Handling Missing Data Techniques: A Meta-Analysis

RAED ALAZAIDAH
Department of Data Science and Artificial Intelligence,
Faculty of Information Technology,
Zarqa University,
Zarqa,
JORDAN

Abstract: - The predictive performance of any classification or regression model highly depends on the quality of the collected data. Most of time datasets suffer from the problem of missing values, and hence, several techniques have been proposed to handle the problem of missing values. Consequently, this paper aims to quickly survey the most well-known techniques that handle missing data, and identify the best one to use concerning several issues such as the ratio of missing values, type of attributes in the dataset, number of instances, and number of class labels. Hence, seven different and well-known missing values handling techniques have been evaluated and compared using five datasets with different characteristics concerning the Accuracy metric. The results revealed that the K- Means technique is the most appropriate technique to handle the problem of missing data and the SMO classifier is the best choice to use as a classification model in case of missing data.

Key-Words: - Classification, learning, meta-analysis, missing values, pre-processing steps, prediction.

Received: April 26, 2024. Revised: November 25, 2024. Accepted: December 23, 2024. Published: February 14, 2025.

1 Introduction

Classification is one of the main and core tasks in machine learning and data science which aims to accurately predict the class label for a new unseen case or example, [1]. Classification has attracted many scholars in the last few decades due to its wide range of real-life applications such as fraud detection, medical diagnosis, spam filtering, document identification, speech recognition, and several other applications, [2].

Classification is categorized into Single Label Classification(SLC) and Multi Label Classification (MLC) [2]. In SLC, all instances must be attached and linked to only one class label, while in MLC instances may linked to one class label or even more. SLC is divided into two subcategories: Binary Classification (BC) and Multi Class Classification (MCC), [3]. In the former subcategory (BC), the number of classes are only two, while in the later subcategory (MCC) several classes in the dataset is more than two classes. MLC is the most complicated type and follows an exponential growth that equals 2m, where m is the total number of class labels, [4]. Class labels in SLC are always mutually exclusive in the contrary of class labels in MLC, [4].

Many classification algorithms (classifiers) have been proposed to handle the classification task in machine learning and data science. The predictive performance of any of these classifiers varies according to the characteristics of the dataset being processed as well as the quality of this dataset. In many cases, the low quality of the dataset badly affects the predictive performance of the classifiers, [5].

The low quality of the dataset arises due to several reasons such as high dimensionality, inconsistency formatting, data integration errors, and missing values, [6]. This research is more interested in the last reason for data low quality, that is, missing values. The main reasons for missing values in datasets are data entry errors, incorrect measurements, and equipment errors.

According to several studies, [7], [8], [9], missing values badly affect the accuracy of any prediction algorithm. Hence, several techniques have been proposed to handle the problem of missing values before training the classification algorithm on the considered dataset. Consequently, this research aims to identify the most suitable technique to handle missing values concerning several characteristics such as the percentage of missing values in the dataset, types and number of attributes,

number of instances, and number of classes in the dataset.

Hence, seven techniques for handling missing values have been evaluated and compared using five datasets with different characteristics, with respect to fifteen classifiers that belong to five well-known learning strategies. Another implicit objective of this research is to identify the most appropriate classifier to use with datasets that suffer from the problem of missing values.

The rest of this paper is organized as follows: Section 2 surveys the related work while Section 3 provides the methodology and datasets. Section 4 provided the empirical results and the main finding of this research. Finally, Section 5 concludes and provides some future directions.

2 Related Work

This research considers seven different techniques for handling missing data. The simplest technique is the ignore technique. This technique ignores any missing values in the dataset and proceeds with the analysis, [10]. This technique is good to adopt with datasets that have a low percentage of missing values or with classifiers that implicitly handle the problem of missing data like decision trees and random forest classifiers.

The second technique for handling missing data is called K-Means, [11]. This technique combines the task of clustering with the approach of soft computing, and uses a Fuzzy based clustering algorithm, in order to estimate the missing values. K-means has been tested on two datasets using the Root Mean Squared Error (RMSE) where it showed a better performance than the original K-Means algorithm, [11].

The Most Common technique for handling missing data has been proposed in [12]. This technique considers several variables to estimate the missing value such as the closet fit for symbolic attributes, the most common global value, and the average of the numerical attributes. Three datasets have been used to test the proposed technique concerning several evaluation metrics such as Sensitivity, Error-rate, Area Under Curve (AUC), and Receiver Operating Characteristic (ROC). The evaluation phase of the Most Common technique showed a good performance compared with other traditional techniques for handling missing data, especially in the case of having a high ratio of missing data, [12].

The fourth technique for handling the missing data is the All-Possible technique, [13]. This technique replaces the unknown value for an attribute with all possible values known in that attribute. This technique suffers from the high complexity especially with high dimensional datasets that have a large number of possible values for its attributes.

The Bayesian Principal Component Analysis (BPCA), [14] handles missing values by estimating them using Principal Component Analysis (PCA). BPCA has been tested on the DNA microarray dataset where it showed a remarkable performance compared with singular value decomposition and K-Means techniques.

The Local Least Squares Imputation (LLSI) technique is another technique to handle the problem of missing data, [15]. LLSI utilizes the local similarity structure in the dataset in order to estimate the missing value. LLSI Showed a competitive performance compared with other missing values handling techniques on several different datasets.

The Single Vector Decomposition Imputation (SVDI) technique works by obtaining a set of orthogonal expression patterns which are then linearly combined to estimate the missing values in the dataset, [16]. SVDI has been evaluated using datasets with different ratios of missing values that range from 1-20%. SVDI has been compared against other traditional techniques and showed an acceptable performance.

3 Methodology, Results and Discussion

This section represents the core of the research. At first, the methodology of the research is provided in Section 3.1. Then, the considered datasets are described in Section 3.2.

3.1 Methodology of the Research

Figure 1 depicts the main steps of the adopted methodology. The first step begins with collecting secondary datasets with different characteristics and features. The second step considers applying several missing values techniques on the considered datasets. The third step trains fifteen different classifiers that belong to five learning strategies on the pre-processed datasets. The final step analyses the results of the trained classifiers considering the missing values techniques being used concerning the Accuracy metric. More information regarding these steps is provided in the following subsections.

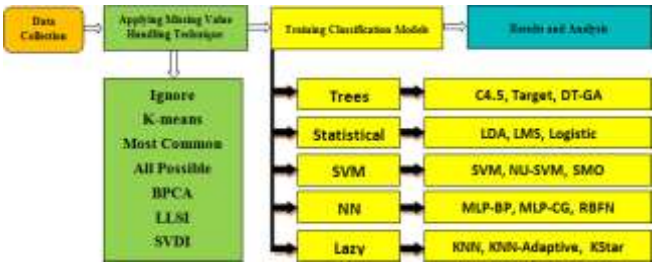

Fig. 1: Main steps in research methodology

3.2 Description of the Considered Datasets

Five datasets are considered in this research as depicted in Table 1. These datasets have been chosen to reflect different characteristics and features. For example, the number of instances(examples) varies in the datasets from 155 examples to 8993 examples. Also, the ratio of the missing values (MV's Ratio) starts at 1.98% as in the Cleveland dataset, and increases up to 48.39% as in the Hepatitis dataset. Moreover, the considered datasets have different types and numbers of attributes as well as different numbers of class labels (Classes). In Table 1, "R" refers to the number of "Real" attributes in the dataset, "I" refers to the number of "Integer" attributes in the dataset, and "N" refers to the number of "Nominal" attributes in the dataset.

Table 1. DatasetsDescription

Name	Attributes			Examples	Classes	MV's Ratio
	R	I	N			
Cleveland	13	0	0	303	5	1.98%
Marketing	0	13	0	8993	9	23.54%
Bands	13	6	0	539	2	32.28%
House votes	0	0	16	435	2	46.67%
Hepatitis	2	17	0	155	2	48.39%

The considered datasets are all available on the KEEL website. KEEL is short for Knowledge Extraction based on Evolutionary Learning, [17]. KEEL is a well-known tool that has been used extensively in the domains of machine learning, data mining, and data science.

All Missing values techniques and all classifiers considered in this research have been used with their default settings and parameters as implemented in KEEL.

4 Evaluation Results of the Considered Missing Values Handling Techniques and Classifiers

The next two subsections provide the evaluation results that will help in achieving the main two objectives of this research. First, the evaluation results for the considered missing values handling techniques are provided in Section 4.1. Then, the evaluation results of the considered fifteen classifiers are provided in Section 4.2.

4.1 Identifying the Best Missing Values Handling Technique

This section provides the evaluation results for the considered missing values techniques as well as the considered classification models. These results help to achieve the main two objectives of this research where the first objective is to identify the best missing value handling technique, and the second objective is to identify the best classification model that can effectively suit datasets with missing values. Consequently, seven different and well-known techniques for handling missing values have been chosen and compared. These techniques are: ignore, K-Means, most common, All Possible, BPCA, LLSI, and SVDI. More information regarding these techniques has been provided in Section 2.

Moreover, fifteen different classification models have been chosen to be trained on datasets with missing values and using one of the previously mentioned missing values techniques. These classification models are C4.5, DT-GA, and Target from Trees learning strategy, LDA, Linear-LMS, Logistic from Statistical learning strategy, C-SVM, NU-SVM, and SMO classifiers from SVM strategy, MLP-PB, MLP-CG, and RBFN from NN learning strategy, KNN, KNN-Adaptive, and K Star from Lazy learning strategy, [18], [19], [20], [21], [22].
All these techniques and classifiers have been used with their default parameters as implemented in KEEL. These techniques and classifiers have been compared using the Accuracy metric which is computed using the following equation, [23], [24], [25], [26], [27]:

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

Where:
TP=True Positive predictions. TN=True Negative predictions. FP=False Positive predictions, and FN=False Negative predictions, [28], [29], [30], [31], [32].

Table 2 depicts the evaluation results for the seven considered missing values techniques concerning fifteen different classifiers on the Cleveland dataset.

Table 2. Evaluation Results Using Cleveland Dataset

Classifier	Ignore	K-Means	Most Common	All Possible	BPCA	LLSI	SVDI
C4.5	0.600	0.600	0.600	0.600	0.466	0.600	0.600
Target	0.566	0.533	0.566	0.466	0.400	0.533	0.566
DT-GA	0.500	0.500	0.566	0.500	0.366	0.566	0.500
LDA	0.600	0.600	0.600	0.633	0.400	0.600	0.600
Linear-LMS	0.633	0.633	**0.633**	0.511	0.400	0.633	0.633
Logistic	**0.700**	0.633	**0.633**	0.511	0.433	**0.666**	0.633
C-SVM	0.533	0.533	0.466	0.511	0.400	0.533	0.533
NU-SVM	0.533	0.533	0.533	0.533	0.400	0.533	0.533
SMO	0.633	**0.737**	**0.633**	0.633	0.500	**0.666**	**0.666**
MLP-BP	0.600	0.533	0.466	0.633	0.466	0.533	0.400
MLP-CG	0.533	0.666	0.533	**0.700**	0.400	0.600	0.500
RBFN	0.366	0.400	0.400	0.433	0.266	0.400	0.400
KNN	0.533	0.533	0.533	0.533	0.433	0.533	0.533
KNN-Adaptive	0.600	0.600	0.600	0.600	0.466	0.600	0.600
KStar	0.566	0.566	0.566	0.533	**0.566**	0.566	0.566

According to Table 2, K-means is the best technique to handle missing values on the Cleveland dataset since the highest Accuracy value has been achieved by the SMO classifier using K-means technique. The second-best value for the Accuracy metric has been achieved by the Logistic classifier when utilizing the Ignore technique. Considering the classification model, SMO is the best one as it manages to achieve the highest Accuracy value four times with four different missing values handling techniques. The logistic classifier is the second-best classification model to handle missing values.

Table 3 depicts the evaluation results for the seven considered missing values techniques with respect to fifteen different classifiers on the Marketing dataset.

Based on Table 3, the highest Accuracy value has been achieved by the SMO classifier when utilizing the K-Means technique. Also, the accuracy of the SMO classifier ranges from 0.175% when utilizing the BPCA technique to 0.444% with the K-Means technique. This indicates the significance of using the most appropriate missing value handling technique. Therefore, it is concluded from Table 3 that K-Means is the best choice to handle missing values on the Marketing dataset. Considering the classification model, the SMO classifier achieved the highest Accuracy value five times, and thus, it is the best classifier to handle missing data among the fifteen considered classifiers.

Table 3. Evaluation Results Using Marketing Dataset

Classifier	Ignore	K-Means	Most Common	All Possible	BPCA	LLSI	SVDI
C4.5	0.279	0.307	0.307	0.218	0.155	0.297	0.211
Target	0.211	0.274	0.290	0.105	0.153	0.295	0.211
DT-GA	0.250	0.276	0.274	0.127	0.138	0.295	0.211
LDA	0.303	0.301	0.298	0.133	0.137	0.300	0.211
Linear-LMS	0.312	0.304	0.304	0.143	0.163	0.301	0.211
Logistic	0.317	0.308	0.313	0.162	0.135	0.314	0.211
C-SVM	0.335	0.412	0.331	**0.269**	0.145	0.350	0.320
NU-SVM	0.292	0.256	0.283	0.208	0.123	0.253	0.200
SMO	**0.350**	**0.444**	0.330	0.230	**0.175**	**0.360**	**0.333**
MLP-BP	0.242	0.194	0.177	0.104	0.104	0.174	0.212
MLP-CG	0.338	0.318	**0.334**	0.154	0.135	0.329	0.212
RBFN	0.212	0.200	0.165	0.046	0.107	0.148	0.212
KNN	0.266	0.261	0.270	0.186	0.132	0.270	0.160
KNN-Adaptive	0.285	0.275	0.277	0.160	0.147	0.286	0.160
KStar	0.303	0.350	0.284	0.128	**0.175**	0.291	0.160

Table 4 depicts the evaluation results for the seven considered missing values techniques with respect to fifteen different classifiers on the Bands dataset.

Table 4. Evaluation Results using Bands Dataset

Classifier	Ignore	K-Means	Most Common	All Possible	BPCA	LLSI	SVDI
C4.5	0.485	0.703	0.666	**0.690**	0.555	0.648	0.611
Target	0.628	0.574	0.574	0.574	**0.574**	0.574	0.648
DT-GA	0.542	**0.759**	0.611	0.600	0.518	0.629	0.611
LDA	**0.657**	0.648	0.611	0.633	0.407	0.666	0.648
Linear-LMS	**0.657**	0.648	0.611	0.623	0.407	0.666	0.648
Logistic	**0.657**	0.648	0.611	0.600	0.407	0.666	0.629
C-SVM	**0.657**	0.648	0.685	0.685	0.555	0.685	**0.703**
NU-SVM	**0.657**	0.648	0.685	0.685	0.555	0.685	0.606
SMO	**0.657**	0.629	0.611	0.611	0.500	0.629	0.648
MLP-BP	0.371	0.425	0.537	0.412	0.481	0.518	0.555
MLP-CG	0.628	0.685	0.666	0.644	0.555	0.500	0.611
RBFN	0.600	0.574	0.629	0.566	0.555	0.611	0.629
KNN	0.628	0.685	0.648	0.601	0.462	**0.740**	0.629
KNN-Adaptive	**0.657**	**0.759**	**0.740**	0.644	0.481	0.703	0.685
KStar	0.628	0.574	0.648	0.587	**0.574**	0.574	0.648

From Table 4, it is clear that K-Means is the best technique to handle missing values since the highest Accuracy has been achieved by the KNN-Adaptive classifier when utilizing the K-Means technique. Also, the DT-GA classifier achieved the

highest Accuracy value with the same missing values handling technique (K-Means). Considering the highest Accuracy achieved value, KNN-Adaptive and DT-GA are the best classification models among the fifteen considered classifiers. KNN classifier which belongs to the Lazy learning strategy showed the second-best predictive performance when utilizing the LLSI missing values handling technique as shown in Table 4.

Table 5 depicts the evaluation results for the seven considered missing values techniques with respect to fifteen different classifiers on the House votes dataset.

Table 5. Evaluation Results on House Votes Dataset

Classifier	Ignore	K-Means	Most Common	All Possible	BPCA	LLSI	SVDI
C4.5	0.900	**0.977**	**0.977**	**0.974**	0.454	0.648	0.840
Target	0.900	0.954	0.954	0.954	0.454	0.574	0.863
DT-GA	0.900	**0.977**	**0.977**	0.900	0.454	0.629	0.863
LDA	0.900	0.954	0.954	0.954	0.454	0.666	0.863
Linear-LMS	0.900	0.954	0.954	0.954	0.454	0.666	0.863
Logistic	0.850	**0.977**	**0.977**	0.879	0.454	0.666	0.863
C-SVM	0.900	0.954	0.954	0.954	0.454	0.685	0.886
NU-SVM	0.900	0.954	0.954	0.954	0.454	0.685	0.886
SMO	0.900	**0.977**	**0.977**	**0.974**	**0.477**	0.629	0.863
MLP-BP	0.900	0.954	0.954	0.954	0.454	0.518	0.818
MLP-CG	**0.950**	0.952	**0.977**	0.844	0.431	0.500	0.818
RBFN	0.900	**0.977**	0.954	0.917	**0.477**	0.611	**0.909**
KNN	0.900	0.931	0.931	0.900	0.454	**0.740**	0.863
KNN-Adaptive	0.900	**0.977**	**0.977**	0.900	0.431	0.703	0.886
KStar	0.800	0.888	0.863	0.878	**0.477**	0.574	0.750

For dataset, House votes and as shown in Table 5, two missing values handling techniques showed superior performance compared with the other five techniques. These techniques are K-Means and Most common since the highest Accuracy value has been achieved using these techniques, and, six classifiers managed to achieve the highest Accuracy value when utilizing these two techniques. Considering the best classification model, several classifiers managed to achieve the highest Accuracy results. These classifiers are C4.5, DT-GA, Logistic, SMO, RBFN, KNN-Adaptive, and MLP-CG.

Table 6 depicts the evaluation results for the seven considered missing values techniques with respect to fifteen different classifiers on the Hepatitis dataset.

Table 6. Evaluation Results on Hepatitis Dataset

Classifier	Ignore	K-Means	Most Common	All Possible	BPCA	LLSI	SVDI
C4.5	**0.875**	0.875	0.750	0.750	0.687	0.875	0.812
Target	**0.875**	0.875	0.750	0.750	**0.750**	0.875	0.750
DT-GA	**0.875**	0.812	0.750	0.750	**0.750**	0.875	0.750
LDA	0.750	0.875	0.812	0.800	**0.750**	0.750	0.812
Linear-LMS	**0.875**	0.812	0.812	**0.812**	**0.750**	0.750	0.812
Logistic	0.750	0.812	0.750	**0.812**	0.687	0.750	0.812
C-SVM	**0.875**	0.812	0.750	0.777	0.687	0.750	0.750
NU-SVM	**0.875**	0.812	0.750	0.777	0.687	0.750	0.750
SMO	**0.875**	**0.937**	0.812	**0.812**	0.687	**0.937**	**0.937**
MLP-BP	0.750	0.875	**0.875**	0.750	0.562	**0.937**	0.687
MLP-CG	**0.875**	0.812	0.687	**0.812**	0.687	0.812	0.750
RBFN	0.625	0.687	0.812	**0.812**	**0.750**	0.875	0.750
KNN	0.750	0.875	**0.875**	0.767	**0.750**	0.812	0.750
KNN-Adaptive	0.750	0.812	0.812	0.699	**0.750**	0.812	0.750
KStar	**0.875**	0.750	0.750	0.750	0.687	0.750	0.750

From Table 6, the highest Accuracy value has been achieved when utilizing three techniques for handling missing data (K-Means, LLSI, and SVDI). SMO classifier showed a superior performance compared with the other considered classifiers. It managed to achieve the highest Accuracy value five times concerning the missing values handling techniques being used.

Table 7 summarizes the results for Table 2 to Table 6 considering the highest achievable Accuracy result concerning the techniques of handling missing values being used.

Table 7. Best Missing Value Handling Technique Based on the Highest Achievable Accuracy

Dataset	Ignore	K-Means	Most Common	All Possible	BPCA	LLSI	SVDI
Cleveland	0.700	**0.737**	0.633	0.700	0.566	0.666	0.666
Marketing	0.350	**0.444**	0.334	0.269	0.175	0.360	0.333
Bands	0.657	**0.759**	0.740	0.690	0.574	0.740	0.703
House votes	0.950	**0.977**	**0.977**	0.974	0.477	0.740	0.909
Hepatitis	0.875	**0.937**	0.875	0.812	0.750	**0.937**	**0.937**

Without any doubt and based on Table 7, the K-Means technique is the best one in handling missing values compared with the other six techniques. The highest Accuracy achieved by the different classifiers has been always using the K-Means technique. For datasets with large ratios of missing values such as House votes and Hepatitis, Most Common, LLSI, and SVDI techniques also showed an excellent performance. K-Means is the best choice for datasets with different ratios of missing

values as well as different types of attributes and examples as well.

Moreover, and for more assurance of the best technique to handle missing values, Table 8 summarizes the Average Accuracy for the fifteen classifiers concerning the missing values technique being used.

Table 8. Accuracy Average for the Fifteen Classifiers with Respect to Missing Value Technique Being Used

Dataset	Ignore	K-Means	Most Common	All Possible	BPCA	LLSI	SVDI
Cleveland	0.566	**0.573**	0.555	0.555	0.424	0.571	0.551
Marketing	0.286	**0.299**	0.282	0.158	0.142	0.284	0.216
Bands	0.607	**0.640**	0.636	0.610	0.506	0.633	0.634
House votes	0.893	**0.957**	0.956	0.926	0.456	0.633	0.856
Hepatitis	0.817	**0.829**	0.783	0.775	0.708	0.821	0.775

From Table 8, it is totally clear that the performance of the considered fifteen classifiers is maximized when utilizing K-Means as a missing value handling technique with respect to the Accuracy metric and considering the five different datasets.

4.2 Identifying the Best Classifier to Handle Datasets with Missing Values

Table 9 shows the Accuracy Average for the considered classifiers on the Cleveland dataset and considers the seven techniques grouped by learning strategy.

From Table 9, the SMO classifier achieved the highest Accuracy Average on Cleveland, Marketing, House votes, and Hepatitis datasets, while KNN-Adaptive achieved the highest Accuracy Average on the Bands dataset. SMO classifier showed an excellent performance on datasets with a low ratio of missing values as well as on datasets with a high ratio of missing values. Hence, it can be concluded that SMO is the best classifier to handle missing values compared with the other classifiers that belong to different learning strategies. Moreover, the effect of the missing values handling technique on the SMO classifier is limited, and hence, it is more flexible than the other considered classifiers in this research.

Table 10 shows the classifier(s) with the highest Accuracy on the five considered datasets regardless of the missing values handling technique being used.

Table 9. Accuracy Average for the Considered Classifiers on the Cleveland Dataset

Classifier	Average				
	Cleveland	**Marketing**	**Bands**	**House votes**	**Hepatitis**
C4.5	0.581	0.253	0.623	0.824	0.803
Target	0.519	0.221	0.592	0.808	0.804
DT-GA	0.500	0.227	0.610	0.814	0.795
LDA	0.576	0.245	0.610	0.821	0.793
Linear-LMS	0.582	0.255	0.609	0.821	0.803
Logistic	0.601	0.258	0.603	0.809	0.768
C-SVM	0.501	0.309	0.660	0.827	0.772
NU-SVM	0.514	0.231	0.646	0.827	0.772
SMO	**0.638**	**0.317**	0.612	**0.828**	**0.857**
MLP-BP	0.519	0.172	0.471	0.793	0.777
MLP-CG	0.562	0.268	0.613	0.782	0.776
RBFN	0.381	0.146	0.595	0.821	0.759
KNN	0.519	0.221	0.628	0.817	0.797
KNN-Adaptive	0.581	0.238	**0.667**	0.825	0.769
KStar	0.561	0.255	0.605	0.747	0.759

Table 10. Classifier(s) with the Highest Accuracy on the Considered Five Datasets

Dataset	Best Classifier						
Cleveland	SMO						
Marketing	SMO						
Bands	KNN-Adaptive						
House votes	SMO	DTGA	Logistic	C4.5	RBFN	KNN-Adaptive	MLP-CG
Hepatitis	SMO	MLP-BP					

Based on Table 10, the SMO classifier is the best choice when there are missing values in the dataset, and especially when the percentage of the missing values is low. SMO classifier showed superior performance on four datasets out of five. Also, Table 10 indicates that with datasets that have a high ratio of missing values, several classifiers may achieve good performance. KNN-Adaptive is the second-best choice after the SMO classifier for handling datasets with missing values. MLP-PB classifier is a good choice with datasets that have a high ratio of missing values and no nominal attributes. For datasets where all its attributes are nominal, SMO, DT-GA, Logistic, C4.5, RBFN, KNN-Adaptive, and MLP-CG are all good choices.

To summarize this section, the K-Means technique showed the best performance in handling missing values among the seven considered datasets with respect to the Accuracy metric. Moreover, the SMO classifier is the most suitable classifier to use with datasets that have missing values among the fifteen classifiers considered in this research.

5 Conclusion and Future Work

Handling missing values is one of the most significant pre-processing steps because it highly affects the predictive performance of any prediction system. Several techniques have been proposed to handle this step. This research investigated the performance of seven well-known missing values techniques using fifteen classifiers belonging to five learning strategies. Results revealed that the K-Means technique is superior in handling the problem of missing values compared with the other considered techniques, regardless of the characteristics of the dataset and the percentage of the missing values in the dataset. Also, the SMO classifier showed the best performance in handling datasets with missing values compared with the other fourteen classifiers from different learning strategies. In future work, more investigation should be conducted considering more datasets and more missing values handling techniques concerning other evaluation metrics than the Accuracy metric.

References:

[1] E. Elbasi, & A. I. Zreikat, (2023). Heart Disease Classification for Early Diagnosis based on Adaptive Hoeffding Tree Algorithm in IoMT Data. *International Arab Journal of Information Technology*, *20*(1), 38-48.

[2] E. Gibaja, & S. Ventura, S.(2015).A tutorial on multi label learning. *ACM Computing Surveys (CSUR)*, *47*(3), 1-38.

[3] R. Alazaidah, G. Samara, S. Almatarneh, M. Hassan, M. Aljaidi, & H. Mansur, (2023). Multi-Label Classification Based on Associations. *Applied Sciences*, *13*(8), 5081.

[4] E. Al Daoud, & G. Samara, (2022, November). Improving the Face Recognition Performance Using Gabor and VGGFace2 Features Concatenation. *In 2022 6th International Conference on Information Technology (InCIT)* (pp. 187-190). IEEE.

[5] R. Alazaidah, & M. A. Almaiah, (2021). Associative classification in multi-label classification: An investigative study. *Jordanian Journal of Computers and Information Technology*, 7(2).

[6] M. Alzyoud, R. Alazaidah, M. Aljaidi, G. Samara, M. Qasem, M. Khalid, & N. Al-Shanableh, (2024). Diagnosing diabetes mellitus using machine learning techniques. *International Journal of Data and Network Science*, 8(1), 179-188.

[7] E. A. Alhenawi, R. Al-Sayyed, A. Hudaib, & S. Mirjalili, (2023). Improved intelligent water drop-based hybrid feature selection method for microarray data processing. *Computational Biology and Chemistry*, 103, 107809.

[8] X. Zhu, J. Li, J. Ren, J. Wang, & G. Wang, (2023). Dynamic ensemble learning for multi-label classification. *Information Sciences*, 623, 94-111.

[9] M. S. Alzboon, A. F. Bader, A. Abuashour, M. K. Alqaraleh, B. Zaqaibeh, & M. Al-Batah, (2023, November). The Two Sides of AI in Cybersecurity: Opportunities and Challenges. *In 2023 International Conference on Intelligent Computing and Next Generation Networks (ICNGN)* (pp. 1-9). IEEE.

[10] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, & O. Tabona, (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1), 1-37.

[11] S. Wang, M. Li, N. Hu, E. Zhu, J. Hu, X. Liu, & J. Yin, (2019). K-means clustering with incomplete data. *IEEE Access*, 7, 69162-69171.

[12] J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse, & X. Zheng, (2005). Handling missing attribute values in preterm birth datasets. *In Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing: 10th International Conference, RSFD GrC2005*, Regina, Canada, August 31-September 3, 2005, Proceedings, PartII10 (pp.342-351). Springer Berlin Heidelberg.

[13] J. W. Grzymala-Busse, (1991). On the unknown attribute values in learning from examples. In *Methodologies for Intelligent Systems: 6th International Symposium, ISMIS'91 Charlotte, NC, USA, October 16–19,1991Proceedings 6* (pp.368-377).Springer Berlin Heidelberg.

[14] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. I. Matsubara, & S. Ishii, (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16), 2088-2096.

[15] H. Kim, G. H. Golub, & H. Park, (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2), 187-198.

[16] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, & R. B. Altman, (2001).Missing value estimation methods for DNA microarrays. *Bio informatics*, 17(6), 520-525.

[17] I. Triguero, S. González, J. M. Moyano, S. García López, J. Alcalá Fernández, J. Luengo Martín, & F. Herrera Triguero, (2017). KEEL 3.0: an open source software for multi-stage analysis in datamining.

[18] Z. Salah & E. A. Elsoud, (2023). Toward Effective Framework for Wireless Intrusion Detection System in Detecting Krack and kr00k attacks in IEEE 802.11.

[19] Z. Salah, K. Salah & E. Elsoud, (2024). Spatial domain noise removal filtering for low-resolution digital images. *Indonesian Journal of Electrical Engineering and Computer Science*, 34(3), 1627-1642.

[20] G. Samara, "Intelligent reputation system for safety messages in VANET." *Int J Artif Intell*., 9, no. 3 (2020): 439-447..

[21] H. A. Owida, B. A. H. Moh'd, N. Turab, J. Al-Nabulsi, & S. Abuowaida, (2023). The Evolution and Reliability of Machine Learning Techniques for Oncology. *International Journal of Online & Biomedical Engineering*, 19(8).

[22] H. A. Owida, O. S. M. Hemied, R. S. Alkhawaldeh, N. F. F. Alshdaifat, & S. F. A. Abuowaida,(2022). Improved deep learning approaches for covid-19 recognition in ct images. *Journal of* theoretical *and* applied information technology, 100(13).

[23] A. Mughaid, I. Obeidat, S. AlZu'bi, E. A. Elsoud, A. Alnajjar, A. R. Alsoud, & L. Abualigah, (2023). A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks. *Multimedia Tools and Applications*, 82(17), 26353-26378.

[24] G. Samara, (2020, November). Wireless sensor network MAC energy-efficiency protocols: a survey. *In 2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1-5). IEEE.

[25] I. Hussain, G. Samara, I. Ullah, and N. Khan, 2021, December. Encryption for end-user privacy: a cyber-secure smart energy management system. *In 2021 22nd International Arab Conference on Information Technology (ACIT)* (pp. 1-6). IEEE.

[26] A. Ghaben, M. Anbar, I. H. Hasbullah, and S. Karuppayah, Mathematical Approach as Qualitative Metrics of Distributed Denial of Service Attack Detection Mechanisms, In September 2021, date of current version September 13, 2021. *IEEE Access.* DOI: 10.1109/ACCESS.2021.3110586.

[27] M. R. Al-Mousa, A. S. Al-Sherideh, A. Ghaben, M. Arabiat, M. Alqudah, H. Almimi, A. Al-Shaikh. 2024, "Applicability of Iot-Aware Models In Health-Care Systems: Potential and Challenges", *Journal of System and Management Sciences*, 2024.

[28] A. Sheta,W. El-Ashmawi, A. Baareh, "Heart Disease Diagnosis Using Decision Trees with Feature Selection Method", *The International Arab Journal of Information Technology (IAJIT)*, Vol. 21, Number 03, pp. 427 - 438, May 2024, doi: 10.34028/iajit/21/3/7.

[29] M. Maree, M. Eleyat, E. Mesqali, "Optimizing Machine Learning-based Sentiment Analysis Accuracy in Bilingual Sentences via Preprocessing Techniques", *The International Arab Journal of Information Technology (IAJIT)*, Vol. 21, Number 02, pp. 257 - 270, March 2024, doi: 10.34028/iajit/21/2/8.

[30] J. Li, R. Wang, "An Anomaly Detection Method for Weighted Data Based on Feature Association Analysis", *The International Arab Journal of Information Technology (IAJIT)*, Vol. 21, Number 01, pp. 117 - 127, January 2024, doi: 10.34028/iajit/21/1/11.

[31] V. Gancheva, I. Georgiev, & V. Todorova, (2023). X-Ray Images Analytics Algorithm based on Machine Learning. *WSEAS Transactions on Information Science and Applications*, vol. 20, pp.136-145, https://doi.org/10.37394/23209.2023.20.16.

[32] T. R. Rani, W. Srimal, A. Al Shibli, N. Z. S. Al Bakri, M. Siraj, & T. S. L. Radhika, (2023). Quantile Loss Function Empowered Machine Learning Models for Predicting Carotid Arterial Blood Flow Characteristics. WSEAS Transactions on Biology and Biomedicine, vol. 20, pp.155-170, https://doi.org/10.37394/23208.2023.20.16.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research has been funded by Zarqa University.

Conflict of Interest

The authors have no conflicts of interest to declare.