Application of Data Mining Algorithms for Feature Selection and Prediction of Diabetic Retinopathy

Chapter · June 2019

DOI: 10.1007/978-3-030-24308-1_56

5 authors, including:

Roseline Ogundokun
Landmark University

Aderonke Anthonia Kayode
Landmark University

Adekanmi Adegun
Landmark University

Marion O. Adebiyi
Covenant University Ota Ogun State, Nigeria

Application of Data Mining Algorithms for Feature Selection and Prediction of Diabetic Retinopathy

Tinuke O. Oladele[1], Roseline Oluwaseun Ogundokun[2(✉)], Aderonke Anthonia Kayode[2], Adekanmi Adeyinka Adegun[3], and Marion Oluwabunmi Adebiyi[2]

[1] Department of Computer Science, University of Ilorin, Ilorin, Kwara State, Nigeria

[2] Department of Computer Science, Landmark University, Omu Aran, Kwara State, Nigeria
ogundokun.roseline@lmu.edu.ng

[3] Discipline of Computer Science, University of KwaZulu-Natal, Durban, South Africa

Abstract. Diabetes Retinopathy is a disease which results from a prolonged case of diabetes mellitus and it is the most common cause of loss of vision in man. Data mining algorithms are used in medical and computer fields to find effective ways of forecasting a particular disease. This research was aimed at determining the effect of using feature selection in predicting Diabetes Retinopathy. The dataset used for this study was gotten from diabetes retinopathy Debrecen dataset from the University of California in a form suitable for mining. Feature selection was executed on diabetes retinopathy data then the Implementation of k-Nearest Neighbour, C4.5 decision tree, Multi-layer Perceptron (MLP) and Support Vector Machines was conducted on diabetes retinopathy data with and without feature selection. There was access to the algorithms in terms of accuracy and sensitivity. It is observed from the results that, making use of feature selection on algorithms increases the accuracy as well as the sensitivity of the algorithms considered and it is mostly reflected in the support vector machine algorithm. Making use of feature selection for classification also increases the time taken for the prediction of diabetes retinopathy.

Keywords: Data mining · Feature selection · Diabetic retinopathy · Prediction · Classification

1 Introduction

Diabetic Retinopathy is a disease that is common in adults and it occurs when diabetes is not treated for a long period of time. The dataset used in this research is the Diabetes Retinopathy Debrecen dataset from the University of California, Irvine (UCI) repository of machine learning databases. The dataset was provided by some researchers from the University of Debrecen, Hungary containing features extracted from the Messidor image set to predict whether an eye image contains signs of diabetic retinopathy or not.

S. Misra et al. (Eds.): ICCSA 2019, LNCS 11623, pp. 716–730, 2019.
https://doi.org/10.1007/978-3-030-24308-1_56

The research paper on the work done which generated the dataset is (Antal and Hajdu 2014). The dataset contains 1151 instances with 19 attributes each and a binary outcome feature as to whether the instance has signs of diabetic retinopathy or not.

This study is poised to help in the automatic prediction of diabetes retinopathy so as to help diagnose it quickly and to protect those having it from becoming totally blind. It also seeks to discover the effect of using the same algorithm for both feature selection and classification with a view to understanding whether it will reduce the size while maintaining accuracy. The feature selection technique used is the wrapper feature selectors such that the algorithms to be used in the classification are the same one used in feature selection. The algorithms will be compared in terms of their accuracy and sensitivity.

2 Literature Review

2.1 Related Work

Rathi and Sharma (2017) performed a review on the prediction of Diabetes Retinopathy using data mining techniques, Algorithms, techniques and approaches used in literature were compared and it was discovered that SVM and kNN performed best on Diabetes Retinopathy data. Ramesh and Padmini (2017) conducted a study on risk level prediction of diabetes retinopathy using classification algorithms. They collected data from patients in different hospitals through questionnaires. The information in the questionnaires was organized and mined using Naïve Bayes, Multilayer perceptron, Random forest, Bayesian networks and decision stump. Multi-layer perceptron was found most suitable in predicting risk level as it has the highest accuracy.

Elibol and Ergin (2016) extracted time domain features from retina images and those features are used to classify the stage of diabetic retinopathy in which an image is in. The algorithms used are Fisher's linear discriminant analysis, Linear Bayes Normal classifier, Decision tree and k-Nearest Neighbour. The classification accuracies show that when all the extracted features were used in the classification, kNN gives the highest average accuracy of 92.22% while when 7 features were carefully selected, Linear Bayes Normal classifier gave the highest average accuracy. The dataset used in this research is the publicly available retina images DIARETDB1.

Bhaisare et al. (2016) proposed a model for a web-based system for the diagnosis of diabetic retinopathy using eye images, such that people can upload their eye images and the system will mine the images for signs of diabetic retinopathy. This system has the potential of saving time and money for patients.

Mankar and Rout (2016) presented a method for automatic detection of diabetic retinopathy; SVM was used to classify the retina data into normal, having non-proliferative diabetic retinopathy or having proliferative diabetic retinopathy. The presence and amount of Hemorrhages and Exudates in the retina data was used as the classifying feature. The source of the dataset used in the research was not reported.

Jalan and Tayade (2015) proposed a method for diagnosing diabetic retinopathy by combining kNN and SVM algorithms together. The method is to detect the presence or absence of diabetic retinopathy and the severity of the disease.

Sujatha and Divya (2015) proposed a method for identifying people with Diabetes Mellitus and non-proliferative diabetes retinopathy samples from images of the tongue of individuals. The images were pre-processed using a median filter and classified using the proximal support vector machine. The results were said to achieve high performance and can handle a large number of data. The tongue images used in this research were captured by the researchers.

Antal and Hajdu (2014) presented a method for the screening of images to investigate the presence or absence of diabetic retinopathy. Several features were extracted from retinal images using image processing algorithms. The extracted features are based on three components which are Image level components (quality assessment, pre-screening and multi-scale Amplitude-Modulation Frequency-Modulation), lesion-specific components (microaneurysms, exudates) and anatomical components (macula, optic disc). The anatomical components were introduced by the researchers as components to be considered in the determination of the presence of diabetic retinopathy. The extracted features were then classified using ensembles of machine learning classifiers; eight machine learning classifiers were stated as potential members of the ensemble and were combined in different ways. Using an ensemble of a lot algorithm can be time-consuming and have very high complexity in real life scenarios and some algorithms might not contribute to an increase in accuracy when the algorithms are much. Aravind et al. (2013) presented a method for automatic detection of microaneurysms and classification of diabetic retinopathy images by removing the optic disk and similar blood vessels from the eye image so as to reduce the size and the memory space, the eye images take. The pre-processed image was used for feature selection and the features selected were used for classification. Support vector machine was reported to have an average accuracy of 90%. The retina data used in this research were gotten from patients in an eye care hospital.

Evirgen and Çerkezi (2004) presented a model for the prediction of diabetes retinopathy using Naïve Bayes using a dataset obtained from a hospital and reported that Naïve Bayes gave an accuracy of 89%.

3 Methodology

3.1 The Approach

The proposed system is to determine the effectiveness of using the same algorithm for feature selection and classification at the same time. The data mining algorithms considered are k-nearest neighbor, J48 (decision tree algorithm), support vector machines and multilayer perceptron. The algorithms are applied individually on the dataset as classifiers without feature selection and then each algorithm is applied on the dataset first as a feature selector and then as a classifier. The algorithms will be compared in terms of their accuracy and sensitivity. The data mining software to be used for carrying out this research is "WEKA" – (Waikato Environment for Knowledge Analysis) tool.

3.2 Data Collection

The dataset used in this research is the Diabetes Retinopathy Debrecen dataset from the University of California, Irvine (UCI) repository of machine learning databases. The dataset was provided by some researchers from the University of Debrecen, Hungary containing features extracted from the Messidor image set to predict whether an eye image contains signs of diabetic retinopathy or not. The research paper on the work done which generated the dataset is (Antal and Hajdu 2014). The dataset contains 1151 instances with 19 attributes each and a binary outcome feature as to whether the instance has signs of diabetic retinopathy or not.

3.3 Dataset Pre-processing

The dataset is already in arff format which is one of the required formats for dataset used with WEKA. No pre-processing was done for the application of the same algorithm with and without feature selection.

3.4 System Architecture

In Fig. 1, the dataset of diabetes retinopathy is classified with and without feature selection on each algorithm namely k-nearest neighbor, J48 (decision tree algorithm), support vector machines and multilayer perceptron. The labels with feature selection using NN, feature selection using J48, feature selection using SVM, feature selection using MLP connote, applying feature selection on those algorithms before classifying each, to predict the presence of diabetes retinopathy. The label with applying classifiers (J48, KNN, MLP, SVM) denote performing classification on each algorithm without feature selection.

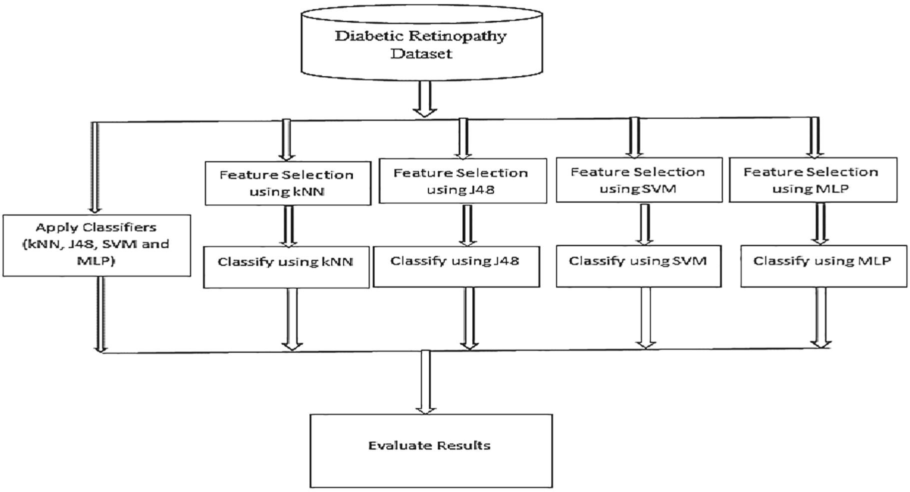

Fig. 1. System architecture

3.5 System Pseudocode

Step 1: Collect dataset
Step 2: Classify using kNN algorithm and record the result
Step 3: Classify using J48 algorithm and record the result
Step 4: Classify using SVM algorithm and record the result
Step 5: Classify using MLP algorithm and record the result
Step 6: Select features and classify using KNN algorithm and record the result
Step 7: Select features and classify using J48 algorithm and record the result
Step 8: Select features and classify using SVM algorithm and record the result
Step 9: Select features and classify using MLP algorithm and record the result
Step 10: Evaluate and compare results.

In step 1, the data are collected from diabetes retinopathy debrecen dataset from the University of California which contain features extracted from the Messidor image set to predict whether an eye image contains signs of diabetes retinopathy or not. The dataset contains 1151 instances with 19 attributes each.

In step 2, 3, 4, 5 involve the use of J48, kNN, SVM, MLP algorithms respectively to classify he 1151 instances without feature selection and the results are recorded. In step 6, 7, 8, 9 involve the use of feature selection to classify the 1151 instances using J48, kNN, SVM, MLP algorithms respectively and the result of each is recorded. In step 10, the results obtained from classification without feature selection are compared with their respective classification with feature selection using the same algorithm.

KNN Algorithm
Training
Step 1: Build the set of training examples D.
Classification
Given a query instance x_q to be classified,
Let $x_1 \ldots x_k$ denote the k instances from D that are nearest to x_q
Step 2: Return

$$F(x_q)=\arg\left[\max \sum_{i=1}^{k} \delta(v,f(x_i))\right]$$

Where (a, b) = 1, if a = b and –(a, b) = 0 otherwise

Decision Tree Algorithm

Training
DecisionTreeTrain(data, remaining features)
guess ← most frequent answer in data
If the labels in data are unambiguous then
  return LEAF(guess)
else if remaining features is empty then
  return LEAF(guess)
else
for all f ϵ remaining features do
NO ← the subset of data on which f=no
YES ← the subset of data on which f=yes
Score[f] ← # of majority vote answers in NO
    # of the majority answers in YES
end for
f ← the feature with maximal score(f)
NO ← the subset of data on which f=no
YES ← the subset of data on which f=yes
left ← DecisionTreeTrain(NO, remaining features \ {f})
right ← DecisionTreeTrain(NO, remaining features \ {f})
return NODE(f, left, right)
end if

Testing
DecisionTreeTest(tree, test point)
If tree is of the form LEAF(guess) then
Return guess
else if tree is of the form NODE(f, left, right) then
  if f = yes in test point then
    return DecisionTreeTest(left, test point)
  else
    return DecisionTreeTest(right, test point)
  end if
end if

SVM Algorithm
Let $(x^{(i)}, y^{i})$ be training data points
Step 1: Compute matrix $H = [H_{i,j}]$ where $H_{i,j=} y^{(i)} y^{(j)} (x^{(i)}.x^{(j)})$
Step 2: Select value β that controls misclassification.
Step 3: Obtain $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ by solving the following quadratic optimization problem
Maximize $(\sum_i \alpha_i + \frac{1}{2}\alpha^T H\alpha)$ subject to the constraints $\sum_i \alpha_i y^i = 0, 0 \le \alpha_i \le \beta$
Step 4: Calculate $\alpha = \sum_i \alpha_i y^{(i)} x^{(i)}$
Step 5: Identify the supporting vectors. These are all the points for which $0 < \alpha_i \le \beta$
Step 6: Compute $b = \frac{1}{n_s} \sum_{s'} (y^s - \sum_s a_i\, y^{(i)} x^{(i)}.x^{(s)})$
Step 7: Compute $sign(\alpha^T x' + b)$for the classification of the given point x'.

MLP Algorithm
Algorithm (forward pass)
Require: pattern $\vec{x}$MLP, enumeration of all neurons in topological order
Ensure: Calculate output of MLP
1: **for all** input neurons *i* **do**
2: set $a_i \leftarrow x_i$
3: **end for**
4: **for all** hidden and output neurons *i* in topological order **do**
5: set $net_i \leftarrow w_{i0} + \sum_{j \in} Pred(i)^{W_{ij}} a_j$
6: set $a_i \leftarrow f_{log}(net_i)$
7: **end for**
8: **for all** output neurons *i* **do**
9: assemble in output vector $\vec{y}$
10: **end for**
11: return $\vec{y}$

3.6 Parameters Used for Evaluation

1. Correctly and Incorrectly Classified instances: The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified while the unclassified instances show the percentage of test instances incorrectly classified. The percentage of correctly classified instances are often called accuracy and the percentage of incorrectly classified instances are gotten by subtracting the correctly classified instances from 100.

   f TP = True positive
   FP = False positive
   TN = True negative
   FN = False negative

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\% \tag{2}$$

2. Sensitivity: This is the proportion of people who have the disease and was rightly classified as having the disease. It is also known as recall or true positive rate.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{2}$$

3. Time taken to build model: This is the time taken by the classifier to build the model to be used for classification.

4 Results and Discussions

Simulations were done by applying the four classification algorithms namely KNN, C4.5, SVM and MLP on diabetes retinopathy dataset, the same algorithms were also used in evaluating features using wrapper feature selection method and then used to classify after feature selection. The results are evaluated and presented as follows.

4.1 Experimental Results

Figure 2 shows how to load the data set into WEKA application. The data set is already in arff format, no pre – processing was carried out.

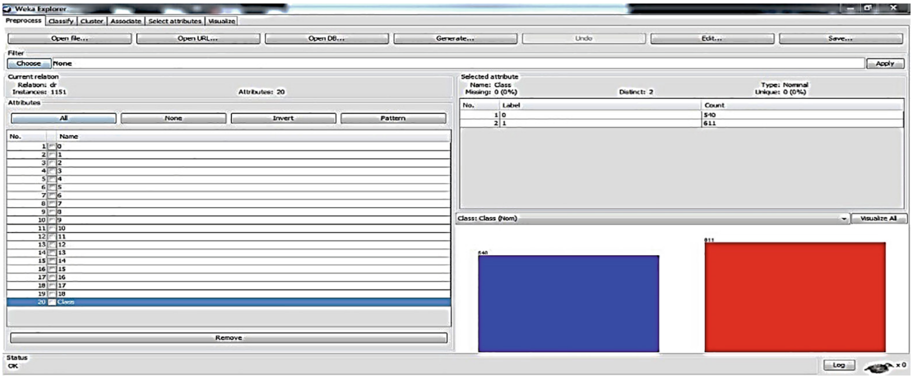

Fig. 2. Loading the data into WEKA

Figure 3 illustrates the output or the results obtained when KNN algorithm is applied as a classifier on the data set which consists of 1151 instances with 19 attributes each. The results are shown in the Table 1.

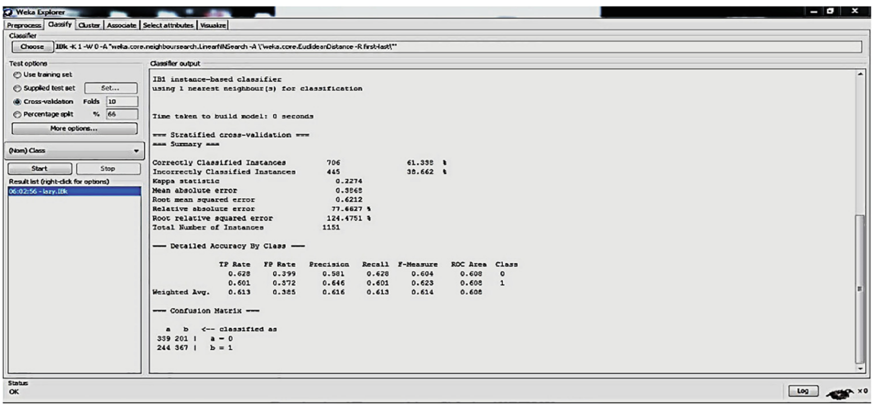

Fig. 3. Applying KNN classification algorithm

Figure 4 illustrates the output or the results obtained when J48 algorithm is applied as a classifier on the data set which consists of 1151 instances with 19 attributes each. The results are shown in the Table 1.

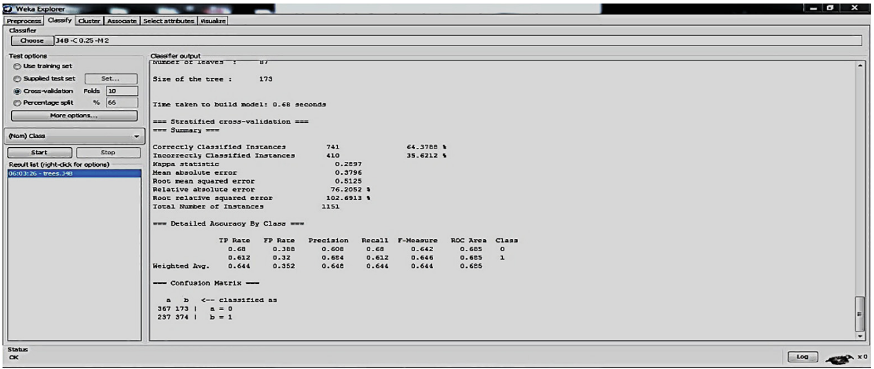

Fig. 4. Applying J48 classification algorithm

Figure 5 illustrates the output or the results obtained when SVM algorithm is applied as a classifier on the data set which consists of 1151 instances with 19 attributes each. The results are shown in the Table 1.

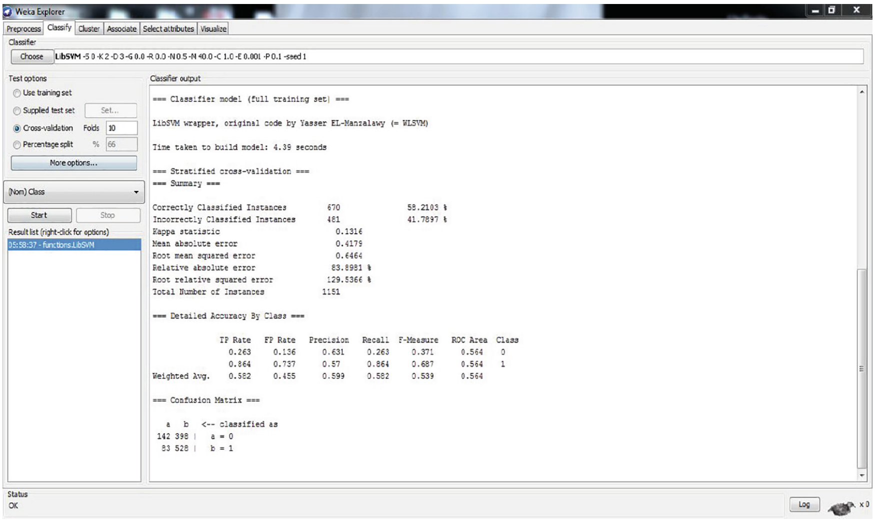

Fig. 5. Applying SVM classification algorithm

Figure 6 shows the output or the results obtained when MLP algorithm is applied as a classifier on the data set which consists of 1151 instances with 19 attributes each. The results are shown in the Table 1.

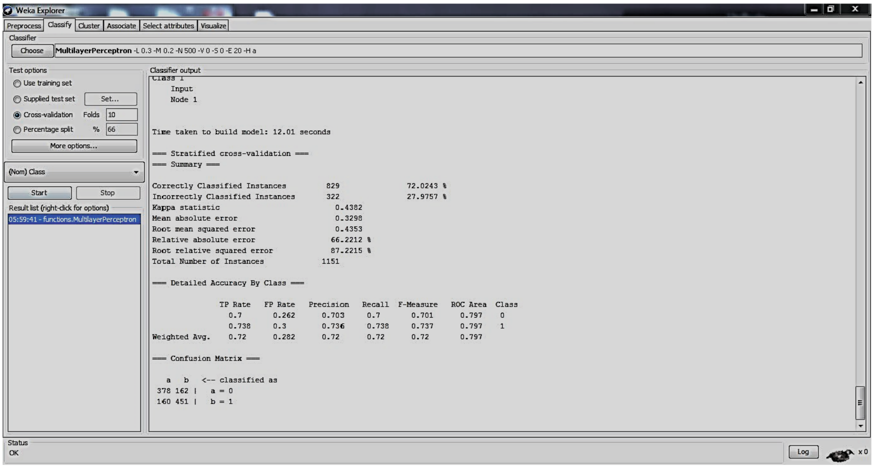

Fig. 6. Applying MLP classification algorithm

Table 1. Summary of confusion matrix for classification algorithms

Algorithms	True positive	False negative	False positive	True negative	Accuracy	Sensitivity
KNN	339	201	244	367	61.32%	62.78%
J48	367	173	237	374	64.38%	67.96%
SVM	142	398	83	528	58.21%	26.3%
MLP	378	162	160	451	72.02%	70%

Figure 7 illustrates the output or the results obtained when feature selection was first performed before KNN is applied as a classifier on the data set. The results are shown in the Table 2.

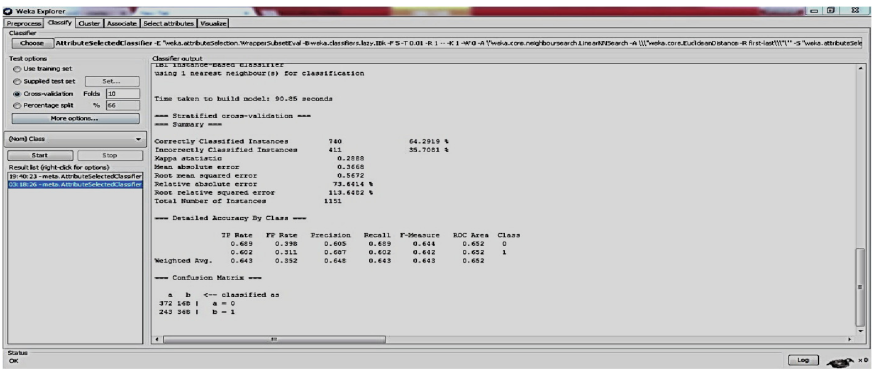

Fig. 7. Applying feature selection and classification on KNN algorithm

Figure 8 illustrates the output or the results obtained when feature selection was first performed, before J48 is applied as a classifier on the data set. The results are shown in the Table 2.

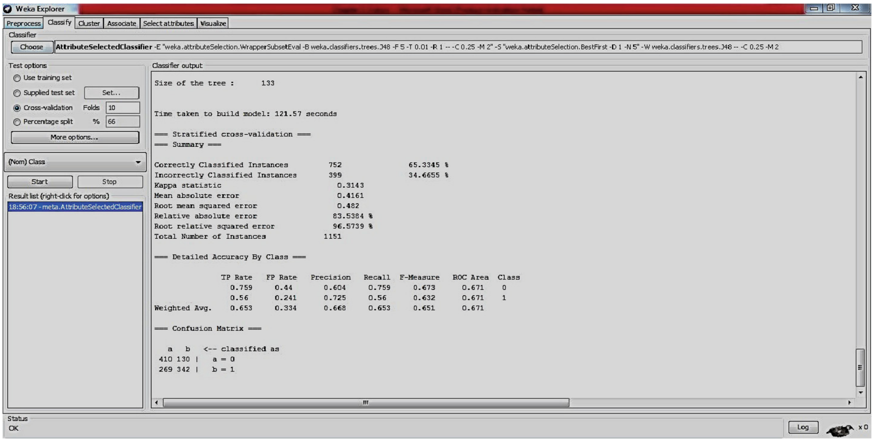

Fig. 8. Applying feature selection and classification on J48 algorithm

Figure 9 illustrates the output or the results obtained when feature selection was first performed before SVM is applied as a classifier on the data set. The results are shown in Table 2.

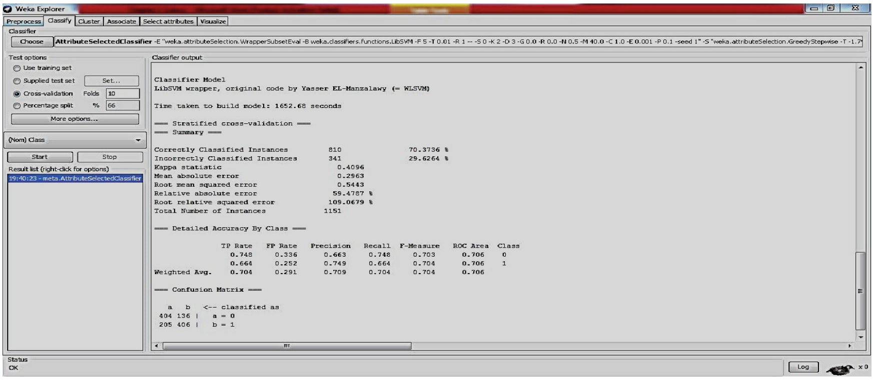

Fig. 9. Applying feature selection and classification on SVM algorithm

Figure 10 illustrates the output or the results obtained when feature selection was first performed before MLP is applied as a classifier on the data set. The results are shown in the Table 2.

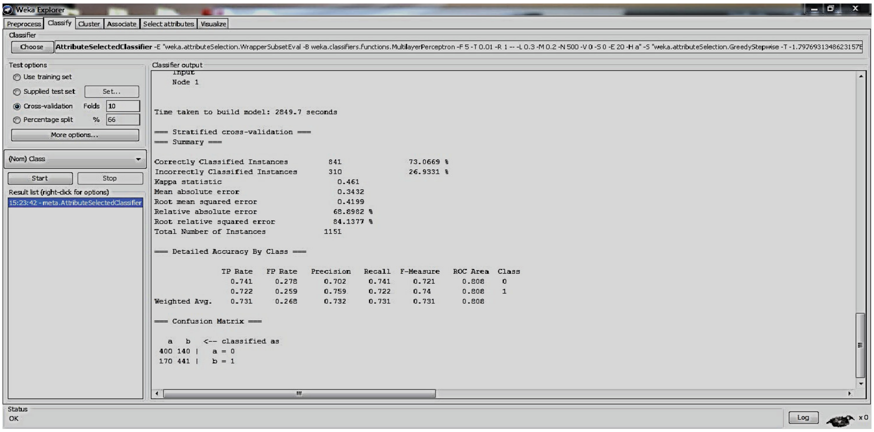

Fig. 10. Applying feature selection and classification on MLP algorithm

Table 2. Summary of confusion matrix for feature selection and classification on the algorithms

Algorithms	True positive	False negative	False positive	True negative	Accuracy	Sensitivity
KNN	372	168	243	368	64.29%	68.8%
J48	410	130	269	342	65.33%	75.90%
SVM	404	136	205	406	70.37%	74.8%
MLP	400	140	170	442	73.7%	74.07%

As depicted in Fig. 11, Comparing the accuracy of each algorithm with its feature selected version, it was discovered that the feature selected version achieves a better accuracy, while the difference is just about 1% in J48 and MLP, and about 3% in KNN, the effect of feature selected SVM was highly pronounced in SVM of which it's accuracy at predicting the presence of diabetes retinopathy was increased by 12.16%. This shows that while using the same algorithms for feature selection and classification have a positive influence in the accuracy of prediction, its influence is emphasized more on some algorithms than others.

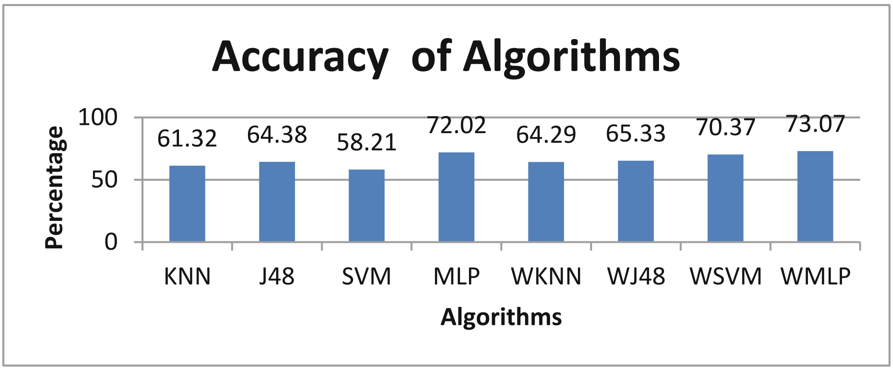

Fig. 11. Accuracy of algorithms in predicting diabetes retinopathy

As depicted in Fig. 12, when the individual algorithms were applied, SVM performed poorly as it was only able to achieve a sensitivity level of 26%, which is not effective in predicting diabetes retinopathy, while the other algorithm at least achieved a 62% sensitivity level. After the algorithms were wrapped, KNN's sensitivity in predicting algorithms was increased by 6.1%, J48 increased by 7.96%, SVM by 48.5% and MLP by 4.07%. Feature selected SVM thus performed much better and only came behind in sensitivity to J48, this shows that using feature selected for algorithms have a positive impact on the sensitivity of prediction. It was also observed by the researcher that the time taken by algorithms to build classification model when using feature selection is considerably longer than applying the algorithm simply by itself. The same applies in the prediction process

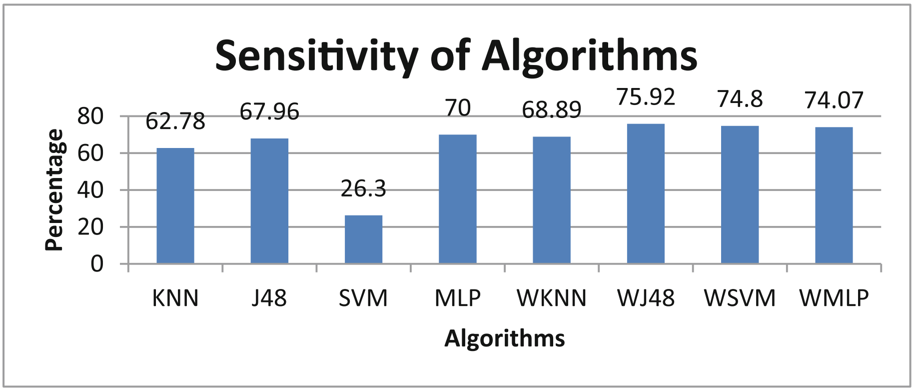

Fig. 12. Sensitivity of algorithms in predicting diabetes retinopathy

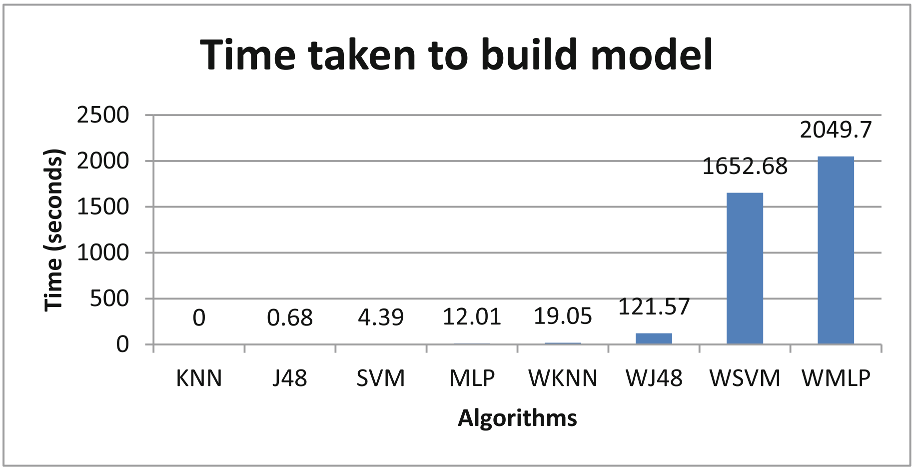

Fig. 13. Time taken to build classification model

As depicted in Fig. 13, the time taken to build the classification model was quite small, the highest time taken was by MLP which is 12 s, but for the use of feature selection for classification model, the time taken increased drastically with KNN that built its classification model initially now taking as much as 19 s. The models that took the longest time to build were wrapped SVM and wrapped MLP. It is worthy to note that the researcher observed that both the time it takes to build classification model and the time taken to classify are directly proportional, thus feature selected MLP took the longest time to classify.

5 Conclusion

Using the same algorithm for feature selection in predicting diabetes retinopathy disease has been shown to positively influence the accuracy and sensitivity of prediction having greater effect on support vector machines in comparison with other algorithms considered, but was also discovered to increase the time taken to build and apply classification model considerably. Thus apart from SVM in which it increases performance considerably, the time-performance trade off in other algorithms might not be worth it except in areas where it is not applied real time and any little increase in the accuracy of prediction is of great importance. This study shows that while using the same algorithm for feature selection and classification improved the performance of algorithms than using the same algorithm for classification without feature selection. In most algorithms considered, the improvement was most pronounced for support vector machines.

6 Recommendations

The results obtained from this research work, show vividly that the use of the same algorithm for feature selection and classification, improve the accuracy and sensitivity in predicting diabetes retinopathy. Therefore, the use of the same algorithm for feature selection and classification should be encouraged.

7 Future Work

Further research can be carried out in order to ascertain the effects; the use of the same algorithm will have in the classification in term of accuracy and sensitivity on other diseases. The use of feature selection and classification can also be applied on other data mining algorithms apart from the algorithms used in this research work, so as to discover those ones that it will enhance their performance greatly.

References

Aravind, C., PonniBala, M., Vijaychitra, S.: Automatic detection of microaneurysms and classification of diabetic retinopathy images using SVM Technique. In: International Conference on Innovations in Intelligent Instrumentation, Optimization and Signal Processing, pp. 18–22 (2013)

Antal, B., Hajdu, A.: An ensemble-based system for automatic screening of diabetic retinopathy. Knowl.-Based Syst. **60**, 20–27 (2014)

Bhaisare, A., Lachure, S., Bhagat, A., Lachure, J.: Diabetic retinopathy diagnosis using image mining. Int. Res. J. Eng. Technol. **3**(10), 858–861 (2016)

Elibol, G., Ergin, S.: The assessment of time-domain features for detecting symptoms of diabetic retinopathy. Int. J. Intell. Syst. Appl. Eng. **4**(Special Issue), 136–140 (2016)

Evirgen, H., Çerkezi, M.: Prediction and diagnosis of diabetic retinopathy using data mining technique. Online J. Sci. Technol. **4**(3), 32–37 (2004)

Jalan, S., Tayade, A.A.: Review paper on diagnosis of diabetic retinopathy using KNN and SVM algorithms. Int. J. Adv. Res. Comput. Sci. Manag. Stud. **3**(1), 128–131 (2015)

Mankar, B.S., Rout, N.: Automatic detection of diabetic retinopathy using morphological operation and machine learning. ABHIYANTRIKI Int. J. Eng. Technol. **3**(5), 12–19 (2016)

Ramesh, V., Padmini, R.: Risk level prediction system of diabetic retinopathy using classification algorithms. Int. J. Sci. Dev. Res. **2**(6), 430–435 (2017)

Rathi, P., Sharma, A.: A review paper on prediction of diabetic retinopathy using data mining techniques. Int. J. Innov. Res. Technol. **4**(1), 292–297 (2017)

Sujatha, S., Divya, D.: A narrative approach for analyzing diabetes mellitus and non proliferative diabetic retinopathy using PSVM classifier. Int. J. Adv. Res. COmput. Eng. Technol. **4**(8), 3341–3345 (2015)