Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI

(Rekayasa Sistem dan Teknologi Informasi)

Vol. 6 No. 5 (2022) 832 - 839 ISSN Media Electronic: 2580-0760

Exploring feature selection techniques on Classification Algorithms for Predicting Type 2 Diabetes at Early Stage

Mila Desi Anasanti[1], Khairunisa Hilyati[2], Annisa Novtariany[3]
[1]Department of Information Studies, University College London, London, United Kingdom
[2,3]Ilmu Komputer, Universitas Nusa Mandiri, Jakarta, Indonesia
[1]mila.mld@nusamandiri.ac.id, [2]14210217@nusamandiri.ac.id*, [3]14210181@nusamandiri.ac.id

Abstract

Predicting early Type 2 diabetes (T2D) is critical for improved care and better T2D outcomes. An accurate and efficient T2D prediction relies on unbiased relevant features. In this study, we searched for important features to predict T2D by integrating ML-based models for feature selection and classification from 520 individuals newly diagnosed with diabetes or who will develop it. We used standard machine learning classifications, such as logistic regression (LR), Gaussian naive Bayes (NB), decision tree (DT), random forest (RF), support vector machine (SVM) with linear basis function, and k-nearest neighbors (KNN). We set out to systematically explore the viability of main feature selection representing each different technique, such as a statistical filter method (F-score), an entropy-based filter method (mutual information), an ensemble-based filter method (random forest importance), and a stochastic optimization (simultaneous perturbation feature selection and ranking (SpFSR)). We used a stratified 10-fold cross-validation technique and assessed the performance of discrimination, calibration, and clinical utility. We attained the highest accuracy of 98% using RF with the full set of features (16 features), then used RF as a classifier wrapper to select the important features. We observed a combination of SpFSR and RF as the best model with a P-value above 0.05 (P-value = 0.26), statistically attaining the same accuracy as the full features. The study's findings support the efficiency and usefulness of the suggested method for choosing the most important features of diabetic data: polyuria, gender, polydipsia, age, itching, sudden weight loss, delayed healing, and alopecia.

Keywords: *Type 2 diabetes, machine learning, feature selection, feature importance*

1. Introduction

Type 2 diabetes (T2D) is one of the fastest growing life-threatening chronic diseases, increasing by 108 million in 1980 [1][2]. T2D killed 1.6 million people in 2016, an increase of more than 1 million from the previous year [3]. In 2018, it increased even more by 422 million people, and the symptoms increased with age 20 to 60. T2D has risen fourfold in the last 40 years, and its risk of death has never decreased [4]. The increasing number of T2D mortalities is due to the delay in diagnosing the disease, which can cause complications [5].

T2D develops when the body's metabolic processes cannot adequately digest meals, leading to increased blood sugar. It may erode blood vessels, raising the possibility of serious health issues that harm the heart, kidneys, eyes, and nerves. There are two common T2D concerns grouped as microvascular (artery damage) and macrovascular (tiny blood vessel damage). The group of microvascular diseases is the organ that is attacked by T2D, which are the eyes (retinopathy), kidneys (nephropathy), and nerve damage (neuropathy) [6][7]. The common T2D symptoms include weight loss, itchy skin, polydipsia, polyuria, and polyphagia [8][9].

Since the symptoms are similar to common illnesses, many people are unaware that they may have T2D, leading to complications. A blood test from a doctor is necessary to determine if someone has T2D or not. However, routine blood tests are rarely performed in developing countries because of their high cost. In this case, machine learning (ML) methods can be served as an alternative to predict T2D outcomes by analyzing several risk factors and symptoms. ML offers many processing algorithms that can be efficiently used to predict disease at a low cost. These algorithms are divided into supervised, unsupervised, and reinforcement learning, where each type is further divided into several types of algorithms.

An efficient data collection stage and computational time can be achieved by incorporating fewer important

Accepted: 10-08-2022 | Received in revised: 27-09-2022 | Published: 31-10-2022

features. Several feature selection (FS) techniques have been proposed to determine the most important features. These techniques can avoid overfitting, poor performance, and excessive computational times, especially when the data is a high-dimensional dataset with many features. Dependent and independent classifiers are the two main divisions of the FS approach. There are two ways to use a dependent classifier: the wrapper technique and the embedding method. [10]. The wrapper technique assesses a subset of variables to look for potential interactions between variables while guaranteeing the classifier's prediction accuracy. [11]. The FS technique involves choosing, either automatically or manually, the features that have the greatest impact on the target variable. The FS is created, in essence, using the feature importance score. [12].

This study aimed to develop an efficient T2D prediction approach using the important features by integrating the ML and FS techniques. Research on T2D has been carried out in recent years using various classification methods; however, only a few studies focused on investigating the important features. Several studies used complete features and achieved high accuracies in predicting T2D using standard ML approaches. In 2016, a support vector machine (SVM) and K-nearest neighbor (KNN) were used by J.A. Putra et al. that achieved an accuracy of 92% [13]. In 2017, Manimaran et al. used the decision tree (DT) method to classify T2D with an accuracy of 85.02% [14]. In 2020, D.A. Agatsa et al. built a classification model using the SVM method with an accuracy of 77.92% [15]. M. M. F. Islam et al. in 2020 compared the performances of several ML methods gaining the highest accuracy of 97% by random forest (RF) [16]. Several studies applied FS techniques in T2D prediction, as shown in Table 1; however, most only focused on one FS technique. Their accuracies are not as high as the previous studies using the complete features.

Table 1. Comparison of different feature selection and classification measures of type 2 diabetes (T2D) prediction

Authors	Feature selection	Classification method	Accuracy
M. Pradhan et al.	GA	ANN	73.83%
S.Y. Rubaiat et al.	RF	ANN	77.08%
B. Sarojini, N. Ramaraj	FCBF	SVM	77.47%
L.W. Astuti et al.	BWOA	ANN, NB	70%
R. Saxena et al.	RF	DT, KNN, RF	79.8%
K.C. Tan et al.	GA	SVM	78.26%

GA: genetic algorithm; RFI: random forest importance; FCBF: fast correlation-based filter; BWOA: binary wheal optimization algorithm; ANN: artificial neural network; SVM: support vector machine; NB: naïve Bayes; DT: decision tree; KNN: k-nearest neighbor.

Our study tried to improve the prediction performance by investigating several FS techniques to select the most important features to predict T2D. Applying the feature selection technique was expected to be more accurate, effective, and efficient than full features. Comparing several FS techniques will optimize the ML classification model to search for the highest accuracy. In addition, stratified cross-validation and a-paired t-test were performed to guarantee that the best performance was attained statistically and not merely by chance.

2. Research Methods

To perform the predictive analysis of T2D, the steps taken to get the results are shown in Figure 1.

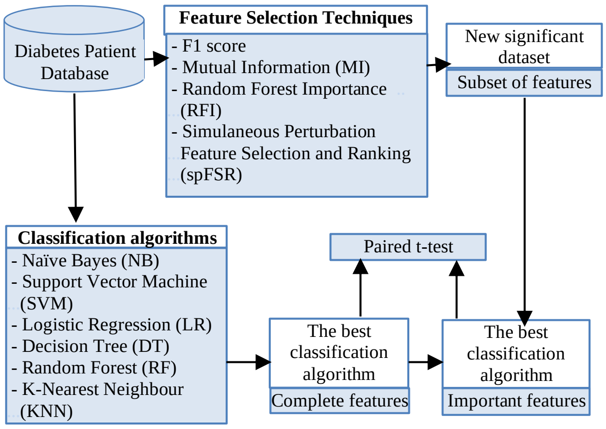
Diabetes Patient Database
Feature Selection Techniques
- F1 score
- Mutual Information (MI)
- Random Forest Importance (RFI)
- Simulaneous Perturbation Feature Selection and Ranking (spFSR)
New significant dataset
Subset of features
Classification algorithms
- Naïve Bayes (NB)
- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- K-Nearest Neighbour (KNN)
Paired t-test
The best classification algorithm
Complete features
The best classification algorithm
Important features

Figure 1. Research Methodology

All the analyses were carried out using Python version 3.9.7 by employing several modules, including 'scikit-learn', a Python ML module built on top of "scipy' package, and 'featurewiz', a Python library for creating and selecting the most important features in data.

2.1 Data Collection

An open dataset from the UCI Machine Learning Repository was used in this project [17]. The dataset contains information on periodic health checks from 520 subjects aged between 20 to 65 years, collected in 2020. There are 16 attributes consisting of 15 numerical variables, one categorical variable, and a categorical response variable (class), as described in Table 2.

We preprocessed the data by encoding the target labels with a value between 0 and 1. It calculates the likelihood of a particular event, such as voting or not voting, using a given dataset of independent variables. We also performed a normalization technique, such as in the variable age, using Min-Max scaling to scale the dataset to a specific range between 0 and 1 by using each attribute's minimum and maximum values.

DOI: https://doi.org/10.29207/resti.v6i5.4419

Table 2. The early stage T2D risk prediction dataset

No	Variable	Description	Class
1	Sex	Gender	Categorical
2	Age	Age	Numerical
3	polyphagia	Polyphagia	Numerical
4	polyuria	Polyuria	Numerical
5	polydipsia	Polydipsia	Numerical
6	Sudden _weight _loss	Sudden Weight Loss	Numerical
7	genital_ thrush	Genital Thrush	Numerical
8	visual_ blurring	Visual Blurring	Numerical
9	weakness	Weakness	Numerical
10	Itching	Itching	Numerical
11	irritability	Irritability	Numerical
12	Delayed _healing	Delayed Healing	Numerical
13	Partial _parises	Partial Parises	Numerical
14	Muscle _stiffness	Muscle Stiffness	Numerical
15	alopecia	Alopecia	Numerical
16	Obesity	Obesity	Numerical
17	Class	Class	Categorical

2.2 Feature selection techniques

We employed different FS techniques during model development. We employed an ensemble-based filter method (random forest importance), an FS using stochastic optimization, an entropy-based filter, and a statistical filter method (F-score).

2.2.1 F-Score feature selection

It is a technique with a statistical approach, sorting the relevant features by assessing each feature [18]]. The F-score formula is:

$$F(i) = \frac{\left(\bar{x}_\iota^{(+)} - \bar{x}_\iota\right)^2 + \left(\bar{x}_\iota^{(-)} - \bar{x}_\iota\right)^2}{\frac{1}{n+-1}\sum_{k=1}^{n+}\left(\overline{x_{k,\iota}}^{(+)} - \bar{x}_\iota^{(+)}\right)^2 + \frac{1}{n--1}\sum_{k=1}^{n-}\left(\overline{x_{k,\iota}}^{(-)} - \bar{x}_\iota^{(-)}\right)^2}$$

Where:

$\overline{x_i}^{(+)}, \overline{x_i}^{(-)}$ = the average of each ith feature across positive and negative datasets, $\overline{x_{k,\iota}}^{(+)}$ = the ith feature of the kth positive instance, $\overline{x_{k,\iota}}^{(-)}$ = the ith feature of the kth negative instance

2.2.2 Mutual information (MI)

MI is one of the FS techniques that has been used since 1990. It is often used to measure the dependence between two features and how much information from a feature can be used for others [19].

The dependency metric between two variables is applied to information-theoretic ranking criteria. Starting with Shannon's definition of entropy:

$$H(X) = \sum_i P(y)\, logP(y)$$

The uncertainties in output Y are represented by the equation above. If a variable X is given, the conditional entropy is calculated using the following formula:

$$H\frac{X}{Y} = -\Sigma_i \Sigma_y P(x,y) logP\frac{x}{y}$$

The equation above says that the uncertainty in the result Y is reduced by looking at the variable X. The reduction in uncertainty is denoted by:

$$I(Y,X) = H(Y) - H(Y|X)$$

Thus, the MI between X and Y is produced, which, if X and Y are interdependent, will be bigger than zero and, if they are not, will be zero. This equation suggests that a dependent relationship may exist whereby one variable can provide insight into another. The conditions provided above are for discrete variables, but they can also be used for continuous variables by changing the sum for the integral [18].

2.2.3 Random Forest importance (RFI)

RFI is an FS technique that calculates the ensemble's average of the selected variables [20]. Assuming there are two columns, [0] and [1], finding the nodes where the split occurred as a result of column [0] is necessary to determine the feature importance of [0]. This collection has only one node per column [0] and column [1]. The formula to determine the importance of a node is as follows:

$$n_i = \frac{N_t}{N}\left[impurity - \left(\frac{N_{t(right)}}{N_t}\, x\ right\ impurity\right) - \left(\frac{N_{t(left)}}{N_t}\, x\ left\ impurity\right)\right]$$

Where:

N_t is the number of rows in that specific note, N is the total number of rows present in the data, Impurity is a Gini index value, $N_{t(right)}$ is the number of nodes in the right node, $N_{t(left)}$ is the number of notes in the left node

2.2.4 Simultaneous perturbation feature selection and ranking (spFSR)

SpFSR is a new method for FS and ranking, an extension of the general-purpose black box stochastic optimization algorithm. SpFSR starts with the initial solution w0, and there is a recursion to find the local minimum $\hat{w}$:

$$\hat{w}_k + 1 := \hat{w}_k - a_k \hat{G}(\hat{w}_k)$$

Where a_k is the order of iteration gain; $a_k \geq 0$ and $\hat{G}(\hat{w}_k)$ are estimates of the gradient at k.

2.3 Machine learning classification methods

Several ML classification methods were used in this study which is detailed below:

DOI: https://doi.org/10.29207/resti.v6i5.4419

2.3.1 Logistic Regression (LR)

It is also known as a sigmoid function, which measures the relationship between the dependent variables that are independent and categorical [21]. A collection of independent variables determines the chance of a specific occurrence, such as voting or not voting. Since the outcome is a probability, the range of the dependent variable is 0 to 1.

$$log\frac{p}{1+p} = a+\beta_1 + x_1 + \beta_2 + x_2.....+\beta_i + x_i$$

p = probability of an outcome, a = intercept
β_1 = associated coefficient, $_i$ = value of the predictor variable

2.3.2 Naïve Bayes (NB)

It is a straightforward probabilistic classifier that determines a set of probabilities by adding up the incidences and mixtures of values from the dataset. The Gaussian naïve Bayes_assumes that the data from each label took from the Gaussian distribution [22][23].

$$p(x) = \frac{p(h) \cdot p(h)}{p(x)}$$

X = proof X, H = hypothesis, P(H|X) = probability that hypothesis H is valid for the proof X.

P(H|X) is the posterior probability of H on the condition of X. P(H|X) is the probability that X is true for hypothesis H or probability X on the condition that H. P(H) is the prior probability of the proof X.

2.3.3 Support Vector Machine (SVM)

It can handle high-dimensional datasets, linear and non-linear kernel classification, and regression, making SVM reliable for classification and regression algorithms. The objective is to identify the most effective classification function to separate the training data in the two classes [24][25]. The separation of the two classes is done using a separator line, called a hyperplane equation, which can be calculated as follows:

H: wT(x) + b = 0

b = the bias term and intercept of the hyperplane equation

The hyperplane is always formatted to be a D-1 operator in D-dimensional space. A hyperplane, for instance, is a linear line in 2-D space (1-D).

2.3.4 Decision Tree (DT)

It is used to classify objects which are usually described as leaves and branches. The leaves themselves are identified by class, while the branches represent the condition of the object attribute being measured. DT is a common classifier in ML that works by dividing into two or more regression tree models and subdividing it into smaller ones so that the decision tree will grow [26].

$E(S) = \sum_{i=1}^{c} -P_1 log_2 P_i$

S = initial condition, i = set class on S, Pi = probability or portion of class *i* in a node

2.3.5 Random Forest (RF)

It is a classification algorithm with more than one decision tree that is formed depending on the value of a random vector sampled independently and identically. In addition, RF is also included in the supervised learning group, which can be used to make predictions. It is a classification method with high accuracy that can handle significant input variables without overfitting [27].

$Entropy(S) = \sum_{i=1}^{n} -P_1 log_2(P_i)$= Set of cases
n = Number of partitions S, Pi = portion of S to S

2.3.6 K-nearest neighbor (KNN)

It is a technique for classifying unknown cases by searching for the nearby patterns in the pattern space [28] [29]. Euclidean distance is used by KNN to predict class:

$$d(x, y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

To locate the nearest example in the pattern space, Euclidean distance $d(x, y)$ is utilized to calculate the distance. A majority vote of its neighbors identifies the class of the unknown examples.

2.4 Evaluation

We evaluated the best performance of the ML method with the complete features by comparing the accuracies of each method. The accuracy can be calculated using the formula [30]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

TP = True Positive, TN = True Negative,
FP = False Positive, FN = False Negative

We used a stratified 10-fold cross-validation technique for the performance assessment with three repetitions to reduce the variability while still managing efficient computational time. To ensure future replications and independent confirmations of our findings, we set the random state to 999. All FS approaches were fitted and tested on the same data partitions since we kept the random state constant across all cross-validation procedures. The technique implies that our trials were done in pairs with significantly lower variability than separately.

However, since the cross-validation method involves a random process, statistical tests are necessary to determine whether any performance difference between any two FS or ML methods is statistically significant or merely a result of sample variation. We used paired t-tests to see the statistically significant differences between the whole set of features-based ML approaches and the ML-FS techniques.

We performed the paired t-test using the 'stats.ttest' function from the 'scipy' package in Python, and then we looked at the p-values. If the p-value is less than 0.05, we can conclude that the difference is statistically significant at a 95% confidence level.

3. Results

3.1 Performances of machine learning classification methods using the full features

The performance comparison of the ML methods used in this study using the complete features is presented in Table 2.

Table 2. Values of different feature selection techniques using the full features

Algorithms	Accuracy	Precision	Recall	F1-Score
Decision Tree	96%	98%	97%	97%
Random Forest	98%	98%	99%	98%
Support Vector Machine	92%	94%	93%	93%
Naïve Bayes	89%	90%	92%	91%
Logistic Regression	93%	95%	93%	94%
KNN	96%	99%	94%	96%

It can be observed that RF outperformed the other ML methods followed by DT. To determine whether the different accuracies are significant statistically and not happened by chance, we conducted several paired t-tests. We obtained all significant P-values < 0.05 with the greatest P-value = 0.003, comparing the performances of RF and DT. Thus, RF was the best method to predict T2D with 98% accuracy

We subsequently used RF as a classifier wrapper in the ML-FS framework to determine which minimal number of features can be used to obtain the same values as using the complete features. We started using four and six features and increased the number by two until we reached the same performance accuracy as those from the full features.

3.2 Performance of random forest classification method using four features

The results of FS performances using four important selection features are shown in Table 3.

Table 3. Values of different feature selection techniques using four features

Feature selection techniques	Accuracy	Precision	Recall	F1-Score
F-Score Feature Importance	89%	88%	95%	91%
Mutual Information	92%	96%	91%	93%
RFI	91%	95%	91%	93%
SpFSR	92%	96%	91%	91%

The results indicated that spFSR-RF and MI-RF using six features outperformed the other FS methods. However, we observed P-value below 0.05 after comparing the values of these two techniques with the values of the full features using paired t-test. Thus, an increasing number of features would be recommended.

3.3 Performance of random forest classification method using six features

The results of FS performances using six important selection features are shown in Table 4.

Table 4. Values of different feature selection techniques using six features

Feature selection techniques	Accuracy	Precision	Recall	F1-Score
F-Score Feature Importance	91%	94%	90%	92%
Mutual Information	94%	96%	94%	95%
RFI	94%	96%	93%	95%
SpFSR	96%	97%	96%	96%

The results indicated that spFSR using six features outperformed the other FS methods. However, the accuracy using the full set of features is still statistically significantly higher after conducting a paired t-test (P-value = 0.0028). Thus, spFSR with six features performed slightly poorly compared to the full features. Therefore, FS using a higher number of features should be examined.

3.4 Performance of random forest classification method using eight features

The results of FS performances using eight important selection features are shown in Table 5.

Table 5. Values of different feature selection techniques using eight features

Feature selection techniques	Accuracy	Precision	Recall	F1-Score
F-Score Feature Importance	94%	95%	95%	95%
Mutual Information	96%	97%	96%	97%
RFI	97%	97%	98%	97%
SpFSR	98%	98%	98%	98%

DOI: https://doi.org/10.29207/resti.v6i5.4419

Again, we observed that spFSR outperformed the other FS techniques, which was confirmed after performing paired t-tests. We observed P-value = 0.26 for spFSR vs. the complete features, indicating the same significant values. Therefore, the performance of the spFSR with eight features is statistically equivalent to that of the complete features. With eight features, we reached the aim of this study to reduce the number of features and find the important ones. The top eight features selected from spFSR are shown in Figure 2 with their respective important scores, showing polyuria is the main vital feature in predicting T2D.

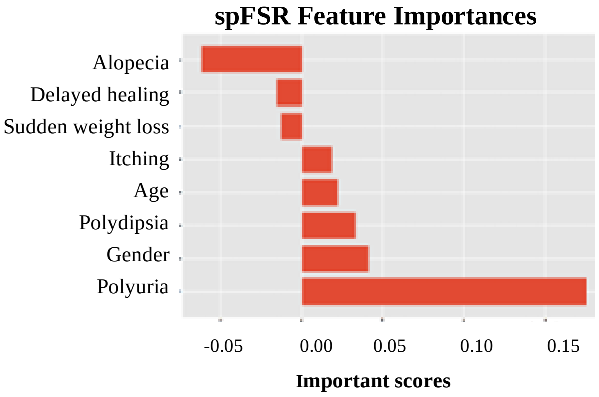

Figure 2. Top Eight spFSR Feature Importances

4. Discussion

This study used several FS techniques on the available T2D dataset. FS reduced the number of features in diagnosing T2D, providing an efficient and low-cost prediction technique. The highest prediction performance on all attributes was obtained using the RF algorithm with an accuracy of 98%, statistically the same as RF using the subset of features (P-value=0.26). Remarkably, we can achieve similar results by only using eight features selected by spFSR, integrating spFSR-RF.

Our study produced the highest 98% accuracy in predicting T2D using the FS technique surpassing the previous studies [31][32][33][34][35] that only achieved accuracies between 70% and 80%. However, the reported performance cannot be directly compared due to the different datasets and validation methodologies. The previous study conducted by M. M. F. Islam et al., using the same dataset with all features, attained the highest accuracy of 97% using RF with the train-test splitting [16]. We still achieved a slightly higher accuracy, but only incorporating half of the features showing our method's efficiency in integrating the FS-ML approach. We also used a stratified cross-validation technique to handle the uncertainty of the prediction. In the train split approach conducted M.M.F. Islam et al., each instance receives only one check and only makes an exact appearance in a test set, producing uncertainty in the result. Even repeated random splits are likely to have less reliable results since the variance will rise while the average will remain relatively constant.

Moreover, our study proposed an ML classification method using fewer features that would be preferred since collecting the complete set of features would require more time, effort, additional cost, and computational complexity. The FS technique aimed to reduce parameters and adequately represent some relevant data. Integrating the FS technique into ML methods can serve as an efficient and low-cost prediction of T2D.

Based on our study, the recommended eight features in predicting T2D attributes are polyuria, gender, polydipsia, age, itching, sudden weight loss, and delayed healing. The feature served as the most feature importances, especially the remarkable polyuria feature where the sugar level in diabetics is high enough that it affects the condition in patients who can produce up to 8 liters of urine per day. Although the sugar level is controlled by diet, the symptoms of polyuria persist, and this is the main symptom of T2D that is relevant to the literature [30]. The results also showed that gender is a significant risk factor for T2D, in line with the medical literature, which states that the percentage of diabetic patients in women is greater than that of men [36]. Women have a higher body fat composition compared to men, which makes them prone to be obese, which is associated with the risk of T2D [36]. The following important feature is polydipsia, a condition of continuous thirst experienced by people with diabetes. This condition is linked to patients' polyuria, resulting from a lack of bodily fluids [34]. It is a risk factor for all degenerative diseases and affects how the body functions, mainly how well the hormone insulin functions, which results in high blood sugar levels [36]. Another significant T2D symptom brought on by the skin drying out is itching (impaired regulation of body fluids) [37].

Similarly, sudden weight loss is when people with diabetes experience hypoglycemia, a decrease in sugar levels due to taking a medication that can cause symptoms of sudden weight loss [38]. In contrast, delayed healing occurs in people with T2D due to high glucose levels, as the body has difficulty processing glucose effectively. And the last is alopecia caused by the lack of nutrition of diabetics due to lifestyle changes associated with treatment and comorbidities and complications that often accompany it [39]. Based on our study, the eight attributes described had been statistically proven to represent the entire dataset used to predict T2D, which aligns with medical literature.

Despite the great benefit, our study has a limitation. Since the dataset is not high-dimensional, we used all the data to train in the FS trials and then tested them on the entire dataset using a repeated cross-validation

DOI: https://doi.org/10.29207/resti.v6i5.4419

technique. Despite its simplicity, this method may lead to overfitting. A better approach would be suggested to perform this comparison within a combined train/test split approach. The data can be divided into two parts: train data and test data, then use a cross-validation technique to identify the most important features in the train data. The performance of these features on the test data can then again be evaluated using a repeated cross-validation technique. Another suggestion is replicating the method on another dataset to ensure the same high accuracy can still be achieved.

5. Conclusion

Integrating spFSR into the RF method would reduce the computational complexity of diagnosing a disease. The best performance in this study was achieved using eight features, obtaining an accuracy of 98%. This study demonstrated that using half of the features would have the same results as using the full set of features to predict T2D on early onset with reliable high accuracy. This approach would be recommended for future studies to apply to a larger dataset or other disease datasets.

Reference

[1] S. A. Mahmoudinejad Dezfuli, S. R. Mahmoudinejad Dezfuli, S. V. Mahmoudinejad Dezfuli, And Y. Kiani, "Early Diagnosis Of Diabetes Mellitus Using Data Mining And Classification Techniques," *Jundishapur Journal Of Chronic Disease Care*, Vol. 8, No. 3, Jul. 2019, Doi: 10.5812/Jjcdc.94173.

[2] F. Nasution, A. Azwar Siregar, And S. Tinggi Kesehatan Indah Medan, “Faktor Risiko Kejadian Diabetes Mellitus (Risk Factors For The Event Of Diabetes Mellitus),” *Jurnal Ilmu Kesehatan*, Vol. 9, No. 2, 2021, Accessed: Sep. 13, 2022. [Online]. Available: Https://Doi.Org/10.32831/Jik.V9i2.304

[3] C. Zhu, C. U. Idemudia, And W. Feng, "Improved Logistic Regression Model For Diabetes Prediction By Integrating Pca And K-Means Techniques," *Inform Med Unlocked*, Vol. 17, Jan. 2019, Doi: 10.1016/J.Imu.2019.100179.

[4] "New Who Global Compact To Speed Up Action To Tackle Diabetes," Apr. 14, 2021. Https://Www.Who.Int/News/Item/14-04-2021-New-Who-Global-Compact-To-Speed-Up-Action-To-Tackle-Diabetes (Accessed Aug. 06, 2022).

[5] R. Saxena, S. K. Sharma, M. Gupta, And G. C. Sampada, "A Novel Approach For Feature Selection And Classification Of Diabetes Mellitus: Machine Learning Methods," *Comput Intell Neurosci*, Vol. 2022, 2022, Doi: 10.1155/2022/3820360.

[6] T. Mahboob Alam *Et Al.*, "A Model For Early Prediction Of Diabetes," *Inform Med Unlocked*, Vol. 16, Jan. 2019, Doi: 10.1016/J.Imu.2019.100204.

[7] Yuhelma, Y. Hasneli, And F. Annis Nauli, “Identifikasi Dan Analisis Komplikasi Makrovaskuler Dan Mikrovaskuler Pada Pasien Diabetes Mellitus”.

[8] W. Apriliah *Et Al.*, “Prediksi Kemungkinan Diabetes Pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest,” 2021. Accessed: Sep. 13, 2022. [Online]. Available: Https://Doi.Org/10.32520/Stmsi.V10i1.1129

[9] Lestari, Zulkarnain, And St. Aisyah Sijid, “Diabetes Melitus: Review Etilogi, Patofisiologi, Gejala, Penyebab, Cara Pemeriksaan, Cara Pengobatan Dan Cara Pencegahan.” Accessed: Sep. 13, 2022. [Online]. Available: Https://Doi.Org/10.24252/Psb.V7i1.24229

[10] J. Piri *Et Al.*, "Feature Selection Using Artificial Gorilla Troop Optimization For Biomedical Data: A Case Analysis With Covid-19 Data," *Mathematics*, Vol. 10, No. 15, P. 2742, Aug. 2022, Doi: 10.3390/Math10152742.

[11] A. Mangal And E. A. Holm, "A Comparative Study Of Feature Selection Methods For Stress Hotspot Classification In Materials," *Integr Mater Manuf Innov*, Vol. 7, No. 3, Pp. 87–95, Sep. 2018, Doi: 10.1007/S40192-018-0109-8.

[12] F. Septianingrum And A. S. Y. Irawan, “Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes: Sebuah Literature Review,” *Jurnal Media Informatika Budidarma*, Vol. 5, No. 3, P. 799, Jul. 2021, Doi: 10.30865/Mib.V5i3.2983.

[13] J. A. Putra And A. Laksita Akbar, “Klasifikasi Pengidap Diabetes Pada Perempuan Menggunakan Penggabungan Metode Support Vector Machine Dan K-Nearest Neighbour,” 2016. Accessed: Sep. 13, 2022. [Online]. Available: Https://Jurnal.Unej.Ac.Id/Index.Php/Informal/Article/View/2719/2515

[14] R. Manimaran And M. Vanitha, "Novel Approach To Prediction Of Diabetes Using Classification Mining Algorithm," *International Journal Of Innovative Research In Science, Engineering And Technology (An Iso*, Vol. 3297, 2007, Doi: 10.15680/Ijirset.2017.0607266.

[15] D. A. Agatsa, R. Rismala, And U. N. Wisesty, “Klasifikasi Pasien Pengidap Diabetes Menggunakan Metode Support Vector Machine,” 2020.

[16] M. M. F. Islam, R. Ferdousi, S. Rahman, And H. Y. Bushra, "Likelihood Prediction Of Diabetes At Early Stage Using Data Mining Techniques," In *Advances In Intelligent Systems And Computing*, 2020, Vol. 992, Pp. 113–125. Doi: 10.1007/978-981-13-8798-2_12.

[17] "Early Stage Diabetes Risk Prediction Dataset | Ieee Dataport." Https://Ieee-Dataport.Org/Documents/Early-Stage-Diabetes-Risk-Prediction-Dataset (Accessed Aug. 10, 2022).

[18] B. Sarojini Ilango, "A Hybrid Prediction Model With F-Score Feature Selection For Type Ii Diabetes Databases," 2010. Accessed: Sep. 13, 2022. [Online]. Available: Http://Dx.Doi.Org/10.1145/1858378.1858391

[19] N. Barraza, S. Moro, M. Ferreyra, And A. De La Peña, "Mutual Information And Sensitivity Analysis For Feature Selection In Customer Targeting: A Comparative Study," *J Inf Sci*, Vol. 45, No. 1, Pp. 53–67, Feb. 2019, Doi: 10.1177/0165551518770967.

[20] C. Strobl, A. L. Boulesteix, A. Zeileis, And T. Hothorn, "Bias In Random Forest Variable Importance Measures: Illustrations, Sources And A Solution," *Bmc Bioinformatics*, Vol. 8, 2007, Doi: 10.1186/1471-2105-8-25.

[21] T. N. Joshi And P. M. Chawan, "Diabetes Prediction Using Machine Learning Techniques," *Computer Engg. And Info. Tech., V.J.T.I*, Vol. 8, Pp. 2248–9622, 2018, Doi: 10.9790/9622-0801020913.

[22] Rakesh S Raj, Sanjay D S, Dr. Kusuma M, And Dr. S. Sampath, *Comparison Of Support Vector Machine And Naive Bayes Classifiers For Predicting Diabetes*. 2019. Accessed: Sep. 13, 2022. [Online]. Available: Https://Doi.Org/10.1109/Icatiece45860.2019.9063792

[23] D. Yuni Utami, E. Nurlelah, And F. Nur Hasan, "Jite (Journal Of Informatics And Telecommunication Engineering) Comparison Of Neural Network Algorithms, Naive Bayes And Logistic Regression To Find The Highest Accuracy In Diabetes," *Jite*, Vol. 5, No. 1, 2021, Doi: 10.31289/Jite.V5i1.5201.

[24] Anita Ahmad Kasim, Muhammad Sudarsono, And M. Sudarsono, “Algoritma Svm Untuk Klasifikasi Ekonomi Penduduk Penerima Bantuan Pemerintah Di Kecamatan Simpang Raya Sulawesi Tengah,” 2019.

[25] N. G. Ramadhan, "Comparative Analysis Of Adasyn-Svm And Smote-Svm Methods On The Detection Of Type 2 Diabetes Mellitus," *Scientific Journal Of Informatics*, Vol. 8, No. 2, Pp. 276–282, Nov. 2021, Doi: 10.15294/Sji.V8i2.32484.

[26] S. Mirza, S. Mittal, And M. Zaman, "Applying Decision Tree For Prognosis Of Diabetes Mellitus," *International Journal Of*

Applied Research On Information Technology And Computing, Vol. 9, No. 1, P. 15, 2018, Doi: 10.5958/0975-8089.2018.00002.7.

[27] Gde Agung Brahmana Suryanegara, Adiwijaya, And Mahendra Dwifebri Purbolaksono, “Peningkatan Hasil Klasifikasi Pada Algoritma Random Forest Untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi,” *Jurnal Resti (Rekayasa Sistem Dan Teknologi Informasi)*, Vol. 5, No. 1, Pp. 114–122, Feb. 2021, Doi: 10.29207/Resti.V5i1.2880.

[28] I. Listiowarni And E. R. Setyaningsih, “Feature Selection Chi-Square Dan K-Nn Pada Pengkategorian Soal Ujian Berdasarkan Cognitive Domain Taksonomi Bloom,” 2018. [Online]. Available: Http://Jurnal.Pcr.Ac.Id

[29] A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja'afari, And S. Abdulwahed, "Diabetes Classification Based On Knn," *Iium Engineering Journal*, Vol. 21, No. 1, Pp. 175–181, 2020, Doi: 10.31436/Iiumej.V21i1.1206.

[30] A. P. Ayudhitama And U. Pujianto, “Analisa 4 Algoritma Dalam Klasifikasi Penyakit Liver Menggunakan Rapidminer”, Accessed: Sep. 13, 2022. [Online]. Available: Https://Doi.Org/10.33795/Jip.V6i2.274

[31] M. Pradhan And G. R. Bamnote, "Design Of Classifier For Detection Of Diabetes Mellitus Using Genetic Programming," In *Proceedings Of The 3rd International Conference On Frontiers Of Intelligent Computing: Theory And Applications (Ficta) 2014*, 2015, Pp. 763–770.

[32] S. Y. Rubaiat, M. M. Rahman, And M. K. Hasan, "Important Feature Selection & Accuracy Comparisons Of Different Machine Learning Models For Early Diabetes Detection," In *2018 International Conference On Innovation In Engineering And Technology (Iciet)*, 2018, Pp. 1–6. Doi: 10.1109/Ciet.2018.8660831.

[33] R. Saxena, S. K. Sharma, M. Gupta, And G. C. Sampada, "A Novel Approach For Feature Selection And Classification Of Diabetes Mellitus: Machine Learning Methods," *Comput Intell Neurosci*, Vol. 2022, P. 3820360, 2022, Doi: 10.1155/2022/3820360.

[34] K. C. Tan, E. J. Teoh, Q. Yu, And K. C. Goh, "A Hybrid Evolutionary Algorithm For Attribute Selection In Data Mining," *Expert Syst Appl*, Vol. 36, No. 4, Pp. 8616–8630, 2009, Doi: Https://Doi.Org/10.1016/J.Eswa.2008.10.013.

[35] L. Widya Astuti, I. Saluza, And E. Yulianti, “Feature Selection Menggunakan Binary Wheal Optimizaton Algorithm (Bwoa) Pada Klasifikasi Penyakit Diabetes,” *Jurnal Ilmiah Informatika Global*, Vol. 13, No. 1, 2022, Doi: 10.36982/Jiig.V13i1.2057.

[36] S. Rahayu And Stik. Jayakarta Pkp Dki Jakarta, “Hubungan Usia, Jenis Kelamin Dan Indeks Massa Tubuh Dengan Kadar Gula Darah Puasa Pada Pasien Diabetes Melitus Tipe 2 Di Klinik Pratama Rawat Jalan Proklamasi, Depok, Jawa Barat,” 2020. Accessed: Sep. 13, 2022. [Online]. Available: Https://Doi.Org/10.34035/Jk.V11i1.412

[37] A. Dewi *Et Al.*, “Pengaruh Minyak Kelapa Terhadap Penurunan Rasa Gatal Pada Pasien Diabetes Mellitus Di Rsud Kota Slatiga.”

[38] I. Wayan, A. Putra, And K. N. Berawi, “Empat Pilar Penatalaksanaan Pasien Diabetes Mellitus Tipe 2,” 2015.

[39] E. Setiyorini And N. A. Wulandari, “Hubungan Status Nutrisi Dengan Kualitas Hidup Pada Lansia Penderita Diabetes Mellitus Tipe 2 Yang Berobat Di Poli Penyakit Dalam Rsd Mardi Waluyo Blitar,” *Jurnal Ners Dan Kebidanan (Journal Of Ners And Midwifery)*, Vol. 4, No. 2, Pp. 125–133, Oct. 2017, Doi: 10.26699/Jnk.V4i2.Art.P125-133.