IEEE International Conference on Bio-Engineering for Smart Technologies (BioSMART), Paris, France, December 2021

Using Machine Learning to Predict Diabetes Complications

Yazan Jian, Michel Pasquier, Assim Sagahyroon, Fadi Aloul
Department of Computer Science and Engineering
American University of Sharjah, UAE
{b00087296, mpasquier, asagahyroon, faloul}@aus.edu

***Abstract*— Diabetes Mellitus (DM) is a chronic disease that is considered to be life threatening. It can affect any part of the body over time, resulting in more serious complications such as Dyslipidemia, Neuropathy and Retinopathy. In this work, different supervised classification algorithms were applied to build several models to predict and diagnose eight diabetes complications. The complications include: Metabolic Syndrome, Dyslipidemia, Neuropathy, Nephropathy, Diabetic Foot, Hypertension, Obesity, and Retinopathy. For this study, a dataset collected by the Rashid Centre for Diabetes and Research (RCDR) located in Ajman, UAE, was utilized. The dataset contains 884 records with 79 features. Some essential preprocessing steps were applied to handle the missing values and unbalanced data problems. Multiple solutions were tested and evaluated.**

Keywords*— Diabetes Prediction, Diabetes Complications, Supervised Learning.*

I. Introduction

Chronic diseases are defined broadly as conditions that last 1 year or more and require ongoing medical attention or limit activities of daily living or both [1]. Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin (Type 1) or when the body cannot effectively use the insulin it produces (Type 2) [2]. According to the World Health Organization (WHO), the number of people with diabetes in 2014 was 422 million. Moreover, in 2016, diabetes was the direct cause of 1.6 million deaths [2].

There are different risk factors for diabetes, especially diabetes Type 2. For instance, age, family history of diabetes, high blood pressure, high level of triglycerides, are all considered as risk factors for diabetes [3]. As mentioned by CDC [4], diabetes can affect any part of the body over time. For example, diabetes can lead to different complications such as hypertension, neuropathy (nerve damage), nephropathy (disease of kidneys), and much more. As a result, it is very important to understand how to deal with diabetes and how to prevent such possible complications.

To reduce the possibility of developing serious complications related to diabetes, several research areas need to be studied. One way of doing so is by applying machine learning and data mining techniques on diabetes-related data sets. This research is making use of several supervised machine learning techniques to predict some of the complications related to diabetes. The dataset in hand consists of various complications such as metabolic syndrome, dyslipidemia, neuropathy, nephropathy, diabetic foot, hypertension, obesity and retinopathy. Furthermore, Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (CART), Random Forest (RF), AdaBoost and XGBoost were all utilized to build and evaluate different classifiers.

II. Literature Review

The applications of data mining in healthcare are an emerging field that enables disease diagnosis, prediction, and deep understanding of medical data [5], [6]. For instance, it increases the chances of better understanding the correlation between different chronic diseases [7], such as Diabetes Mellitus (DM). It is a serious health problem and cause of death. Few existing studies have reported the use of machine learning to develop prediction models of diabetes complications. For instance, the authors in [8] built a model to predict chronic complications, especially eye disease, kidney disease, coronary heart disease and hyperlipidemia. The authors started with a dataset of 455 records. The number of records was decreased through data selection and cleaning. The final number of records as well as the number of features used to build the model were not mentioned in the paper. The authors used Iterative Decision Tree (ID3) algorithm for building the model [9]. A 10-fold cross validation was used to train the model and evaluate its performance, yielding an accuracy of 92.35%. It is worth mentioning that the high accuracy score in this case is not enough to indicate the performance of the model, especially in case of unbalanced data. This is mainly because a model can ignore the minority class by predicting all the instances as the majority class and still achieve good accuracy scores.

Dagliati et al. [10] focused on predicting the onset of retinopathy, neuropathy, or nephropathy in T2DM patients at different time scenarios, at 3, 5 and 7 years from the first visit at the hospital. The first visit at the hospital provides the patient's health status. The selection of patients in this study consists of the following criteria: patient has a follow-up time longer than the corresponding temporal threshold (3, 5 or 7); Patient develops the complication after the first visit; Patient's complication onset date has been registered. The dataset has been collected by Istituto Clinico Scientifico Maugeri (ICSM), Hospital of Pavia, Italy for over than 10 years. It contains 943 records with the following features: gender, age, time from diagnosis, body mass index (BMI), glycated hemoglobin (HbA1c), hypertension, and smoking habit. The classification models used were Naïve Bayes (NB), LR, SVM, and RF. The missing data has been handled using missForest [11], whereas the unbalanced class problem was solved by oversampling the minority class. According to the paper, the maximum accuracy score was reached by LR with 83.8%.

The authors in [12] focus only on studying one complication which is sarcopenia, which is a geriatric syndrome closely related to the prevalence of type 2 diabetes mellitus (T2DM). The goal for this paper is to make risk assessment of sarcopenia easier by building ML models using SVM and RF. The dataset used in the paper, which is clearly limited in size, has 132 records of patients aged over 65 and diagnosed with T2DM. It contains several records for each patient, such as age, duration of diabetes, history of hypertension, smoking and drinking hobbits as well as some medical records like serum albumin and 25-OH-Vitamin D3. The missing value problem has been solved using k-NN imputer with k set to 10. As mentioned in the paper, the area under the receiver operating characteristic curve (AUC) was over 0.7 and the mean AUC of SVM models was higher than that of RF.

From the previous literature, some limitations can be noticed, and especially related to the datasets employed. For

example, the number of studied complications is very limited, as it does not exceed two to three complications in most of the available literature. Moreover, there is a clear limitation when it comes to the number of features used in each study and the nature of these features. For instance, the number of available medical tests in [12] is very limited.

Accordingly, the objective of this research is to achieve reliable and improved results in predicting diabetes complications in diabetic patients using various, state-of-the-art machine learning algorithms by utilizing a decent UAE based dataset. Extensive number of experiments is conducted testing several data imputation methods, balancing techniques, as well as model tuning.

III. Materials and Methods

This section focuses on explaining the entire framework by first analyzing the dataset in hand. After that, several essential preprocessing steps will be discussed as well as clarifying the machine learning algorithms to be used.

A. Dataset

Utilizing a decent dataset plays a significant role for any ML problem. In this research, the dataset in hand is collected from the Rashid Centre for Diabetes and Research (RCDR) located in Ajman, UAE [13]. The dataset collected mainly consists of medical records for patients with diabetes.

The data consists of 884 patients with 79 input attributes and 8 output classes (complications). The input attributes are distributed as follows: 73 are numerical attributes and 6 nominal attributes. From the 73 numerical attributes we have 64 medical tests, including Age, Sex, BMI, HbA1c, Vitamin D Blood Pressure and Diabetes types. For the output (target) attributes, we have the main 8 complications i.e.: metabolic syndrome, dyslipidemia, neuropathy, nephropathy, diabetic foot, hypertension, obesity, and retinopathy.

B. Preprocessing

The given dataset presents issues that require several preprocessing steps. Such steps are important to properly train hence to improve the performance of the models.

1) Data Cleaning

The first step in processing the dataset is cleaning it and removing the unnecessary records and attributes by following a systematic procedure. For a start, the dataset consists of several categorical values that need to be deleted for confidentiality purposes i.e.: Hospital Number, Episode Date and Episode Description. Furthermore, the dataset consists of missing values for the diabetes type for some patients, which is a critical information in this research since we are studying diabetes complications in diabetic patients. Therefore, all the 26 instances suffering from this problem were removed.

Another step found to be needed in this study is checking the total number of missing values per record (or patient). By testing different percentages, it was found that removing all records with > 60% of missing values achieved better performance compared to other experiments where this problem was ignored.

Following the approach in [14], the missing values was also investigated per column (or attribute). Based on several experiments, a threshold of 40% was set for this step. Since this dataset consists of many numerical attributes, it was found that 16 attributes have missing values of more than 40%. More precisely, most of these attributes have more than 90% missing values. This specific threshold was selected experimentally and influenced by the literature [14].

2) Data Imputation

Handling missing values is essential in training ML classifiers since most of the available machine learning algorithms cannot be utilized with missing data. For the categorical values available in our dataset, such issues occur only with the Nationality attribute. The most frequent value in that column was thus used to fill the missing values.

On the other hand, three different methods have been extensively tested and evaluated to solve the missing values problem in numerical attributes. Mean substitution [15], k-NN [16] model and MissForest [11] were all utilized and evaluated on the dataset.

To evaluate and compare the performance of all the three algorithms, RMSE was calculated as per equation (1), for all the three methods as follows. The first step was to simulate the missing value problem by choosing a complete subset of the dataset with no missing values. The total number of records in the complete subset was 217 records. After that, the missing values percentage in the original dataset was calculated and utilized to drop random values from each column in the complete dataset. More precisely, the percentage found was 4.4%, resulting in dropping 9 records per column in the complete dataset. After building the artificial dataset, the three mentioned methods were used to impute the missing values. As noticed in Table I, it was found that MissForest results in the minimum RMSE value, hence it was utilized in this study. It is worth to mention that Table I represents the RMSE for some randomly selected attributes as well as the total RMSE for all columns.

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j-\hat{y}_j)^2} \quad (1)$$

3) Data Balancing

One of the common challenges when building and training machine learning and data mining models is dealing with unbalanced dataset, as it leads to biased learning models, which will not allow to correctly predict the minority class. Unfortunately, this problem is present in our dataset. More precisely, the neuropathy, nephropathy, retinopathy, and diabetic foot attributes all have a severe unbalanced distribution, urging the need to use some effective balancing method to address the problem. Both undersampling and oversampling techniques have been evaluated on the dataset.

One of the tested approaches is Cluster Centroids [17]. This method under samples the majority class by replacing a cluster of majority samples by the cluster centroid of a *k*-Means algorithm. The advantage of following such approach is to preserve any possible loss of data that could happen when removing random instances. Another technique is over-sampling the minority class. The Synthetic Minority Oversampling Technique (SMOTE) [18] was used. Both methods have been evaluated on the data in this research.

Table I: RMSE results for each imputation method

Method	BMI	Triglycerides	Total RMSE
MissForest	0.6264	1.2051	15.962
KNN	0.9711	1.3514	18.560
Mean Substitution	0.8972	1.3378	19.788

After experimenting with the previously mentioned balancing methods, a combination of both SMOTE and Cluster Centroids has been used for the final output. Figure 1 shows the final class distributions for all the complications. Since the severity of imbalance problem varies between the complications, we treated each complication independently.

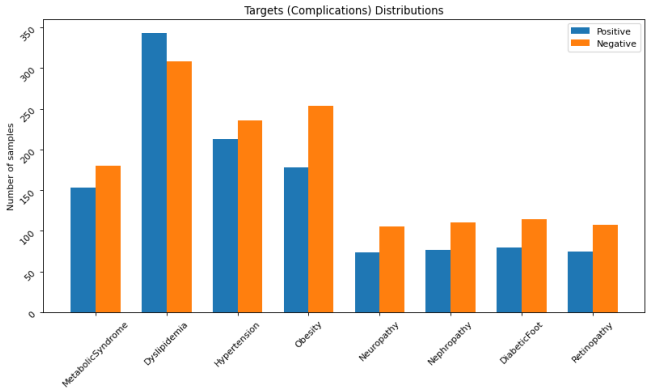

Figure 1: Classes distributions for each complication after handling imbalance problem

4) Other preprocessing steps

To ensure the best performance, three other preprocessing steps applied to the dataset. The first one is to encode the categorical attributes to numerical representation. This step has been done using one-hot encoding technique [19], resulting in 22 attributes for the original 3 categorical columns. One other essential step is to normalize the numerical values, especially because some of these features were recorded with different measurement units. Finally, the dataset has been divided into training and testing sets which helps testing if the model can generalize well to new data and its ability to avoid overfitting. Several percentage splits were evaluated in this research, the best split found to work properly is by using 80% of the total number of records for training and a 20% for testing.

C. ML Models and Evaluation Metrics

After carrying out extensive number of experiments to select the best preprocessing techniques, several ML learning models trained to classify the 8 complications. Logistic Regression [14], SVM [20], entropy-based CART-DT [21], Random Forest [22], AdaBoost and XGBoost [23], all have been utilized. The selected algorithms have proven to be most efficient in similar datasets.

To ensure reaching the best performance of each learning model, selecting the best hyperparameters for an estimator plays a significant role. In this research, GridSearchCV [24] was utilized to test a decent number of possible combinations for each estimator. The GridSearchCV model works by testing the performance of each possible combination of the given sets. To better check the performance of each complication, a cross validation with a value of 5 was utilized. In this study, all the estimators required a suitable tuning. For example, tuning Polynomial SVM model requires selecting a polynomial degree beforehand. For that purpose, different polynomial degrees were tested and evaluated. The values tested for SVM's polynomial degree contains the following list: [2, 4, 6, 8]. Decision Tree (CART) also requires selecting and tuning different parameters. Such parameters contain the max depth of the tree, the minimum samples need to perform a split and the minimum sample per leaf. It is worth mentioning that all the tuning sets used followed the best practices found in the literature.

To test the performance of the built models, several evaluation metrics have been utilized. Accuracy, F1-Score and AUC score are reported for the conducted experiments.

IV. Results and Discussion

Since we have 8 independent attributes (one column for each attribute) in the dataset, and since a patient can suffer from multiple complications at the same time, we decided to build binary classifiers for each complication utilizing all the algorithms mentioned before. Table II shows the extensive experiments accomplished. The Accuracy, F1-Score as well as AUC score are all reported for the top 3 performing models for each complication as shown.

To understand the improvement of performance and the effect of training all the algorithms, a baseline was constructed and compared to the final performance. For that, we applied simple classifiers. The job for each simple classifier is to classify all the instances in the training set as 1 (positive). After accomplishing this step, the accuracy and F1-Score have been calculated for all these basic estimators. The performance of these classifiers is reported in Table III. The reason behind establishing the basic predictors is that the dataset in hand is used for the first time in this research and there is no prior performance scores available to compare against. By comparing the results in Table III with the best results achieved for complications' models, it can be easily noticed that the final trained models almost doubled the performance of the basic classifiers.

Table II: Summary of all extensive experiments for the selection of the best performing classifier for each diabetes complication. The AUC score is reported for the best classifier.

Complication	Algorithms	Accuracy	F1-Score	Best AUC
Metabolic Syndrome	***LR***	***0.7283***	***0.7368***	***0.78***
	RF	0.7174	0.7111	
	XGBoost	0.7283	0.7253	
Dyslipidemia	LR	0.7518	0.8426	
	SVM Poly.	0.7664	0.8571	
	SVM Linear	***0.7664***	***0.8571***	***0.7***
Hypertension	LR	0.7293	0.7049	
	SVM Linear	0.7444	0.7302	
	XGBoost	***0.7444***	***0.7344***	***0.79***
Obesity	CART (DT)	0.7719	0.7234	
	RF	0.8158	0.7961	
	XGBoost	***0.8246***	***0.8148***	***0.87***
Neuropathy	SVM Poly.	0.8387	0.8387	
	AdaBoost	0.871	0.8462	
	XGBoost	***0.871***	***0.8462***	***0.93***
Nephropathy	LR	0.8276	0.7826	
	SVM Linear	0.8621	0.8182	
	RF	***0.8966***	***0.8696***	***0.94***
Diabetic Foot	LR	0.7647	0.6	
	SVM Linear	0.7647	0.6	
	AdaBoost	***0.8235***	***0.7273***	***0.96***
Retinopathy	SVM Linear	0.875	0.8333	
	RF	0.875	0.8333	
	AdaBoost	***0.875***	***0.8571***	***0.97***

Moreover, by comparing our results with the reported accuracy scores in [10], we can notice that our models achieved more than 10% improvement for predicting retinopathy, nephropathy as well as neuropathy.

Table III: Base-Line Performance

Algorithm	Accuracy	F1-Score
Metabolic Syndrome	0.4595	0.6296
Dyslipidemia	0.5269	0.6901
Hypertension	0.4744	0.6435
Obesity	0.4599	0.6301
Neuropathy	0.4134	0.585
Nephropathy	0.4118	0.5833
Diabetic Foot	0.4124	0.5839
Retinopathy	0.4121	0.5837

It is also important to study and compare the performance reached for each complication. For instance, the only complication that achieved a higher F1-Score compared to its accuracy is dyslipidemia. We believe the main reason behind it is the distribution of its classes. In fact, dyslipidemia is the only complication that has more positive instances than the negative ones. Another interesting observation is that the severity of imbalance problem affects the overall performance of a model. For example, the diabetic foot models resulted in achieving the worst performance between all other models. We believe the main reason behind it is the fact that diabetic foot has a very limited number of positive instances, which makes it harder to avoid this problem even with the use of balancing techniques. Finally, we observed that the distribution of the output class itself plays a significant role and can affect the overall performance. This can be noticed since we used the same independent features to predict all the complications, in fact, the only thing changed is the output class (the complications in this case).

V. Conclusion

In this research, data mining and machine learning algorithms were used to prognose and diagnose eight different diabetes complications. The complications' set consists of metabolic syndrome, dyslipidemia, hypertension, obesity, diabetic foot, neuropathy, nephropathy, and retinopathy. All these complications are available in a dataset provided by the Rashid Centre for Diabetes and Research (RCDR). The dataset consists of 884 records and 79 attributes. After cleaning the dataset, multiple experiments have been conducted to solve the missing value problem. For that, simple mean imputation, K-NN as well as MissForest were all tested and evaluated. It was found that MissForest achieved the minimum RMSE score. As a result, it was utilized throughout the rest of this research.

Since the dataset in hand suffers from data imbalance issue, different balancing methods were examined. A combination of SMOTE for oversampling the minority class and cluster centroids for under sampling the majority class was used. The algorithms constructed for this study contains Logistic Regression, SVM, Decision Tree (CART), Random Forest, AdaBoost and XGBoost. Overall, XGBoost was found to be one of the best performing algorithms.

To evaluate the models, simple baseline classifiers have been constructed and compared. Furthermore, our built models found to be able to exceed the performance of other available studies by more than 10% accuracy difference.

VI. References

[1] "About Chronic Diseases | CDC." https://www.cdc.gov/chronicdisease/about/index.htm (accessed Mar. 05, 2021).

[2] "World Health Organization, Diabetes," 2020. https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed Nov. 30, 2020).

[3] "How to Prevent Diabetes: MedlinePlus." https://medlineplus.gov/howtopreventdiabetes.html (accessed Nov. 30, 2020).

[4] "Diabetes, Centers for Disease Control and Prevention - CDC," 2019. https://www.cdc.gov/diabetes/managing/problems.html (accessed Nov. 30, 2020).

[5] R. Sharma, S. N. Singh, and S. Khatri, "Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey," Feb. 2016. doi: 10.1109/CICT.2016.142.

[6] V. Dominic, D. Gupta, S. Khare, and A. Aggarwal, "Investigation of chronic disease correlation using data mining techniques," Dec. 2015. doi: 10.1109/RAECS.2015.7453329.

[7] M. H. Tekieh and B. Raahemi, "Importance of Data Mining in Healthcare," 2015. doi: 10.1145/2808797.2809367.

[8] K. Kantawong, S. Tongphet, P. Bhrommalee, N. Rachata, and S. Pravesjit, "The Methodology for Diabetes Complications Prediction Model," Mar. 2020. doi: 10.1109/ECTIDAMTNCON48261.2020.9090700.

[9] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, Mar. 1986, doi: 10.1007/BF00116251.

[10] A. Dagliati *et al.*, "Machine Learning Methods to Predict Diabetes Complications," *Journal of Diabetes Science and Technology*, vol. 12, no. 2, Mar. 2018, doi: 10.1177/1932296817706375.

[11] D. J. Stekhoven and P. Buhlmann, "MissForest--non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, Jan. 2012, doi: 10.1093/bioinformatics/btr597.

[12] M. Cui *et al.*, "Risk Assessment of Sarcopenia in Patients With Type 2 Diabetes Mellitus Using Data Mining Methods," *Frontiers in Endocrinology*, vol. 11, Mar. 2020, doi: 10.3389/fendo.2020.00123.

[13] "Rashid Centre for Diabetes & Research – SKMCA." https://www.skmca.ae/rashid-centre-for-diabetes-research/ (accessed Jan. 21, 2021).

[14] M. Ashraful Alam Assistant Professor and A. Zaida Khanom, "Prediction of Diabetes Induced Complications Using Different Machine Learning Algorithms," BRAC University, 2018. Accessed: Nov. 30, 2020. [Online]. Available: http://dspace.bracu.ac.bd/xmlui/handle/10361/10945

[15] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2989857.

[16] "Imputation of missing values — scikit-learn 1.0 documentation." https://scikit-learn.org/stable/modules/impute.html#knnimpute (accessed Oct. 04, 2021).

[17] "ClusterCentroids — Version 0.8.1." https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.ClusterCentroids.html (accessed Oct. 10, 2021).

[18] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, Jun. 2002, doi: 10.1613/jair.953.

[19] "sklearn.preprocessing.OneHotEncoder — scikit-learn 1.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html (accessed Oct. 15, 2021).

[20] "1.4. Support Vector Machines — scikit-learn 0.23.2 documentation." https://scikit-learn.org/stable/modules/svm.html (accessed Dec. 05, 2020).

[21] J. R. Quinlan, "Induction of Decision Trees ," vol. 1, pp. 81–106, 1986.

[22] L. Breiman, "Random forests," *Springer*, vol. 45, no. 1, pp. 5–32, 2001.

[23] "1.11. Ensemble methods — scikit-learn 0.24.1 documentation." https://scikit-learn.org/stable/modules/ensemble.html (accessed Mar. 27, 2021).

[24] "sklearn.model_selection.GridSearchCV — scikit-learn 1.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed Oct. 11, 2021).