Missing Value Imputation in Data MCAR for Classification of Type 2 Diabetes Mellitus and its Complications

Anik Andriani[1], Sri Hartati[2*], Afiahayati[3], Cornelia Wahyu Danawati[4]
Doctoral Program, Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia[1]
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia[2, 3]
Department of Public Health, Universitas Gadjah Mada, Yogyakarta, Indonesia[4]
Department of Information System, Universitas Bina Sarana Informatika, Jakarta, Indonesia[1]

***Abstract*—Type 2 diabetes mellitus (T2DM) is a disease that is at risk for many complications. Previous research on the prognosis of T2DM and its complications is limited to the impact of T2DM on one particular disease. Guidebook for T2DM Management in Indonesia has eight categories of T2DM complications. The purpose of this study is to classify T2DM prognosis into eight categories: one controlled class and seven classes of aggravating disorders. The classification was based on medical record data from T2DM patients at Panti Rapih Hospital in Yogyakarta between 2017 and 2022. The problem is that the medical record data has numerous missing values (MV). The dataset had 29% missing values, classified as Missing Completely at Random (MCAR). This study performed imputation on the dataset prior to categorization. For MV imputation, a variety of imputation methods were used, and their accuracy was measured using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The best imputation results were utilized to update the dataset. Subsequently, the dataset was used for classification employing several classification methods. The classification results were compared to determine the method with the highest accuracy in this scenario. The Decision Tree method with stratified k-fold cross-validation emerged as the optimal method for this classification. The results revealed an average accuracy value of 0.8529.**

***Keywords*—Missing value; prognosis of diabetes mellitus; missing completely at random; decision tree**

I. Introduction

Missing data is one of the most crucial problems in research. This tends to happen when collecting data [1]. A large number of missing values (MV) can reduce classification quality due to decreased performance on test data [2]. MV is very vulnerable to data such as weather [3], medical [1], finance [4], employees and salaries [5]. Patient medical record data is one example of data that frequently has MV.

Research regarding the prognosis of T2DM patients and their complications is needed to determine the progress of the patient's disease. The prognosis is a prediction of the development of a disease, including the progress of recovery, the emergence of other diseases, and even death. Detection of the patient's prognosis of the development of the disease complications is needed to determine the type of treatment and proper care [6]. Early detection of the prognosis of T2DM patients for their complications cannot be done medically. If this is done early, it can reduce the risk of complications [7]. This can be done by studying patient medical record data and patient activity data to determine the prevalence of T2DM for several diseases [8].

One of the supervised learning techniques used in medical research, including diagnosis, prognosis, and treatment, is classification [9]. In this study, we used medical record data from T2DM patients at Yogyakarta's Panti Rapih Hospital for classification. There is a 29% missing value rate in this data. There are 700 rows and 21 characteristics in this dataset. When classifying datasets with MV of less than 50% or less than 30%, previous research frequently overlooked MV data and deleted them, producing biased classification findings [10]. Missing values in the medical record data of T2DM patients are randomly present in some features of the dataset. The type of MV in this T2DM dataset is Missing Completely at Random (MCAR). Datasets with MCAR show that MVs appear randomly independent of a feature. In MCAR, the appearance of MV does not depend on another variable [11]. This study proposes a classification model for the prognosis of T2DM patients with this complex disease by first imputing the data.

Classification using a variety of imputation methods and classification methods was done on several datasets at different percentages of MVs. The results demonstrated that accuracy is decreased when the percentage of MVs increased [12]. This study conducted experiments on datasets with a fairly high percentage of MVs with the type of MV in the dataset being MCAR. Imputation on the MCAR dataset requires intransitive imputation techniques, namely the imputation of missing values on observed variables is independent of other variables. MEAN is one of the most frequently used imputation methods in the intransitive imputation type. In addition, Linear Regression is one of the methods often used in this case [13]. In previous studies, the KNN method provided the best performance in imputing missing values on the MCAR Dataset containing numeric data [14]. Therefore, this study uses missing value imputation, including MEAN, K-Nearest Neighbor (KNN), and Linear Regression (LR). Meanwhile, the classification methods employed are Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machine (SVM).

*Corresponding Author.

II. METHODS

A. Proposed Model

This research consists of several stages. Fig. 1 illustrates the stages of the research, which include dataset preparation, calculation of correlation values between features, MV imputation, data validation, classification, and evaluation. Fig. 1 shows the stages of the research. The dataset for T2DM patients' prognoses regarding their complicating diseases comprises 700 rows and 21 features. One of these features is the target feature, which encompasses eight classes.

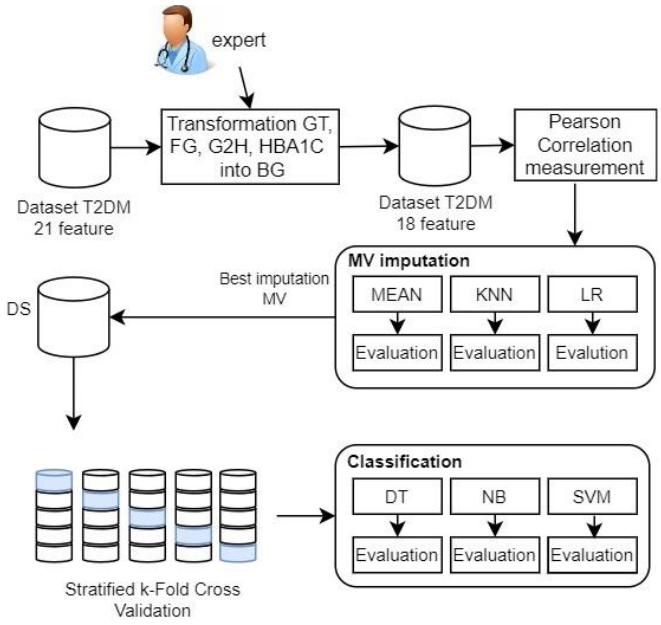

Fig. 1. Research stages.

In the first stage, two datasets were created: DS21 and DS18. DS21 is a T2DM dataset without any transformation, while DS18 is a T2DM dataset that underwent a transformation process. In this study, transformation refers to the process of consolidating features that can be represented by a single feature to determine a value. These features include Blood Glucose at Time (GT), Fasting Glucose (FG), and Blood Glucose 2 Hours after meals (G2H), which are combined into one feature called Blood Glucose level (BG). Doctors do not always rely on all three glucose test values to determine a patient's blood glucose level; sometimes, they only check GT, FG, or G2H. Consolidating multiple features that can substitute for each other into a single feature is also beneficial for reducing the number missing values. The outcome of this transformation is DS18, a dataset containing 18 features.

Dataset imputation involves the use of MEAN, KNN, and LR methods. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were employed to assess the imputation. A comparison of the three methods is conducted to determine the most effective imputation results. An error value close to 0 indicates an imputed value that is close to the true are minimized in the DS dataset. The subsequent step involves classification using DS. The classification methods utilized are DT, NB, and SVM. The classification outcomes from these three methods are assessed based on accuracy values to identify the most suitable approach for classifying the T2DM prognosis and its associated complications.

B. Missing Value Imputation Method

MV is handled using the Imputation Missing Value (IMV) technique. The IMV phases are described in Fig. 2. The correlation value between the features is calculated at the start of the stage. Correlation values for numerical features are computed using Pearson correlation. A relationship between two features is shown by a positive correlation value. Features that show correlation are given regression values. The outcomes are applied to IMV. The dataset (DS) that will be utilized for classification is subsequently created using the imputation findings.

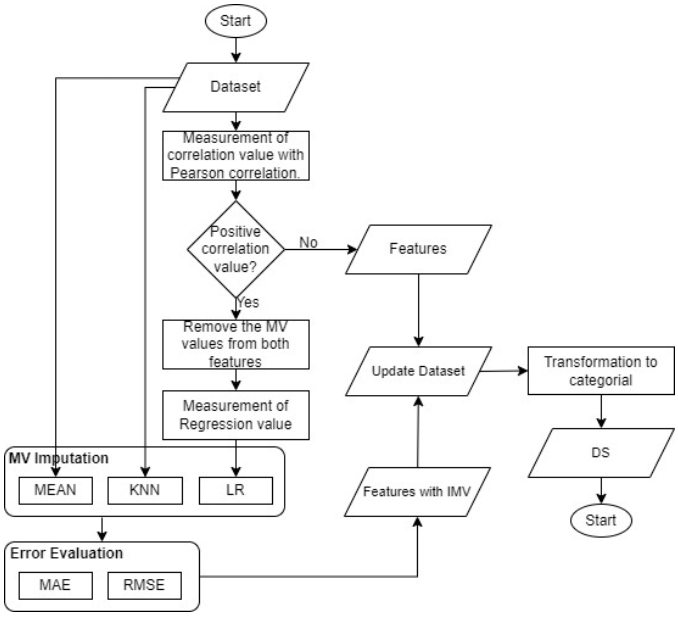

Fig. 2. IMV stages.

The imputation steps shown in Fig. 2 illustrate the three types of imputation that are used: MEAN, KNN, and LR. IMV calculates the average value of the observed feature values in the MEAN method. A mean computation for IMV is assumed in Eq. (1). The MEAN value computation relies on the global average value derived from the summation of all values in the observed features in the dataset, denoted as N. Eq. (1) presents a formula for determining the mean value [14].

$$\mu_i = \frac{1}{N} \sum_{i=1}^{N} X_i \quad (1)$$

Imputation using the KNN methods entails dividing the dataset into complete data and MV data. The method finds the closest value to impute MV data by comparing it with the entire data sample. Eq. (2) is used to calculate the Euclidean distance (d), which is the distance between two points in a two-dimensional space [15].

$$d = \sqrt{\sum_{r=1}^{n} (x_{ir} - x_{jr})^2} \quad (2)$$

The imputation stage employs the LR method to compute the correlation value between features utilizing Pearson correlation. The correlation values are compute among features

to identify those with a positive correlation. Features exhibiting a positive correlation are employed for MV imputation, whereas those with a negative correlation are excluded from MV imputation. The linear correlation between the two properties as shown in Eq. (3) is characterized by the Pearson correlation value. A strong correlation between two dependent qualities is indicated by a Pearson correlation coefficient value of 1, which ranges from -1 to 1.

In the imputation stage, Pearson correlation is used to calculate the correlation value between features using the LR approach. The correlation values are compute among features to identify those with a positive correlation. Features exhibiting a positive correlation are employed for MV imputation, whereas those with a negative correlation are excluded from MV imputation. The linear correlation between the two properties as shown in Eq. (3) is characterized by the Pearson correlation value [16]. A strong correlation between two dependent features (X, Y) is indicated by a Pearson correlation coefficient (ρ_{XY}) value of 1, which ranges from -1 to 1.

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

The process of IMV for features exhibiting a positive correlation commences with the MV values from both features. This step leads to the retrieval of complete values for both features. Subsequently, the regression value is calculated based on complete feature data, and the MV value is estimated based on the regression value through the utilization of Eq. (4), (5), and (6) [17].

$$c = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \tag{4}$$

$$d = \frac{\sum Y - c\sum X}{n} \tag{5}$$

Where, n is the selected amount of data points, even or c, and coefficients or d. They y-intercept is the point on the y-axis where the graph crosses the y-axis. It is the place where the line's slope, which indicates how it step is, is located. Eq. (6) describes to build linear regression.

$$\bar{Y} = c + dX \tag{6}$$

In this formulation, the y-intercept is shown by c, while the slope of the line is represented by d. $\bar{Y}$ notation for represented the expected value of the dependent variable Y for a specific value of the independent variable X. When building a line in algebra, the equation of the line must be found at two locations (x,y).

Implementations MAE and RMSE for quantify the error in IMV result from the three imputation methods. Eq. (7) was employed to determine the MAE value, while Eq. (8) was utilized for calculating the RMSE value [18].

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n} |x_i - \bar{x}_i| \tag{7}$$

$$RMSE = \sqrt{(\sum_{i=1}^{n}(x_i - \bar{x}_i)^2/n)} \tag{8}$$

Evaluation of MV imputation results with the third method was compared to find out which method produced the best MV imputation values. Evaluation with the lowest MAE and RMSE and close to 0 is the best MV imputation result. The best imputation results are then used to build the DS dataset.

C. Classifications Method

Classifications were applied in the DS dataset, utilizing various methods such as Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machine (SVM). These three widely recognized classification techniques are commonly employed in research related to classification.

The Decision Tree (DT) method is a classification algorithm that is widely used in data mining and machine learning. This algorithm predicts target values by learning simple decision rules derived from the features in the dataset. DT segments data into smaller subsets based on existing features, and each subset is then processed recursively. The selection of the most informative features is achieved by minimizing impurities (such as entropy or Gini impurity) at each data division. This method offers advantages, such as being easy to understand and interpret. DT can be depicted in the form of a decision tree structure and can process both categorical and numeric data. The Decision Tree method has been applied in various applications, from pattern recognition and business data analysis to medical diagnosis. In DT, it is necessary to compute the Entropy value first to measure the level of uncertainty or irregularity in a dataset. In the context of DT, entropy is often utilized to assess the quality of splits of attributes used as nodes in a decision tree. Eq. (9) illustrates the calculation of the entropy value for the data. The entropy value obtained is then used to calculate Information Gain. Equation 10 demonstrates the calculation of Information Gain, which is used to classify classes by segregating data based on specific features. The feature with the highest Information Gain values is chosen as the root feature.

$$En_{A_i}(Data) = \sum_{j=1}^{k} \frac{|Attr|}{|Data|} . Attr_j \tag{9}$$

$$InfoG(A_i) = En(Data) - En_{A_i}(Data) \tag{10}$$

Naïve Bayes (NB) uses a Bayesian learning approach that incorporates the concept of probability in classification tasks. One of the most straightforward and widely used Bayesian learning models is Naïve Bayes [19]. Naïve Bayes performs exceptionally well in multiclass classification scenarios with a single label, delivering high accuracy. Multiclass classification with a single label involves categorizing data into more than two classes, with each class assigned only one label [20]. The calculation of probability values in Naïve Bayes is based on Eq. (11), where P is probability.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{11}$$

Support Vector Machine (SVM) works by finding a hyperplane to separate two classes in binary classification. In multiclass classification, multiple binary SVMs are used [21]. Eq. (12) defines vector w_i as the hyperplane coefficient vector, b as the bias, and $f(x) = 0$, for x on the hyperplane.

$$f(x) = \sum_i w_i \times x_i + b \tag{12}$$

III. Result and Discussion

A. Dataset Preparation

Learning to classify the prognosis of T2DM for its complications involves multiclass classification learning. In the dataset, there is a feature called the Target Class, which consists of eight classes. According to the Guidebook for the Management and Prevention of Adult Type 2 Diabetes Mellitus in Indonesia, prepared by the Indonesian Endocrinology Association (PERKENI), there are eight diseases classified as risks and complications of T2DM. Both of these are considered complications of T2DM. Table I presents information regarding complications associated with Type 2 Diabetes Mellitus (T2DM) and the conversion of values into the Target Class. The table delineates data on the various complications. These eight categories serve as the target class labels in the dataset for classification purposes. The data on the complication categories of T2DM was extracted from the 2021 Guide to Management and Prevention of Type 2 Diabetes Mellitus in Indonesia.

TABLE I. Categories of Complications in T2DM

Feature	Categories of Complication	Disease	Class label	Class label quantity
Prognosis of Complication	Controlled	-	0	186
	Nephropathy	CKD, Diabetic NephropathyInsuff Renal	1	96
	Cardiovascular	IHD, CHF, KAD	2	193
	Neuropathy	Neuropathy, Cellulitis	3	95
	Hyperglycemia	Hyperglycemia	4	70
	Macroangiopathy	Macroangiopathy, Ulkus DM	5	40
	Hypoglycemia	Hypoglycemia	6	2
	Retinopathy	Retinopathy Diabetic	7	18

The dataset comprises 21 features, with one of them being the target feature. It consists of 700 rows of data. The features are detailed in Table II. Within the T2DM patient dataset, there are 4321 missing values out of a total 14700 values, representing 29% missing values. The distribution of MV in the dataset is illustrated in Fig. 3.

TABLE II. Dataset Features

Feature	Feature	Type
Gender	GEN	Categorial
Age	AGE	Numeric
Blood Glucose at Time	GT	Numeric
Fasting Glucose	FG	Numeric
Blood Glucose 2 Hours after meals	G2H	Numeric
HbA1C	HBA1C	Numeric
Creatinine	CREAT	Numeric
Ureum	UREUM	Numeric
Systolic	SYST	Numeric
Diastolic	DIAST	Numeric
Cholesterol	CHOL	Numeric
Low-density lipoproteins	LDL	Numeric
High-density lipoproteins	HDL	Numeric
Triglycerides	TGD	Numeric
Uric Acid	UA	Numeric
Nutrition	NUT	Categorial
Treatment	TREAT	Categorial
Early Diagnosis	ED	Categorial
Hypertension	HT	Categorial
Early Complications	EC	Categorial
Prognosis Complications	PC	Class target

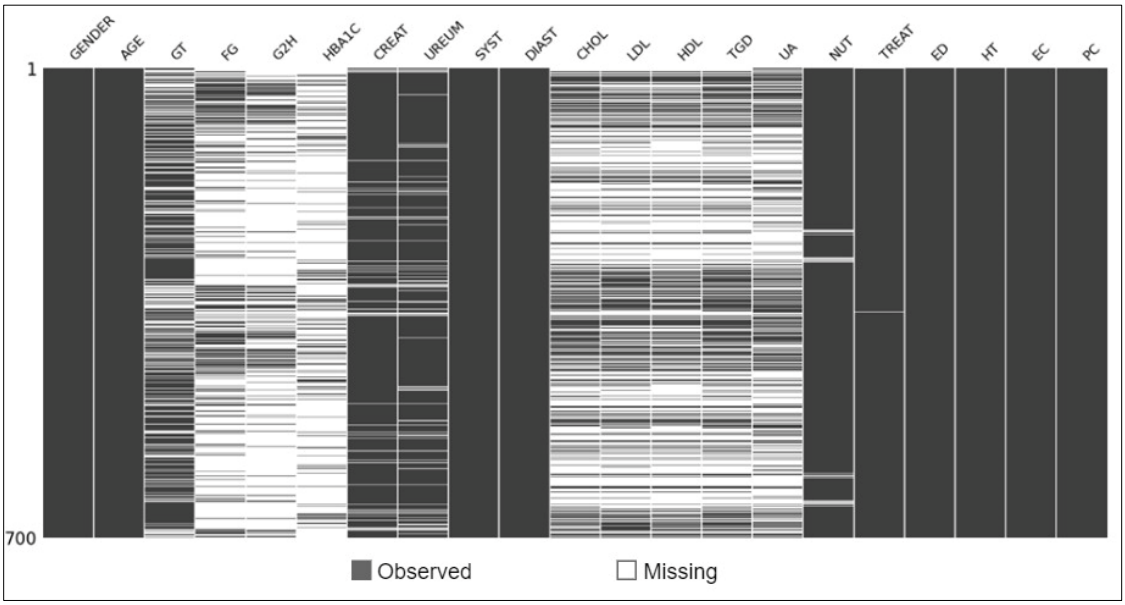

Fig. 3. MV distributions.

The T2DM dataset exhibits 29% MV. Omitting these missing values from the dataset would lead to the loss of a significant portion of the data. Out of the 700 rows in the dataset, only three complete rows of data are available. Therefore, it is essential to perform MV imputation in order to enhance the dataset by increasing the number of usable data rows for classification purposes.

Following consultations with experts, particularly internal medicine physicians, it has been suggested that certain features can be consolidated into a single feature. Specifically, the GT, FG, G2H, and HBA1C features can be amalgamated into a singular feature known as the Diabetes Blood Sugar feature. This consolidation is based on the observation that medical practitioners may not always assess all four features when determining a patient's blood sugar status. Frequently, doctors may only examine one or a combination of features to ascertain blood sugar levels. The transformation of these four features into a single feature is guided by the Blood Sugar category outlined in Table III.

TABLE III. CATEGORIES OF BLOOD GLUCOSE

Value	Categories	Categories Label
HBA1C: <5.7 GT: 70-139 mg/dL FG: 70-99 mg/dL G2H: 70-139 mg/dL	Normal	1
HBA1C: 5.7-6.4 GT: 140-199 mg/dL FG: 100-99 mg/dL G2H: 140-199 mg/dL	Prediabetes	2
HBA1C: >=6.5 GT: 200-299 mg/dL FG: 126-199 mg/dL G2H: 200-299 mg/dL	Diabetes	3
GT: >=300 mg/dL FG: >=200 mg/dL G2H: >=300 mg/dL	Hyperglycemia	4
GT: < 70 mg/dL FG: < 70 mg/dL G2H: < 70 mg/dL	Hypoglycemia	5

The process of consolidating four features into one feature was conducted in collaboration with experts, specifically internal medicine physicians. The outcome of this consolidation effectively decreased the missing values (MVs) by 9%, reducing them from 29% to 20%. Despite this reduction, further MV imputation is deemed necessary to minimize the number of missing values.

B. Missing Value Imputation Result

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

The subsequent missing value (MV) imputation technique employs the K-nearest neighbors (KNN) method. This approach is commonly utilized for MV imputation and involves imputing missing values by computing similarity metrics based on distances. Eq. (2) is employed to determine the imputation value. The distribution of MV data post-KNN imputation is illustrated in Fig. 6.

IMV using Linear Regression has multiple steps that must go in order. These steps involve figuring out regression coefficients, estimating MV values, and computing correlation coefficients between features. In order to determine the interdependencies between all features, a correlation analysis is carried out using numerical data.

	AGE	CREAT	UREUM	DIAST	CHOL	LDL	HDL	TGD	UA
AGE	1.000	-0.149	-0.106	-0.047	-0.077	-0.135	0.118	-0.039	-0.151
CREAT	-0.149	1.000	0.831	0.039	0.108	0.103	-0.273	0.122	0.417
UREUM	-0.106	0.831	1.000	0.025	-0.018	-0.028	-0.169	0.058	0.376
DIAST	-0.047	0.039	0.025	1.000	0.366	0.270	0.032	0.167	0.033
CHOL	-0.077	0.108	-0.018	0.366	1.000	0.721	0.367	0.500	0.091
LDL	-0.135	0.103	-0.028	0.270	0.721	1.000	0.297	-0.020	0.104
HDL	0.118	-0.273	-0.169	0.032	0.367	0.297	1.000	-0.195	-0.244
TGD	-0.039	0.122	0.058	0.167	0.500	-0.020	-0.195	1.000	0.139
UA	-0.151	0.417	0.376	0.033	0.091	0.104	-0.244	0.139	1.000

Fig. 4. Pearson correlation values in numeric features.

Eq. (3) is utilized to calculate Pearson's correlation coefficient, which measures the degree of linear association between features. Fig. 4 shows the results of the correlation analysis.

Based on the results of correlation calculations, several features exhibited significant correlation values. Specifically, CREAT was correlated with UREUM, CHOL with DIAST, CHOL with LDL, CHOL with HDL, CHOL with TGD, and CREAT with UA. The Pearson correlation values for these features are visually represented in Fig. 5.

MV imputation was performed on six features (CREAT, UREUM, AU, CHOL, LDL, TGD) based on correlation values. The imputation reduced missing values from 20% to 2% in the dataset, resulting in 598 rows out of 700. Fig. 6 illustrates the distribution of imputed data in the dataset.

Evaluation of the MV imputation results is conducted to assess their accuracy compared to the actual values. This assessment involves calculating error values using MAE and RMSE equations. The evaluation compares imputation results from three methods: MEAN, KNN, and LR. Table IV displays the MAE and RMSE for the three imputation methods. The evaluation shows that MV imputation with LR yields the smallest errors, close to 0. The dataset imputed using LR will be used for classifying T2DM prognosis and its complications.

TABLE IV. EVALUATION OF ERRORS IN IMPUTATION RESULTS

Feature	MEAN		LR		KNN	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
CREAT-UREUM	13.5	338.4	1.6E-14	4.0E-13	0.1	1.6
CHOL-DIAST	7.4	120.9	1.9E-15	3.1E-14	2.5E-13	4.1E-12
LDL-CHOL	0.1	1.9	1.0E-14	2.0E-13	1.6E-13	2.7E-12
TGD-CHOL	31.5	529.8	5.0E-14	8.0E-13	5.2E-13	8.8E-12
CREAT-UA	0.2	3.4	3.0E-15	5.0E-14	0.02	0.3

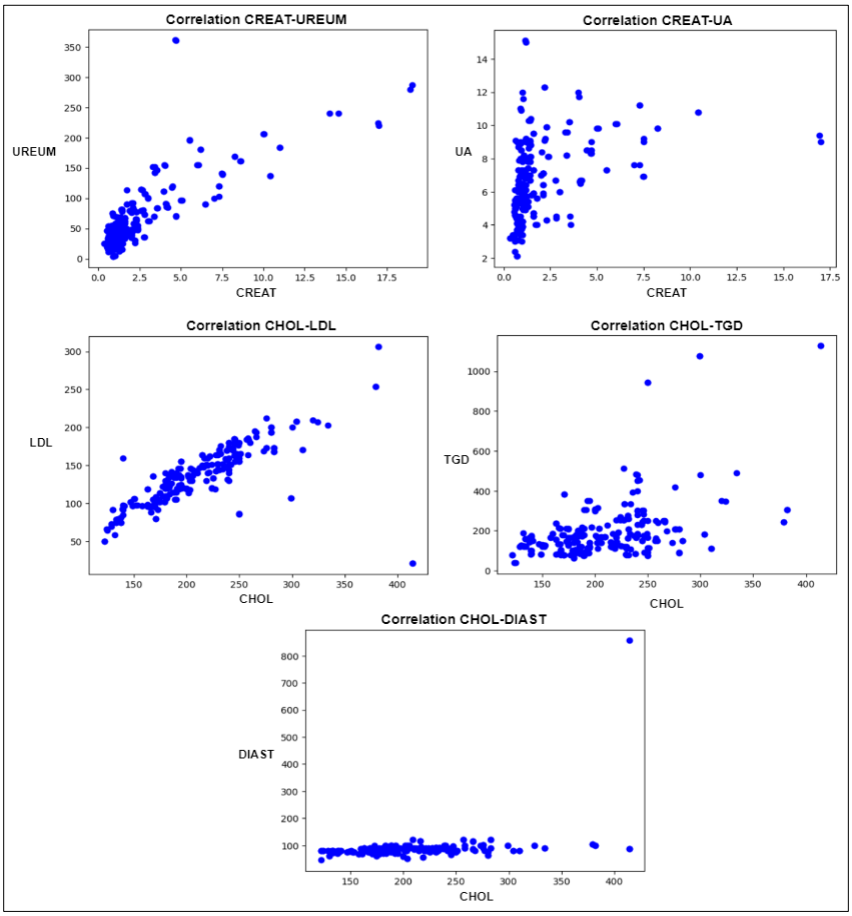

Fig. 5. Features with a high correlation values.

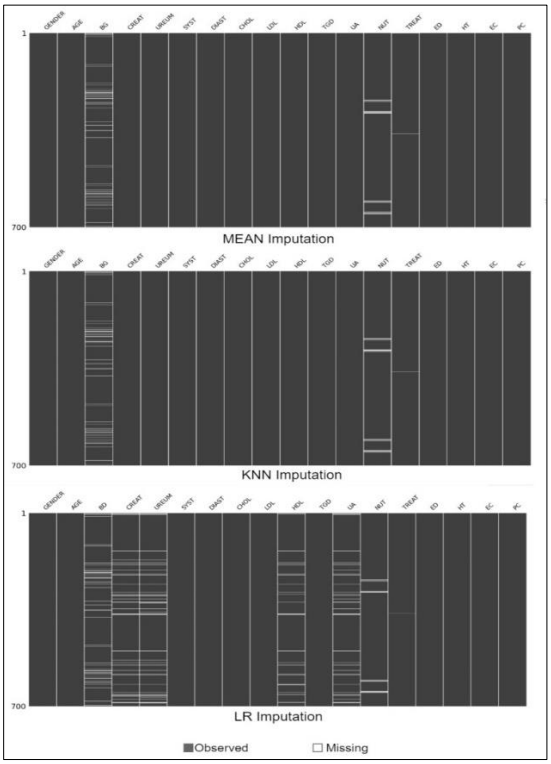

Fig. 6. Distribution of MV.

C. Classification Result

The pre-processing stage after MV imputation is data transformation. The dataset of MV imputation results with LR is used for this study. As seen in Table V, the variables in the dataset are converted into categorical form.

TABLE V. DATASET FEATURE CATTEGORIES FOR TRANSFORMATION DATA

Feature	categories
AGE	1-14 (1), 15-24(2), 25-44(3), 45-64(4), >65(5)
GENDER	Man(1), Women(2)
BG	Normal (1), prediabetes (2), diabetes(3), hyperglicemia (4), hypoglicemia (5)
CREAT	Woman: <0.6 (Low=1), 0.6-1.1(Normal=2), >1.1(High=3) Man: <0.7 (Low=1), 0.7-1.3 (Normal=2), >1.3 (High=3)
UREUM	<6 (Low=1), 6-23(Normal=2), >23(High=3)
SYST	Age <=60: <90(Low=1), 90-120(Normal=2), >120 (High=3) Age >60: <100(Low=1), 100-140(Normal=2), >140(High=3)
DIAST	Age <=60: <60(Low=1), 60-80(Normal=2), >80 (High=3) Age >60: <60(Low=1), 60-90(Normal=2), >90(High=3)
CHOL	<200 (Normal=1), 200-240 (Borderline=2), >240 (High=3)
LDL	<100 (Normal=1), 100-129 (Optimal=2), 130-160(Borderline=3), >160 (High=4)
HDL	>=40 (Normal=1), <40 (Low=2)
TGD	<149 (Normal=1), 150-200 (Borderline=2), >200 (High=3)
UA	Woman: <1.5 (Low=1), 1.5-6(Normal=2), >6 (High=3) Man: <2.5 (Low=1), 2.5-7 (Normal=2), >7 (High=3)
NUT	Good(1), Enough(2), Medium(3), Less(4), Over(5)
TREAT	Routine check-up+medicine(1), Medicine(2), Non-Routine check-up(3), Insulin(4), Insulin+Medicine(5), Non-routine medicine(6)
ED	DM2NO(1), DM2 OBESE(2), DM2 HYPERGLICEMIA(3)
HT	Yes(1), No(2)
EC	No(0), CKD(1), Nephropathy(2), Insuff renall(3), IHD(4), CHF(5), Stroke/Post Stroke(6), KAD(7), Neuropathy(8), Hyperglycemia(9), Ulcus(10),Cellulitis(11), Retinopathy(12)

Following MV imputation and data transformation in the pre-processing stage is the data validation stage. Data validation and classification are the two primary phases of the classification process. To divide the data into training and testing sets, the data validation stage uses the stratified k-fold cross-validation technique, as shown in Fig. 1. By using this technique, it is ensured that the training and testing sets of data have equal class distributions. The method is dividing the data into k folds, or groups, at random [20]. K-fold cross-validation with k=5 is used in this investigation. Five folds are created in the dataset for k-fold cross-validation. One-fold is used for testing data and (k-1) folds are used for training data in each fold. The process is iterated so that each fold serves as the testing data exactly once.

D. Evaluation Result

Classification results are assessed using accuracy values. Table VI provides a comparison of the accuracy values obtained from various classification methods.

TABLE VI. CATEGORIES OF BLOOD GLUCOSE

k-Fold=5	DT	NB	SVM
Fold1	0.9000	0.2750	0.5167
Fold2	0.8417	0.2667	0.5250
Fold3	0.8083	0.2750	0.4250
Fold4	0.8824	0.2185	0.5714
Fold5	0.8319	0.2269	0.5462
Average	0.8529	0.2524	0.5169

Table V presents the classification outcomes, indicating that the Decision Tree method yields the highest average value for classification. Across all folds, consistent accuracy exceeding 0.8 is achieved when employing the Decision Tree approach. Conversely, the Naïve Bayes and Support Vector Machine methods exhibit notably lower classification accuracies, both falling below 60%. These methods do not offer sufficiently accurate results for predicting the prognosis of T2DM patients with their complications.

E. Discussion

Our findings show that MV imputation results on the MCAR type T2DM prognosis dataset are shown in Fig. 6. The MEAN and KNN imputation methods impute more missing values than the LR imputation method. However, based on the evaluation with the MAE and RMSE values in Table IV, the imputation results with the LR method provide error values close to 0. This means that the synthetic data which is the result of imputation is close to the real value. In addition, the application of the dataset that has been imputed with the LR method works well when the classification method with Decision Tree is applied as shown in Table VI. The LR imputation method is an alternative for data imputation with a high percentage of missing values in the MCAR type dataset in addition to the MEAN and KNN imputation methods in previous studies.

IV. CONCLUSIONS

Imputing missing values in medical data is crucial before initiating the classification learning process, especially when working with datasets that frequently have many missing values. In the context of predicting complications in patients with Type 2 Diabetes Mellitus (T2DM), the utilization of the LR method for missing value imputation has demonstrated lower error rates in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) calculations compared to the MEAN and KNN approaches. Furthermore, the fraction of missing values on this specific dataset was significantly reduced from 29% to 2% by using missing value imputation with LR. The imputed dataset generated by the LR technique is then utilized to categorize T2DM patients' prognosis with respect to their problems. Stratified k-fold cross-validation with k-fold 5 has been identified as the best validation approach for getting the most accurate classification results when using the Decision Tree method during the data validation step. The average accuracy achieved through the Decision Tree method, utilizing a blend of training and testing data, is reported to be 0.8529.

Research on the classification of Type 2 Diabetes Mellitus (T2DM) prognosis for its complications is presently constrained to a single complication that manifests first. Subsequent research could focus on the classification of T2DM prognosis based on potential disease composition.

REFERENCES

[1] R. K. Bania and A. Halder, "R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data," Comput. Methods Programs Biomed., vol. 184, p. 105122, 2020, doi: 10.1016/j.cmpb.2019.105122.

[2] U. Bentkowska, J. G. Bazan, W. Rząsa, and L. Zaręba, "Application of interval-valued aggregation to optimization problem of k−NN classifiers for missing values case," Inf. Sci. (Ny)., vol. 486, pp. 434–449, 2019, doi: 10.1016/j.ins.2019.02.053.

[3] A. Aieb, K. Madani, M. Scarpa, B. Bonacorso, and K. Lefsih, "A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria," Heliyon, vol. 5, no. 2, p. e01247, 2019, doi: 10.1016/j.heliyon.2019.e01247.

[4] Q. Lan, X. Xu, H. Ma, and G. Li, "Multivariable data imputation for the analysis of incomplete credit data," Expert Syst. Appl., vol. 141, 2020, doi: 10.1016/j.eswa.2019.112926.

[5] L. Ni, F. Fang, and J. Shao, "Feature screening for ultrahigh dimensional categorical data with covariates missing at random," Comput. Stat. Data Anal., vol. 142, p. 106824, 2020, doi: 10.1016/j.csda.2019.106824.

[6] L. Qiao, Y. Zhu, and H. Zhou, "Diabetic Retinopathy Detection Using Prognosis of Microaneurysm and Early Diagnosis System for Non-Proliferative Diabetic Retinopathy Based on Deep Learning Algorithms," IEEE Access, vol. 8, pp. 104292–104302, 2020, doi: 10.1109/access.2020.2993937.

[7] M. A. H. Shareef, K. Narasimhalu, S. E. Saffari, F. P. Woon, and D. A. De Silva, "Reccurrent vascular events partially explain association between diabetes and poor prognosis in young ischemic stroke patients." pp. 1–4, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38219383/#:~:text=Recurrent vascular events partially explain association between diabetes,J Neurol Sci. 2024 Jan 11%3A457%3A122881. doi%3A 10.1016%2Fj.jns.2024.122881.

[8] B.-Y. Zhou et al., "Association of D-dimer with long-term prognosis in type 2 diabetes mellitus patients with acute coronary syndrome," Nutr. Metab. Cardiovasc. Dis., no. xxxx, 2022, doi: 10.1016/j.numecd.2022.05.013.

[9] Eliyani, S. Hartati, and A. Musdholifah, "Machine Learning Assisted Medical Diagnosis for Segmentation of Follicle in Ovary Ultrasound." Springer, pp. 71–80, 2019. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-15-0399-3_6.

[10] Y. Pin Chen, C. Hua Huang, Y. Hsun Lo, Y. Ying Chen, and F. Lai, "Handle MV on Time Series Data.pdf." pp. 1271–1287, 2022.

[11] T. M. Pham, N. Pandis, and I. R. White, "Pham-MCAR MAR MNAR-2022.pdf." American Journal of Orthodontics and Dentofacial Orthopedics, pp. 138–139, 2022.

[12] C. F. Tsai and Y. H. Hu, "Empirical comparison of supervised learning techniques for missing value imputation," Knowl. Inf. Syst., vol. 64, no. 4, pp. 1047–1075, 2022, doi: 10.1007/s10115-022-01661-0.

[13] S. M. Mostafa, "Imputing missing values using cumulative linear regression," CAAI Trans. Intell. Technol., vol. 4, no. 3, pp. 182–200, 2019, doi: 10.1049/trit.2019.0032.

[14] F. I. Kurniadi, R. C. Rohmana, and L. Taufani, "Kurniadi-Local Mean Imputation for Handling MV-2023.pdf." Procedia Computer Science, pp. 301–309, 2023.

[15] D. Zou et al., "Outlier detection and data filling based on KNN and LOF for power transformer operation data classification." pp. 698–711, 2023.

[16] S. Peng, W. Han, and G. Jia, "Pearson correlation and transfer entropy in the Chinese stock market with time delay," Data Sci. Manag., vol. 5, no. 3, pp. 117–123, 2022, doi: 10.1016/j.dsm.2022.08.001.

[17] J. M. Sangeetha and K. J. Alfia, "Financial stock market forecast using evaluated linear regression based machine learning technique," Meas. Sensors, vol. 31, no. April 2023, p. 100950, 2024, doi: 10.1016/j.measen.2023.100950.

[18] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well," Inf. Sci. (Ny)., vol. 585, pp. 609–629, 2022, doi: 10.1016/j.ins.2021.11.036.

[19] S. Farhana, "Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm," Procedia Comput. Sci., vol. 194, pp. 224–228, 2021, doi: 10.1016/j.procs.2021.10.077.

[20] A. Kag, L. M. Jenila Livingston, L. M. Livingston Merlin, and L. G. X. Agnel Livingston, "Multiclass Single Label Model for Web Page Classification," 2019 Int. Conf. Recent Adv. Energy-Efficient Comput. Commun. ICRAECC 2019, pp. 3–8, 2019, doi: 10.1109/ICRAECC43874.2019.8995087.

[21] B. A. Akinnuwesi et al., "Application of Support Vector Machine Algorithm for Early Differential Diagnosis of Prostate Cancer." pp. 1–12, 2023.