Missing Value Imputation for Diabetes Prediction

Fei Luo[1], Hangwei Qian[1], Di Wang[1], Xu Guo[1,2], Yan Sun[3], Eng Sing Lee[4],
Hui Hwang Teong[5], Ray Tian Rui Lai[5], and Chunyan Miao[1,2]
[1]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly,
Nanyang Technological University, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Health Services & Outcomes Research Department, National Healthcare Group, Singapore
[4]National Healthcare Group Polyclinics, Singapore
[5]Tan Tock Seng Hospital, Singapore
Email: fei.luo@ntu.edu.sg, qian0045@e.ntu.edu.sg, wangdi@ntu.edu.sg, xu008@e.ntu.edu.sg,
yan_sun@nhg.com.sg, eng_sing_lee@nhgp.com.sg, {hui_hwang_teong,ray_tr_lai}@ttsh.com.sg, ascymiao@ntu.edu.sg

***Abstract*—Machine learning (ML) models have been widely used to improve the accuracy and efficiency of various types of disease diagnostic tasks. However, it is still challenging to apply ML models to perform diabetes-related prediction tasks mainly because patients' health records are sparse and have a vast amount of missing values. Missing values often break the diabetes prediction pipelines, posing challenges to existing approaches. Such problem deteriorates significantly when critical attribute values (e.g., blood test results on *HbA1c*, *FPG* and *OGTT2hr*) are missing. In this paper, we introduce a large-scale diabetes-related dataset named Chronic Disease Management System (CDMS) dataset, which collects the clinical records of more than 700,000 visits of over 65,000 patients across eight years. CDMS is anonymously collected and has a high percentage of missing values on several critical attributes for diabetes prediction. If not being dealt with carefully, the missing values will cause significant performance degradation of the applied ML models. In this paper, we also investigate the effectiveness of multiple data imputation methods through conducting extensive experiments using CDMS. Experimental results show that k-Nearest Neighbor Imputation (KNNI) performs better than other methods in this diabetes prediction task. Specifically, with KNNI applied, the diabetes prediction accuracy and precision are both over 0.8 using various ML predictive models.**

***Index Terms*—diabetes-related dataset, diabetes prediction, missing values, data imputation techniques**

I. Introduction

Early identification and intervention of chronic diseases are deemed as critical tasks in reducing burdens of healthcare systems for both individuals and the society. Diabetes mellitus (DM) is one of the most prevalent and challenging chronic diseases, which is imperceptible in its early stage, but leads to severe morbidities from complications if not being well taken care of. According to [1], the number of global diabetes patients is projected to increase from 380 million in 2013 to 590 million by 2035. Patients with diabetes have a higher risk for complications. Such morbidities of diabetes result in substantial societal and economic burdens [2]. To better facilitate the early detection and prevention of DM, various evidence-based and patient-centric approaches have been adopted by clinicians to care for patients across a continuum of patients' lives [3]. In addition, the DM problem can be modeled as a classification problem with the goal being predicting the clinical outcomes of the DM, e.g., {healthy, pre-diabetes, diabetes}. With the emergence of artificial intelligence, various ML models have been leveraged to provide timely prediction on the occurrence of pre-diabetes, diabetes and comorbidities [2].

Despite the popularity of ML models, the prediction of clinical outcomes calls for extra caution when models make automated decisions on health-related tasks [4]. Different from applications such as image classification on handwritten digits, the incorrect prediction of DM can lead to much severe consequences. Specifically, if a person with pre-DM or DM is not identified promptly, we may miss the critical time window for early interventions. Another important limiting factor that requires extra caution is the high prevalence of missing values in real-world DM-related datasets. This is caused by the fact that a patient's data may be collected at irregular time intervals with different subsets of health records at different time points. In addition, different patients typically have different numbers of health records, which correspond to different number of visits to hospitals. All these real-world medical data issues pose a big challenge to ML models because standard ML models require input data of high integrity with a fixed number of dimensionality.

Rather than removing the data entries having missing values, data imputation techniques may increase data quality by replacing missing values by appropriate imputed values. Mean imputation replaces missing values of a certain variable with the mean value of this variable, which is easy to put into practice [5]. However, one defect of this method is the ignorance of population variance. Hot-deck imputation technique groups instances based on several variables, then substitutes the missing values with mean values within each group. For instance, for the k-Nearest Neighbor Imputation (KNNI) approach, each group corresponds to k number of nearest neighbours. Recently, discriminative and generative deep learning (DL) imputation methods [6], [7] have been proposed as well.

Note that there is no perfect imputation strategy for all real-world datasets. Hence, it requires dedicated analysis and experiments to determine the most appropriate data imputation technique. In this paper, we aim to investigate the data imputation strategies using a large-scale medical dataset named Chronic Disease Management System (CDMS) dataset. The CDMS dataset comprises 1,486,746 health records of chronic disease patients who visited five polyclinics being operated by the National Healthcare Group in Singapore from 2010 to 2017. Each health record corresponds to a patient's visit to the polyclinic. Note that the CDMS dataset used in this paper only comprises the health records of diabetes-related patients being archived in the much more comprehensive, continuously updated data repository manged by the National Healthcare Group, Singapore [3].

Other than introducing the CDMS dataset, we also aim to study how to properly conduct analysis on diabetes-related data using ML models. In particular, we observe that the vast amount of missing values in CDMS cannot be ignored, otherwise, it would significantly affect the performance of the downstream ML models. Therefore, we focus on the problem of having a large proportion of missing values in DM-related records, and the investigation on how different data imputation approaches affect the final model performance. In addition, we also conduct multi-facet investigations related to the missing values, such as preprocessing procedures, outlier detection, etc. After imputing missing values, we apply seven ML and DL models for diabetes prediction. Empirical results indicate that KNNI outperforms the other imputation methods on the diabetes prediction task.

The contributions of our paper are summarized as follows: (i) Based on CDMS, a large-scale diabetes-related dataset collected in Singapore from 2010 to 2017, which covers multiple ethnic groups, , we introduce the demographic composition of the diabetes cohort in Singapore. (ii) We thoroughly analyse the data preprocessing procedure, especially the data imputation methods, to handle missing values using CDMS. (iii) We conduct extensive experiments on the imputed dataset using multiple ML predictive models for diabetes prediction.

The rest of the paper is organized as follows. Related work of diabetes prediction and data imputation methods are presented in Section II. The details of the CDMS dataset is introduced in Section III. Experimental results and discussions are delineated in Section IV. Finally, conclusion and future work are reported in Section V.

II. Related Work

In this section, we review related prior studies on diabetes datasets, machine learning algorithms used for diabetes prediction, and relevant missing value imputation methods.

A. Diabetes Prediction

The rapidly increasing number of diabetes cases has posed a huge and imminently growing burden on healthcare systems in many countries. Early detection of diabetes is critical to prevent the patients from developing other chronic diseases and serious complications. In this attempt, governments, hospitals and healthcare providers have put much effort in providing anonymous clinical datasets for diabetes-related research. The commonly used Pima Indians Diabetes dataset [8] comprises 768 records of females who are at least 21 years old. Each record has eight independent medical indicators, such as glucose, blood pressure (BP) and Body Mass Index (BMI), and one outcome indicating whether the patient has diabetes or not. The Jackson Heart Study (JHS) dataset comprises 3,340 participants' health records obtained from interviews during clinic visits including demographics, socioeconomic status, lifestyle data, medication use, and other sociocultural parameters from 2000 to 2004, which were collected to predict Type-2 diabetes [9]. The MarketScan Commercial Claims and Encounter (CCAE) dataset, produced by IBM Watson Health, comprises records of 13,050 Type-2 diabetes patients between 19 and 64 years old from 2011 to 2015. The records cover patients' demographics and medications for predicting twelve kinds of complications after the onset of diabetes [10]. In this paper, we collect a large-scale anonymous diabetes-related dataset named CDMS in Singapore from 2010 to 2017. As shown in Table I, our CDMS dataset is much larger than the existing datasets. Additionally, comparing to the afore-reviewed datasets, CDMS covers multiple ethnic groups, including Chinese, Malay, Indian, Eurasian, Caucasian and other ethnicities, due to the multiracial nature of Singapore.

TABLE I
COMPARISONS OF DIABETES DATASETS

Dataset	# patients	Time span (year)	# input variables
Pima	768	1965-1970	8
JHS	3,340	2000-2004	20
CCAE	13,050	2011-2015	21
CDMS (Ours)	**65,259**	**2010-2017**	**20**

In the attempt to accurately predict diabetes, numerous ML algorithms and data mining techniques have been exploited. Support vector machines (SVM), K-means clustering and decision tree algorithms are among the most widely used methods on the Pima dataset, with SVM being the most competent algorithm for this binary classification problem [11]. Logistic Regression and Random Forest are shown to be effective on the JHS dataset [12]. For the CCAE dataset, the survival analysis approach was adopted to model the longitudinal observations for diabetic complication prediction [10]. However, the insights conveyed by these research results may be limited due to the datasets' constraints, such as small sample sizes, high data density and few predictive indicators in the medical domain. The potential challenges of using huge datasets comprising millions of health records to build a data-driven diabetes predictive model are rarely discussed in details in the literature. In this paper, we mainly study the missing value imputation problem on our collected large-scale dataset.

B. Missing Value Imputation Methods

The presence of missing values is prevalent in the medical domains [13]. The loss of information in data samples, especially, the lack of potential predictive indicators, can be detrimental to the performance of predictive models. While discarding data samples with missing values may lead to inferior results due to the elimination of representative information, Missing Value Imputation (MVI) methods may preserve all the data samples by substituting the missing values with estimated values based on other available information. MVI methods can be categorized into single imputation and multiple imputation methods based on whether the imputed values are treated as stationary or not [14]. The latter in the literature mostly assumes data missing-at-random.

Mean Imputation (MI) is a popular single imputation approach in which the missing values of one attribute or predictive variable are replaced by the mean of other observed values for this variable [5]. However, if a variable varies a lot from patient to patient, MI introduces bias to the model due to the ignorance of large population variance. In contrast, the k-Nearest Neighbours Imputation (KNNI) approach only considers similar data samples characterized by other available attributes [15], and ignores dissimilar samples that may introduce bias. Multiple imputation methods, on the other hand, treat the missing values of an attribute as a dependent variable and it can be iteratively updated throughout the analysis process. Multiple Imputation by Chained Equations (MICE) is one such approach where the imputed values are drawn from a distribution multiple times and determined via statistical analysis iteratively until convergence [16]. Imputing data multiple times may produce more robust values than single imputation methods do. However, when the dataset is large in size and the data features are complex, due to non-linearity and high dimensionality, as is the case with our CDMS dataset, MICE incurs heavy computation and thus is difficult to apply. More recent approaches use deep learning for multiple imputation [17]. Models such as autoencoders have shown better predictive performance than MICE on heterogeneous data [6], [7]. In this paper, we extensively explore the popular MVI methods, aiming to show the benefit of handling missing values in diabetes predictive models.

III. Using CDMS for Diabetes Prediction

In this section, we introduce the details of the collected large-scale diabetes-related dataset. In addition, we present the details of the data preprocessing steps, data imputation methods and predictive models used in this research work.

A. Chronic Disease Management System (CDMS) Dataset

Health records of the following patients who visited the polyclinics managed by the National Healthcare Group, Singapore, and were diagnosed with pre-DM or DM and also diagnosed with hypertension or dyslipidemia from 2010 to 2017 are collected in CDMS. All together, there are 225,051 patients and 9,258,902 clinic visit records in the initially extracted CDMS dataset. Relevant ethic approval (NHG DSRB Ref: 2020/00714) has been obtained to conduct relevant research activities and all the personal identifiable information has been removed when extracting the dataset.

CDMS includes clinic visit records, personal health data, laboratory test results, medical diagnoses, pharmacy records, complications, and other peripheral information of patients with hypertension, dyslipidaemia or DM. Because our point-of-interest is the prediction and diagnosis of DM, we further extract relevant records of only the pre-DM and DM patients from the initial CDMS dataset to collect DM-related data (see first three rows in Table II). As a result, the extracted dataset comprises a total number of 713,968 health records of 65,259 patients who were either diagnosed with pre-DM or DM. Note that we exclude all data samples of Type-1 diabetes patients.

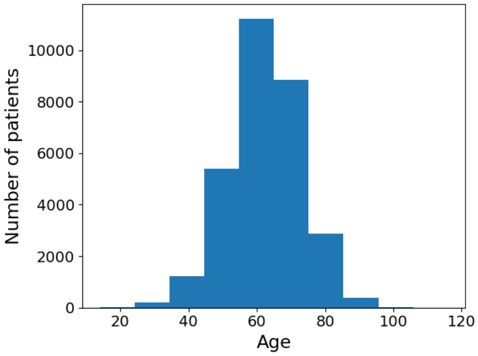

Fig. 1. Age distribution of the diabetes cohort in CDMS.

Variables in CDMS include demographics (age, gender and ethnic group, etc.), physical examination records (weight, BMI, blood pressure, etc.), laboratory test records (glycated hemoglobin test, lipoprotein cholesterol test, glucose tolerance test, etc.), DM-related comorbidities and complications, diagnostic descriptions, etc. In general, there are three types of medical data: numerical data (age, weight, blood pressure, count of visits, etc.), categorical data (disease type, ethnic group, smoking status, etc.), and descriptive data (diagnostic descriptions in natural language) [18]. In this research work, we mainly use a subset of the numerical and categorical variables in CDMS that are closely related to our diabetes prediction study (see Table IV).

B. Demographic Characteristics od CDMS

In this subsection, we analyse the demographic characteristics of pre-DM and DM patients in CDMS. In CDMS, the patients are from six ethnic groups, namely *Chinese*, *Malay*, *Indian*, *Eurasian*, *Caucasian* and *Others*, where *Others* refer to all the other ethnicities apart from the first five.

To better understand the demographic characteristics of the diabetes cohort, we divide the patients into the following four categories: (i) *pre-DM_only* (patients with diagnostic records of only pre-DM), (ii) *pre-DM_to_DM* (patients with diagnostic records of both pre-DM and DM), (iii) *DM_only* (patients with diagnostic records of only DM), and (iv) *non_DM* (patients without any DM-related diagnostic record). It is worth mentioning that a patient is classified as *DM_only* because this patient has no diagnostic records of pre-DM in the dataset, which does not mean this patient was diagnosed with DM without having pre-DM first. We take the records of *pre-DM_only* and *pre-DM_to_DM* patients to support diabetes prediction. As shown in Fig. 1, the age of the extracted

TABLE II
ETHNICITY AND GENDER DISTRIBUTION OF DIFFERENT CATEGORIES

	Ethnicity						Gender		Total
	Chinese	Malay	Indian	Eurasian	Caucasian	Others	Female	Male	
pre-DM_only	10,984	1,172	1,036	42	4	538	7,132	6,644	13,776
pre-DM_to_DM	6,001	827	755	20	1	344	4,262	3,686	7,948
DM_only	29,325	6,067	5,669	150	10	2,314	20,839	22,696	43,535
non_DM	121,926	15,429	14,509	629	73	7,226	84,608	75,184	159,792

TABLE III
GENDER DISTRIBUTION BY ETHNICITY

		Chinese	Malay	Indian
Gender	Female	87,257	12,686	10,974
	Male	80,979	10,809	10,995

diabetes cohort ranges from 14 to 116 while most of patients fall in the range of 50∼80.

Furthermore, we analyse the demographic composition of the diabetes cohort by gender and ethnicity. As shown in Table II, four patient categories have roughly the same ethnic and gender distribution. Ethnic distribution shows the highest proportion of Chinese (around 80%) while Malay and Indian account for about 8∼10%, respectively. Eurasian, Caucasian and other ethnicities make up the rest population in CDMS. Although there is a slightly higher proportion of female than that of male, the gender distribution is overall balanced. The ratio among various ethnic groups having either pre-DM or DM shown in Table II is similar to the ethnic groups' composition in Singapore. Therefore, we do not select patients from any specific ethnic group(s) but use all of their data for experiments because CDMS is a national-wide unbiased dataset.

Due to the fact that the ethnic groups of Chinese, Indian and Malay take up over 95% of all the data samples, we further analyse the statistics within these ethnicities. The gender distribution of these three ethnicities are shown in Table III. Female and male are almost evenly distributed across the three ethnicities in each category.

C. Data Preprocessing

Data preprocessing plays a critical role in the overall prediction pipeline because the quality of the data greatly affects the prediction performance. To apply a ML or DL model for diabetes prediction, we take the records of *pre-DM_only* and *pre-DM_to_DM* patients (167,503 records from 21,724 patients in total). Then, we make the following efforts to preprocess the dataset.

1) Predictor Variables: We analyze the medical implications of each attribute and its relevance to DM. In this study, we identify 21 attributes (see Table IV) based on both the suggestions of clinicians and literature review. The 21st attribute (i.e., *Disease_Type*: {pre-DM, DM}) is taken as the label of each data sample, therefore, a binary classification task can be conducted using this dataset. However, relevant lab tests and physical examinations may be carried out on a date different from the diagnosis date. To ensure the validity of the extracted data samples, laboratory tests and physical examinations are considered to be valid only if they were performed close to the diagnosis date. Specifically, we only take laboratory tests and physical examination results when they were performed within a certain period of time before/after the diagnosis date. For example, a patient was diagnosed with DM on 1st May 2015, but the *OGTT2hr* test result was missing on that date. Because the validity period of *OGTT2hr* is set as "0∼30 days before/after diagnosis date", then the most recent *OGTT2hr* test result between 1st April 2015 and 31st May 2015 is used for this data sample. More details on how the data samples are extracted based on the closest timing are presented in Table IV.

2) Removal of Outliers: It is inevitable that there are errors and inaccuracies in the records due to manual mistakes, machine failures or even database crashes. According to clinicians' guidance, we remove records which are obviously unrealistic, such as inappropriate negative values and values exceeding the reasonable range. The reasonable value restriction ranges are presented in Table IV. Besides, there are unexplained records in the dataset. For instance, "Unknown" was recorded in the field *Ethnicity*. Nonetheless, CDMS only has 21 unreasonable records and 497 unexplained records among the total number of 0.17 million selected records. Thus, we deem removing these 518 records has no discernible impact on the quality of the overall dataset. After the removal of unreasonable and unexplained records, the number of records in CDMS becomes 166,985, which is sufficient for training and testing predictive models.

3) One-hot Encoding: For categorical attributes in CDMS, we apply one-hot encoding on them. One-hot encoding is capable of converting categorical data, especially those having no ranking orders associated with the values, into numerical inputs that can be readily used by machine learning algorithms. For instance, with six categories of *Ethnicity* in CDMS, we correspondingly obtain six binary variables for each record after applying one-hot encoding. Among these six variables, one and only one of them has the value of 1 with the others set to 0, indicating the respective ethnicity.

4) Missing Value Imputation: Missing values (MVs) are prevalent in most real-world datasets, which could lead to increase in errors for algorithms requiring a large number of indicators [18], [19]. In our study, the setting of the validity period (see Table IV) does not alleviate the critical MV issue. As mentioned in Section III-C1, in our study, not all laboratory

TABLE IV
ATTRIBUTES IN THE CDMS DATASET

Index	Attributes	Descriptions	Range	Period of validity
1	*Gender*	female, male	-	-
2	*Age*	age in months when the disease is confirmed	-	-
3	*Race*	Chinese, Malay, Indian, Eurasian, Caucasian, Others	-	-
4	*Weight*	weight in kilograms	(, 300]	0∼30 days before/after diagnosis date
5	*BMI*	Body Mass Index in kg/m^2	[5, 150]	0∼30 days before/after diagnosis date
6	*OGTT2hr*	2 hour oral glucose tolerance test value in mmol/L	[1, 40]	0∼30 days before/after diagnosis date
7	*FPG*	Fasting Plasma Glucose value in mmol/L	[1, 40]	0∼30 days before/after diagnosis date
8	*HbA1c*	glycated hemoglobin test value in %	[3, 20]	0∼30 days before/after diagnosis date
9	*LDLc*	low-density lipoprotein cholesterol test value in mmol/L	[0.1, 10]	0∼30 days before/after diagnosis date
10	*HDLc*	high-density lipoprotein cholesterol test value in mmol/L	[0.1, 10]	0∼30 days before/after diagnosis date
11	*TG*	triglycerides test value in mmol/L	[0.1, 30]	0∼30 days before/after diagnosis date
12	*TC*	total cholesterol test value in mmol/L	[0.1, 30]	0∼30 days before/after diagnosis date
13	*Smoking_Status*	non-smoker, ex-smoker, smoker	-	0∼6 months before/after diagnosis date
14	*SBP*	systolic blood pressure in mmHG	[50, 300]	0∼30 days before/after diagnosis date
15	*DBP*	diastolic blood pressure in mmHG	[30, 300]	0∼30 days before/after diagnosis date
16	*Hypertension*	yes if one diagnosed with hypertension, otherwise no	-	before diagnosis date
17	*Dyslipidaemia*	yes if one diagnosed with dyslipidaemia, otherwise no	-	before diagnosis date
18	*Surgical_procedure*	yes if one has surgical in the past one year, otherwise no	-	0∼1 year before diagnosis date
19	*No_of_complications*	number of complications	-	before diagnosis date
20	*Counts_of_visits*	counts of recent visits to the hospital	-	the same calendar year
21	*Disease_Type*	pre-DM, DM	-	-

tests and physical examinations are considered as valid records to ensure data quality. As a consequence, data samples are fairly sparse in our dataset.

Accurate prediction in the presence of large number of MVs in the dataset has always been a challenging problem [20]. Most hybrid models address this challenge by either removing missing data instances from the dataset (often referred to as case deletion) or using data imputation methods to fill in the MVs with certain values. For the purpose of analyzing the influence of various MV imputation methods, we conduct extensive experiments to identify the most appropriate MV imputation method to better handle our dataset. The evaluated MV imputation methods in this study are listed as follows:

- *Case Deletion*: Data samples containing any MVs in the attributes are removed from the dataset.
- *Zero Imputation (ZI)*: This method inserts zeroes to MVs. Many studies have experimentally confirmed that zero imputation leads to sub-optimal performances on the prediction tasks [21].
- *Concept Most Common (CMC)*: Based on the Most Common (MC) method which replaces numerical MVs with the mean values and replaces categorical MVs with the most common categories, CMC is similar to the MC method for computing MVs, but it only considers missing instances in the same category.
- *k-Nearest Neighbor Imputation (KNNI)*: This method finds the k nearest neighbors of the current data instance having MVs, and then fills in MVs with the average value for the numerical attributes, and the most common value among all the nearest neighbors for the categorical attributes. The nearest neighbors are identified using a predefined distance metric, which is also referred to as the dissimilarity metric. The most commonly used distance metrics are Euclidean and Manhattan distances. Given two data instances $x_i, x_j \in \mathbb{R}^M$ where M denotes the number of feature dimensions, the weighted Euclidean and Manhattan distances are defined as follows:

$$d_{\text{Euclidean}}(x_i, x_j) = \sqrt{\sum\nolimits_{p=1}^{M} w_p (x_{i,p} - x_{j,p})^2}, \tag{1}$$

$$d_{\text{Manhattan}}(x_i, x_j) = \sum\nolimits_{p=1}^{M} w_p \left| x_{i,p} - x_{j,p} \right|, \tag{2}$$

  where w_p denotes the weight of the pth feature. Uniform weights are used in our experiments, i.e., the weights of all the M features are equal.
- *Multivariate Imputation by Chained Equation (MICE)*: This method works by repeatedly filling in MVs. Multiple Imputations (MIs), as opposed to Single Imputations (SIs), are far superior because it takes the statistical uncertainty in the imputations into consideration. This method is quite adaptable, and works well with both numerical and categorical variables [22].
- *DataWig*: It is a robust and scalable method that uses deep neural networks to impute MVs in the dataset. This method can handle both numerical and categorical variables [6].

5) Data Splitting and Standardization: We split the dataset into training/testing sets with the 80%/20% ratio. Then we apply zero-mean, unit-variance standardization (w.r.t the training set) on the input features before feeding them into the predictive models.

D. Machine Learning Models for Diabetes Prediction

In this research work, we apply seven ML/DL models for diabetes prediction, which are listed as follows:

1) Logistic Regression (LR): This classification algorithm trains a statistical model using a logistic function to model a binary dependent variable. Although being simple in its dynamics, LR is widely adopted because it is interpretable and often achieves robust and competitive performance in real-world applications [23].

2) Decision Tree (DT): A DT is a flowchart-like structure in which each node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label.

3) Random Forest (RF): A random forest comprises a large number of independent DTs operating as a union. Each DT in the RF outputs a categorical prediction, and the category with the most votes becomes the overall model's prediction.

4) k Nearest Neighbor (KNN): KNN is a classical classification algorithm, which is built on the assumption that similar/neighboring data instances belong to the same category. Other researchers' prior studies show the feasibility to employ KNN and its extensions for diabetes prediction [24].

5) Support Vector Machine (SVM): SVM is a supervised ML algorithm. In SVM, data instances are projected onto a higher dimensionality, wherein data instances belonging to two classes can be relatively easily separated.

6) Long Short-Term Memory (LSTM): The LSTM architecture, which is an extension of the classical recurrent neural network (RNN). In [25], LSTM was employed to enhance patient health status prediction, especially diabetes prediction.

7) Gated Recurrent Unit (GRU): GRU could be considered as a variation of LSTM. GRU uses two types of gates, namely update gates and reset gates, to model the long-term and short-term memory, respectively. Because GRU has no extra memory cell to store information, it can only control information within the unit. GRU has been shown to outperform classical ML approaches on DM diagnosis [26].

IV. Experimental Results and Discussions

In this section, we present and discuss the extensive experimental results including MVs analysis, and prediction performance of various combinations of data imputation methods and ML/DL models.

A. Missing Value Analysis

As discussed in Section III-C, 166,985 data samples are extracted and used for a binary classification task among which 86,106 are pre-DM and 80,879 are Type-2 DM records. However, laboratory tests and physical examinations are considered to be valid only if they were performed close to the diagnosis date to ensure data validity. As a result, there are a large number of data samples with MVs in CDMS. Before applying data imputation methods to CDMS, we conduct MVs analysis to better understand the data samples having MVs.

In CDMS, the missing percentage is significantly high especially for certain DM-related laboratory test results. To gain a better insight, we split data samples of pre-DM and DM, and show the missing percentage of all attributes having MVs in Table V. We find that although *HbA1c* seems to be a routine examination for DM patients (only 15.69% missing among all hospital visits), while clinicians diagnose pre-DM (92.67% missing) without this test. It is curious to us that *OGTT2hr*, as a blood test frequently used to screen diabetes patients, has a quite high missing percentage in both pre-DM and DM samples (80.20% and 94.57%, respectively). It seems that DM-related laboratory tests, such as *OGTT2hr* and *FPG*, are not routine medical tests even for DM patients, not to mention patients with pre-DM. Moreover, over half of the laboratory test results and less than half of physical examination are missing in CDMS.

TABLE V
Prevalence of MVs in the extracted CDMS dataset

Index	Attribute	pre-DM		DM	
		Missing%	Missing% in ERs	Missing%	Missing% in ERs
1	*OGTT2hr*	80.20	61.74	94.57	51.01
2	*FPG*	48.17	31.54	89.96	27.75
3	*LDLc*	53.78	51.34	55.38	53.68
4	*HDLc*	53.54	50.91	55.06	53.11
7	*HbA1c*	92.67	92.19	15.69	24.79
5	*TC*	53.51	50.90	55.05	53.09
6	*TG*	53.52	50.90	55.05	53.09
8	*Weight*	20.90	22.29	22.44	17.86
9	*BMI*	20.90	22.28	22.43	17.84
10	*SBP*	5.03	7.40	3.97	5.20
11	*DBP*	5.03	7.40	3.98	5.20

ERs: Earliest (diagnostic) records

DM patients could visit a doctor due to various reasons, including but not limited to DM, which could partially explain the overall high missing percentage. Normally, a patient diagnosed with DM for the first time based on the relevant laboratory tests, i.e., the earliest records (ERs) of DM diagnosis, is more likely to include those relevant laboratory tests and physical examinations. In order to ensure the data integrity, we only take the ERs of 21,615 pre-DM samples and 7,947 DM samples. The missing percentage of all records and ERs are given in Table V. It is obvious that there are considerable falls in the missing percentage of *OGTT2hr* and *FPG* in ERs. However, the high proportion of MVs in *HbA1c* still stands as a non-negligible problem (see Section IV-B2).

B. Data Imputation and Classification

It is evident that as a large-scale real-world medical dataset, there are serious data missing issues in CDMS. As shown in Table V, all the attributes with MVs are numerical. Hence, we only present and discuss data imputation methods for numerical MVs in our study. Furthermore, we conduct extensive experiments using the data imputation methods introduced in Section III-C4 and ML/DL models introduced in Section III-D. The model implementation details are described as follows.

1) Case Deletion: If we remove all the data samples having MVs in CDMS, due to the high prevalence of MVs, we will end up with insufficient data to train any ML/DL model. Therefore, we do not investigate the performance of case deletion in our subsequent experiments.

TABLE VI
RESULTS WITH MULTIPLE DATA IMPUTATION METHODS (NO *HbA1c*)

	Model	Accuracy	Precision	Recall	F1-score
ZI	LR	0.756	0.826	0.120	0.209
	DT	0.783	0.599	0.587	0.593
	RF	0.850	0.792	0.597	0.681
	KNN	0.760	0.602	0.321	0.419
	SVM	0.814	**0.875**	0.359	0.509
	LSTM	0.748	0.853	0.075	0.138
	GRU	0.757	0.794	0.128	0.221
	Mean	0.781	0.763	0.312	0.396
CMC	LR	0.825	0.824	0.442	0.575
	DT	0.778	0.585	0.597	0.591
	RF	0.842	0.772	0.587	0.667
	KNN	0.772	0.644	0.338	0.444
	SVM	0.824	0.826	0.437	0.572
	LSTM	0.821	0.813	0.432	0.564
	GRU	0.828	0.805	0.474	0.597
	Mean	0.813	0.753	0.472	0.573
KNNI	LR	0.841	0.851	0.491	0.625
	DT	0.826	0.769	0.506	0.610
	RF	**0.854**	0.832	0.574	0.680
	KNN	0.786	0.694	0.365	0.479
	SVM	0.845	0.844	0.521	0.645
	LSTM	0.839	0.851	0.488	0.620
	GRU	0.849	0.833	0.548	0.661
	Mean	**0.834**	**0.811**	**0.499**	**0.617**
MICE	LR	0.806	0.734	0.439	0.550
	DT	0.756	0.543	0.598	0.569
	RF	0.832	0.715	0.623	0.666
	KNN	0.767	0.628	0.325	0.428
	SVM	0.817	0.749	0.481	0.586
	LSTM	0.796	0.725	0.386	0.504
	GRU	0.812	0.733	0.471	0.574
	Mean	0.798	0.690	0.475	0.554
DataWig	LR	0.826	0.788	0.480	0.597
	DT	0.792	0.611	0.620	0.615
	RF	0.848	0.708	**0.737**	**0.722**
	KNN	0.787	0.682	0.391	0.497
	SVM	0.840	0.735	0.632	0.679
	LSTM	0.797	0.847	0.332	0.477
	GRU	0.797	0.847	0.332	0.477
	Mean	0.812	0.745	0.503	0.581

2) ZI: We insert zeroes to all MVs in CDMS, and then use the zero-imputed data samples for diabetes prediction. Intriguingly, we find that almost every model could achieve a quite high accuracy on the binary classification task. The key reason is due to the high proportion and notable imbalance of MVs. As shown in Table V, there are notably a large number of MVs among all the presented attributes, especially for the DM-related laboratory test results. Moreover, the missing percentage of *HbA1c* is extremely high and imbalanced in pre-DM and DM categories (92.19% and 24.79%, respectively). If we impute zeroes to MVs in *HbA1c*, the predictive model will be trained erroneously, i.e., it could simply classify data samples with zeroes in *HbA1c* as pre-DM, which would result in a high classification accuracy. As such, we remove the *HbA1c* attribute in all the subsequent experiments because imputing zeroes implicitly reveals label information to data samples with fairly imbalanced MVs.

As shown in Table VI, ZI achieves acceptable performance on accuracy and precision, but its recall is relatively low among all imputation methods, especially for LR (0.120) and LSTM (0.075). Similar to ZI, we also conduct experiments on imputing constant values to MVs. As shown in Fig. 2, imputing MVs using 5 or 10 considerably improves the recall and F1-score, while the performance is worse when using 15 or 20. Imputing MVs using constant values seems feasible, but it remains as an open problem on how to determine the optimal value for imputation, which we leave for future work.

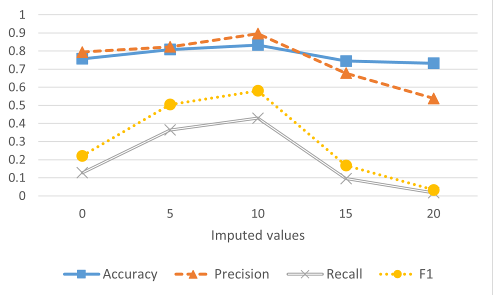

Fig. 2. Results of constant imputation using GRU.

3) CMC: In our study, according to the data distribution presented in Section III-B, we select *Gender*, *Age* and *Ethnicity* as the common concepts used for data imputation. Although CMC is straightforward and easy to implement, it achieves quite good performance on all predictive models. As shown in Table VI, most models can achieve the accuracy of around 0.8 using CMC. Compared with other imputation methods, CMC achieves the competitive performance on Precisoin and Recall partially because the CDMS dataset is large in size, which makes CMC relatively robust by having more reference values from the same concepts.

4) KNNI: As shown in Table V, over half of the attributes have MVs in CDMS. Therefore, we employ KNNI over those attributes without MVs. Specifically, we adopt the Euclidean distance metric (see (1)) in our study. As shown in Table VI, KNNI outperforms the other imputation methods. In terms of accuracy, KNNI obtains competitive performance (0.834 on average), which is better than that of the other four imputation methods. KNNI also outperforms all the other imputation methods on precision, recall and F1-score. It is worth mentioning that, k, as the only hyper-parameter of this algorithm, is set to 100 in our experiments. As shown in Table VII, the prediction performance is not sensitive to different values of k.

5) MICE & DataWig: MICE imputes MVs repeatedly considering statistical uncertainty in multiple imputations while DataWig trains neural networks over the data. Therefore, these two imputation methods require more expensive computations, especially for large-scale datasets like CDMS. In spite of this, these two methods obtain inferior performance. For MICE and DataWig, the accuracy mostly lies between 0.7∼0.8 and recall lies between 0.3∼0.6. In terms of precision, the overall performance of these two methods (around 0.7∼0.75) is worse than KNNI (over 0.8 on average) as well.

In summary, KNNI achieves the best prediction performance among all imputation methods for our CDMS dataset, while CMC achieves decent performance. Despite the relatively high computation cost, MICE and DataWig obtain inferior results comparing with KNNI and CMC. ZI is shown as not suitable

TABLE VII
RESULTS WITH DIFFERENT k VALUES IN KNNI USING LR

k	Accuracy	Precision	Recall	F1-score
50	0.819	0.796	0.439	0.566
100	0.825	0.824	0.442	0.575
500	0.834	0.866	0.425	0.594
1000	0.832	0.860	0.449	0.589

for data imputation in CDMS due to the imbalance distribution of MVs between pre-DM and DM samples.

V. Conclusion and Future Work

In this paper, we introduce a large-scale diabetes-related dataset CDMS, which was collected in Singapore from 2010 to 2017. The large proportion of missing values caused by irregular and distinct clinic visits poses a great challenge to machine learning models. Hence, we conduct extensive investigations on the missing value problem and analyze multiple data imputation techniques. The experimental results show that KNNI is capable of handling MVs in CDMS and predictive models using KNNI achieve the best performance on the task of classifying pre-DM and DM patients.

In this study, the advantages of the multiracial nature of CDMS are not well exploited. Going forward, we plan to investigate the complex intrinsic relationships among various ethnicities in diabetes-related tasks.

VI. Acknowledgements

This research is supported, in part, by the National Research Foundation (NRF), Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-003). H. Qian thanks the support from the Wallenberg-NTU Presidential Postdoctoral Fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

References

[1] M. Ravaut, H. Sadeghi, K. K. Leung, M. Volkovs, K. Kornas, V. Harish, T. Watson, G. F. Lewis, A. Weisman, T. Poutanen *et al.*, "Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–12, 2021.

[2] P. Dworzynski, M. Aasbrenn, K. Rostgaard, M. Melbye, T. A. Gerds, H. Hjalgrim, and T. H. Pers, "Nationwide prediction of type 2 diabetes comorbidities," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[3] B. H. Heng, Y. Sun, J. T. Cheah, and M. Jong, "The Singapore National Healthcare Group diabetes registry—descriptive epidemiology of type 2 diabetes mellitus," *Annals Academy of Medicine Singapore*, vol. 39, no. 5, p. 348, 2010.

[4] D. Wang, C. Quek, and G. S. Ng, "Ovarian cancer diagnosis using a hybrid intelligent system with simple yet convincing rules," *Applied Soft Computing*, vol. 20, pp. 25–39, 2014.

[5] S. Wang, M. E. Celebi, Y.-D. Zhang, X. Yu, S. Lu, X. Yao, Q. Zhou, M.-G. Miguel, Y. Tian, J. M. Gorriz *et al.*, "Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects," *Information Fusion*, vol. 76, pp. 376–421, 2021.

[6] F. Biessmann, T. Rukat, P. Schmidt, P. Naidu, S. Schelter, A. Taptunov, D. Lange, and D. Salinas, "Datawig: Missing value imputation for tables." *Journal of Machine Learning Research*, vol. 20, pp. 175–1, 2019.

[7] P.-A. Mattei and J. Frellsen, "MIWAE: Deep generative modelling and imputation of incomplete data sets," in *International Conference on Machine Learning*, 2019, pp. 4413–4423.

[8] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 1988, p. 261.

[9] V. S. Effoe, A. Correa, H. Chen, M. E. Lacy, and A. G. Bertoni, "High-sensitivity C-reactive protein is associated with incident type 2 diabetes among African Americans: The Jackson Heart Study," *Diabetes Care*, vol. 38, no. 9, pp. 1694–1700, 2015.

[10] B. Liu, Y. Li, S. Ghosh, Z. Sun, K. Ng, and J. Hu, "Simultaneous modeling of multiple complications for risk profiling in diabetes care," *arXiv preprint arXiv:1802.06476*, 2018.

[11] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[12] R. Casanova, S. Saldana, S. L. Simpson, M. E. Lacy, A. R. Subauste, C. Blackshear, L. Wagenknecht, and A. G. Bertoni, "Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning," *PloS One*, vol. 11, no. 10, p. e0163942, 2016.

[13] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: Role of missing value and outliers," *Journal of Medical Systems*, vol. 42, no. 5, pp. 1–17, 2018.

[14] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. New York: Wiley, 1987.

[15] G. E. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 519–533, 2003.

[16] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?" *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011.

[17] T. M. Whitehead, B. W. Irwin, P. Hunt, M. D. Segall, and G. J. Conduit, "Imputation of assay bioactivity data using deep learning," *Journal of Chemical Information and Modeling*, vol. 59, no. 3, pp. 1197–1204, 2019.

[18] J. Paetz, "Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activation functions," *Artificial Intelligence in Medicine*, vol. 28, no. 2, pp. 207–230, 2003.

[19] S.-F. Huang and C.-H. Cheng, "A safe-region imputation method for handling medical data with missing values," *Symmetry*, vol. 12, no. 11, p. 1792, 2020.

[20] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5621–5631, 2015.

[21] J. Yi, J. Lee, K. J. Kim, S. J. Hwang, and E. Yang, "Why not to use zero imputation? Correcting sparsity bias in training neural networks," *arXiv preprint arXiv:1906.00150*, 2019.

[22] S. Van Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, pp. 1–67, 2011.

[23] D. Wang, X. Qian, C. Quek, A.-H. Tan, C. Miao, X. Zhang, G. S. Ng, and Y. Zhou, "An interpretable neural fuzzy inference system for predictions of underpricing in initial public offerings," *Neurocomputing*, vol. 319, pp. 102–117, 2018.

[24] Y. Du, A. R. Rafferty, F. M. McAuliffe, L. Wei, and C. Mooney, "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus," *Scientific Reports*, vol. 12, no. 1, pp. 1–14, 2022.

[25] A. Massaro, V. Maritati, D. Giannone, D. Convertini, and A. Galiano, "LSTM DSS automatism and dataset optimization for diabetes prediction," *Applied Sciences*, vol. 9, no. 17, p. 3532, 2019.

[26] Z. Alhassan, A. S. McGough, R. Alshammari, T. Daghstani, D. Budgen, and N. Al Moubayed, "Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models," in *International Conference on Artificial Neural Networks*, 2018, pp. 468–478.