International Journal of Innovative
Computing, Information and Control
Volume **18**, Number **4**, August **2022**

ICIC International ©2022 ISSN 1349-4198
pp. 1117–1132

ENHANCE WEAK LEARNER MODEL OF ADABOOST (EWDM) FOR DIABETES MELLITUS CLASSIFICATION

Ployphan Sornsuwit

Faculty of Management Science, Digital Business Technology Program
Kamphaeng-Phet Rajabhat University
69, Nakhorn-chum District, Muang, Kamphaeng-Phet 62000, Thailand
ployphan@kpru.ac.th

Received January 2022; revised May 2022

Abstract. *Diabetes is a disorder in which the body is unable to use blood sugar as energy normally. There are also many factors that can cause diabetes such as age, weight, and blood sugar levels. Detection of disease is not easy, and prognosis is also expensive. Our research has therefore developed the Enhance Weak Learner Model of AdaBoost (EWDM), which is an improvement of original AdaBoost, EWDM uses a correlation-based feature selection, reduces the number of features to only those that are related to each other to increase the efficiency of classification and also reduce the burden of processing, and builds a model of weak learner using three algorithms: k-nearest neighbor, Naïve Bayes and support vector machine by which EWDM will improve the process of weak learner modeling that will create the best hypothesis for each iteration while managing to remove as many errors as possible before creating the best final hypothesis. Our experiments used the Pima Indian Diabetes dataset. By analyzing the performance compared to supervised learning and ensemble learning, the results showed that EWDM outperformed all comparative methods both supervised learning and ensemble learning, with an accuracy of 88.26%. In addition, compared to other diabetes datasets, EWDM was the highest effective model as well compared to other methods, The Early Stage diabetes risk prediction dataset shows accuracy up to 100% and Type 2 diabetes dataset as high as 96.84%. The obtained results indicate EWDM is well suited to classifying diabetes in both blood results and health data.*

Keywords: CFS, Diabetes mellitus, Ensemble learning, EWDM

1. **Introduction.** Diabetes is a disease in which the body's cells malfunction in the process of converting blood sugar into energy. When sugar is not used, blood sugar levels rise above abnormal levels. The condition can be found commonly with people all over the world. It is a silent threat that is the main cause of other diseases such as blindness, renal failure, heart attack, stroke and lower limb amputation [1]. According to International Diabetes Federation, approximately 537 million adults (20-79 years) were living with diabetes in 2019. It is estimated that by 2002 this figure will be increased to 783 million. It also costs at least $966 billion in health costs, and the 541 million adult population is at increased risk of developing Type 2 diabetes (Diabetes Type 2) [2].

There are three types of diabetes. Diabetes Type 1, also known as Insulin Dependent Diabetes Mellitus (IDDM), is a type of diabetes that occurs when cells in the pancreas are damaged and unable to produce insulin. Therefore, there is a problem of lack of insulin secretion by beta cells of the pancreas. Diabetes Type 2, also called Non-Insulin Dependent Diabetes Mellitus (NIDDM), is a type of diabetes caused by the body's poor use of insulin and unable to maintain normal blood sugar levels. And Diabetes Type 3, is

DOI: 10.24507/ijicic.18.04.1117

a type of diabetes that occurs with elevated blood sugar levels while women are pregnant [3].

Currently, there have developed a variety of machine learning techniques including in the application of advanced technology in daily life [4,5] and health to assist in the diagnosis and detection of diabetes with accuracy, for example, used in the diagnosis of diabetic retinopathy, clinical diagnosis support or predictive population risk stratification [6]. Each method uses a variety of machine learning models and increases the efficiency of forecasting to be more accurate. In particular, the incidence of Type 2 diabetes, a silent disease, has been increasing year by year. If there is a tool that can analyze symptoms quickly and accurately, it will help reduce the number of people with chronic diabetes. It also prevents the occurrence of other diseases in the future as well. There is a lot of research to develop an efficient algorithm to do this classification which will have different accuracy but analyzing a large number of feature patients can affect the calculation accuracy. It also requires a lot of processing as well in this paper.

Therefore, the objective is to develop and improve the AdaBoost algorithm for diabetes to classify between diabetic and non-diabetic patients. By our method, the number of features is reduced to only the features that are related to each other. And the new algorithm will focus on improving the process of weak learner modeling in any iteration to have the best hypothesis. By using this algorithm, it will be assured by the information that it classifies correctly and focuses on correcting any misdiagnosed data in each iteration in order to create the most accurate final hypothesis. In this experiment, Pima Indian Diabetes dataset was used. The results were then compared with other machine learning and ensemble learning efficacy results, analyzed for efficacy with other diabetes datasets, and finally compared with other previous studies.

The rest of the paper is structured as follows: the second section discusses the related work, the third section addresses material and methods, the fourth section is about experiments and results, the fifth section is a discussion and the sixth section implies the conclusions and suggestions for future work.

2. **Related Work.** The machine learning technique has become a useful tool for analyzing diabetes in medicine which has done a lot of research. It has experimented with different patient datasets, and has continued to use a variety of algorithms over the years [7]. Each method has a different technique and algorithm in order to achieve the most efficient method and suitable for further development as a medical device in the future such as breast cancer [8], heart disease [9], as well as diabetes [10-14] which will be useful in prediction, screening, diagnosis, treatment and other recommendations for treatment.

Predict diabetes with machine learning as a new diagnostic tool including supervised learning [10], ensemble learning [11,15-17], deep learning [18], etc. The goal of each research is to increase the efficiency of classification with a more efficient method, aiming to develop new algorithms to increase the accuracy of the previous methods and is suitable for the dataset to be tested as well. In order to improve the efficiency, ensemble algorithm is one of the highly efficient methods to classify diabetes datasets, as Kumari et al. [11] presented an experimental ensemble soft voting classifier binary classification and uses of the ensemble of three machine learning algorithms. The proposed method was found to have high accuracy for both the Pima Indian Diabetes Dataset (PIDD) and Breast Cancer dataset, with PIDD having an accuracy of 79.08%, similar to [15-17]. An ensemble algorithm is used to classify diabetes such as bagging, and boosting hybrid classifiers although it is more efficient than other comparative approaches. However, most of them have an accuracy that is less than 90%.

In addition, disease datasets often contain incomplete data which may be one of the reasons that when doing classification there are not very high efficiency values such as missing data, class imbalance, or features that may not be related. Especially diabetes datasets, which are often researched to improve efficiency in doing classification in preprocessing to solve the problems mentioned above, for example, Nomura et al. [6] proposed a novel method which is PMSGD for classification of diabetes mellitus using genetic algorithm and decision tree, majorly due to the existence of class imbalance and missing values in the data. CFSGA was also used in feature selection, highlighting how to deal with incompleteness. A complete PIDD dataset, like [8-10], features selections so that only related features are selected for better forecasting [10]. We used Correlation-based Feature Selection (CFS) and PCA to compare the efficiency of classification using machine learning algorithms as decision tree classifier, random forest, k-nearest neighbor, AdaBoost classifier, J48graft classifier and logistic regression. Comparing the efficiency of classification, it was found that CFS improved performance across all the algorithms compared and increased efficiency higher than PCA in every algorithm as well.

From the research mentioned above, it was found that developing and improving an ensemble algorithm that is highly effective for predicting diabetes is important today. It is also necessary to deal with the data in the process. Preprocess the diabetes dataset so that it can improve efficiency better than previous methods. This will be a guideline for further development as a model for the healthcare system in the future.

3. **Material and Methods.** The algorithm we propose will be applied according to the data mining technique with the overall structure as shown in Figure 1.

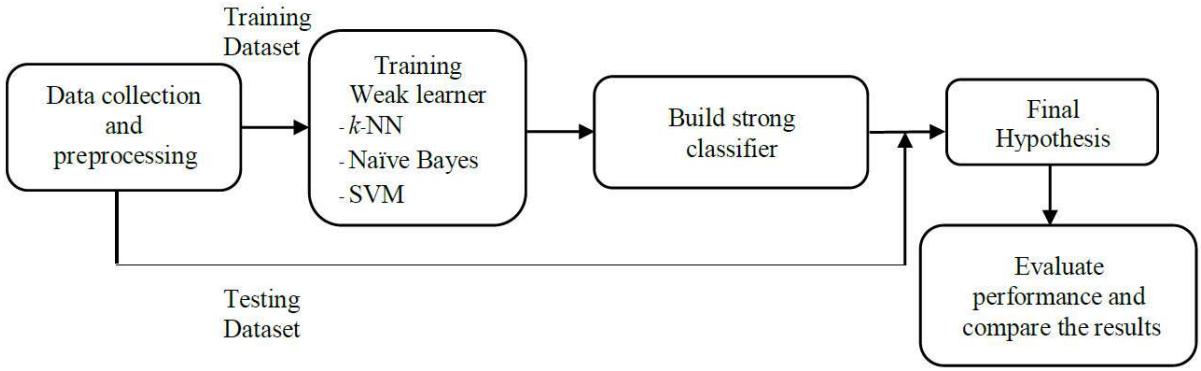


FIGURE 1. This proposed model

From Figure 1, our proposed model has a process of making classification the first step which is to do data collection and preprocessing to collect diabetes data and do preprocessing to divide the data into the training dataset and testing dataset to be appropriate and then use only the training dataset. In the training weak learner phase, the EWDM we developed is an improvement over the original training weak learner phase and leads to the next step in building the most efficient final hypothesis. It will test with the testing dataset and then measure and evaluate performance in the final step. Show details of various steps as in Sections 3.1-3.4.

3.1. Data collection and preprocessing.

3.1.1. *Data collection.* Our research used the Pima Indian Diabetes Dataset (PIDD) [19]. This dataset was developed by the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of the dataset is to predict a diagnosis of diabetes. The

TABLE 1. Detailed description of PIDD features

No.	Name of feature	Meaning	Average mean	sd.	min/max
1	Pregnancies	Number of times pregnant	3.85	3.37	0/17
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	120.89	31.97	0/199
3	Blood pressure	Diastolic blood pressure (mm Hg)	69.11	19.36	0/122.00
4	Skin thickness	Triceps skin fold thickness (mm)	20.54	15.95	0/99
5	Insulin	2-hour serum insulin (mu Uml)	79.8	115.24	0/846.00
6	Age	Age (years)	33.24	11.76	21/81
7	BMI	Body mass index (weight in kg/(height in m)^2)	31.99	7.88	0/67.1
8	Diabetes pedigree function	Diabetes pedigree function	0.47	0.33	0.078/2.42
9	Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0	–	–	–

datasets consist of several medical predictor variables and one target variable, including 8 features: pregnancy times, glucose, blood pressure, skin thickness, insulin, BMI, diabetic pedigree function and age. This dataset has 768 total number of instances including 768 instances divided into 268 non-diabetic instances and 500 diabetic instances. Details and characteristic of the dataset are shown in Table 1.

3.1.2. *Preprocessing.* In our experiment, we perform checking and handling missing values by filling any missing data using previous data in the same field, then the dataset is split into 70% training and 30% testing because these proportions are sufficient for effective learning and a suitable one for training the ML models for the use of PIDD diabetes dataset [20-22], so the number of training datasets is 538 records and the number of testing datasets is 230 records.

In the feature selection, we used Correlation-based Feature Selection (CFS) [23] to select correlated features. CFS is widely used in machine learning applications in biomedical engineering and others, which is a recognized effective method [24-28]. The equation for CFS is given below.

$$Merit_s = \frac{k\bar{r}_{cf}}{\sqrt{k} + k(k-1)\bar{r}_{ff}} \tag{1}$$

$Merit_s$ is the correlation between the summed components and the outside variable. k is the number of components. $\bar{r}_{cf}$ is the mean feature-class correlation. $\bar{r}_{ff}$ is the average inter-correlation between components.

3.2. **Train weak learner.**

3.2.1. *Classification technique.* AdaBoost algorithm or adaptive boost algorithm [29], which is an effective ensemble method algorithm, has high accuracy in making classification and it is popularly applied to a wide variety of data formats. Both in the biomedical dataset and many other fields including information about the current COVID-19 disease that is still spreading [30,31].

AdaBoost generates a strong learner by iteratively adding weak learners, weak learners can be used with any machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), and Multilayer Perceptron (MLP), where several models from weak learner are created and the hypothesis is obtained from all models through a combination process, then it predicts the best outcomes as the final hypothesis, as shown of traditional AdaBoost architecture in Figure 2.

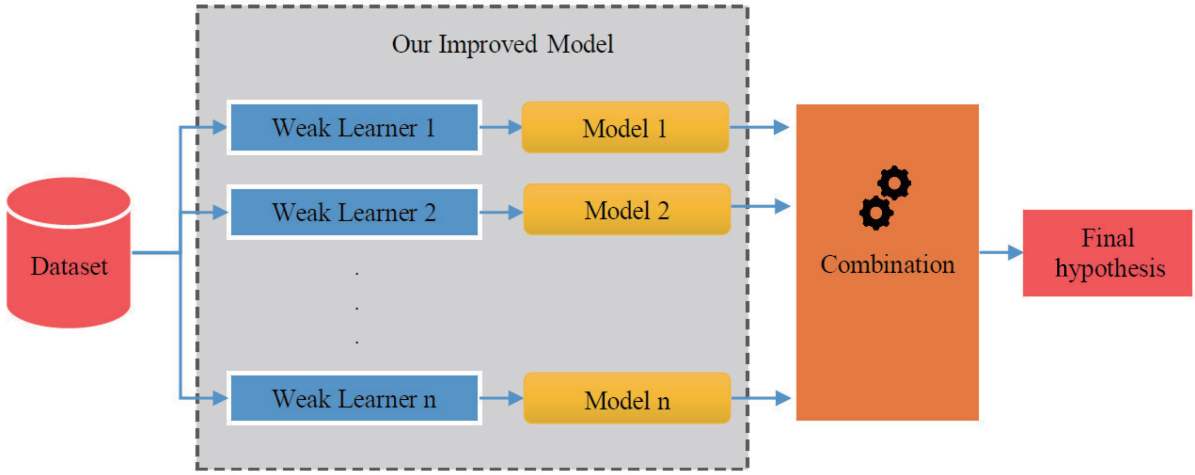


FIGURE 2. AdaBoost architecture

From Figure 2, our research has improved AdaBoost.m1, an algorithm developed from the original AdaBoost concerning its ability to reduce the training error, in the modeling process for each weak learner. Weak learners used include k-NN, Naïve Bayes and SVM.

The principle key that our research focuses on is that if we have a weak learner that performs well in classification, and can develop a learning algorithm to get the best model out of every weak learner, then we can combine models when obtaining that can predict the most accurate answer. The process of improving models and combination of EWDM is shown in detail in Section 3.2.2.

3.2.2. *Weak learner.* The method for improving the modeling process shows the structure of the improved method as shown in Figure 3.

In Figure 3, given 10 iterations, training with a weak learner starts with the first weak learner k-NN. Once the model is created and the hypothesis is obtained from k-NN, the algorithm searches for the correct sample. In any array and sample that is misdiagnosed it is sent to the second weak learner, Naïve Bayes. Once a model is created and a hypothesis is obtained from Naïve Bayes, it looks for additional correct samples in the original array and any samples that were misdiagnosed will be sent to the 3rd weak learner. That is, the SVM generates the model iterates and finds correct and incorrect samples. Then it is sent to store in the original array as an answer to the last guess, thus getting the hypothesis from iteration 1.

Iterate for 10 iterations, and then vote for the final hypothesis, which we designed to achieve the most accurate final hypothesis. This is because we can create the best hypothesis for each iteration as if managing to remove as many errors as possible before creating the final hypothesis in the final step.

k-Nearest Neighbor (k-NN) methods. The k-NN algorithm, is a machine learning algorithm which is simple and widely used. The k-NN algorithm is mainly based on the distance calculation in which the calculation uses the principle of comparing the data of interest with other data. How similar or distant are they? If the information you are

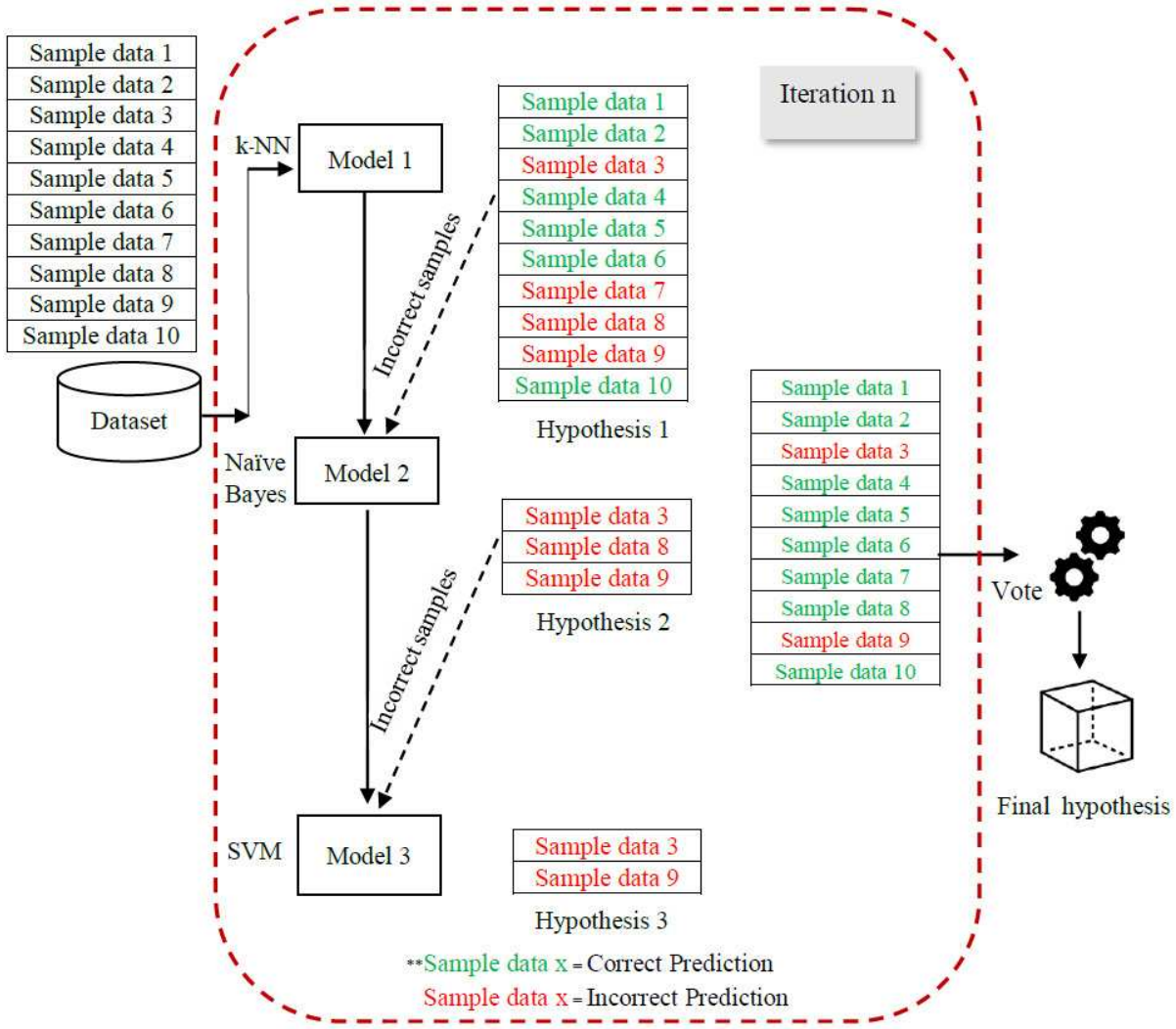


FIGURE 3. Our improved model architecture

interested in is closest to the information, the system will give an answer as an answer to the closest information with the most popular method for calculating within the Euclidian distance. Therefore, it is necessary to apply the KNN algorithm in numerical datasets [32].

The main issue is how to measure the similarity between records. The most popular measurement of similarity is the Euclidean distance between two records $(x_1, x_2, \ldots, x_p)$ and $(y_1, y_2, \ldots, y_p)$ defined by the following equation [33].

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

Naïve Bayes. Naïve Bayes is a method based on the Bayes theorem principle, which is predicted for independent speculation. Because it is elegantly simple and robust, it is widely used for classifying purposes.

It uses the probability theory to classify data. Naïve Bayes helped develop models that provide predictive capabilities. Features of all data should be relatively independent. It is easy to model Naïve Bayes and does not have complex refutation parameters [34]. To classify the unseen data, Naïve Bayes computes the posterior probability for each class C_i as follows [35].

$$P(C_i|X) = \frac{P(C_i) \prod_{k=1}^{n} P(X_k|C_i)}{P(X)}$$

where C_i is the class label and X is an instance to be classified. After calculating the posterior probability for each class, it assigns the class label that has the highest probability to the unseen data.

Support Vector Machine (SVM). SVM is considered to be one of the most powerful statistical learning techniques. SVM is one of the algorithms that can learn to find the highest accuracy from classification and is widely used in many applications such as handwriting digits, character and text recognition, anomaly detection and more recently to satellite image classification. The essence of SVM generates a straight line that divides the data (Hyperplane) [34] and finds the best line to classify it as a positive class or negative class.

For all x that is a member of class $+1$, they satisfy the following constraints [36]:

$$w^T x + b \geq +1$$

For all x that is a member of class -1, they satisfy the following constraints:

$$w^T x + b \leq -1$$

We want to find the optimal hyperplane $w^T x + b = 0$ that maximizes the margin of the two conditions above. After finding the optimal hyperplane, the decision function can be defined as

$$f(x) = sign\left(w^T x + b\right)$$

3.3. **Build strong classifiers and final hypothesis.** In the build step of the strong classifier, a combination of hypothesis in each learning cycle is used to get the best model, in which case each iteration has an error value greater than 0.5 and must iterate to training again until the error value is less than 0.5, if all values are less than 0.5 then the error value is calculated α and then updated with the distribution for use in training to achieve the most accurate final hypothesis. Our proposed algorithm is shown in Algorithm 1.

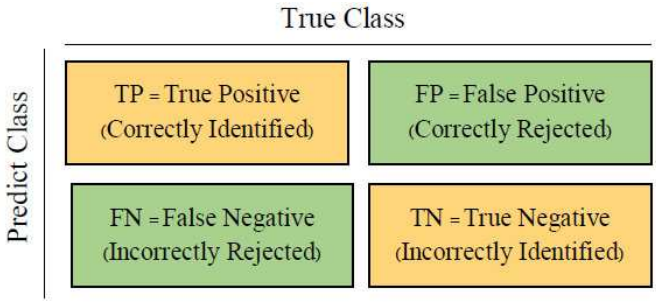
Algorithm 1

Input: sequence of example $((x_1, y_1), \ldots, (x_m, y_m))$ with labels $y_i \in Y = \{1, \ldots, k\}$
Weak learning algorithm **Weaklearn**
Integer T specifying number of iterations
Integer J specifying number of weak learners
Integer G specifying incorrect sample
Initialize $D_1(i) = 1/m$ for all i
Do for $t = 1, 2, \ldots, T$
1. Call **Weaklearn** providing it with the distribution D_t
2. Get back hypothesis $h_t = X \to Y$
if $\varepsilon_j \neq 0$, get $G(m_i)$
Do for $j = 1, 2, \ldots, J$
then call $weaklearn(j + 1)$
3. Calculate the error of h_t: $\varepsilon_t = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$ if $\varepsilon_t > 1/2$ then set $T = T - 1$ and abort loop
4. Set $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$
5. Update distribution D_t: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t & \text{if } h_t(x_i) = y_i \\ 1 & \text{otherwise} \end{cases}$
where Z_t is a normalization constant (Chosen so that D_{t+1} will be a distribution)
Output the final hypothesis: $h_{fin} = \arg\max_{y \in Y} \sum_{t:h_t(x)=y} \log \frac{1}{\beta_t}$

From Algorithm 1, in each iteration, EWDM calculates ε_j from each weak learner. It selects only incorrect sample $G(m_i)$ and then passes it on to the weak learner. Complete the number of weak learners, and then calculate the error ε_t from the h_t of that iteration; thus each hypothesis has the lowest error value, which leads to an update distribution and a vote up to the best final hypothesis.

3.4. **Evaluate performance and compare the results.** Our research uses the confusion matrix to analyze the efficiency of classification, which has 4 values for analysis, as shown in Table 2.

TABLE 2. Confusion matrix



Performance analysis includes accuracy, precision, recall and f-Measure. The calculations are as follows.

Precision is the ratio of correctly predicted positive samples to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall or Sensitivity is the ratio of correctly predicted positive sample to the total predicted in actual class.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

f-Measure is the weighted average of Precision and Recall.

$$f\text{-}Measure = \frac{2 \times (Precision + Recall)}{Precision + Recall} \tag{4}$$

Accuracy is ratio of the correctly predict label to the total sample label.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

4. **Experimental Results.** Our research used a personal computer Intel core, i5-4258U CPU @2.4GHz, 8 GB memory without GPU acceleration and our algorithm was implemented in MATLAB R2017. The results of the experiment are as follows.

4.1. **Feature selection.** In the feature selection process with CFS, features were selected out of a total of 8 features, leaving only 4 features as shown in Table 3.

TABLE 3. The selected features

Feature No.	Feature name
1	Pregnancies
2	Glucose
6	Insulin
8	BMI

In the feature selection, the feature count can be reduced from 8 features to 4, which is half the total feature number.

4.2. **Performance evaluation.** In the experimental results, the comparative efficiency is done by using efficiency: accuracy, precision, recall and f-Measure displaying the analyzed confusion matrix table. The feature selection is not performed as shown in Table 4, and the case where the feature selection is performed is shown in Table 5.

Table 4. The confusion matrix method in case the feature selection is not performed

Predict class	Actual class		Accuracy
	Diabetic instance	Non-diabetic instances	
Diabetic instance	**80**	33	85.65%
Non-diabetic	0	**117**	

Table 5. The confusion matrix method in case the feature selection is performed

Predict class	Actual class		Accuracy
	Diabetic instance	Non-diabetic instances	
Diabetic instance	**80**	27	88.26%
Non-diabetic	0	**123**	

From Table 4 and Table 5, after the feature selection, the TP and FP values have a higher number of valid classifications, where the number of features is only half of the total number of features. When analyzing the efficiency values, it was found with higher efficiency before feature selection at 85.65% accuracy.

After feature selection, accuracy was 88.26%. Performance was also compared with other methods, where performance analysis was performed with supervised learning as shown in Table 6 and the results were compared with ensemble learning as shown in Table 7.

Table 6. Comparison of the effectiveness of the proposed method with supervised learning method

Method	Class	Precision (%)	Sensitivity (%)	Specificity (%)	f-Measure (%)	Accuracy (%)
k-NN	Diabetic	61.25	59.76	79.05	60.50	72.17
	Non-diabetic	78.00	79.05	59.76	78.52	
Naïve Bayes	Diabetic	62.50	67.57	80.77	64.94	76.52
	Non-diabetic	84.00	80.77	67.57	82.35	
Decision Tree	Diabetic	68.75	63.95	82.64	66.26	75.65
	Non-diabetic	79.33	82.64	63.95	80.95	
SVM	Diabetic	73.75	68.60	85.42	71.08	79.13
	Non-diabetic	82.00	85.42	68.60	83.68	
MLP	Diabetic	69.57	20.00	95.33	31.07	69.13
	Non-diabetic	69.08	95.33	20.00	80.11	
Our proposed without feature selection	Diabetic	78.70	100.00	78.00	88.08	85.65
	Non-diabetic	100.00	78.00	100.00	87.64	
Our proposed with feature selection	**Diabetic**	**74.77**	**100.00**	**82.00**	**85.56**	**88.26**
	Non-diabetic	**100.00**	**82.00**	**100.00**	**90.11**	

TABLE 7. Comparison of the effectiveness of the proposed method with the ensemble learning method

Method	Class	Precision (%)	Sensitivity (%)	Specificity (%)	f-Measure (%)	Accuracy (%)
AdaBoost	Diabetic	100.00	0.00	78.00	0.00	78.00
	Non-diabetic	0.00	78.00	0.00	0.00	
Logitboost	Diabetic	0.00	88.00	0.00	0.00	88.00
	Non-diabetic	100.00	0.00	88.00	0.00	
Bagging	Diabetic	0.00	86.67	0.00	0.00	86.67
	Non-diabetic	100.00	0.00	86.67	0.00	
Our proposed without feature selection	Diabetic	78.70	100.00	78.00	88.08	85.65
	Non-diabetic	100.00	78.00	100.00	87.64	
Our proposed with feature selection	**Diabetic**	**74.77**	**100.00**	**82.00**	**85.56**	**88.26**
	Non-diabetic	**100.00**	**82.00**	**100.00**	**90.11**	

The performance analysis results shown in Tables 6 and 7 showed that EWDM had higher f-Measure and accuracy than all other methods of comparison, especially precision and specificity of 100% in the non-diabetic class whole experiment in the part that does feature selection and without feature selection.

Compared to supervised learning as shown in Table 6, EWDM is a combination of a powerful weak learner and then voted the best final answer, thus having a markedly higher accuracy of 85.65% since classification without feature selection. And when doing feature selection, it has an accuracy of 88.26%, which is the highest compared to all methods and uses only 4 features to calculate, reducing computing resources and getting the highest efficiency as well. And in comparison with the ensemble learning method, it was found that other methods compared even with similar efficiency but the EWDM has improved the process of choosing a model from the weak learner with the lowest error to create the final hypothesis, a new method that has not yet been ensemble learning, so EWDM has the highest accuracy of all. When bringing our research to comparisons with other diabetes datasets, The Early Stage diabetes risk prediction dataset [37] and Type 2 diabetes [38,39], which collect diabetes data, were created by the Department of Computer Science and Engineering, BIT Mesra, Ranchi-835215.

The results show a comparison graph of f-Measure and accuracy. The analysis shows The Early Stage diabetes risk prediction dataset as shown in Figures 4 and 5, and the analysis results of the Type 2 diabetes dataset as shown in Figures 6 and 7.

From Figure 4, it is found that the method we propose has an accuracy of up to 100% where other methods are presented. The descending value is that the decision tree was 94.87%, Naïve Bayes was 92.31%, SVM was 89.74%, k-NN was 65.38%, and MLP was 60.26%, respectively.

From Figure 5, it is found that the method we propose brings f-Measure values up to 100% in both classes, while the other methods presented have descending f-Measure values of the non-diabetics class and the diabetics class as follows: decision tree 95.75% and 93.55%, Naïve Bayes 93.82% and 89.83%, SVM 90.91% and 88.24%, k-NN 78.05% and 18.18%, and finally MLP 75.20% and 0%, respectively in which MLP has the lowest efficiency especially classifying diabetics class.

Based on the efficiency of the classification The Early Stage diabetes risk prediction dataset, the EWDM was astonishingly good at recognizing diabetes without error. Other studies have been able to classify accuracy to 100% with ensemble learning as well [40], although that study uses all features in its experiments, but EWDM still uses fewer features that provide the highest accuracy.

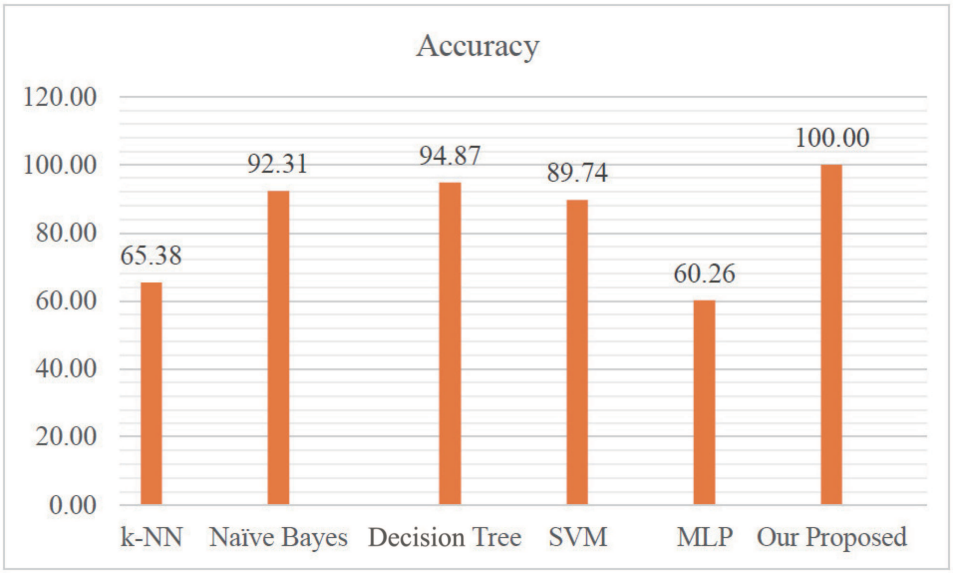


FIGURE 4. Comparison of accuracy of The Early Stage diabetes risk prediction dataset

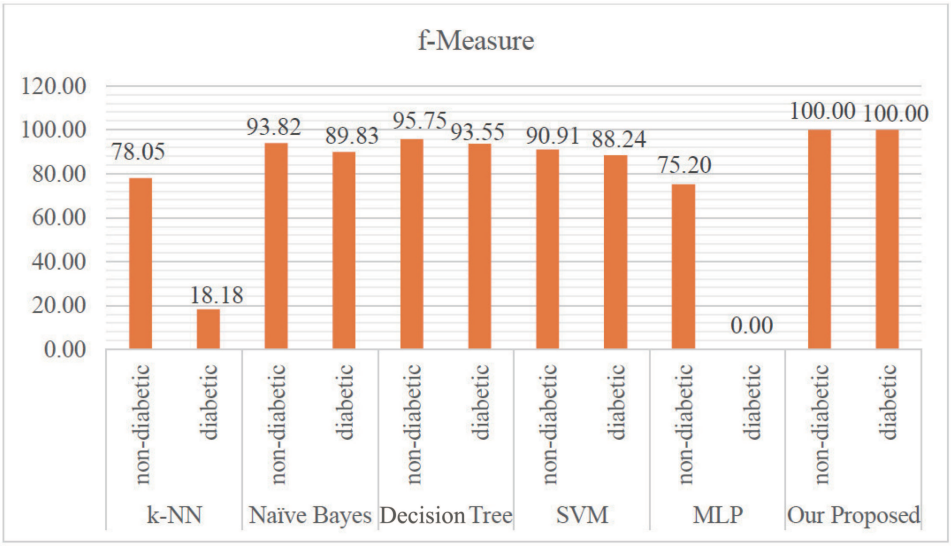


FIGURE 5. Comparison f-Measure of The Early Stage diabetes risk prediction dataset

From Figure 6, it is found that the method we offer has an accuracy of up to 96.84% compared to the other methods which are presented descending of decision tree and k-NN was equal at 91.58%, Naïve Bayes was 89.47%, SVM was 87.37% and MLP was 71.23%, respectively.

From Figure 7, our proposed methods resulted in f-Measure values for both non-diabetics class and diabetics class as high as 97.85% and 94.04%, with the other methods having descending f-Measure values. The decision tree was 94.26% and 84.21%, k-NN 94.47% and 82.35%, Naïve Bayes 92.72% and 81.01%, SVM 90.95% and 79.07%, and finally MLP was 80.84% and 42.25%, respectively in which MLP has the lowest efficiency especially classifying of diabetics class as well.

The Type 2 diabetes classification found this dataset to be a new dataset published in 2020 and contains key information elements related to health, lifestyle and family

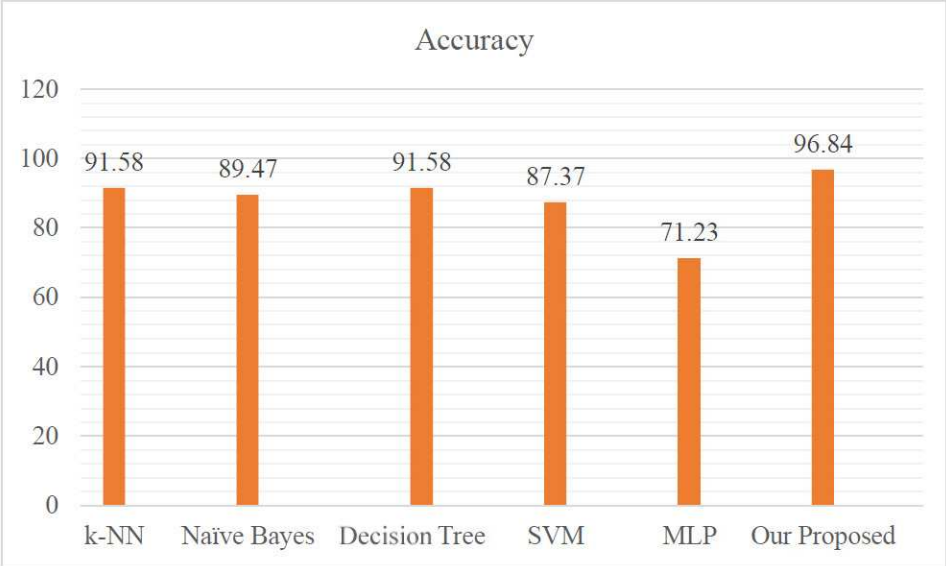


FIGURE 6. Comparison of accuracy of Type 2 diabetes dataset

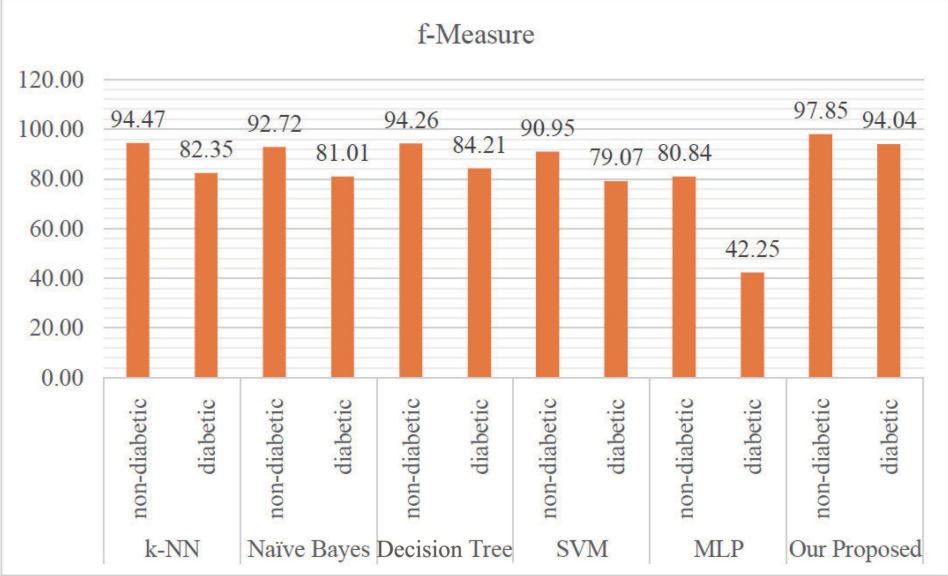


FIGURE 7. Comparison of f-Measure of Type 2 diabetes dataset

background. Based on its efficiency, EWDM is still the most effective compared to other studies and higher than the research the developers of this dataset have compared. It is like EWDM is an aid in the analysis of health data without the need for blood results which has good accuracy in classifying diabetic patients as well.

In addition, we have compared the efficiency of classification in the methods we have presented with other studies using the same dataset which is shown in Table 8.

From Table 8, our proposed method has higher accuracy than other methods compared especially, for machine learning, ensemble learning and for some deep learning as well.

5. **Discussion.** From the experimental results, the method we proposed is more efficient than other methods because we reduce the number of features to only 4 features. These include Pregnancies, Glucose, Insulin and BMI, which are related and important features

TABLE 8. Comparison of the effectiveness of our proposed method and other studies

Reference	Method	Year	Accuracy (%)
[12]	Re-RX with J48graft	2016	83.83
[13]	PCA and PSO	2019	79.57
[40]	CNN	2019	76.81
[41]	Backward Elimination and SVM	2021	85.71
[10]	PMSGD	2021	82.13
[42]	Neural Network based on Particle Swarm Optimization	2021	81.25
[11]	Eensemble combined soft voting classifier	2021	79.04
[43]	SVM	2021	87.01
[44]	Enhanced Naïve Bayes with mice	2021	83.33
[15]	Hybrid classifier of Random Forest – Bayes Net	2021	83.91
Our proposed	**EWDM without feature selection**	**2021**	**85.65**
Our proposed	**EWDM**	**2021**	**88.26**

in the diagnosis of diabetes [32]. And the process of creating hypothesis from weak learners is effective in modeling each iteration with low error. This will make the final hypothesis with the lowest error even more possible. Looking at Figure 3, it can be seen that it is different from the general AdaBoost. If AdaBoost uses an inefficient weak learner, this will ultimately lead the algorithm to help each other to vote the answer, the final hypothesis with no high error accordingly. However, the algorithm we proposed will assure you of any data that has the correct answer immediately, without re-learning, where only the wrong answer is sent, learning to the last weak learner in each iteration, just like giving the weak learners to help each other to find the correct answer only for each iteration.

6. **Conclusion.** In this paper, we developed EWDM where the main concern reduces the number of features with CFS to only correlated features and improved the process of modeling any iteration to have the best final hypothesis. We used weak learners including k-NN, Naïve Bayes and SVM. The experiment was done using the Pima Indian Diabetes dataset. The results showed that EWDM reduced the number of features from 8 features to 4 features. From the results of classification, it was found to be the most effective than other methods compared for both supervised learning and ensemble learning, with an accuracy of 88.26%. And when compared to other diabetes datasets it was also found to be the most efficient, with The Early Stage diabetes risk prediction dataset with 100% accuracy and Type 2 diabetes with 96.84% accuracy. The attained results indicate EWDM is suitable for classifying diabetes in blood results and health data.

In the future, we want to solve the world pandemic crisis by developing real-time AdaBoost with precision for detecting COVID-19 to be effective and practical.

REFERENCES

[1] World Health Organization, *Diabetes*, https://www.who.int/news-room/fact-sheets/detail/diabetes, Accessed on 05 January 2022.

[2] International Diabetes Federation, *Facts & Figures*, https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html, Accessed on 05 January 2022.

[3] O. Ozougwu, The pathogenesis and pathophysiology of Type 1 and Type 2 diabetes mellitus, *J. Physiol. Pathophysiol.*, vol.4, no.4, pp.46-57, 2013.

[4] C. Zhang, S. Chai, L. Cui and B. Zhang, Road condition recognition in self-driving cars based on classification and regression tree, *ICIC Express Letters, Part B: Applications*, vol.10, no.12, pp.1115-1122, 2019.
[5] A. C. B. Monteiro, R. P. França, R. Arthur and Y. Iano, A look at machine learning in the modern age of sustainable future secured smart cities, in *Data-Driven Mining, Learning and Analytics for Secured Smart Cities: Trends and Advances*, C. Chakraborty, J. C.-W. Lin and M. Alazab (eds.), Cham, Springer International Publishing, 2021.
[6] A. Nomura, M. Noguchi, M. Kometani, K. Furukawa and T. Yoneda, Artificial intelligence in current diabetes management and prediction, *Curr. Diab. Rep.*, vol.21, no.12, p.61, 2021.
[7] V. Jaiswal, A. Negi and T. Pal, A review on current advances in machine learning based diabetes prediction, *Primary Care Diabetes*, vol.15, no.3, pp.435-443, 2021.
[8] G. Chugh, S. Kumar and N. Singh, Survey on machine learning and deep learning applications in breast cancer diagnosis, *Cogn. Comput.*, vol.13, no.6, pp.1451-1470, 2021.
[9] S. Verma and A. Gupta, Effective prediction of heart disease using data mining and machine learning: A review, *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp.249-253, 2021.
[10] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh and A. Khamparia, Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus, *Multimedia Systems*, 2021.
[11] S. Kumari, D. Kumar and M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, *International Journal of Cognitive Computing in Engineering*, vol.2, pp.40-46, 2021.
[12] Y. Hayashi and S. Yukita, Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of Type 2 diabetes mellitus in the Pima Indian Dataset, *Informatics in Medicine Unlocked*, vol.2, pp.92-104, 2016.
[13] D. K. Choubey, P. Kumar, S. Tripathi and S. Kumar, Performance evaluation of classification methods with PCA and PSO for diabetes, *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol.9, no.1, 2019.
[14] J. Manjula, S. Radharani, N. H. Rao and Y. Madhulika, An ensemble classification techniques based on 'ML' model for automatic diabetic retinopathy detection, *Turkish Online Journal of Qualitative Inquiry (TOJQI)*, vol.12, no.3, pp.1002-1010, 2021.
[15] N. A. B. S. Noor, I. Elamvazuthi and N. Yahya, Classification of diabetes mellitus using ensemble algorithms, *2020 8th International Conference on Intelligent and Advanced Systems (ICIAS)*, pp.1-6, 2020.
[16] K. Akyol and B. Aen, Diabetes mellitus data classification by cascading of feature selection methods and ensemble learning algorithms, *International Journal of Modern Education and Computer Science*, vol.10, no.6, 2018.
[17] G. T. Reddy et al., An ensemble based machine learning model for diabetic retinopathy classification, *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp.1-6, 2020.
[18] H. Gupta, H. Varshney, T. K. Sharma, N. Pachauri and O. P. Verma, Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction, *Complex Intell. Syst.*, 2021.
[19] UCI Machine Learning, *Pima Indians Diabetes Database*, https://kaggle.com/uciml/pima-indians-diabetes-database, Accessed on 08 January 2022.
[20] R. Patil and S. Tamane, A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes, *International Journal of Electrical & Computer Engineering*, vol.8, no.5, pp.3966-3975, 2018.
[21] U. Ahmed et al., Prediction of diabetes empowered with fused machine learning, *IEEE Access*, vol.10, pp.8529-8538, 2022.
[22] E. K. Hashi, M. S. U. Zaman and M. R. Hasan, An expert clinical decision support system to predict disease using classification techniques, *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp.396-400, 2017.
[23] M. A. Hall, *Correlation-Based Feature Selection for Machine Learning*, Ph.D. Thesis, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, 2000.
[24] M. Fahim, S. Islam, S. Noor, M. Hossain and M. Satu, Machine learning model to analyze telemonitoring dyphosia factors of Parkinson's disease, *International Journal of Advanced Computer Science and Applications*, vol.12, 2021.

[25] S. Larabi-Marie-Sainte, Outlier detection based feature selection exploiting bio-inspired optimization algorithms, *Applied Sciences*, vol.11, no.15, 2021.
[26] A. Maru, A. K. Sharma and M. Patel, Hybrid machine learning classification technique for improve accuracy of heart disease, *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pp.1107-1110, 2021.
[27] Y. Liu, F. Bai, Z. Tang, N. Liu and Q. Liu, Integrative transcriptomic, proteomic, and machine learning approach to identifying feature genes of atrial fibrillation using atrial samples from patients with valvular heart disease, *BMC Cardiovasc. Disord.*, vol.21, no.1, 2021.
[28] I. S. Thaseen, J. S. Banu, K. Lavanya, M. R. Ghalib and K. Abhishek, An integrated intrusion detection system using correlation-based attribute selection and artificial neural network, *Trans. Emerging Telecommunications Technologies*, vol.32, no.2, 2021.
[29] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, Chapman and Hall/CRC, New York, 2012.
[30] E. Sevinç, An empowered AdaBoost algorithm implementation: A COVID-19 dataset study, *Computers & Industrial Engineering*, vol.165, DOI: 10.1016/j.cie.2021.107912, 2022.
[31] B. Thilagavathi, K. Suthendran and K. Srujanraju, Evaluating the AdaBoost algorithm for biometric-based face recognition, *Data Engineering and Communication Technology*, pp.669-678, DOI: 10.1007/978-981-16-0081-4_67, 2021.
[32] O. Altay and M. Ulas, Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children, *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, pp.1-4, 2018.
[33] G. Shmueli, N. R. Patel and P. C. Bruce, *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, 2nd Edition, Wiley, Hoboken, NJ, 2010.
[34] P. Scholar, Prediction of loan risk using Naïve Bayes and support vector machine, *Int. Conf. Adv. Comput. Technol. (ICACT)*, vol.4, no.2, pp.110-113, 2018.
[35] V. Jensen, *Introduction to Bayesian Networks*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.
[36] R. G. Brereton and G. R. Lloyd, Support vector machines for classification and regression, *Analyst*, vol.135, no.2, pp.230-267, 2010.
[37] UCI Repository, *Index of /ml/machine-learning-databases/00529*, https://archive.ics.uci.edu/ml/machine-learning-databases/00529/, Accessed on 11 January 2022.
[38] N. Tigga, *Diabetes Dataset 2019*, https://kaggle.com/tigganeha4/diabetes-dataset-2019, Accessed on 11 January 2022.
[39] N. P. Tigga and S. Garg, Prediction of Type 2 diabetes using machine learning classification methods, *Procedia Computer Science*, vol.167, pp.706-716, 2020.
[40] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, A decision support system for diabetes prediction using machine learning and deep learning techniques, *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, pp.1-4, 2019.
[41] F. Maulidina, Z. Rustam, S. Hartini, V. V. P. Wibowo, I. Wirasati and W. Sadewo, Feature optimization using Backward Elimination and Support Vector Machines (SVM) algorithm for diabetes classification, *J. Phys.: Conf. Ser.*, vol.1821, no.1, DOI: 10.1088/1742-6596/1821/1/012006, 2021.
[42] H.-L. Yang and B.-Y. Li, A hybrid neural network based on particle swarm optimization for predicting the diabetes, *2021 10th International Conference on Software and Computer Applications*, New York, NY, USA, pp.302-306, 2021.
[43] Y. Miao, Using machine learning algorithms to predict diabetes mellitus based on Pima Indians Diabetes dataset, *2021 5th International Conference on Virtual and Augmented Reality Simulations*, New York, NY, USA, pp.47-53, 2021.
[44] P. B. K. Chowdary and R. U. Kumar, An enhanced Naïve Bayes classification algorithm to predict Type II diabetes, *Journal of Engineering Science and Technology*, vol.16, no.4, pp.2927-2937, 2021.

Author Biography

Ployphan Sornsuwit received a Ph.D. in computer science from King Mongkut's Institute of Technology Ladkrabang (KMITL) in 2019. She is a Lecturer at Faculty of Management Science, Digital Business Technology Program at the Kamphaeng-Phet Rajabhat University. Her research interests are machine learning, data mining, intrusion detection and biomedical engineering.