Evaluating the Effectiveness of Wrapper Feature Selection Methods with Artificial Neural Network Classifier for Diabetes Prediction

M. A. Fahmiin(✉) and T. H. Lim

Universiti Teknologi Brunei, Bandar Seri Begawan, Brunei Darussalam
fahmiinabdullah96@gmail.com, lim.tiong.hoo@utb.edu.bn

Abstract. Feature selection is an important preprocessing technique used to determine the most important features that contributes to the classification of a dataset, typically performed on high dimension datasets. Various feature selection algorithms have been proposed for diabetes prediction. However, the effectiveness of these proposed algorithms have not been thoroughly evaluated statistically. In this paper, three types of feature selection methods (Sequential Forward Selection, Sequential Backward Selection and Recursive Feature Elimination) classified under the wrapper method are used in identifying the optimal subset of features needed for classification of the Pima Indians Diabetes dataset with an Artificial Neural Network (ANN) as the classifying algorithm. All three methods manage to identify the important features of the dataset (Plasma Glucose Concentration and BMI reading), indicating their effectiveness for feature selection, with Sequential Forward Selection obtaining the feature subset that most improves the ANN. However, there are little to no improvements in terms of classifier evaluation metrics (accuracy and precision) when trained using the optimal subsets from each method as compared to using the original dataset, showing the ineffectiveness of feature selection on the low-dimensional Pima Indians Diabetes dataset.

Keywords: Feature selection · Wrapper methods · Diabetes classification

1 Introduction

The application of Artificial Intelligence (AI), Machine Learning (ML) and Internet of Things (IoT) encompasses a wide range of industrial fields that have benefitted from the results of data mining, data acquisition and accurate predictions brought up by said technological advances. The healthcare sector is no different worldwide. The use of AI and ML into early detection and prediction of harmful diseases, notably non-communicable diseases such as diabetes, have greatly improved the diagnostics accuracy for healthcare professionals which conversely improves the standard of living for their patients through the means of prevention over treatment [1, 2].

An important stage of performing machine learning for classification and detection is to determine the specific feature that would help to speed up and improve the

H. Gao et al. (Eds.): TridentCom 2019, LNICST 309, pp. 3–17, 2020.
https://doi.org/10.1007/978-3-030-43215-7_1

detection rate. Additional machine learning algorithms are usually applied to determine the most important or relevant features that contributes the most towards performing correct classifications. By selecting the correct features, overall training time is reduced as well removing the problem of overfitting to the diabetes dataset, enabling better generalizability for new inputs of patient data. However, another issue arises from the case of low dimensional datasets, defined as datasets with low number of features relative to the number of instances. In this case, feature selection arguably does not contribute much towards improving the classification algorithm [2].

In this paper, we evaluated a multi-layer wrapper feature selection methods using Sequential Forward Selection (SFWS), Sequential Backward Selection (SBS) and Recursive Feature Elimination (RFE)) and proposed the use of Artificial Neural Network (ANN) to classify diabetic patients trained on the Pima Indians Diabetes dataset. The main contribution of the paper is to evaluate the reliability of the different feature selection methods statistically and compare the performance of the algorithms in selecting the relevant features for classification of diabetes. The relevant subset of features selected by these methods is compared against the results of previous literatures on the same diabetic dataset. The results have shown that all methods are able to identify the two most important features with varying other additional features. The second contribution is showing the effectiveness of feature selection on low-dimensional diabetic dataset, where the results concluded with little to no improvement on the classification model evaluation metrics. We believe this is the time that the wrapper feature selections methods have been evaluated statistically.

The organization of the paper is as follows. Section 2 of the paper explains previous works related to the current study. Section 3 introduces the feature selection methods used in this paper, followed by an introduction to ANN in Sect. 4. In Sect. 5, the proposed methodology is explained and with its results discussions found in Sect. 6. Statistical analysis of the results obtained is done in Sect. 7, while Sect. 8 concludes the paper with a discussion of the contributions and prospects for future work.

2 Literature Review

In current literatures, there are numerous studies done towards the application of AI and ML in the case for diabetes mellitus, Vijayan et al. [3] used a combination of multiple algorithms to produce a model that can classify patients likely to contract diabetes mellitus for up to 80.7% accuracy rate, while Wei et al. [4] obtained the highest accuracy of 77.9% amongst five different individual algorithms (Neural Net-work, Support Vector Machine, Decision tree, Logistic regression and Naïve Bayes). Sowjanya et al. [5] and Duke et al. [6] created their own unique web and mobile inter-faces for diabetes diagnostics in addition to constructing the machine learning models. From the above studies, data preprocessing beforehand is proven to be an important factor that the authors have acknowledged when it comes to achieving better results.

In order to extract the correct feature, Gacav et al. [7] propose the Sequential Feature Selection (SFS) to extract the most important subset of distance vectors on facial expressions for classification purposes. Their method yields an 89.9% mean class recognition accuracy. Zheng et al. [8] also applied the backward variation of SFS with the addition of Information Gain for a hybrid approach to determine an optimal subset of diabetic patient risk factors from the Korean National Health survey. 10 out of their 33 initial features were determined to be optimal for classification, yielding 95.6% in accuracy. SFS and its variances showed positive effects in selecting the most optimal features necessary for accurate prediction results.

The recursive feature elimination (RFE) method involves fitting a model and evaluating the contribution of each feature towards the accuracy of prediction. The least important feature is then removed, and the process is repeated until the desired number of features is reached. Lv et al. [9] made use of a Support Vector Machine (SVM) based RFE to construct a low dimensional face image feature for face recognition which obtained 93.5% recognition accuracy with feature reduction from 720 to only 60. Zhang et al. [10] proposed a Random Forest (RF) based RFE to extract the key feature subset of transient stability assessment of New England 39-bus power system. They have obtained a 99.1% accuracy score with reduction of features from 263 to 45 using said method. From the studies mentioned, the combination of RFE with different classifier models can improve on the feature selection methods.

Similar works have also been done on the publicly available Pima Indians Diabetes dataset [11]. Dutta et al. [12] performed an in-depth analysis of feature importance for the using Random Forest, the algorithm in which they obtained the best result from. They have determined that five of the eight given features having the most importance towards classifying diabetics and non-diabetics. In the work done by Balakrishnan et al. [13], they have used a classification algorithm known as the Support Vector Machine with feature reduction technique, where the accuracy of prediction is assessed after each subsequent elimination of the least important feature of the Pima Indians Diabetes dataset. They have obtained a 1.88% increase in accuracy when removing 37.5% of the features in the dataset. These studies have shown the usage of feature selection in areas of diabetes prediction but with minimal contribution to improvement of the classifying model.

3 Feature Selection Wrapper Methods

Different Feature Selection Wrapper (FSW) methods have been evaluated for its capabilities in obtaining the important features of the dataset and determining the effectiveness of improving the evaluation scores for a classifying algorithm such as Artificial Neural Network (ANN) using the optimal subsets obtained from the tests. Under feature selection techniques, wrapper methods consider the different combinations of subset of features to determine the best combination which resulted in the overall improvement of the evaluation metrics for the specific classifying algorithm. This would result in a more accurate selection than other methods [14]. In this paper, three wrapper methods, proven to be effective based on previous literatures mentioned in Sect. 2, are used for the tests in selecting relevant features from the Pima Indians Diabetes dataset.

Sequential Feature Selection (SFS)

This method of feature selection makes use of a classifier's performance to determine the most optimum feature subset that gives the best result. SFS have two variances which are Sequential Forward Selection (SFWS) and Sequential Backward Selection (SBS). In SFWS, the feature subset started off empty and features from the collection, which results in the best classifier performance, is added until a terminating condition has been reached.

```
let complete dataset: $D = \{d_1, d_2, d_3, \ldots, d_n\}$
let new subset: $S = \{\ \}$
for k iterations do
    $s_{add} = \text{best}\, F(S + s)$, where $s \in D - S$
    $S = S + s_{add}$
    $k = k + 1$
```

Similarly, in SBS, the process works in reverse where a feature, which contributes to the best result for the classifier performance upon removal, is removed from the feature subset. The final optimal subset is then fed into the ANN where its scoring metrics can be determined.

```
let complete dataset: $D = \{d_1, d_2, d_3, \ldots, d_n\}$
let new subset: $S = D$
for k iterations do
    $s_{minus} = \text{best}\, F(S - s)$, where $s \in S$
    $S = S - s_{minus}$
    $k = k + 1$
```

Recursive Feature Elimination (RFE)

In this method, features of least importance is iteratively removed, and the model reconstructed until the desired number of inputs is reached. The dataset is put through RFE using four common estimators (Logistic Regression, Support Vector Machine, Gradient Boosting and RandomForest) and its subset of classified important features is then be fed into the ANN and its scoring metrics determined after. The number of optimal features to be selected is determined by introducing cross-validation into the RFE and scoring different feature subsets before selecting the best scoring collection.

```
let complete dataset: $D = \{d_1, d_2, d_3, \ldots, d_n\}$
let new subset: $S = D$
for k iterations do
    train F(S), rank S according to importance
    $S = S - s_{minus}$, where $s_{minus} = \text{least important feature}$
    $k = k + 1$
```

4 Artificial Neural Network

In order to evaluate the FWS, it is necessary to feed the features extracted into an machine learning algorithm. Artificial Neural Network (ANN) is an machine learning algorithm that attempts to emulating the inner workings of the human brain in which the model learns through its experience and taking corrective measures in reducing the errors in prediction over each cycle. Supervised learning will be used in this paper where outputs are known during training and the weights of each neuron in the hidden layer to be adjusted iteratively to bring about the lowest difference in measurement between network output and desired output. The output being a single Boolean neuron that represents either the patient is diabetic or not (Fig. 1).

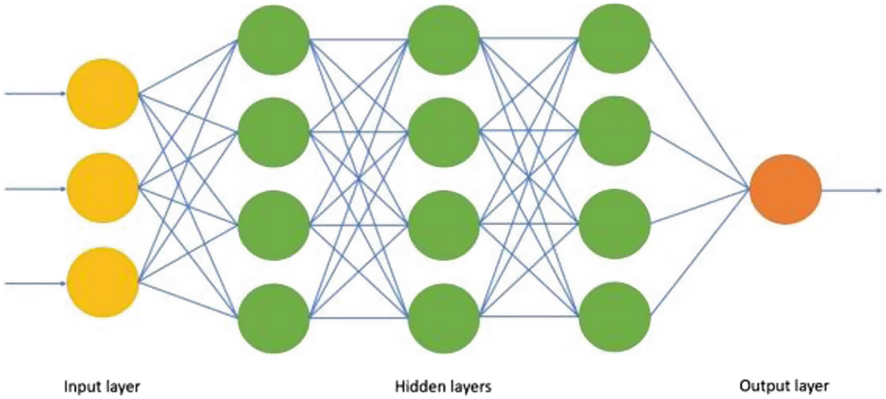

Fig. 1. Representation of an Artificial Neural Network (ANN)

In other literatures, Jayalakshmi et al. [15] made use of ANN as the predicting model for diabetes mellitus of the same dataset while Dey et al. [16] obtained a high accuracy score when predicting diabetics from their local dataset. In both cases, the dataset was preprocessed but does not implement any feature selection techniques. Accounting for this, ANN is chosen in this paper for the classifying algorithm.

5 Methodology

In this paper, three distinct steps are proposed to evaluate the performance of the FWS and predict the diabetes patients. Initially, the data is preprocess before the feature selection algorithm is applied. Finally, classification using the chosen predictor model is executed as shown in Fig. 2.

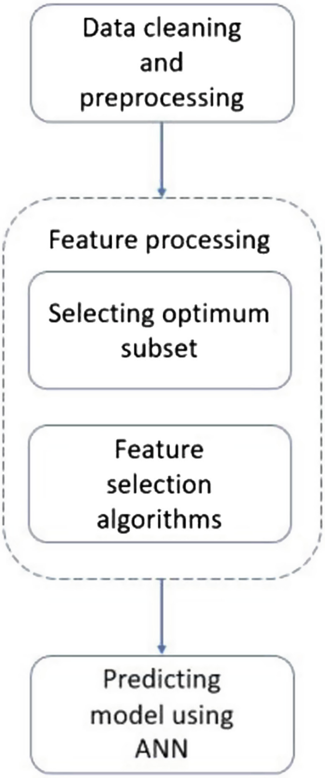
Data cleaning and preprocessing
Feature processing
Selecting optimum subset
Feature selection algorithms
Predicting model using ANN

Fig. 2. Flow diagram of proposed approach

Data Preprocessing

As most real-world data comes with its own errors due to external influences, the datasets procured are generally considered as unclean which would result in a worse performance for the classifying model. From inspecting the Pima Indians diabetes dataset, there are missing data for some feature columns. A common method to clean this dataset would be to replace all zero or null values in each column with the median of that feature column. Another important consideration to note would be the different ranges of values contained in each feature column. Preferably, these values would need to be rescaled in order better the performance of the predicting model which entails transforming the ranges of values of each column to be in the span of 0 and 1.

Feature Preprocessing

The next step of would be the implementation of the wrapper methods for selecting the relevant features from the dataset. For the four common estimators used, each of them would have a different optimal number of features, k, that gives the best classification results. In SFWS and SBS, the terminating condition is only reached when a certain number of k features is reached where k < N (number of features in the dataset). A similar number of k is also used for the RFE method for picking up the top k features from the N number of features in the dataset.

This can be determined by obtaining a cross-validation score against the number of features used to obtain said score. Logistic Regression classifier requires the greatest number of features (7) to perform optimum classification, followed by Random Forest (5), Support Vector Machine and Gradient Boosting (both 4) (see Figs. 3, 4, 5 and 6).

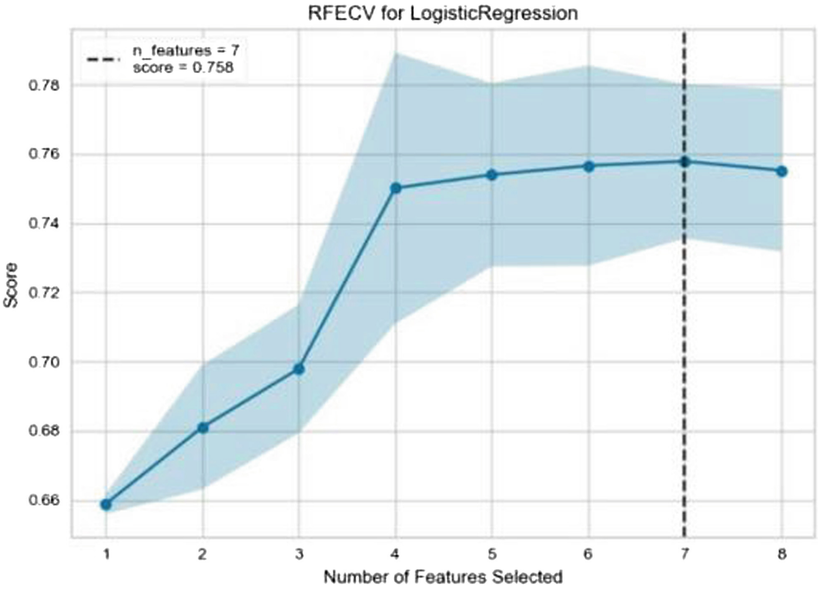

Fig. 3. Optimum number of features for Logistic Regression

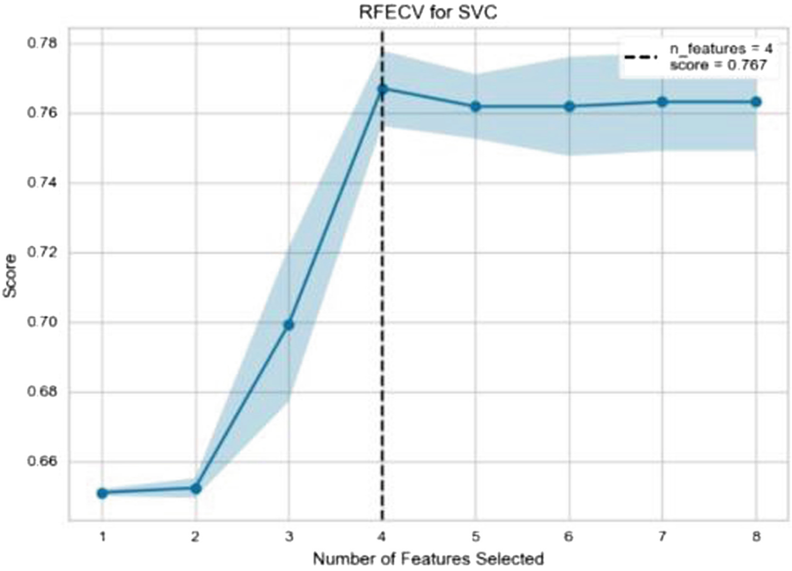

Fig. 4. Optimum number of features for Support Vector Machine

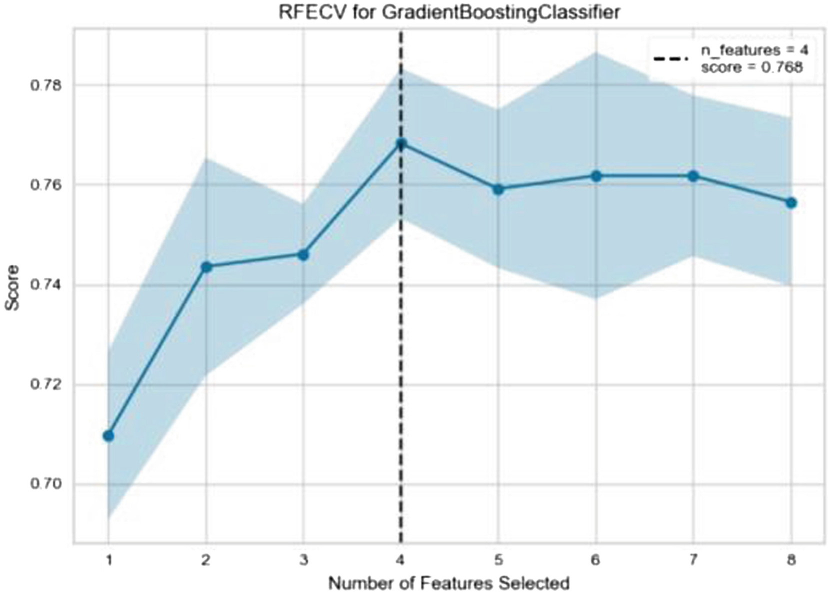

Fig. 5. Optimum number of features for Gradient Boosting Classifier

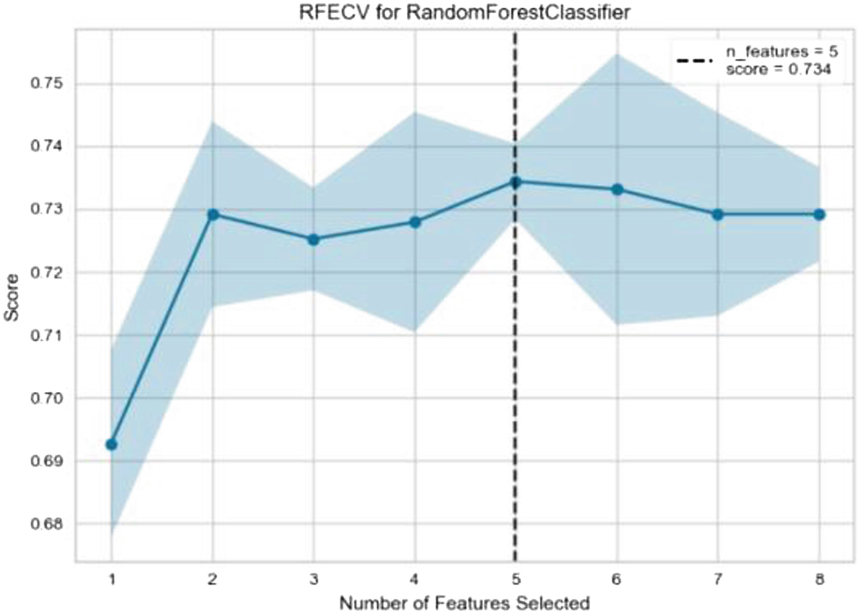

Fig. 6. Optimum number of features for Random Forest Classifier

Scoring Evaluation Metrics

The evaluation metrics of a classifying model are commonly dictated by the accuracy and prediction scores of the classifying model. This establishes the effect of each subset of features have on the model, which in turn determines the effectiveness of the wrapper methods.

The ANN model is built on the Keras Sequential model [17]. A suitable amount of dense neural network layers is added to the model with the Rectified Linear Unit (RELU) activation function and Adam optimizing algorithm before the end classification. This process of training will be repeated for at least 1000 times with an added condition of stop training if the error loss value difference does not change for 5 consecutive iterations.

The training test split is 7:3 respectively and evaluated based on the accuracy and precision values. A 10-fold cross validation is integrated to the predicting mode as well to remove the problems of overfitting and bias. For the purpose of the experiments, the hyperparameters are kept as a constant, removing its effect of prediction improvement unrelated to feature selection.

6 Results and Discussion

In SFWS and SBS, the scoring metric are based on accuracy scores of each classifier, while in RFE, the features are chosen based on its importance ranking. Features selected 75% of the time throughout all classifiers is accounted as one of the chosen features to be used in the ANN. The following Tables 1, 2 and 3 shows the results obtained from the tests.

Table 1. Feature ranking through SFWS.

Features	Importance ranking				Chosen features
	LR	SVM	GB	RF	
Pregnancies	✓		✓		
Glucose	✓	✓	✓	✓	✓
Blood pressure		✓		✓	
Skin thickness	✓			✓	
Insulin	✓				
BMI	✓	✓	✓		✓
DiabetesPedigreeFunction	✓		✓	✓	✓
Age	✓	✓		✓	✓

Table 2. Feature ranking through SBS.

Features	Importance ranking				Chosen features
	LR	SVM	GB	RF	
Pregnancies	✓			✓	
Glucose	✓	✓	✓	✓	✓
Blood pressure	✓	✓			
Skin thickness	✓	✓		✓	✓
Insulin	✓		✓		
BMI	✓	✓	✓	✓	✓
DiabetesPedigreeFunction	✓				
Age			✓	✓	

Table 3. Feature ranking through RFE.

Features	Importance ranking				Chosen features
	LR	SVM	GB	RF	
Pregnancies	✓	✓			
Glucose	✓	✓	✓	✓	✓
Blood pressure	✓			✓	
Skin thickness	✓				
Insulin				✓	
BMI	✓	✓	✓	✓	✓
DiabetesPedigreeFunction	✓	✓	✓	✓	✓
Age	✓		✓		

The model is run for 50 10-fold cross validation iterations and the scores averaged out. Table 4 tabulates the results obtained from the neural network evaluations with its own feature subsets. Figures 7 and 8 shows the boxplot distribution of the readings

Table 4. Evaluation of predicting model.

Feature ranking method	Mean accuracy (%)	Mean precision (%)
Unfiltered (UNF)	78.85	73.34
SFWS	78.41	72.07
SBS	76.18	70.42
RFE	77.13	71.19

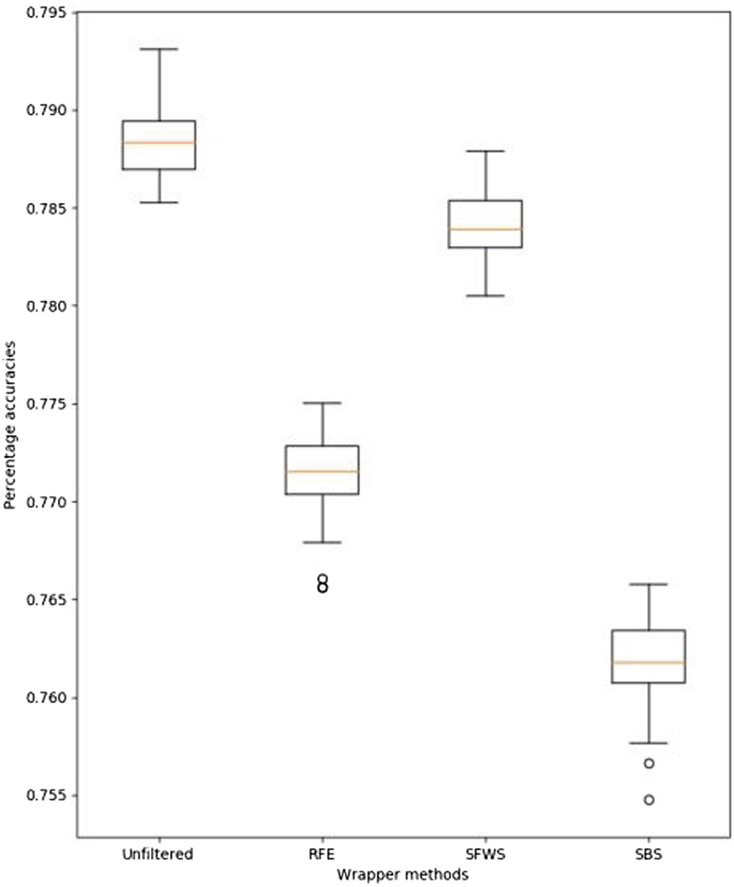

Fig. 7. Accuracy scores distribution for unfiltered and for each wrapper methods

From the above tables and figures, the best performing wrapper method would be the SFWS in both accuracy and precision scores of the ANN model. However, none of the feature selection methods yielded any significant improvement compared to using the existing dataset – in fact, the metrics have decreased, though only slightly.

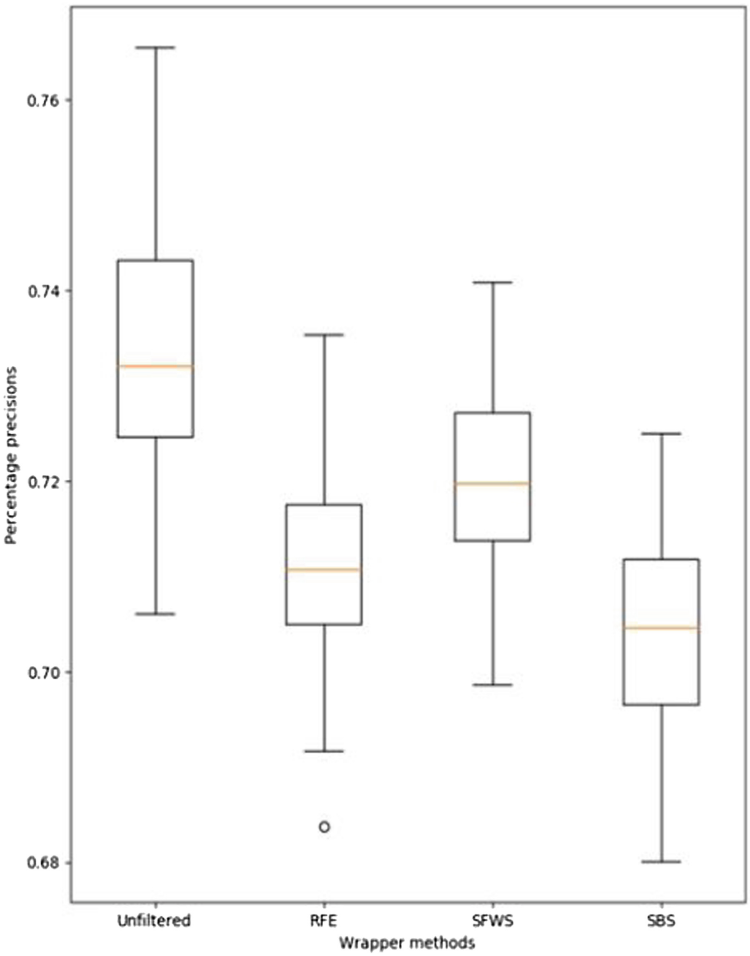

Fig. 8. Precision scores distribution for unfiltered and for each wrapper methods

7 Statistical Analysis

In order to evaluate the statistical significant of our result, a t-test statistical test is applied. According to Lim et al. it is necessary to apply significant testing in order to ensure that the test results are scientifically and statistically significant [18]. Using the t-test, we can determine the difference between the means of the results obtained to one another, and the significance of these differences [19]. A null hypothesis stating that

There is no difference between the accuracy and precision scores of each wrapper method

is formulated, while an alternate hypothesis states that the results obtained are unique form one another. The hypothesis can be rejected if the p-value $\leq \alpha$, where the common value of α is 0.05, indicating a 95% confidence of a valid conclusion for the test. The *t*-score is the scale of difference between the two groups, a larger value of *t* indicates the repeatability probability of the results. The results obtained are tabulated in Table 5.

Table 5. t-score and *p*-values for the ANN readings

Wrapper methods	Metrics used for the ANN model			
	Accuracy		Precision	
	t-score	*p*-value	*t*-score	*p*-value
UNF-RFE	43.144436	1.484971e−65	9.456317	1.832930e−15
UNF-SFWS	12.404437	8.354766e−22	5.701209	1.257081e−07
UNF-SBS	65.519795	9.891198e−83	12.012385	5.647437e−21
SFWS-RFE	34.657911	8.362349e−57	4.365991	3.140681e−05
RFE-SBS	22.706234	8.068926e−41	3.469553	0.000776444
SFWS-SBS	58.863079	2.773827e−78	7.542704	2.366585e−11

The resulting p-values from the tests fall way below the threshold of $\alpha = 0.05$, concluding the validity of readings from the ANN model for each wrapper method to be statistically significant and the test results to be valid.

8 Conclusion

The paper concluded that feature selection using wrapper methods is effective at determining the important features from the Pima Indians Diabetes dataset, that is being Plasma Glucose Concentration and BMI reading of the patients, as proven according to previous literatures seen in the works of Choubey et al. [20] and Rubaiat et al. [21]. As this paper only evaluates the wrapper methods separately, future improvements to produce a more robust feature selection method would be through an introduction of multi-stage process which incorporates all the above methods or in combination with filter and embedded methods. For the latter, SFWS is a prime candidate to be used as the wrapper component of the hybrid technique, considering that this method performs the best overall.

The paper also shows the ineffectiveness of the optimal subset of features derived from the feature selection methods in improving the evaluation scores of the ANN. Considerations for future research would include comparing different classification algorithms using the same methodology discussed on different low-dimensional datasets to further determine the true benefits of feature selection.

References

1. Norhafizah, D., Pg, B., Muhammad, H., Lim, T.H., Binti, N.S., Arifin, M.: Non-intrusive wearable health monitoring systems for emotion detection. In: 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), Siem Reap, pp. 985–989 (2017)

2. Adenin, H., Zahari, R., Lim, T.H.: Microcontroller based driver alertness detection systems to detect drowsiness. In: Proceedings of SPIE 10615, Ninth International Conference on Graphic and Image Processing (2018)
3. Veena Vijayan, V., Anjali, C.: Prediction and diagnosis of diabetes mellitus—a machine learning approach. In: IEEE Recent Advances in Intelligent Computational Systems (RAICS), Trivandrum, India, 10–12 December 2015 (2015)
4. Wei, S., Zhao, X., Miao, C.: A comprehensive exploration to the machine learning techniques for diabetes identification. In: IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, Singapore, 5–8 February 2018 (2018)
5. Sowjanya, K., Singhal, A., Choudhary, C.: MobDBTest: a machine learning based system for predicting diabetes risk using mobile devices. In: IEEE International Advance Computing Conference (IACC), Bangalore, India, 12–13 June 2015, pp. 297–402 (2015)
6. Duke, D.L., Thorpe, C., Mahmoud, M., Zirie, M.: Intelligent diabetes assistant: using machine learning to help manage diabetes. In: IEEE/ACS International Conference on Computer Systems and Applications, Doha, Qatar, 31 March–4 April 2008, pp. 913–914 (2008)
7. Gacav, C., Benligiray, B., Topal, C.: Sequential forward feature selection for facial expression recognition. In: 24th Signal Processing and Communication Application Conference, Zonguldak, Turkey, 16–19 May 2016 (2016)
8. Zheng, H., Park, H.W., Li, D., Park, K.H., Ryu, K.H.: A hybrid feature selection approach for applying to patients with diabetes mellitus: KNHANES 2013–2015. In: 5th NAFOSTED Conference on Information and Computer Science, Ho Chi Minh City, Vietnam, 23–24 November 2018 (2018)
9. Lv, X., Wu, J., Liu, W.: Face image feature selection based on gabor feature and recursive feature elimination. In: Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 26–27 August 2014 (2014)
10. Zhang, C., Li, Y., Yu, Z., Tian, F.: Feature selection of power system transient stability assessment based on random forest and recursive feature elimination. In: IEEE PES Asia-Pacific Power and Energy Engineering Conference, Xi'an, China, 25–28 October 2016 (2016)
11. Pima Indians Diabetes Dataset. https://www.kaggle.com/mehdidag/pimaindians/home
12. Dutta, D., Paul, D., Ghosh, P.: Analysing feature importances for diabetes prediction using machine learning. In: IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, Vancouver, BC, Canada, 1–3 November 2018 (2018)
13. Balakrishnan, S., Narayanaswamy, R., Savarimuthu, N., Samikannu, R.: SVM ranking with backward search for feature selection in type II diabetes databases. In: IEEE International Conference on Systems, Man and Cybernetics, Singapore, Singapore, 12–15 October 2008 (2008)
14. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
15. Jayalakshmi, T., Santhakumaran, A.: A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. In: International Conference on Data Storage and Data Engineering, Bangalore, India, 9–10 February 2010 (2010)
16. Dey, R., Bajpai, V., Gandhi, G., Dey, B.: Application of Artificial Neural Network (ANN) technique for diagnosing diabetes mellitus. In: IEEE Region 10 and the Third international Conference on Industrial and Information Systems, Kharagpur, India, 8–10 December 2008 (2008)
17. Keras Sequential Model. https://keras.io/models/sequential/
18. Hoo, T., Lim, I.B., Timmis, J.: A self-adaptive fault-tolerant systems for a dependable Wireless Sensor Networks. Des. Autom. Embedded Syst. **18**(3–4), 223 (2014)

19. Lim, T., Lau, H., Timmis, J., Bate, I.: Immune-inspired self healing in wireless sensor networks. In: Coello Coello, C.A., Greensmith, J., Krasnogor, N., Liò, P., Nicosia, G., Pavone, M. (eds.) ICARIS 2012. LNCS, vol. 7597, pp. 42–56. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33757-4_4
20. Choubey, D., Paul, S., Kumar, S., Kumar, S.: Classification of Pima indian diabetes dataset using Naive Bayes with genetic algorithm as an attribute selection, pp. 451–455 (2016)
21. Rubaiat, S.Y., Rahman, Md.M., Hasan, Md.K.: Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. In: International Conference on Innovation in Engineering and Technology (2018)