Date of publication SEP-30, 2023, date of current version FEB-19, 2023.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.22.3.3230

Predicting Life Style of Early Diabetes Mellitus Using Machine Learning Technique

Salliah Shafi Bhat[1], Venkatesan Selvam[2], Gufran Ahmad Ansari[3]
[1]B.S Abdur Rahman Crescent Institute of Science &Technology
[2]Department of Computer Application, B.S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, India
[3]Faculty of Science, MIT World Peace University (MIT-WPU) Pune- 411 038, India

Corresponding author: Venkatesan Selvam (selvamvenkatesan@gmail.com).

ABSTRACT A branch of artificial intelligence called Machine Learning (ML) enables machines to learn without having to be emphatically instructed. Machine Learning Techniques (MLT) have been used to forecast a variety of chronic diseases in the healthcare sector. Improvement in clinical approaches is necessary for early diabetes prediction to prevent complications and prolong the diagnosis of diabetes. Diabetes is growing fast in this world. In this paper MLT based Framework is recommended for early prediction of Diabetes Mellitus (DM). In this Paper the authors make use of PIDD data set. Different MLTs are used including Support Vector Classification (SVC), Logistic Regression (LR), K Nearest Neighbor (KNN) and Random Forest (RF). Data analysis is the first step in our method after which the information is transferred for data pre-processing and feature selection methods. RF performed better than other models with a 92.85 % accuracy rate followed by SVC (91.5%), LR (83.11) and KNN (89.6). K-fold cross-validation technique is utilized to verify the outcomes. The contribution of lifestyle characteristics is calculated using a feature engineering process. As a result, comprehensive overall comparative assessments of all the algorithms are performed taking into account variables such as accuracy, precision, sensitivity, recall, F1 score and ROC-AUC. The medical field can use the proposed framework to make early diabetes predictions. Additionally, it can be applied to other datasets that have data in common with diabetes.

KEYWORDS Machine Learning Techniques, Diabetes Mellitus, Feature Engineering, SVC, LR, KNN, RF.

I. INTRODUCTION

ONE of the most common serious diseases in the globe is "Diabetes Mellitus"(DM). The World Health Organization (WHO) estimates that diabetes kills 1.6 million people worldwide and affects 8.5% of individuals over the age of 18 (World Health Organization 2022). Despite the fact that many developing countries had a decrease in the rate of diabetes-related premature death from 2000 to 2010 the data increased once more between 2010 and 2016. Over 18% of people die from the four main diseases which are heart disease, cancer, chronic infections and diabetes making them a major health problem. For instance, in 2000, mortality from diabetes was 70% and it was predicted that mortality among men would grow to 80% by 2020. Obesity, old age, poor diet, genetic diabetes, high blood pressure, inadequate food, etc. all can cause the diabetes mellitus. Diabetes raises the chance of developing conditions like heart disease, cancer, renal failure, nerve damage, vision problems, etc. over time. The recent treatment approach requires gathering the information required to identify diabetes by a variety of tests followed by the administration of a proper diagnostic medication (Florez, J. C., & Pearson, E. Ret et al.). MLTs both supervised and unsupervised are used in the healthcare industry to identify a variety of disorders. With the use of MLT research hidden patterns in data sets were examined to predict outcomes and lower the cost of diagnosing difficult illnesses (Ahmad, G. N., Fatima, H. et al). Practical health datasets with a variety of attributes and external variables are used to train Machine Learning Techniques. With the appropriate medical care, major problems from diabetes can be avoided if they are detected early. MLT can help with the early detection of these diseases. Since MLT enable computers to learn from past knowledge or a defined dataset they are used to develop prediction models (Leite, A. F., Vasconcelos, K. D. F. et al). The ability of the predictive model to know and understand new data enables it to make more precise decisions. Machine Learning (ML) and Ensembles Training have been used in recent methods that are important for predicting various chronic disorders like diabetes

and have achieved success using various outputs. The earlier diagnosis of DM is crucial for achieving a higher accuracy rate and the use of machine learning algorithms is essential (Kaur, H., & Kumari, V.). The healthcare profession has been significantly affected by technological progress. For individuals who are unable to visit a clinic or seek emergency care there could be negative effects on the health (Hamnvik, O. P. R., Agarwal, S. et al.). For the welfare of all it eliminates the difference between resources and distance due to technology. A vast amount of databases with structured data is used for data collecting in the healthcare industries likewise unstructured data. The paper gives us contribution to develop a computational model for the early prediction of DM disease based on Machine Learning Techniques. In this paper the authors have used four MLTs such as RF, KNN, LR and SVC. For Experimental Data Analysis (EDA) many statistical and machine learning tasks are carried out to enhance the evaluation of the data quality. The framework is created using advanced MLT and numerous statistical measurements are used to assess how well these classifiers perform.

The paper is organized as follows: first section is Introduction. Section 2 discusses Literature Review on Early Diabetes Mellitus Using Machine Learning Techniques. Section 3 is about Framework and methodology. Section 4 tells us about Machine Learning Techniques. In Section 5, statistical and machine learning metrics for disease prediction are utilized; the findings of an experimental investigation are presented and analyzed. Finally Section 6 concludes this research program and provides conclusion and future scope.

II. LITERATURE REVIEW

This section presents some earlier studies that used MLT to predict and detect DM techniques. There is a lot of research being done in the field of diabetic mellitus. Therefore some related literature review works are presented. Every patient has different problems and risk factors related to the diabetic condition. The use of ML has grown in popularity in recent years for predicting various diseases. Researchers have created a variety of python scripts and algorithms. These features have shown enormous possibilities exist in the health care industry. The well-known dataset known as the PIMA Indians Diabetes Dataset (PIDD) from the "Kaggle" or "data World" repository was used by numerous researchers (Suyanto, S., Meliana et al.). Primary challenges had been mainly to classification such as precision of the adaptability for two or more predictions integrating two ML models and using more datasets at once PIDD data the logistic regression approach and modified K-means regression formula and programed the study of 89.42 percent accurate Waikato (WEKA) tool [8[. A total of four ML models including logistic regression, RF, Artificial Neural Network and NB use bagging and ensemble together. The author applied this methodology with 30,122 occurrences collected from Sawanpracharak Regional Hospitals 26 basic care departments. The random forest outperformed other algorithms in the testing, scoring 85.55% accuracy (Miriyala, N. P., Kottapalli et al.). The author obtained the primary data for the modelling process from a private medical source in Guangzhou, China. Three ML training algorithms were employed by the author to create the diagnostic system LR, ANN and DT. Out of all the ML models, the decision tree (C5.0) which the author built using the R environment has the highest accuracy (80.68%), according to experimental analysis (Su, S. W., & Wang, D). On the PIDD dataset which was constructed using the R language and divided into training and testing (Tigga and Garg) the logistic regression algorithm was employed. The model metrics yielded an accuracy rating of 75.32 percent. Moreover, the rate inaccuracy was 24.68% (Laila, U. E., and Mahboob et al.). Previously Neural Networks and RF to visualize and predict PIDD dataset were used for diabetes. The random forest scored with the highest accuracy based on previously published experimental findings and the best degree of accuracy was observed to be restricted by the lowest amount of features. Maniruzzaman et al. selected four categories for the prediction of diabetes patients including Naives Bayes (NB), DT, AdaBoost, and RF. Different partition protocols are used in these techniques (K2, K5, and K10). Precision (ACC) and curve surface metrics are used to evaluate the effectiveness of these classifiers (AUC) and a hybrid strategy is applied to identify disease. We employ K-mean clustering then Random Forest and XGBoost classifiers to extract unknown, hidden properties from the dataset and to produce more precise results. A. Yahyaoui et al., proposed a Machine Learning (ML) for diabetes prediction. They contrasted deep learning methods with conventional machine learning techniques. In [22] used SVM and the Random Forest two classifiers that are frequently applied in traditional machine learning techniques (RF). To forecast and identify people who have diabetes, therefore, these authors used a full-scale neural network (CNN) for deep learning (DL). The study (Wei, S., Zhao et al.) combined several data pre-processing methods, including principal component analysis (PCA) and linear discriminant analysis, with prominent algorithms for predicting diabetes, such as deep neural networks, support vector machines, etc. (LDA). In order to observe the metrics based on the 10-fold cross validation, [22] employed PIDD data. The deep neural network (DNN), which had the best accuracy of 77.86% during the experiment, beat other techniques. The purpose of predicting diabetic disease using soft computing is to offer insights into the enormous amount of information using Learning algorithms. Bhat, S. S., & Ansari, G. A. employed the PIDD dataset, and as data imputation had to be removed they used data pre-processing technique that combined the Decision Tree (CART) technique and Genetic Algorithm. According to [23] in the world there are 232 million people who even do not know that they have diabetes mainly due to ignorance and an underfunded healthcare system. The PIMA diabetic prediction dataset was utilized by [24] to test a variety of MLT. Four advanced MLTs were used: Support Vector Classification, Logistic Regression, K nearest Neighbor, Random Forest and an accuracy of 92.85 was achieved by RF SVC (91.5%), LR(83.11) and KNN (89.6). The collection of classifiers was implemented using a Jupyter notebook (integrated development environment), Python 3.8, and Windows PC with the open-source framework Anaconda 2020.

III. FRAMEWORK AND METHODOLOGY

The conceptual framework for developing MLT algorithms for forecasting various chronic diseases with diabetes as an illustration is shown in Figure 1.

In this framework the PIDD dataset is utilized. ML models for predicting diabetes have been constructed. To improve the performance analysis of the data, several statistical measurements across the dataset have been performed using pre-processing techniques. To evaluate the contribution of the parameters to the disease prediction, feature engineering has been done. The various ML models have been built after data

transformation. The performance evaluation of various statistical indicators of ML models has been K-fold cross-validation, with K = 10 used to validate.

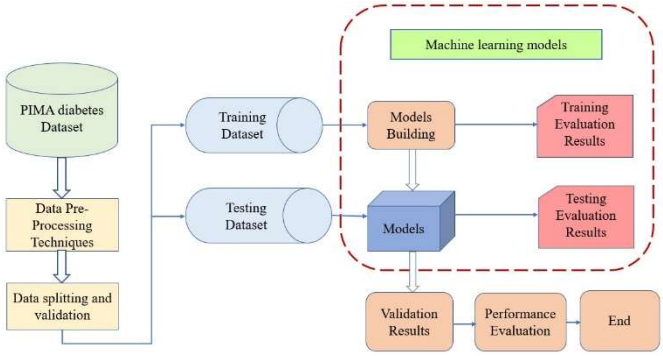

Figure 1. Proposed Framework for Prediction of Early Risk

Prediction of diabetes mellitus. It shows the sequential steps required to create a real framework utilizing machine learning methods based on the proposed methodology.

Step 1: Diabetes life style data set

The experiment work has made use of the PIMA diabetes dataset, which is openly accessible. The UCI Irvine Machine Learning Repository is where the dataset was obtained. To determine whether or not a patient has diabetes it used 768 instances, 10 characteristics, 9 predicted variables, and 1 target variable. The dataset includes a number of contributing characteristics that can be examined to evaluate the severity of the diabetic condition. A realistic diabetes health management system has been constructed by many ML classifiers. Table 1 lists the attribute names along with their range, data type, and description as used in the PIMA dataset.

Table 1. Data set Description

S.No	Attribute Name	Description	D Type
1	BMI	Body Mass Index	Float 64
2	Smoking	Whether the person is smoking or not	object
3	Sex	Gender of person(Male/Female)	object
4	Age Category	Age of persons in years	object
5	Thirst	The quantity of water a person consumes each day and night	Int 64
6	Glucose	Plasma Glucose concentration	Int 64
7	Diastolic Blood Pressure	Blood Pressure low or high	Int 64
8	Triceps Skin fold Thickness	Skin Thickness of the person	Int 64
9	2 Hours Serum Insulin	Insulin level of person low or high	Int 64
10	Diabetes Pedigree Function	Pedigree Function of a person	Float 64
11	Family History of Diabetes	Whether the person is suffering from diabetes or not	Int 64
12	Outcome	Binary variable 0 or 1	Int 64

Step 2: Data Pre-processing Technique

Pre-processing data is an essential first step that is given to raw data to get them ready for the analysis. In order to achieve better outcomes before developing the MLT, data pre-processing is an important step. The gathered dataset is pre-processed employing the Integrated Development Environment (IDE) Spyder and Python (3.9.1) as the programming language, as well as a number of statistical libraries. Techniques like resampling and discretization are used. By averaging particular attribute values in the required libraries for data quality assessment, missing values are filled such as BMI, Smoking, Sex, Age Category, Thirst, Glucose, Diastolic Blood Pressure, triceps Skinfold Thickness, 2 Hours Serum Insulin, Diabetes Pedigree Function and Family History of Diabetes. The mean and median range approach is used to replace the outlier with the values that would normally be found after the outlier having been detected using a boxplot. Before creating the machine learning models, data transformation was carried out to increase the efficiency of the data. Additionally, utilizing various data exploration analysis methods, duplicate, inconsistent and invalid files have been removed from the dataset. Now, the imputation pre-processing and data cleaning have been done.

- **Data Cleaning:** The process of eliminating or formatting undesired material from the gathered raw data is known as data cleaning. Because the data we obtained cannot be stored in a CSV file, we cleaned the data by deleting some symbols from the dataset and converting it to a CSV file.
- **Imputation:** Imputation is the process of substituting another data value for the missing one. Some of the properties in this dataset lack values, thus we created those values using the replace missing value function in the Weka tool. This work is based on the mean imputation technique.

Step 3: Data Splitting and K-fold cross validation

Researchers and practitioners frequently utilize the K-fold cross validation technique to create models and eliminate dataset skew. With a k value of 10, the K-fold cross-validation approach has been applied. Ten partitions of equal size are created by randomly dividing the full dataset. One partition out of the ten is kept to serve as the model validation (testing set) while the remaining ten partitions minus one are utilized as training data. Each is used only once as the validation out of the 10 partitions data during the entire 10-time process. By using the accumulation function the results of all iterations have been combined. The issue of overfitting and underfitting has been reduced in the dataset in order to meet the performance of both the training and testing dataset. The benefit of this method is that it eliminates data bias in order to create ML models producing accurate outcomes. Every sample of data is used for both testing and training, and each testing data bin has been used exactly once to analyze the results.

Step 4: Dataset Distribution

The distribution of predicate parameters has been visualized using the Facet Grid technique (Seaborn package) such as BMI, Smoking, Sex, Age Category, Thirst, Glucose, Diastolic Blood Pressure, triceps Skinfold Thickness, 2 Hours Serum Insulin, Diabetes Pedigree Function and Family History of Diabetes. In this method, the distribution of the observations in the dataset has been shown using the Kernel Density Estimate (KDE) plot function. It uses a continuous probability curve in one or more dimensions to represent the data samples. The vertical or y-axis depicts the probability density function (frequencies) of a random variable while the horizontal or x-axis reflects the range of data samples in a dataset. Then place a kernel function K on each data point xi, and the total shaded area of the curve under the two point's y1 and y2 will be the likelihood of value. The kernel density is calculated as shown in equation 1.

$$P(y) = \frac{1}{Mh}\sum_{i=1}^{m} k\frac{(y-y1)}{h}\ , \quad (1)$$

where P is the density at the specified place y,
kernel non-negative function is denoted by "*K*",
"*M*" – required steps,
the smoothing parameter is represented by "*h*",
the highest random value is "*Y*",
the variable "Y_i" stands for the variable rate of data samples.

IV. MACHINE LEARNING TECHNIQUES

The research main goal is to provide well-known machine learning algorithms that are frequently used in healthcare analytics to forecast various diseases. Depending on the symptoms, it is possible to determine whether a patient has diabetes or not. ze have created a number of machine learning models for the binary classification problem in this research. Four ML classifiers have been utilized for the development of a model to diagnose early risk prediction of DM using the lifestyle *dataset* ?. The MLTs are SVC, LR, KNN and RF.

Support Vector Classification (SVC): The supervised machine learning algorithm Support Vector Classification (SVC) divides the data into positive and negative classes with the greatest possible distance or margin.

Logistic Regression (LR): A classified dependent variable and a set of independent factors making up LR are compared. It is a statistical model that creates a binary-dependent variable by using logical function (Wang, C., Lin, Q et al.). Based on probabilities, the link between the dependent and independent variables is estimated. In this approach the dependent variable is categorical. It is depicted mathematically as follows:

$$\text{h}\,\emptyset(y) = p(k = 1|x; theta). \quad (1)$$

The likelihood that k=1 given y, denoted by "theta"

$$P\,(k{=}1|x;\, theta){+}p(k{=}0|x;theta){=}1. \quad (2)$$

K-Nearest Neighbor (KNN): When using the KNN the algorithm investigates K instances of the dataset that are close to the observation. After that, the algorithm itself will use its output to assess the variable y of the evaluation that should be expected (Li, Y., and Yang. et al.). The following equation is used to determine the Euclidean distance between two observations:

$$\text{D}\,(x_i, y_i) = \sqrt{(xi1, yi1)^2} + \cdots \ldots + \sqrt{(xim - yim)^2}. \quad (3)$$

Because K nearest neighbor does not initially require training and mostly learns from the data set while making predictions, it takes very little processing time. Due to the fact that it only needs two values, this algorithm is simple to implement: i) the K value and (ii) the distance function value. However, it has issues with large data sets and performs poorly with data that has several dimensions.

Random Forest (RF): To get a single result the RF mixes the output of different DT. Row sample and column sampling are done using the DT as a base. If the base learner population grows, the variance may decline or vice versa. K is a reasonable solution to cross-validation. It is regarded as a crucial bagging technique.

Random Forest=Bagging (Row sampling with replacement) + DT (base learner) + feature.

Bagging is the combination of column sampling and aggregation (mean/median, majority vote). Variability and little bias are qualities of decision trees and by averaging decision trees, the variance component of the model is minimized. It is possible to create the unknown samples by averaging the prediction.

$$\text{I}{=}\frac{1}{M}\sum_{M=1}^{M} f(x), \quad (4)$$

where uncertainty is

$$\sigma = \frac{\sqrt{\sum_{M=1} M((f(x)-f)2}}{M-1}.$$

The Random Forest (RF) algorithm analyzes data using a variety of decision trees, collecting predictions from each one, and determining the most effective course of action (Asif, M., Nishat, et al.). Additionally, it is built on an ensemble learning method that uses the bagging algorithm and can handle missing values for data.

V. RESULTS AND DISCUSSION

This section discusses and presents the results of the experimental design by using MLT for the early risk prediction of DM based on lifestyle factors. HP Z60 was the PC (Work Station) used for the experimental procedure. The hardware technical specifications include an Intel XEON processor running at 2.4 GHz (12 CPUs) and a NV Quadra K2200 GPU. Machine RAM and display RAM both have 4 GB of memory. The installed operating system is Windows 10 pro-64-bit and the machine storage capacity is 1TB [25]. The fundamental detailed statistics on lifestyle factors and associated metrics such as in Figure 2 are displayed the mean, standard deviation, minimum and maximum values, respectively. For example, in BMI parameters values are count, mean, Std, min, 25%, 50%, 75% and max are as 768, 291.102201, 6.612376, 14.69000, 24.80000, 28.12000, 32.63000 and 75.82000 respectively. In order to show the dataset major characteristics, the same calculations have been done for additional parameters as well.

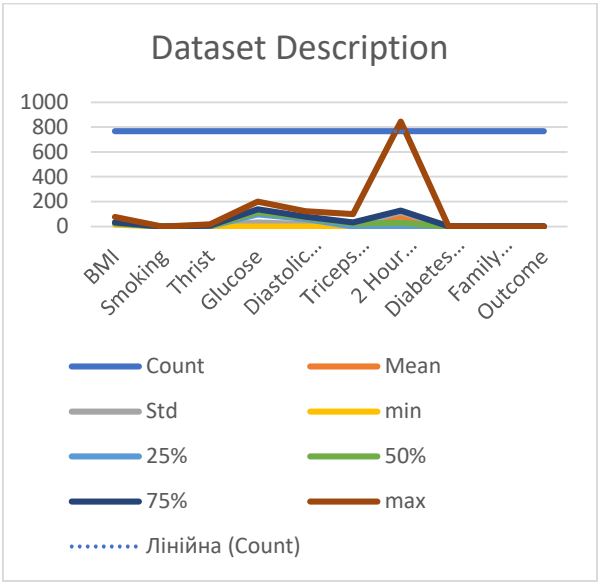

Figure 2. Data Set description

A. Correlation Coefficient and Analysis

The dataset two variables in relation to one another are discovered using the Correlation Coefficient Analysis (CCA) technique. Identifying the significance of the dataset features is the primary objective of the CCA. There is a strong correlation between the factors that are dependent and independent. The feature set is taken into account to be suitable for developing machines learning models. The correlation matrix between the set of features is shown in Figure 3.

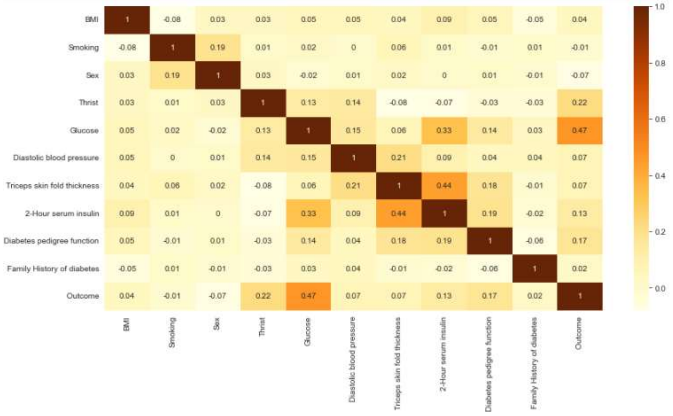

Figure 3. Correlation Coefficient matrix

A finite integer is in the range of +1 and -1 indicates the range of the values that are determined numerically. The values of the CCA matrix are represented by a finite number in the range of +1 and -1:

- r +1 indicates an entirely positive correlation;
- -0.08 represents negative correlation;
- +0.07 indicates entirely positive correlation;
- +0 represents no correlation;
- +0.08 indicates positive correlation;
- -0.07 indicates entirely negative correlation;
- -0.01 indicates complete negative correlation.

The equation below analyses the attribute.

$$\text{CCA} = \frac{g\ avg(corrfc)}{\sqrt{g+g(g-1)avg(corrff)}} \qquad (10)$$

When building the rank transformation, CCA is used to assess the set of parameters, features, and determine the relevance between dependent and independent biological aspects. The mean correlation between the independent and dependent variables of the predicate is represented by the (Avg (corr $_{fc}$)) while the average correlation between parameters is defined by the (corr $_{ff}$)). K also indicates how many features are included in the dataset. The parameters such as BMI, Smoking, Sex, Thrist, Glucose, Diastolic blood Pressure, Triceps Skinfold thickness,2- Hours serum insulin, Diabetes pedigree Function, Family History of diabetes show a good correlation with the outcome class for the diagnosis of early prediction of DM condition. For instance, relations between independent and dependent variables include BMI, smoking, and sex, with a correlation coefficient of 0.04, 0.1, and 0.07, respectively. Correlation between many parameters, such as glucose weighted average to diastolic Blood pressure is 0.05, sex in relation to heart rate is 0.03, etc.

B. Performance Evaluation of ML Algorithms

Table 2 shows the summarization of the results obtained using the cross validation method for several MLT models. It presents the outcomes of various evaluation metrics, including Accuracy, Precision, Recall, fi score and ROCAUC. Out of all the concepts RF has achieved the highest accuracy of 92.85% on the basis of accuracy followed by SVC, LR, KNN, which have shown 91.5%, 83.11% and 89.6% respectively. However, in terms of precision RF has achieved 98%, SVC (95%), LR (88%) and KNN (78%). In recall analysis RF has shown 93%, SVC (92%), LR (72%) and KNN (89%). In fi score RF has achieved 95%, SVC (94%), LR (88%), KNN (84%). In addition, in ROCAUC RF has shown 94%, SVC (90%), LR (80%) and KNN (89%).

Table 2. Classification Performance of the dataset

Algorithms	Accuracy	Precision	Recall	Fi Score	ROC AUC
SVC	91.5	95	92	94	90
LR	83.11	88	72	88	80
KNN	89.6	78	89	84	82
RF	92.85	98	93	95	94

C. Feature Engineering

Feature engineering is crucial to the process of constructing ML models. Insignificant or incorrect characteristics may have a negative impact on how well a model runs. The training time is reduced and accuracy is improved with proper feature selection. Embed, filter, wrapper, embedded, and hybrid strategies are a few of the methods for feature selection used by machine learning models (Ferhatoglu, C., & Miller, B. A). The features are selected in this work using the Information Gain and Correlation techniques. The majority of the identified characteristics, with the exception of "Sex," are found to contribute significantly to the early prediction of DM as shown in Figure 4. The importance/ranking of all features, in order of highest to lowest, are Outcome, 2 Hours Serum Insulin, Diastolic Blood Pressure, Thirst, Triceps Skinfold Thickness, BMI, Family History of Diabetes, Sex and Smoking. However, it has a strong link with unrelated factors and is a key aspect of lifestyle.

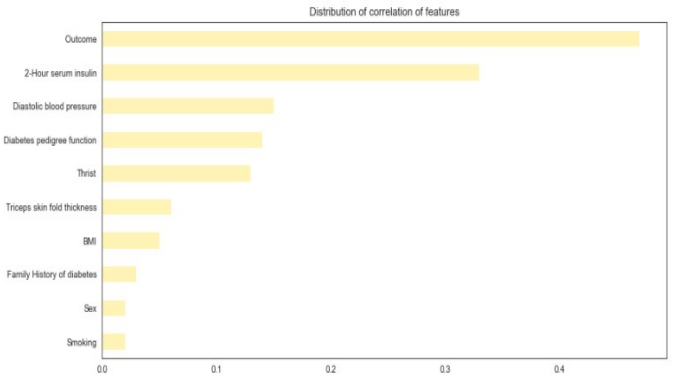

Figure 4. Correlation of feature importance

D. Comparative Analysis with the existing work

Table 3 compares the effectiveness of our proposed framework to a variety of relevant types of literature in terms of the methods utilized, the dataset and the analysis. Most lifestyle factors are universal across all studies conducted for use in opposition to the work being proposed. It is shown that our framework, when taken into consideration, has produced positive results in terms of numerous evaluation criteria, including accuracy for DM early prediction. K-fold cross-validation is employed to obtain results that are more

trustworthy than those from comparable studies when developing the suggested framework.

Table 3. Comparison of existing works

Authors	Technique used	Data set	Analysis
[22]	LR, KNN, DT, NB, SVM, and ANN	Medical data set of Bangladesh	87% for SVM
[23]	SVM, RF and LR	PIMA Data set	90% for RF
[24]	LR, KNN, DT, NB, SVM, and ANN	PIMA Data set	89% for DT
Our Proposed System	LR ,RF,KNN,RF	Lifestyle diabetes prediction	92.85 for RF.

VI. CONCLUSION AND FUTURE SCOPE

Millions of people all over the world are being badly affected by diabetes chronic condition at an alarming rate. MLT is used for early DM prediction based on lifestyle variables in this research work. The data on patient's lifestyle has been thoroughly examined in order to create the framework. Finally, four MLTs based on RF were implemented using 10 fold cross validation. To our knowledge, this is the first time a framework has been given that yields prediction results that are so much better than those from earlier research. By metric analysis, the findings of this proposed work are accurate and reliable. This framework can be used to predict disease likelihood in patients and prospective patients at an early point in the future. Different patient types can be encouraged to adjust their lifestyles appropriately (using dietary changes and exercise routines) depending on the disease's tendency. Furthermore mobile applications can be created to assist medical professionals in diabetes detection and prediction. Additionally, it will help users prevent hospital readmissions and reduce diabetes complications at an early stage.

References

[1] World Health organization, 2022. Online. Available at: https://www.who.int/newsroom/factsheets/detail/diabetes

[2] M. Fang, D. Wang, J. Coresh, E. Selvin, "Trends in diabetes treatment and control in U.S. adults, 1999-2018," *N Engl J Med*, vol. 384, no. pp. 2219–2228, 2021. https://doi.org/10.1056/NEJMsa2032271

[3] E. Ahlqvist, P. Storm, A. Karajamaki et al, "Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables," *Lancet Diabetes Endocrinal*, vol. 6, issue 5, pp. 361–369, 2018. https://doi.org/10.1016/s2213-8587(18)30051-2

[4] A. F. Leite, K. D. F. Vasconcelos, H. Willems, R. Jacobs, "Radiomics and machine learning in oral healthcare," *PROTEOMICS–Clinical Applications*, vol. 14, issue 3, 1900040, 2020.

[5] R. C. Slieker, L. A. Donnelly, H. Fitipaldi et al, "Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: an IMI-RHAPSODY study," *Diabetologia*, vol. 64, issue 9, pp. 1982–1989, 2021. https://doi.org/10.1007/s00125-021-05490-8

[6] O. P. R. Hamnvik, S. Agarwal, C. G. AhnAllen, A. L. Goldman, S. L. Reisner, "Telemedicine and inequities in health care access: the example of transgender health," *Transgender Health*, vol. 7, issue 2, pp. 113-116, 2022.

[7] S. Suyanto, S. Meliana, T. Wahyuningrum, S. Khomsah, "A new nearest neighbor-based framework for diabetes detection," *Expert Systems with Applications*, vol. 199, 116857, 2022.

[8] M. Ishi, J. Patil, V. Patil, "An efficient team prediction for one day international matches using a hybrid approach of CS-PSO and machine learning algorithms," *Array*, vol. 14, 100144, 2022.

[9] N. P. Miriyala, R. L. Kottapalli, G. P. Miriyala, G. Lorenzini, C. Ganteda, V. A. Bhogapurapu, "Diagnostic analysis of diabetes mellitus using machine learning approach," *Revue Intelligence Artificielle*, vol. 36, issue 3, pp. 347-352, 2022.

[10] S. W. Su, D. Wang, "Health-related quality of life and related factors among elderly persons under different aged care models in Guangzhou, China: a cross-sectional study," *Quality of Life Research*, vol. 28, issue 5, pp. 1293-1303, 2019.

[11] N. P. Tigga, S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Procedia Computer Science*, vol. 167, pp. 706-716, 2020.

[12] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, W. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study," *Sensors*, vol. 22, issue 14, 5247, 2022.

[13] M. Maniruzzaman, M. Rahman, B. Ahammed, M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health information science and systems*, vol. 8, issue 1, pp. 1-14, 2020.

[14] A. Yahyaoui, A. Jamil, J. Rasheed, M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," *Proceedings of the 2019 IEEE 1st International Informatics and Software Engineering Conference (UBMYK)*, November 2019, pp. 1-4.

[15] S. Wei, X. Zhao, C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," *Proceedings of the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, February 2018, pp. 291-295.

[16] S. S. Bhat, G. A. Ansari, "Predictions of diabetes and diet recommendation system for diabetic patients using machine learning techniques," *Proceedings of the 2021 IEEE 2nd International Conference for Emerging Technology (INCET)*, May 2021, pp. 1-5.

[17] E. Bonora, M. Trombetta, M. Dauriz et al, "Chronic complications in patients with newly diagnosed type 2 diabetes: prevalence and related metabolic and clinical features: the Verona newly diagnosed type 2 diabetes study (VNDS) 9," *BMJ Open Diabetes Res Care*, vol. 8, issue 1, e001549, 2020. https://doi.org/10.1136/bmjdrc-2020-001549

[18] M. Battaglia, S. Ahmed, M. S. Anderson et al, "Introducing the endotype concept to address the challenge of disease heterogeneity in type 1 diabetes," *Diabetes Care*, vol. 43, issue 1, pp. 5–12, 2020. https://doi.org/10.2337/dc19-0880

[19] Y. Li, Y. Yang, J. Che, L. Zhang, "Predicting the number of nearest neighbor for kNN classifier," *IAENG International Journal of Computer Science*, vol. 46, issue 4, pp. 662-669, 2019.

[20] E. R. Pearson, "Type 2 diabetes: a multifaceted disease," *Diabetologia*, vol. 62, issue 7, pp. 1107–1112, 2019. https://doi.org/10.1007/s00125-019-4909-y

[21] C. Ferhatoglu, B. A. Miller, "Choosing feature selection methods for spatial modeling of soil fertility properties at the field scale," *Agronomy*, vol. 12, issue 8, 1786, 2022.

[22] M. Kowsher, M. Y. Turaba, T. Sajed, M. M. Rahman, "Prognosis and treatment prediction of type-2 diabetes using deep neural network and machine learning classifiers," *Proceedings of the 2019 IEEE 22nd International Conference on Computer and Information Technology (ICCIT)*, December 2019, pp. 1-6.

[23] R. Patil, K. Shah, "Assessment of risk of type 2 diabetes mellitus with stress as a risk factor using classification algorithms," *Int. J. Recent Technol. Eng.*, vol. 8, issue 4, pp. 11273-11277, 2019.

[24] R. Sivakani, M. Syed Masood, "Analysis of COVID-19 and its impact on Alzheimer's patient using machine learning techniques," *International Journal of Computing*, vol. 21, issue 4, pp. 468-474, 2022. https://doi.org/10.47839/ijc.21.4.2782.

[25] G. A. Ansari, S. S. Bhat, "Exploring a link between fasting perspective and different patterns of diabetes using a machine learning approach," *Educational Research*, vol. 12, no. 2, pp. 500-517, 2022.

[26] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, M. H. Rahman, "Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district Bandipora," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2789760, 2022.

***SALLIAH SHAFI BHAT** received her B.Sc. from Kashmir University. She has also received MSC IT from BGSBU Rajouri. She is currently pursuing her PhD degree in B.S Abdur Rahman Crescent Institute of Science &Technology Chennai Tamil Nadu.*

DR. VENKATESAN SELVAM *is a Professor in B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nādu, India. He received his M.E, from Anna University in Computer Science and Engineering. He received his Ph.D. degree from Anna University. He has more than 19 years' of experience in teaching. He has contributed more than 30 research articles in reputed journals and conferences. His research area are Ant Colony Optimization, Soft Computing Tools, Pattern Recognition, Genetic Algorithm, Digital Image Processing, Optimization Techniques and Advanced Computer Networks.*

DR. GUFRAN AHMAD ANSARI *is a Professor at MIT World Peace University Pune, Maharashtra India. He received his Master's degree in MCA from DR. B.R. Ambedkar University Agra in 2002 and Ph.D. (Computer Science) from Babasaheb Bhimrao Ambedkar (A Central) University, Lucknow, U.P., India in 2009. He has more than 20 years of experience in teaching. He has contributed more than 55 research articles in reputed journals and conferences. His research area is Software Engineering, UML, Machine Learning, Modelling, Testing, Artificial Intelligence, Data Mining, Software Security, Testing etc.*