INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Feature Minimization for Diabetic Disorders High Performances Prediction System-based on Random Forest Tree

Sahar Jasim Mohammed [a,*], Ali Mohammed Saleh Ahmed [a], Mohammed Sami Mohammed [a]

[a] *Computer Department-College of Education for Pure Science, University of Diyala , Baquba, Diyala, 32001, Iraq*

Corresponding author: [*]m.sc.sahar.jasim.m@uodiyala.edu.iq

***Abstract*—Human organ failure due to high blood sugar is considered a chronic disease. Early prediction might reduce or prevent complications due to such disorders, especially with recent machine-learning improvement techniques and the availability of electronic data from different sources. The number of diabetic patients roughly increased and may reach more than 600 million by twenty years. Transforming data into valuable and helpful information is an effort for researchers to improve the performance of ML techniques. This paper applies several types of sampling to predict 1000 samples with attributes and three diabetes class types (Random Forest tree, Hoeffding tree, LWL, NB updatable, and support vector Machine). This paper focused on most parameters that affected overall prediction accuracy. ML performances have been measured depending on the accuracy and mean absolute error for several cross-validation values before Feature reduction and after feature minimization by applying feature selection methods. It shows that Gender, Age, Blood Sugar Level (HbA1c), Triglycerides (TG), and Body Mass Index (BMI) are the most impact attributes applied. It is also shown that the Random Forest tree was the best method (97.7 and 98.6 %) with and without feature minimization, respectively, but it has a higher performance by omitting some unbalanced features from the diabetic dataset. Weight minimization has also been applied to techniques like SVM to obtain a better-searching plane and a robust model. In addition, this study specifies which parameters have weight minimization with the required analysis. Also, the feature selection method was applied to gain memory and reduce time.**

***Keywords*— Decision tree; naïve bayes; support vector machine; diabetic disease.**

Manuscript received 3 Jun. 2023; revised 8 Sep. 2023; accepted 16 Oct. 2023. Date of publication 30 Nov. 2023.

I. INTRODUCTION

Many countries have suffered from severe health issues like heart failure, Kidney failure, cancer, and lastly, COVID-19. Diabetic disorders are not less than other health issues, especially with solid complications that appear after a simple period if not predicted early. A healthy type and disease identification might delay or prevent this progression. In addition, the Diabetic database is available but needs processing to obtain good analysis and provide a will pattern recognition. Extracting information from images such as tongue features is a technique researchers use to collect information and upload it to deep learning and classical ML (Machine learning) methods for a prediction process. A study focused on the relationship between tongue images and diabetic occurrences revealed that 1140 samples in China were divided equally into diabetic/non-diabetic samples with image Collection by an instrument [1]. In [2], researchers depend on a dataset of 800 samples with ten features, applying different approaches and obtaining the highest methods based on the highest accuracy. In this paper, logistic regression methods provided the best accuracy compared to the super/unsupervised algorithms that were applied in this paper.

Early prediction of diabetic disorder prevents several types of human part's failure like liver and heart. DT and two other ML techniques are applied and compared for 767 pregnant and non-pregnant females with seven attributes in India to predict diabetic disorders [3]. Among the applied techniques, Naïve Bayes got the lead in accuracy with 74-28%, followed by decision tree and SVM, respectively. Negative results because of the complications that diabetes causes must be recorded and analyzed to create a prediction system that can hold the future similar or class to the uploaded dataset [4], [5]. Data was studied and tested for more than one million samples from Canada for three years of risk outcomes due to diabetic issues. In [6] and [7], the authors also depend on the Pima Indian database but apply

deep learning methods with two different cross-validations (5 and 10 cross-validations) and compare them. The classification of a dataset such as the Pima diabetes dataset is a widespread issue for medical data due to several parameters and features available, such as BMI, skin characteristics, and other features. The R program presents faster decision-making for internet users with 3 ML types [8]. In [9], authors focus on applying KNN to these datasets to discover the features that most influence prediction accuracy. KNN was utilized to focus on the main effective classes based on the k parameter. Modifying KNN towards accuracy enhancement, authors in [10] utilized standard deviation principles with KNN to replace the common distance calculation of classical KNN.

In the same way as applying ML but with an adjustment on the PID dataset, SVM has the best performance by comparing two types of datasets (before and after dataset enhancement) [11]. In [12], authors focused on the style mode of like in India with heath family history to predicate types of diabetes. The study applied a question aired to obtain 952 patients' information relevant to their family history and lifestyle and delivered it to different types of ML-like (KNN, SVM, etc.). The collected dataset was compared to 768 samples in Pima/India to train and test the prediction system. Attributes in this paper were dependent and independent based on life and health history data by dividing data into training sets triple times the testing set. The random forest tree had the best performance among the six applied techniques, with 94% and 75% applying it on the second dataset.

II. Materials and Methods

A. Decision Tree

Different applications deal with Decision tree classification types; Cache replacement planning to enhance the duty of proxy servers has been introduced and studied [13] by modifying a fast version of the decision tree. Some of these disorders are complicated, such as cancer, because it is hard to predict and needs fast detection. Therefore, others ought to use multiple algorithms for such a disease. DT depends on features and is considered one of the non-parametric processes. Therefore, features and attributes should be evaluated well, and then the type that fits these data types and properties should be chosen when the classification is based on a decision tree algorithm [14], [15]. It was applied for different fields with data and as an Image processing with data extraction. Compared to weighted methods, DT has no nodes to deal with and does not need to define the number of layers or neurons per layer. DT is a much simpler mathematical process that leads the path to the right way of detection. The activation function in DT depends on features at each step, which define and create branches. The decision tree has main steps [16]:

- Specify at each step which entropy should start with and its branches.
- Move towards two branches after starting with nodes, then determine the value ranges.
- According to the probabilistic of any events, DT moves forward until it reaches a particular class.
- Make the right path for each class it finds.

In order to reduce the prediction time with DT enhancement, an omitting feature of some non-useful features or independent attributes before applying it to show the impact of selecting features is provided [17]. The steps shown in Fig. 1 explain the node paths and conditions toward diabetic classes using C4.5, which also defines the number of total tree leaves and their size.

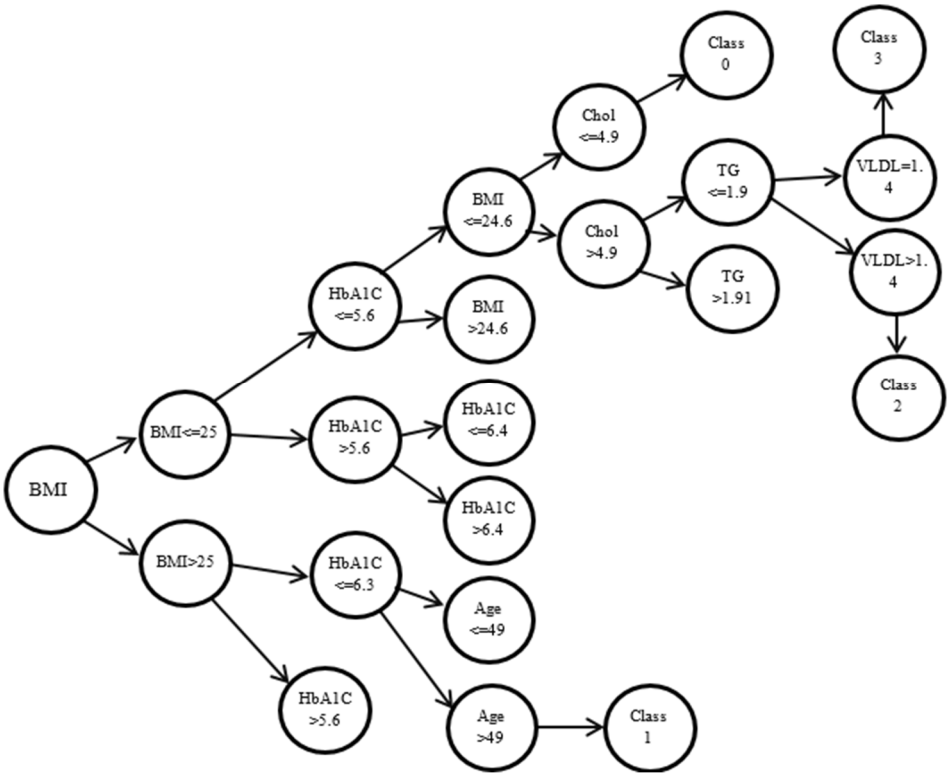


Fig. 1 The C4.5 node paths and conditions toward diabetic classes.

B. Hoeffding Tree (HT)

This Version of the decision tree used the boundary conditions of Hoeffding, which determine attributes that process at a time. HT works significantly with streaming data, even if it is large when they do not alter at a time. Coming features that HT processes are not limited like DT. On the other hand, new data supplied to HT can be easily dealt with. HT can reduce the memory size because it works at any step with incoming data, which does not need to start over the starting node [18] and [19]. This paper presented a model consisting of attributes and classes in the format (D, A), where A refers to 11 attributes and D refers to three types of Diabetic. The main objective of this model is to generate D=f (A) with high precision for more incoming data. The main steps that this paper works on are explained and followed.

- The root node is initialized according to the highest gain of attributes at first.
- B leaves can get branches according to enough data to measure the gain of each attribute to continue with.
- Applying the H condition to decide which attributes provide better classification accuracy by testing these paths till the end.
- Repeating the previous step until speckling the optimal attributes assisting in tree growth.

For example, after applying HT, one of the diabetic attribute conditions is shown in (Table I).

TABLE I
CLASSES ACCURACY CALCULATION DEPENDS ON ONE ATTRIBUTE WITH CONDITION SPECIFICATIONS

Attributes name	Condition	Class 0	Class 1	Class 2
HbA1C	<=6.45	43%	22%	34%
HbA1C	>=6.45	2%	6%	92%
HbA1C	neglected	10%	5%	85%

C. LwL

Unlike DT and HT, linear regression is an effortless regression model. It does not work on the overall data at once unless creating neigh boon line detection for each point. By moving in this detection road, LwL provides a weighting factor based online creation for each point. It is widely applied to dataset prediction with shortest line creation according to dependent points (such as a diabetic class type) and independent points (such as 11 dataset attributes). In this paper, the proposed methods of a diabetic dataset are complex because they deal with multiple dependent and independent variables, as in the equation below.

$$D = K+SA+E \quad (1)$$

Where D is the dependent diabetic class, A = independent diabetic attributes, K is the intersection Point, S is the line's loss, and E is LwL error process. Fig. 2. shows HbA1c with line determination according to independent diabetic attributes for some parameters. As mentioned previously, mining techniques have plenty of application fields related to regression problems, such as in programming solvers, performances of educational studying as in a school or university, intruder detection of deferent types network and calibration models, or even some activities like seismic as explant in details in [20] and [21]. Most cooperative fields can use lazy classification, especially with a moderate volume of data, such as disease prediction.

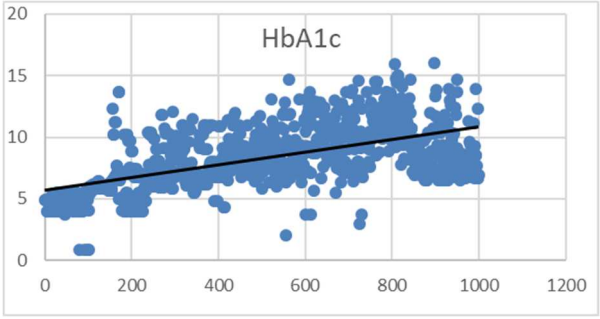


Fig. 2 HbA1c with line determination.

D. Naïve Bayes Updatable

Simple Version of probabilistic methods, while independent attributes are related to multiple class types. The updatable Version of Naïve Bayes is mainly moved in the right path if multiple wends (attributes) occur with the same type of diabetic classes. Multiple probability values omitted the one occurrence of a single attribute related to one class. In (Table II), shows the SVM probabilistic of three types of diabetic classes equal to one (class 0 = 0.1, class 1 = 0.05, class 2 = 0.84) and the mean of each attribute related to each class. In [22], authors view a prediction for breast for the non-cancer dataset by trying to predict the cancer activities using a hybrid type of Naïve Bayes were also presented for academic performances based on parson's activities and their features as well as in [23]. Authors viewed a study to reduce the late detection of cancer disorders by using a modified Naïve Bayes and compared it with artificial intelligence [24]. A new paper was also recently presented for the workers in Japan to investigate the people next to COVID-19 patients as in [24].

E. Support Vector Machine

Detection of the outlier's points belonging to a dataset with a regression process is the primary step of SVM. It is considered with different techniques of big datasets, especially when the number of attributes is more significant than the number of patients. It is also a subset-creating machine that creates multiple vectors in the dataset. Treatment of an early disorder prediction provides better chances for patients to get well and reduce any vital symptoms. Cancer is one of these disorders with either low or high value for the same feature. This makes the maker dataset more complicated and needs a detecting system to record the critical pattern between patients' datasets, as in [25] and [26]. A hybrid technique has been utilized by authors in [27] by combining ML and DL for diabetic image classification. Several types of SVM were considered the main approach for detecting many types of disease, such as heart disease with a diabetic disease, as an automated system, as explained in [28]. The authors in [29] presented a modified version of SVM called Mayfly-SVM, which depends on Mayfly optimization to be applied to the feature selection. It showed promising results with about 94.5% by applying it to PID dataset. Data types, size, and the ratio between training and testing parts make M a preference in a

never disease prediction system sash in COVID-19. Crises could be prevented with an early prediction system for such a disease, especially when contact between samples or between samples and therapists leads to vital spread. Table III shows the weight reduction using SVM for the 1st and the 10th cross-validation for the applied dataset. In this table, weight has been minimized for some attributes like (Age, Gender, Urea, and Chol.) after ten full cross-validations. Other attributes are not changed with cross-validation movement.

TABLE II
ATTRIBUTES PROBABILITY RELATED TO EACH CLASS

Attributes	Class Types		
	Diabetic_01	Diabetic_02	Diabetic_03
	0.1	0.06	0.84
Gender mean	0.6214	0.3208	0.4194
AGE mean	44.3757	43.4605	55.4106
Urea mean	4.6722	4.52	5.2176
Cr mean	62.84	64.7399	69.8009
HbA1c mean	4.5487	6.0141	8.8785
Chol mean	4.2697	4.59	4.9509
TG mean	1.6769	2.1688	2.4884
HDL mean	1.2062	1.1293	1.1896
LDL mean	2.6272	2.5019	2.6126
VLDL mean	0.9476	0.9822	2.0802
BMI mean	22.3877	23.9885	30.816
HDL mean	1.2062	1.1293	1.1896
LDL mean	2.6272	2.5019	2.6126
VLDL mean	0.9476	0.9822	2.0802
BMI mean	22.3877	23.9885	30.816

TABLE III
WEIGHT MINIMIZATION AFTER 10 CROSS-VALIDATION

Features	Weight for the 1 Cross Validation	Weight for the 10 Cross Validation
Gender	0.2445	0.056
AGE	0.2974	0.2537
Urea	0.311	0.2652
Chol	0.585	0.5467

F. Features Selection Techniques

The large set of factors is not generally useful for machine learning due to some outlier's points with some attributes that will impact the prediction accuracy. A well-feature selection will reduce memory occupation and time spent by ML technique. Also, these features are unrelated to some of them, with a need for redundancies to maximize prediction accuracy. Multiple types of selection features are applied and embedded with the ML system design to gain higher precision and are compared to other methods, as in [30], [31], and [32]. Applying multiple types of feature selection in this paper for a diabetic dataset that is used for this prediction is shown better in the table below, which shows the main features that should be kept for the prediction system. A study has been presented to compare applying three types of these methods to define the most affected features on DDoS detection [33]. In the same field, three MLs have been utilized to be combined with feature selection as explained in [34]; authors compared their studies with past works to demonstrate the benefits of feature minimization for the different applied datasets related to DDoS. For the first time, researchers may select Age or Gender as a secondary affection feature's and might be omitted from the system. While (Table IV) shows that some of the viewed features have the majority affection on prediction accuracy and should not be canceled, or at least it must be kept for the system.

TABLE IV
AFFECTED FEATURES ON PREDICTION ACCURACY

	Attributes	Feature Selection Method ID					
		Correlation based Feature Selection	Classifier Attribute	Correlation Attribute	Gain Ratio	Relief Attribute	Symmetrical uncertainty
1	Gender	H	L	H	H	L	L
2	Age	L	H	H	H	H	H
3	Urea	L	H	L	L	L	L
4	Cr	L	L	L	L	L	L
5	HbA1C	H	H	H	H	H	H
6	Chol	L	L	L	L	H	L
7	TG	H	L	H	H	H	H
8	HDL	L	L	L	L	L	L
9	LOL	L	L	L	L	L	L
10	VLDL	L	L	L	L	L	H
11	BMI	H	H	H	H	H	H

However, the other irrelevant attributes will be highlighted to be tested before neglecting the process and according to the selection techniques. At the same time, attribute selectors should be agreed upon and matched to be similar feature determination. It means that when two or more techniques are demonstrated, the attribute names should be similar to other techniques even if one or two attributes are different from other techniques, as suggested in [35], [36], [37], [38], and [39]. Diabetic disease needs a unique study to improve the availability of investigation or reduce the variance and spread of this vital disease. Studies are becoming more sensitive to such an issue due to its causes and symptoms, as discussed and explained in [40] and [41].

G. Dataset

The diabetic dataset is available but needs an accurate prediction system that analyzes data and studies its attributes well to fit and match the required applied ML. The dataset is downloaded from [42] for 1000 samples with 11 attributes and three different cases (Diabetic, Non-Diabetic, and Predicted). Which was collected from Kidney Teaching Hospital by anthers and available for research on the site. (Table V) shows the data description with attribute names.

TABLE V
DATA DESCRIPTION WITH ATTRIBUTE NAMES

Item Name	Min	Max	Mean	Std. Dev.
Gender	/	/	0.435	0.496
Age	20	79	53.5	8.7
Urea	0.5	38.9	5.125	2.935
Creatine Ratio	6	800	68.94	59.985
Sugar Level in Blood HbA1C	0.9	16	8.281	2.534
Cholesterol Chol	0	10.3	4.86	1.302
Triglycerides TG	0.3	13.8	2.35	1.401
HDL Cholesterol	0.2	9.9	1.205	0.66
Total LDL	0.3	9.9	2.61	1.115

Item Name	Min	Max	Mean	Std. Dev.
Total VLDL	0.1	35	1.855	3.664
Body Mass Index BMI	19	47.75	29.578	4.962
Classes	/	/	/	/

Data has been transformed from strings to numbers like (o for male, I for female) as well as class names (Diabetic-0 for no diabetic sample, Diabetic-1 for diabetic sample, Diabetic-2 for predated samples) to be suitable for processing with ML techniques. In Fig. 3, two parameters are related to three selected classes.

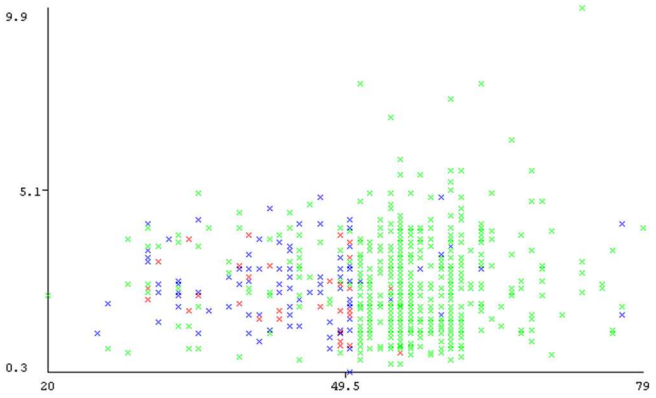


Fig. 3 Two parameters (LDL + Age) relation.

III. Results and Discussion

In machine learning, dealing with modern datasets related to recent diseases needs more data to observe the requirement. A more extensive dataset provides a better division sample to train and test simultaneously. Also, data can be divided into multiple subsets to work with and extract enough information to design a prediction system. The training dataset should not be small compared to the overall use dataset. Otherwise, the prediction system will suffer from predicted unsimilarity classes and lead to unsteady techniques. At the same time, healthy samples supply a huge capacity for sample evaluation with a better prediction compared to the deflected samples. Dataset types, sizes, dividing data ratio into trains and testing parts, missing data, and the number of healthy or affected samples have been considered in this paper before selecting any machine learning method.

In addition, selecting the most affected features also impacted these techniques, which can provide a robust model or not. SVM was applied to a diabetic dataset with weight minimization to obtain the optimal searching plane to cover all the required available and incoming data. In SVM, (Gender, Age, Urea, and Chol.) has been weight minimization after 10-cross validation, leading to some good perdition performances. While other features have unchanged weights with crass validation changed. The number of dataset features is not beneficial for prediction accuracy, so ML needs a relevant dataset with its attributes.

For this reason, features are also studied and selected in this paper to gain memory and time reduction. In addition, features omitted should be analyzed before applying them with ML because some features have the majority of influence on model prediction and should not be canceled. Feature selection methods showed that (Gender, Age, HbA1C, TG, and BMI) are the most affected features and should not omitted from the design. At the same time, it did not mean that other attributes could be canceled before testing prediction without accuracy. The outcome was calculated before and after feature selection. Random forest provided the best results for both (without and with feature sedation), but the differences were close to about 98.6% and 98.8%, respectively. Random forest has a robust system due to (well processing with static and continual variables, missing values covariance, no affection of non-linear parameters on RF performances, outlier points detection, and RF has a low affection with incoming data or the noisy data. Still, the mathematical operations done by RF have sub-trees trees reaching 100 in some applications, leading to a higher training data period. The Hoeffding tree also obtained a high accuracy before and after risk factor specification, with 94.2% and 94.5%, respectively.

However, it is still less than the RFT technique due to boundary conditions related to DT. HT was useful for storing memory space while working on streaming or incoming datasets step by step. The reusing mechanism was replaced with attributes gain preferring of the incoming data. The best attribute based on gain evaluation will be evaluated at each step, reducing accuracy compared to a random forest tree. However, even when an approximate change happened when feature minimization was accorded, it still affected the accuracy as in LwL. It provided accuracy with 89.3% and 90.96% for without/with features selection. As mentioned, LWL deals with a linear and non-linear relation between dependent and independent features by smoothing weights. LwL accuracy was low compared to RFT and HT for many reasons: At each step, all data required for LwL weight evaluation and cast and memory size are higher than previously applied techniques.

Otherwise, LwL still provided better results with well-impact feature extraction. In Naive Bayes updatable techniques was close to HT accuracy due to multiple classification of problems. It requires a lower training data size than other techniques and may save time and memory. NBU had a 94.4% and 94.9% for not/with applying feature selections due to using a probability mechanism to reduce the final classpath.

In this paper, the probability of each related attribute with the related or non-related classes was calculated, studied, and explained in the methodology section. NBU, considers an issue related to dissimilar prediction attributes that will be set to zero. For that reason, some smoothing or getting the closest point mechanism must be applied to omit this problem. SVM also showed a good result near LwL but still after the RFT with 91.5% and 91.4% for both (non/using the feature selection technique). SVM operates very sufficiently with a well class's type separation, and needs more training data. It is also related to the relation between attributes and samples number, noisy data is impact on SVM accuracy, all of the previous reasons are considered as negative points lead to minimize SVM performance. (Table VI) shows the comparison between all 5 applied techniques before and after feature selection applied.

TABLE VI
COMPARISON BETWEEN APPLIED TECHNIQUES BEFORE/AFTER FEATURE SELECTION.

Technique ID*	Correctly Classified Samples		Incorrectly Classified Samples		Mean Absolute Error	
	Before	After	Before	After	Before	After
RFT	986	988	14	12	0.0219	0.0202
HT	942	945	58	55	0.0459	0.0403
LWL	893	909	107	91	0.0904	0.0822
NB	944	949	56	51	0.0458	0.0415
SVM	915	914	85	86	0.2516	0.2522

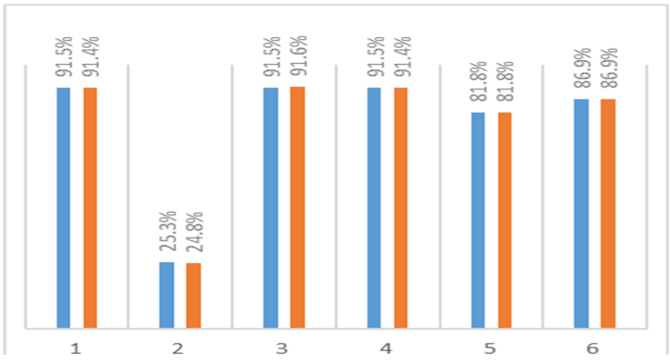
*Where RFT is Random Forest tree, HT is Hoeffding Tree, LWLW is Locally Weighted Regression, NBU is Naïve Bayes Updatable and SVM is Support Vector Machine.

Also, the confusion matrix was calculated for all 5 applied techniques, as shown in (Table VII) which determined the false prediction samples and their positions related to each class (noted in red color).

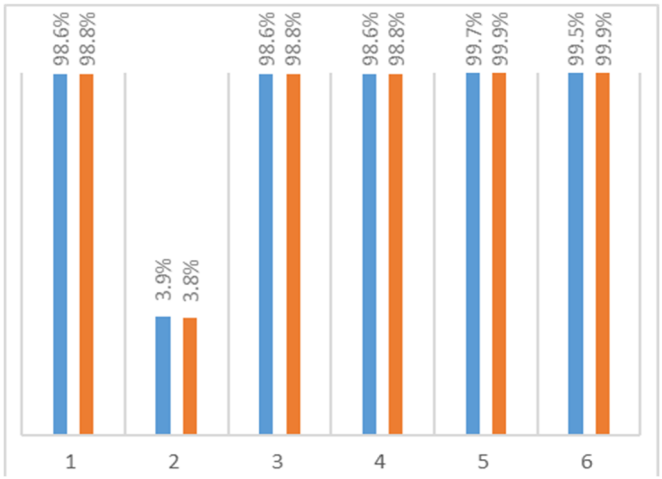
TABLE VII
THE FPT AND ITS POSITIONS RELATED TO CLASS (IN RED COLOR).

Methods	Before Feature Selection			After Feature Selection		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
	98	0	5	99	0	4
RFT	0	51	2	0	50	3
	7	0	837	4	1	839
	49	1	8	92	1	10
HT	1	48	4	3	43	7
	38	6	800	33	1	810
	103	0	0	103	0	0
LwL	51	0	2	52	0	1
	54	0	790	38	0	806
	94	1	8	95	0	8
NBU	0	49	4	3	44	6
	37	6	801	33	1	810
	92	0	11	93	0	10
SVM	18	0	35	18	0	35
	12	0	823	23	0	821

In Fig. 4. the true/false predicated samples with the prediction method performances have been calculated and presented. In this figure, RFT and SVM are explained well by showing their performances with and without feature minimization to demonstrate that minimization of data and risk factors specification shows a promising result. Which assist healthcare and specialist to provide a minimum dataset with better results as well to minimize cost and prevent patients from disease affection.



(a) Support Vector Machine (SVM) performances.



(b) Random Forest Tree (RFT) performances.

Fig. 4 Prediction method performances. 1: Accuracy, 2: FPT, 3: Precision, 4: Recall, 5: Precision of converge and 6: PRC area.

IV. CONCLUSION

It should involve several working and repeating processes to obtain high performance in any disease prediction and make a significant decision. In addition, these processes require multiple and different features with model- or algorithm-specific selection to make a compatible work step leading to a successful system. The knowledge of any feature selection process plus the representativeness of the dataset are considered important points of a model to hit the objective of dealing with features. Model complexity and the problem environment make choices to select which size of the dataset is more adequate and provides more effectiveness for the utilizing methods. Minimizing features of the selected Diabetic dataset provides an easy way to deal with and manage the dataset. In addition, complex techniques have faster training steps and are less effective in overfitting issues if compared to larger datasets. Also, collecting od dataset will be easier and make patients less exposed to dangerous or even vital circumstances.

Machine Learning techniques need an adequate sample to avoid false ratings on both sides (positive or negative). At the same time, features are essential to identify in such types of diseases to be selected carefully and reduce patients moving or even reduce the effect of getting more difficult issues due to a diabetic person's weak immunity system. Some of the features such as Gender and Age, in addition to some test results such as HbA1C and IG were shown to have a higher impact on the overall prediction system. After studying several ML types and with feature selection, RFT showed the best results for both studies with a high prediction of about 98% and a va very low difference. This was due to RFT ability to cover missing values and dependent on the the nonlinearity characteristics of the the utilized dataset. HT also provided a high accuracy for both studies, with about 94% and low differences of about 0.3%. The independent and dependent relation privilege of LwL, provided a high accuracy as well but still not similar to RFT due to the evaluation of the overall dataset again at each step of prediction levels. RFT has a learning type with building a several DT and then combining their decisions with diversity between the created trees.

Focusing on the most compelling features, such as Gender, Age, and HbA1C enhances RFT diversity and prohibits any strong feature from dominating the overall process unless all features are tested. Also, as shown in figure 3, RFT can capture any relationship between features, even if it has a complex version, such as between LDL plus Age. The relations between the most effective features and other features must be specified well so neglecting any other features may not affect the overall process. RFT shows a perfect handling of some features outlier values such as Creatine, Urea and HDL shown in Table V is due to its robustness in dealing with outlier's values. Table VI and Table VII showed the results related to feature affection on ML in correctly classified samples and class types. As conclusion, the size of dataset and making choices need to specify the problem objectives to provide a balance among the model, the quality of the selected dataset, and its size.

REFERENCES

[1] J. Li et al., "A tongue features fusion approach to predicting prediabetes and diabetes with machine learning," J. Biomed. Inform., vol. 115, Mar. 2021, doi: 10.1016/j.jbi.2021.103693.

[2] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," in Procedia Computer Science, 2019, vol. 165, pp. 292–299, doi: 10.1016/j.procs.2020.01.047.

[3] S. Shafi and G. Ahmad Ansari, "Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach." [Online]. Available: https://ssrn.com/abstract=3852590.

[4] M. Ravaut et al., "Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data," npj Digit. Med., vol. 4, no. 1, Dec. 2021, doi: 10.1038/s41746-021-00394-8.

[5] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," Sci. Rep., vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-68771-z.

[6] S. Islam Ayon and M. Milon Islam, "Diabetes Prediction: A Deep Learning Approach," Int. J. Inf. Eng. Electron. Bus., vol. 11, no. 2, pp. 21–27, Mar. 2019, doi: 10.5815/ijieeb.2019.02.03.

[7] F. M. Aswad, A. M. S. Ahmed, N. A. M. Alhammadi, B. A. Khalaf, and S. A. Mostafa, "Deep learning in distributed denial-of-service attacks detection method for Internet of Things networks," J. Intell. Syst., vol. 32, no. 1, 2023, doi: 10.1515/jisys-2022-0155.

[8] Chang V, Bailey J, Xu QA, Sun Z. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms". Neural Comput Appl. 2022 Mar 24:1-17. doi: 10.1007/s00521-022-07049-z.

[9] P. Aziz, A. Hermawan, and D. Avianto. "Analyze Important Features of PIMA Indian Database for Diabetes Prediction Using KNN." Jurnal Sisfokom (Sistem Informasi dan Komputer) 12.1 (2023): 70-75.

[10] Patra, Radhanath. "Analysis and prediction of Pima Indian Diabetes Dataset using SDKNN classifier technique." IOP Conference Series: Materials Science and Engineering. Vol. 1070. No. 1. IOP Publishing, 2021.

[11] Miao Y. "Using machine learning algorithms to predict diabetes mellitus based on Pima Indians Diabetes dataset". In2021 the 5th International Conference on Virtual and Augmented Reality Simulations 2021 Mar 20 (pp. 47-53).

[12] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," in Procedia Computer Science, 2020, vol. 167, pp. 706–716, doi: 10.1016/j.procs.2020.03.336.

[13] P. Julian Benadit and F. Sagayaraj Francis, "Improving the performance of a proxy cache using very fast decision tree classifier," in Procedia Computer Science, 2015, vol. 48, no. C, pp. 304–312, doi: 10.1016/j.procs.2015.04.186.

[14] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[15] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458.

[16] J. F. Magee, "Decision Making: Decision Trees for Decision Making", Harvard Business Review, 1964.

[17] Zhou H, Zhang J, Zhou Y, Guo X, Ma Y. A feature selection algorithm of decision tree based on feature weight. Expert Systems with Applications. 2021 Feb 1;164:113842.

[18] P. K. and P. S. Arvind Kumar, "A Survey on Hoeffding Tree Stream Data Classification Algorithms," in Proceedings of the National Conference on Recent Innovations in Science and Engineering (RISE-2016, 2015, vol. 1, no. 2, pp. 28–32.

[19] A. Muallem, S. Shetty, J. W. Pan, J. Zhao, and B. Biswal, "Hoeffding Tree Algorithms for Anomaly Detection in Streaming Datasets: A Survey," J. Inf. Secur., vol. 08, no. 04, pp. 339–361, 2017, doi: 10.4236/jis.2017.84022.

[20] H. Hanafi, A. Hendi Muhammad, I. Verawati, and R. Hardi. "An intrusion detection system using sdae to enhance dimensional reduction in machine learning." JOIV: International Journal on Informatics Visualization 6, no. 2 (2022): 306-316.

[21] Sarmini, A. Alhabeeb , M. M. Abusharhah , T. Hariguna , A. R. Hananto, "An Investigation into Indonesian Students' Opinions on Educational Reforms through the Use of Machine Learning and Sentiment Analysis "JOIV : International Journal on Informatics Visualization 6, no. 3 (2022), PP: 604-609.

[22] E. Bahmani, J. Mojtaba, and S. Abdusalam. "Breast cancer prediction using a hybrid data mining model." JOIV: International Journal on Informatics Visualization 3.4 (2019): 327-331.

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

[24] H. Yoshikawa, "Can naive Bayes classifier predict infection in a close contact of COVID-19? A comparative test for predictability of the predictive model and healthcare workers in Japan: Infection Prediction in a Close Contact of COVID-19," J. Infect. Chemother., vol. 28, no. 6, pp. 774–779, Jun. 2022, doi: 10.1016/j.jiac.2022.02.017.

[25] D. Keerthana, V. Venugopal, M. K. Nath, and M. Mishra, "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer," Biomed. Eng. Adv., vol. 5, p. 100069, Jun. 2023, doi: 10.1016/j.bea.2022.100069.

[26] F. J. Shaikh and D. S. Rao, "Prediction of Cancer Disease using Machine Learning Approach," in Materials Today: Proceedings, 2021, vol. 50, pp. 40–47, doi: 10.1016/j.matpr.2021.03.625.

[27] Mohanarathinam, A., et al. "Diabetic Retinopathy Detection and Classification using Hybrid Multiclass SVM classifier and Deeplearning techniques." Mathematical Statistician and Engineering Applications 71.3 (2022): 891-903.

[28] Khan, Asfandyar, et al. "Cardiovascular and Diabetes Diseases Classification Using Ensemble Stacking Classifiers with SVM as a Meta Classifier." Diagnostics 12.11 (2022): 2595.

[29] Patil R, Tamane S, Rawandale SA, Patil K. A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus. Int. J. Electr. Comput. Eng. 2022 Feb 1;12(1):524-33.

[30] M. Shaheen, N. Naheed, and A. Ahsan, "Relevance-diversity algorithm for feature selection and modified Bayes for prediction," Alexandria Eng. J., Mar. 2022, doi: 10.1016/j.aej.2022.11.002.

[31] Yab, L. Y., Wahid, N., & Hamid, R. A., Inversed Control Parameter in Whale Optimization Algorithm and Grey Wolf Optimizer for Wrapper-based Feature Selection: A comparative study. JOIV: International Journal on Informatics Visualization, 2023, 7(2), 477-486.

[32] D. H. Jeong, B. K. Jeong, N. Leslie, C. Kamhoua, and S.-Y. Ji, "Designing a supervised feature selection technique for mixed attribute data analysis," Mach. Learn. with Appl., vol. 10, p. 100431, Dec. 2022, doi: 10.1016/j.mlwa.2022.100431.

[33] M. T. Kurniawan, S. Yazid, and Y. G. Sucahyo. "Comparison of Feature Selection Methods for DDoS Attacks on Software Defined Networks using Filter-Based, Wrapper-Based and Embedded-Based." JOIV: International Journal on Informatics Visualization 6.4 (2022): 809-814.

[34] M. A. H. Azmi, C. F. M. Foozy, K. A. M. Sukri, N. A. Abdullah, I. Rahmi, A. Hamid, H. Amnar,. "Feature Selection Approach to Detect DDoS Attack Using Machine Learning Algorithms." JOIV: International Journal on Informatics Visualization 5.4 (2021): 395-401.

[35] M. S. M. and S. J. M. S. KURNAZ, "A High Efficiency Thyroid Disorders Prediction System with Non-Dominated Sorting Genetic Algorithm NSGA-II as a Feature Selection Algorithm," in 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1–6, doi: 10.1109/INCET49848.2020.9154189.

[36] S. J. M. and M. S. Mohammed, "COVID-19 risk factors specification using Decision Tree based on the degree of redundancy between features," in 022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), 2022, pp. 1–11, doi: 10.1109/GCAT55367.2022.9971950.

[37] D. Fahrudy, & S. 'Uyun "Classification of Student Graduation using Naïve Bayes by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection," JOIV : International Journal on Informatics Visualization, vol. 6, no. 4, , pp. 798-808, Dec. 2022.

[38] M. Azmi, C. Foozy, K. Sukri, N. Abdullah, I. Hamid, & H. Amnur "Feature Selection Approach to Detect DDoS Attack Using Machine Learning Algorithms," JOIV : International Journal on Informatics Visualization, vol. 5, no. 4, , pp. 395-401, Dec. 2021.

[39] Y. Nataliani "Feature-reduction Fuzzy c-means Clustering for Basketball Players Positioning," JOIV : International Journal on Informatics Visualization, vol. 5, no. 4, , pp. 415-421, Dec. 2021.

[40] Minarno, A. E., Mandiri, M. H. C., Azhar, Y., Bimantoro, F., Nugroho, H. A., & Ibrahim, Z. (2022). Classification of Diabetic Retinopathy Disease Using Convolutional Neural Network. JOIV: International Journal on Informatics Visualization, 6(1), 12-18.

[41] Toresa, D., Shahril, M. A. E., Harun, N. H., Bakar, J. A., & Amnur, H. (2021). Automated Detection and Counting of Hard Exudates for Diabetic Retinopathy by using Watershed and Double Top-Bottom Hat Filtering Algorithm. JOIV: International Journal on Informatics Visualization, 5(3), 242-247.

[42] A. Rashid, "Diabetes Dataset," Mendeley Data, 2020. https://data.mendeley.com/datasets/wj9rwkp9c2/1.