (majority voting):

$$\hat{y} = mode(T_1(x), T_2(x), \ldots, T_B(x))$$

5.4 Gradient Boosting – itcreate sequential tree. And every new tree corrects the error of previoustree.

Where:

- $h_m(x)$= weak learner (tree)
- γ_m= learning rate weight

5.5 Support Vector Machine: it finds hyper plane that maximizes margin between classes. and SVM decision boundary defined by following function-

The optimization objective is:

$$min_{w,b,\xi} \frac{1}{2} || w ||^2 + C \sum_{i=1}^{N} \xi_i$$

Subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

Where:

- $C$is regularization parameter,
- ξ_iAre slack variables.

For nonlinear cases, the RBF kernel is applied:

$$K(x_i, x_j) = exp(-\gamma || x_i - x_j ||^2)$$

S.NO.	Imputation	Model	Accuracy	Recall	F1 Score
1	Mean	RF	0.986915	0.986690	0.986660
2	MICE	RF	0.986915	0.986690	0.986660
3	Mean	Gradient Boosting	0.979818	0.979141	0.979411
4	MICE	Gradient Boosting	0.979818	0.979141	0.979411
5	Mean	KNN	0.970503	0.969393	0.969893
6	MICE	KNN	0.970503	0.969393	0.969893
7	Mean	SVM	0.599246	0.535071	0.436726
8	MICE	SVM	0.599246	0.535071	0.436726

But mainly works on random forest classifier due to its robustness and accuracy. And it handles high volume of data, resistance to overfitting and performs better with ids dataset.

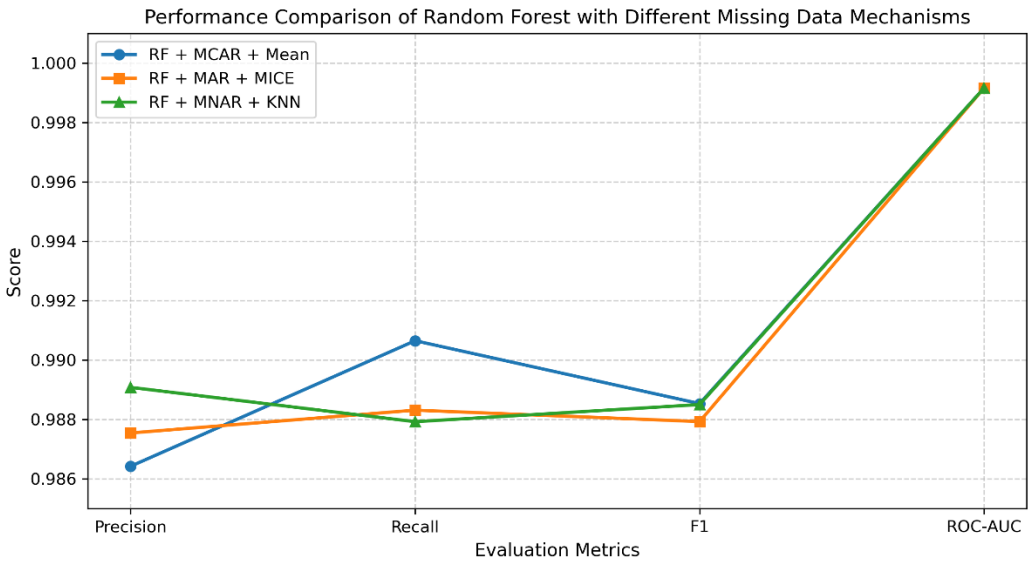


Fig. Performance comparison of the Random Forest classifier under different missing data mechanisms and imputation strategies.

The results indicate that MNAR with KNN imputation achieves the highest precision, while MCAR with mean imputation provides the best recall and F1-score

There are various supervised machine learning algorithm apply on NSL-KDD dataset and predict the missing value on various missingness pattern. After that compare all models based on the accuracy and find out the best model based on missing percentage robustly. And give the result based on accuracy and f1-macro score.

6. **Evaluation Metrics**: Both accuracy and the macro-averaged F1-score were utilized to evaluate the model's performance. While accuracy offers a general indicator of classification correctness, it can be misleading in datasets which are imbalanced like intrusion detection benchmarks. Because it provides each class the same weight and offers an accurate evaluation of detection performance across majority and minority attack categories, the macro-averaged F1-score was chosen as the primary metric.

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision (Macro)

$$Precision = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c}$$

Recall (Macro)

$$Recall = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}$$

F1-Score (Macro)

$$F1 = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c}$$

Motivation: This study performs an in-depth study of data missing techniques in systems that detect breaches to address these limitations. The proposed work does not based on a one missing scenario but itworks on different scenario based conditions with different missing rates and examines it performance on different missing rates. This mechanism-aware approach provides us a better idea of how missing data affects the reliability of IDS and gives us useful guidance on how to decide on the best ways to deal about values that are missing.

4. Result

The findings show that MAR missed values always produces the most stable performance in classification among all missing rates, while the use of MCAR results in a major performance decrease. MNAR is works effectively when there are not lots of missing values, but it becomes a little less reliable as the amount of values that is missing goes up.

S.No.	Missing Mechanism	Rate	Accuracy	F1_Macro
1	MCAR	0.05	0.986250	0.985980
2	MAR	0.05	0.987137	0.986885
3	MNAR	0.05	0.986472	0.986206
4	MCAR	0.10	0.987137	0.986873
5	MAR	0.10	0.986693	0.986434
6	MNAR	0.10	0.986250	0.985979
7	MCAR	0.20	0.984697	0.984366
8	MAR	0.20	0.987137	0.986883
9	MNAR	0.20	0.986472	0.986206
10	MCAR	0.30	0.982479	0.982084
11	MAR	0.30	0.985584	0.985303
12	MNAR	0.30	0.983367	0.983040

Table presents IDS performance under varying missing rates and missing data mechanisms using Random Forest and macro F1-score.

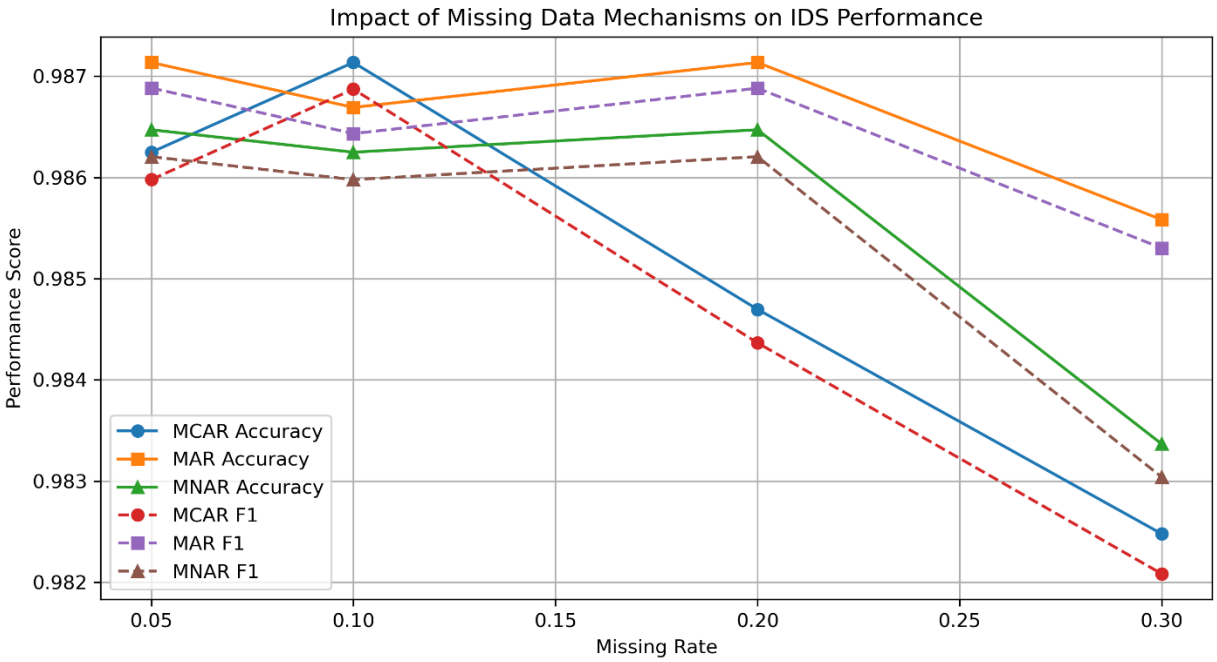


Fig. shows Impact of missing rate on IDS accuracy and f1 score across different missing data mechanisms

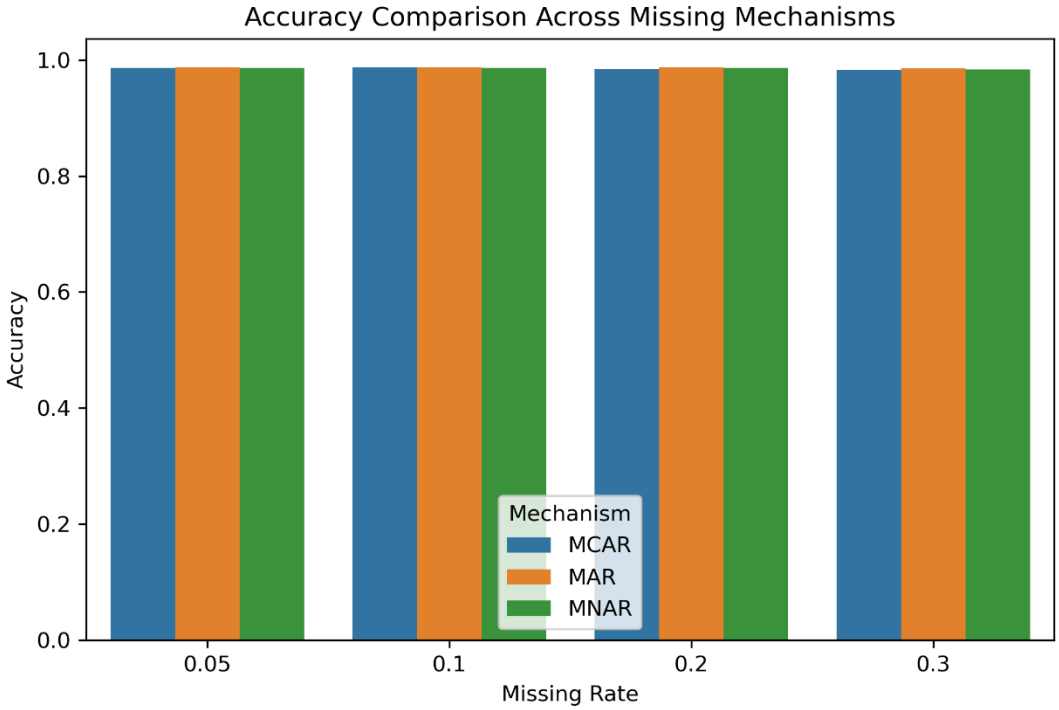


Fig. shows Comparison of classification accuracy under MCAR, MAR, and MNAR mechanisms.

.

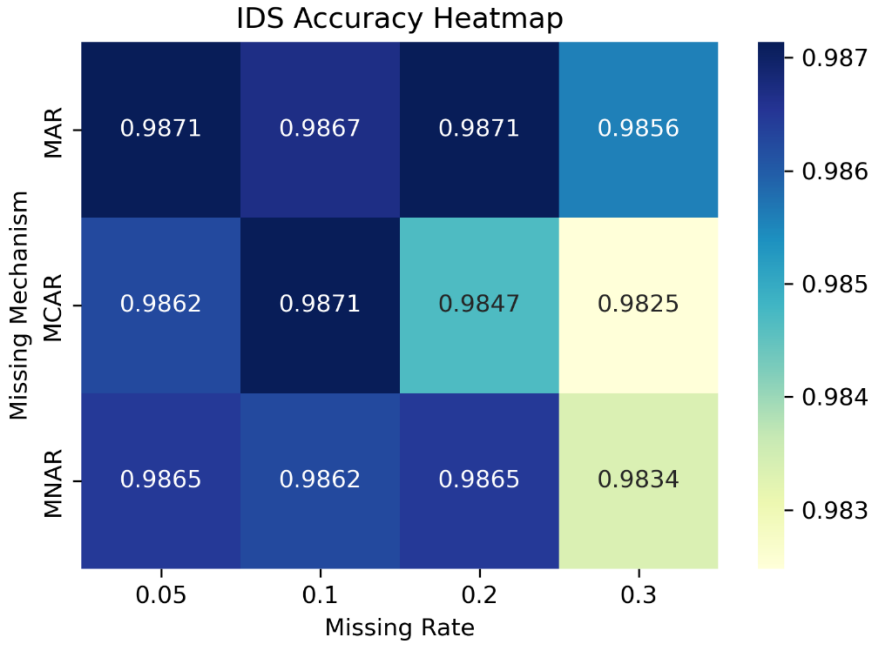


Fig. shows Heatmap representation of IDS accuracy showing the performance variation across missing rates and mechanisms.

Experimental Design: The method by which the experiment performed:

1. Load NSL_KDD dataset and preprocess it anddivided into trainingset (70%)and testing set (30%).
2. Add missing values to prepare the data based on MCAR, MAR, as well as MNAR Mechanism.
3. After that apply mechanism specific imputation.
4. Train the classifier on imputed data.
5. After that evaluate the performance.
6. Repeat 2-5 steps across multiple missing rates.
7. Analyze and compare results.

5. Conclusion

This paper mainly workin a comprehensive analysis of missing data in machine learning–based IDS. Unlike most existing IDS studies that implicitly assume randomly missing data, but this paper mainly focus on evaluated three fundamental missing data mechanisms like MCAR, MAR, and MNAR—under various level of missing percentage. Missing value injected on nsl-kdd set and evaluates the classification performance on different ranging 5% to 30% on training data. The experimental results based on both the performance of intrusion detection vary by the amount of missing data and pattern of missing data. If it has some missing value all the methods gave same result but missing rate increases various performance trends to begin to show up.MCAR showed the most visible degredation.which implies that it erased data evenly across all features. on the other hand MAR had the most stable performance. This shows that right imputation technique can help the missingness that depends on features.

MCAR demonstrated the most distinct degradation, which is a sign of its uniform information loss across features. On the reverse side, MAR had the most stable performance, which means that using the right imputation method can help with the missing information. MNAR displayed a competitive performance, but it did not completely stable, and this indicates the value dependent missing data can lead to systematic bias. These findings confirms that the missing information mechanism is more important than the missing data volume for evaluate IDS effectiveness.

This research can be enhancing in future in various direction. One good piece of advice is use advanced imputation technique those based on, deep learning or generative approaches, for better deal with complex patterns of

missing data. Another direction is for searching toward the way mechanism-aware missing value handling performs in real-time.

References:

1. Buczak, Philip, Jian-Jia Chen, and Markus Pauly. "Analyzing the effect of imputation on classification performance under MCAR and MAR missing mechanisms." *Entropy* 25.3 (2023): 521.
2. Gabr, Menna Ibrahim, Yehia Mostafa Helmy, and Doaa Saad Elzanfaly. "Effect of missing data types and imputation methods on supervised classifiers: An evaluation study." *Big data and cognitive computing* 7.1 (2023): 55.
3. Ge, Yingfeng, Zhiwei Li, and Jinxin Zhang. "A simulation study on missing data imputation for dichotomous variables using statistical and machine learning methods." *Scientific Reports* 13.1 (2023): 9432.
4. Tahir, Mahjabeen, et al. "A novel approach for handling missing data to enhance network intrusion detection system." *Cyber Security and Applications* 3 (2025): 100063.
5. Emmanuel, Tlamelo, et al. "A survey on missing data in machine learning." *Journal of Big data* 8.1 (2021): 140.
6. Kaur, Kamaljeet, Amrit Kaur, and Navjot Kaur. "HANDLING MISSING VALUES IN NUMERIC DATASET USING MACHINE LEARNING TECHNIQUES: A REVIEW." *JOURNAL PUNJAB ACADEMY OF SCIENCES* 23 (2023): 217-227.
7. Tiwaskar, Shweta, Sandip Thite, and Rashid Mamoon. "A Comparative Analysis of Machine Learning Imputation Techniques for MAR Missingness." *Advances in Nonlinear Variational Inequalities* 28 (2025): 46-57.
8. Samad, Manar D., Sakib Abrar, and Norou Diawara. "Missing value estimation using clustering and deep learning within multiple imputation framework." *Knowledge-based systems* 249 (2022): 108968.
9. Rizvi, Syed Tahir Hussain, et al. "Analysis of machine learning based imputation of missing data." *Cybernetics and Systems* 56.6 (2025): 818-832.
10. Fouad, Khaled M., et al. "Advanced methods for missing values imputation based on similarity learning." *PeerJ Computer Science* 7 (2021): e619.