RISK PREDICTION ANALYSIS FOR CLASSIFYING TYPE 2 DIABETES OCCURRENCE USING LOCAL DATASET

M. HAFIZ FAZREN ABD RAHMAN[1], WAN WARDATUL AMANI WAN SALIM[1], M. FIRDAUS ABD-WAHAB[1*]

[1]Dept. of Biotechnology Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, Selangor, Malaysia

**Corresponding author: firdaus@iium.edu.my*

(Received: Day Month Year; Accepted: Day Month Year; Published on-line: Day Month Year)

https://doi.org/10.31436/iiumej.vxxix.x

ABSTRACT: The steep rise of cases pertaining to Diabetes Mellitus (DM) condition among global population has encouraged extensive researches on DM, which led to exhaustive accumulation of data related to DM. In this case, data mining and machine learning applications prove to be a powerful tool in transforming data into meaningful deductions. Several machine learning tools have shown great promise in diabetes classification. However, challenges remain in obtaining an accurate model suitable for real world application. Most disease risk-prediction modelling are found to be specific to a local population. Moreover, real-world data are likely to be complex, incomplete and unorganized, thus, convoluting efforts to develop models around it. This research aims to develop a robust prediction model for classification of type 2 diabetes mellitus (T2DM), with the interest of a Malaysian population, using three different machine learning algorithms; Decision Tree, Support Vector Machine and Naïve Bayes. Data pre-processing methods are utilised to the raw data to improve model performance. This study uses datasets obtained from the IIUM Medical Centre for classification and modelling. Ultimately, the performance of each model is validated, evaluated and compared based on several statistical metrics that measures accuracy, precision, sensitivity and efficiency. This study shows that the random forest model provides the best overall prediction performance in terms of accuracy (0.87), sensitivity (0.9), specificity (0.8), precision (0.9), F1-score (0.9) and AUC value (0.93) (Normal).

KEY WORDS: *Type 2 Diabetes Mellitus, Prediction Model, Random Forest, Support Vector Machine, Naïve Bayes.*

1. INTRODUCTION

Diabetes Mellitus (DM) was considered to be one of the leading causes of death worldwide. According to a study conducted by International Diabetes Federation in 2015, around 415 million people were diagnosed with diabetes, 75% out of which are from second and third world countries. Overall, the disease has contributed to over 5 million recorded deaths. It was predicted that by 2040, the number of people with diabetes will rise to 642 million worldwide [1]. In Malaysia alone, type 2 diabetes has shown a 21 percent increase in 2015, affecting 2.8 million of its population [2, 3].

DM is characterized by chronic hyperglycaemia, caused by defective insulin secretion or impaired biological effects that lead to several life-threatening complications. The increase in obesity rate, poor dietary option, sedentary lifestyle and poverty are believed to be the leading causes of this health problem [3]. Therefore, in order to decrease the morbidity rate of DM, focus has been placed on high risk groups of DM that falls within certain profile: Age ≥ 45, BMI ≥ 24kg/m2, impaired glucose tolerance (IGT) or impaired fasting glucose (IFG), family with a history of DM, low high-density lipoprotein cholesterol or hypertriglyceridemia (HTG), existent of hypertension or cardiovascular and cerebrovascular disease and female with gestation age ≥30 [4]. Early detection of diabetes is vital to ensure extended life expectancy of individuals as well as improvement in the quality of life. One of the ways of early DM diagnosis is through data mining and machine learning application.

Since the past decade, healthcare-related data has exponentially grown in abundance due to the rapid increase in development and use of electronic healthcare devices. Due to this, health-related data mining has also grown in parallel [5, 6]. Data mining is a computational process that involves discovering valuable information from large datasets using methods such as artificial intelligence, machine learning, statistics, and database systems [7]. Machine learning provides a system that is able to learn and adapt automatically from past experiences without being explicitly programmed. It consists of complex models and algorithms used to recognise pattern and understand input data, which are later used to make predictions or decisions for new data [8]. Applying a prediction model into real clinical settings can help with detection and screening in undiagnosed high-risk subjects, allowing for early intervention or prevention by physicians and other healthcare providers. Furthermore, the design of machine learning models has the possibility of predicting and diagnosing future disease.

Most diabetes-related prediction studies utilise the Pima Indian diabetic dataset from UCI machine learning repository for developing prediction models [9, 18, 19], such that the dataset has become the standard in studies where accuracy and algorithm parameter tuning are primary objectives. However, training classifier constructed using Pima dataset does not necessarily translate well when a new dataset from a different population is introduced. For machine learning model to be applicable in clinical setting, the classifiers must be trained under targeted population data to ensure an accurate representation of the patients' demographics [10].

Additionally, diabetes prediction studies in Malaysia are typically based on risk score model frameworks [11, 12]. However, recent studies have raised concerns on adopting risk scoring models for disease prediction due to limitations on their applicability to local populations, calibration capacity and discrimination the models [13].

The aim of this study is to design a site-specific prediction model that is closely relevant to Malaysian context. The proposed model is expected to predict the likelihood of diabetes in high risk patients with high accuracy. Therefore, three machine learning classification algorithms namely Random Forest, SVM and Naïve Bayes are used in this study. Algorithm testing is conducted on datasets collected from the International Islamic University Malaysia Medical Centre (IIUMMC). The output performances from the three algorithms are evaluated based on precision, accuracy, F-measure, sensitivity and, specificity. Accuracy is measured over correctly and incorrectly classified instances.

2. METHOD

2.1. Research Framework

The framework of this study comprises of a standard knowledge discovery in databases (KDD) process with slight modification being made in accordance to the research objectives. All mathematical model and performance analyses are done using R software. The first step involves collection of raw data from IIUM Medical centre. Then, the data is processed through a series of exploratory data analysis (EDA) tool to better understand it in terms of attribute type, class distribution, mean, standard deviation and confidence interval of the data before proceeding to the machine learning aspect. The raw data is then pre-processed. This process includes missing value imputation, data scaling, data sampling and feature selection. Moreover, the pre-processed data was split into training and testing set with 70:30 ratio while also being validated using a 10-cross fold method. Machine learning algorithms or classifiers such as Naïve Bayes, Random Forest and SVM were used to construct the prediction models from the training dataset. Once constructed, the prediction models were assessed using the testing dataset to determine the predictability of each model. The performance of each model was evaluated in terms of accuracy, precision, sensitivity, specificity, F1 score and AUC values. Fig. 1 shows the flow diagram of the research framework**.**

2.2. Dataset

Patients' diabetes dataset was collected from the Medical Records department at IIUM Medical Centre Kuantan, Pahang. The dataset comprises health profile from a total of 200 patients diagnosed with type 2 diabetes and 100 non-diabetic patients. The dataset consists of 149 females and 151 males with the majority of them were from single ethnic group (Malays) within the age range of 16 to 81 years old. Moreover, the collected data were made up of 15 input features such as age, gender, blood pressure, fasting blood glucose level, body mass index, etc. The responding variable in this study were categorized as binary classification (diabetes and non-diabetes). Status was the outcome variables which determine whether patients was diagnosed with diabetes or not. A more detailed description of IIUMMC dataset was listed in Table 3.

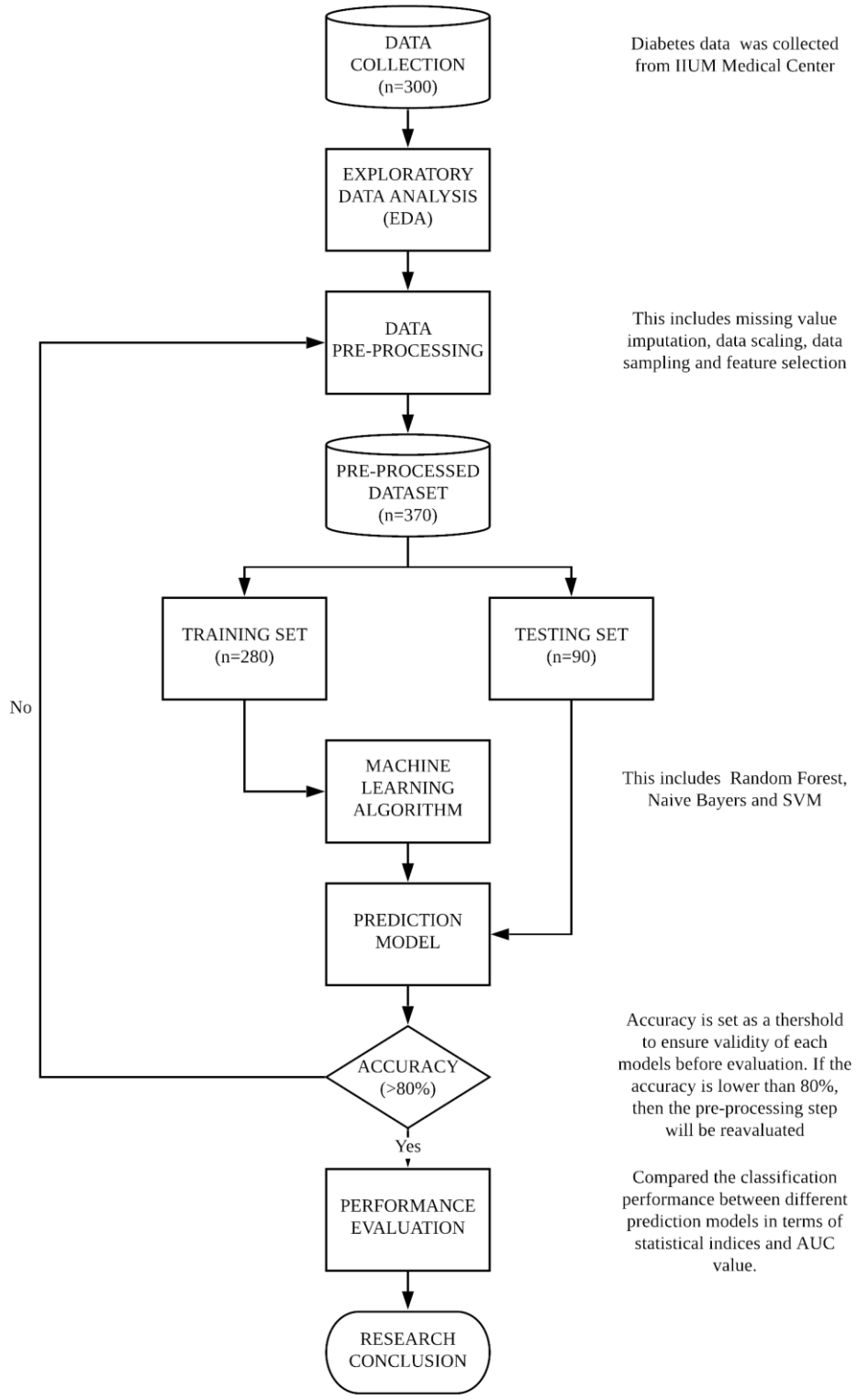

Fig. 1. Research Framework illustrating the strategy adopted in data pre-processing, as well as the different machine learning algorithm used in this study. An accuracy of 80% was set as the minimum threshold for model acceptance.

2.3. Pre-processing

Several data features that was collected from the IIUMMC medical records were flawed and absent, giving rise to dataset imperfection. Missing values were found for some of the variables; Uric acid data had 42 samples labelled as NA (Not Available). The combination of missing values from all variables in the dataset constituted to about 30% of the observation in the dataset, meaning that removing these values would result in significant information loss [14]. To address this, pre-processing missing values imputation was done. The missing values were replaced with mean or K-nearest neighbour method (KNN) values from the corresponding value of k nearest-neighbour column in Euclidean distances. In this study, the k-value was chosen through cross fold validation (KCFV), where the dataset was divided into k-subset and tested multiple times against its reserved data. The best k-value was chosen based on the accuracy obtained from various testing of dataset. A k-value of 5 was selected for this dataset.

The IIUMMC dataset also showed imbalanced data distribution ratio of diabetic over non-diabetic patients. This could introduce model biasness towards the majority classes and result in poor prediction accuracy [15]. It is thus necessary to resolve dataset imbalance by implementing either under sampling or oversampling. In the IIUMMC dataset, the existing imbalanced distribution is skewed towards the diabetic class with twice the number of samples compared to the non-diabetic class, as shown in fig. 2. Thus, the non-diabetic class was oversampled using SMOTE to match the diabetic class training set. In this process, oversampling was implemented to avoid information losses from data elimination method via under sampling.

Prior to feature selection and data fitting, data normalization were carried out within the range of -1 to 1 for all predictor variables except for diabetic status where it will be classified as 1 and 0 for diabetic and non-diabetic, respectively. This is done to eliminate variation of variables range in a dataset [16]. The formula used for feature normalization is stated in Eq. (1)

$$x'_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Following normalization, the dataset were screened for feature identification and selection. This is done to optimize the number of input variables when developing a predictive model, thereby improving the dimensionality, downtime, and memory requirements for model training [17]. In this study, correlation matrix and feature importance were the feature selection processes applied to the IIUMMC dataset by implementing data correlation method between continuous variables as shown in fig. 3. A correlation plot was drawn to observe the relationship between input features. A cutoff of the absolute Pearson's correlation coefficient of $|r| \leq 0.5$ was set for rejecting weakly correlated input feature, which is then excluded from the modelling parameter. Next, feature importance was implemented using the *varimp* function in R, which determines the best features for the output based on variance threshold. Thus, the final selected features are haemoglobin a1c, age, blood pressure (systolic), blood pressure (diastolic), body mass index, uric acid, cholesterol, creatine and high-density lipoproteins.

Fig. 1. Bar plot of diabetic and non-diabetic distribution in IIUMMC dataset before (a) and after (b) SMOTE oversampling was applied.

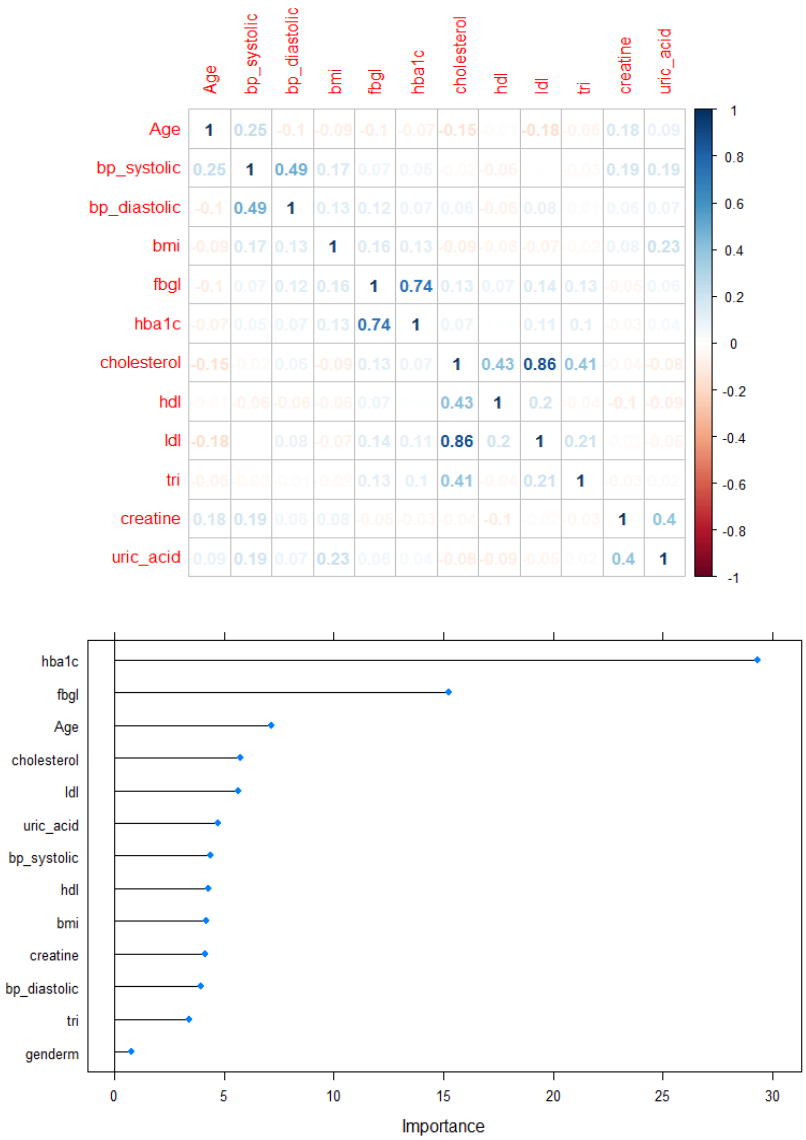

Fig. 3. Correlation plot among independent variables of the IIUMMC dataset (top) and feature importance plot constructed using random forest function (bottom)

2.4. Model Building

R software was used to build the prediction models based on Naïve Bayes, Random forest and SVM classifiers. In order to validate the capabilities of the predictive models, percentage split and cross-validation method were used. Percentage split method refer to the splitting of dataset into training set that was used by classifier to build prediction models while testing set was used to validate the performance of the constructed models [18]. In this study, the dataset was split to 70% training, 30% testing which translates into 210 samples for the training set (before data sampling) and 90 samples for the testing set. The models were also validated using k-fold cross method. In k-fold cross method, the whole dataset was used to train and test the classifier. K refer to the number of random splits that occur within the dataset. K= 10 is chosen for this study meaning that with each k-folds, the algorithm was trained on 90% of the dataset and the other 10% was used to test it. The model was tested to check the effectiveness for k^{th} fold. This was repeated until each of the k-folds

was used as the test set. The average of k recorded errors called the cross-validation error was used as performance metric for the model [19]. The advantage of this method was that the whole samples in the dataset were trained and tested, avoiding the possibilities of high variance within the dataset.

2.5. Performance Evaluation

The classification performance of each classifiers was compared based on the statistical indices obtained from the confusion matrix analysis. True Positive (TP) and True Negative (TN) values represent predicted values that are congruently classified with the actual values. False Positive (FP) indicates the a prediction that has falsely classified a value to be true, when in actuality it is not. Analogously, False Negative (FN) are predictions that falsely classifies a decoy to be true from the prediction model. The data obtained from the confusion matrix will then be used to calculate various statistical performance indices as shown in Eq. (2) to (8). The accuracy in the confusion matrix does not represent standard deviation and mean, but instead represent the total count of correctly predicted value by the model over the total number of predicted sample. Accuracy is one of the most important indices of this study as it determines the classification capability of each prediction model. Other performance criteria are precision, sensitivity, specificity, and F1 scoring. Precision represents the fraction of positive predictive values against actual positive case. Sensitivity, or true positive rate, measures the proportion of actual positives that were correctly identified, whereas specificity (true negative rate) shows the actual negatives that were correctly identified. F1 score represents the mean of recall and precision were used to compare the performance of each parameters for overall efficiency of the classifier [20]. The target of predictive models were to maximize accuracy, specificity, sensitivity, precision, and F1-score. The purpose of this study is to develop a prediction model for classifying whether an individual is diabetic or non-diabetic.

Table 1. Example of confusion matrix of train data from SVM model, where the performance classification measurements were done to compare between the qualities of the statistical prediction values against the actual values reported.

		Predicted Value	
		Diabetic	**Non-diabetic**
Actual Value	**Diabetic**	126	7
	Non-diabetic	14	133

By plotting a Receiver Operating Characteristic (ROC) curve, the performance of a binary classifier can be visualised and further analysed. An ROC plots the True Positive Rate (TPR) against the False Positive Rate (FPR). Values for both rates are calculated using Eq. (7) and (8). ROC was a probability curve that shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). This measures how well is a prediction model able to output a result that corroborates with the actual data. The closer the plot is to the top-left corner of the y-axis implies the better the classifier is in predicting the output. A diagonal line where FPR = TPR serves as a baseline to an ROC curve. This indicates that if the prediction model results in a plot on the ROC that coincides with, or comes close to, the diagonal line, then the classifier is doing no better than a random guess [21, 22]. ROC is

useful in evaluating the performance of binary classification as it is not affected by class distribution within the dataset, nor dependant of disease prevalence, since it is based on specificity and sensitivity. Furthermore, the accuracy of each classifier can be calculated using the area under the curve (AUC). A classifier with an AUC value close to 1.0 indicates optimal predictive capabilities in distinguishing non-diabetic patients from diabetic ones.

$$\boldsymbol{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\boldsymbol{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\boldsymbol{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

$$\boldsymbol{Sensitivity\ or\ Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$\boldsymbol{F1\ score} = \frac{2TP}{2TP + FP + FN} \quad (6)$$

$$\boldsymbol{True\ Positive\ Rate} = \frac{TP}{TP + FN} \quad (7)$$

$$\boldsymbol{False\ Positive\ Rate} = \frac{FP}{FP + TN} \quad (8)$$

3. RESULTS AND DISCUSSION

3.1. Classification Performance

From table 2 and fig. 4(a), the model built from random forest algorithm showed the highest accuracy performance of 91%., followed by models from SVM and Naïve Bayes with an accuracy of 86% and 83% respectively. This illustrates the percentage of each classifier in accurately predicting diabetic and non-diabetic groups in their actuality. This also mean that incorrect classification of diabetic group to non-diabetic group and vice versa are the lowest in random forest classifier.

Table 2. Comparison table in performances metric between classifiers using test data. ($p \leq 0.10$)

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-score
Random Forest	0.91	0.93	0.86	0.93	0.93
SVM	0.86	0.81	0.96	0.98	0.89
Naïve Bayes	0.83	0.83	0.83	0.91	0.86

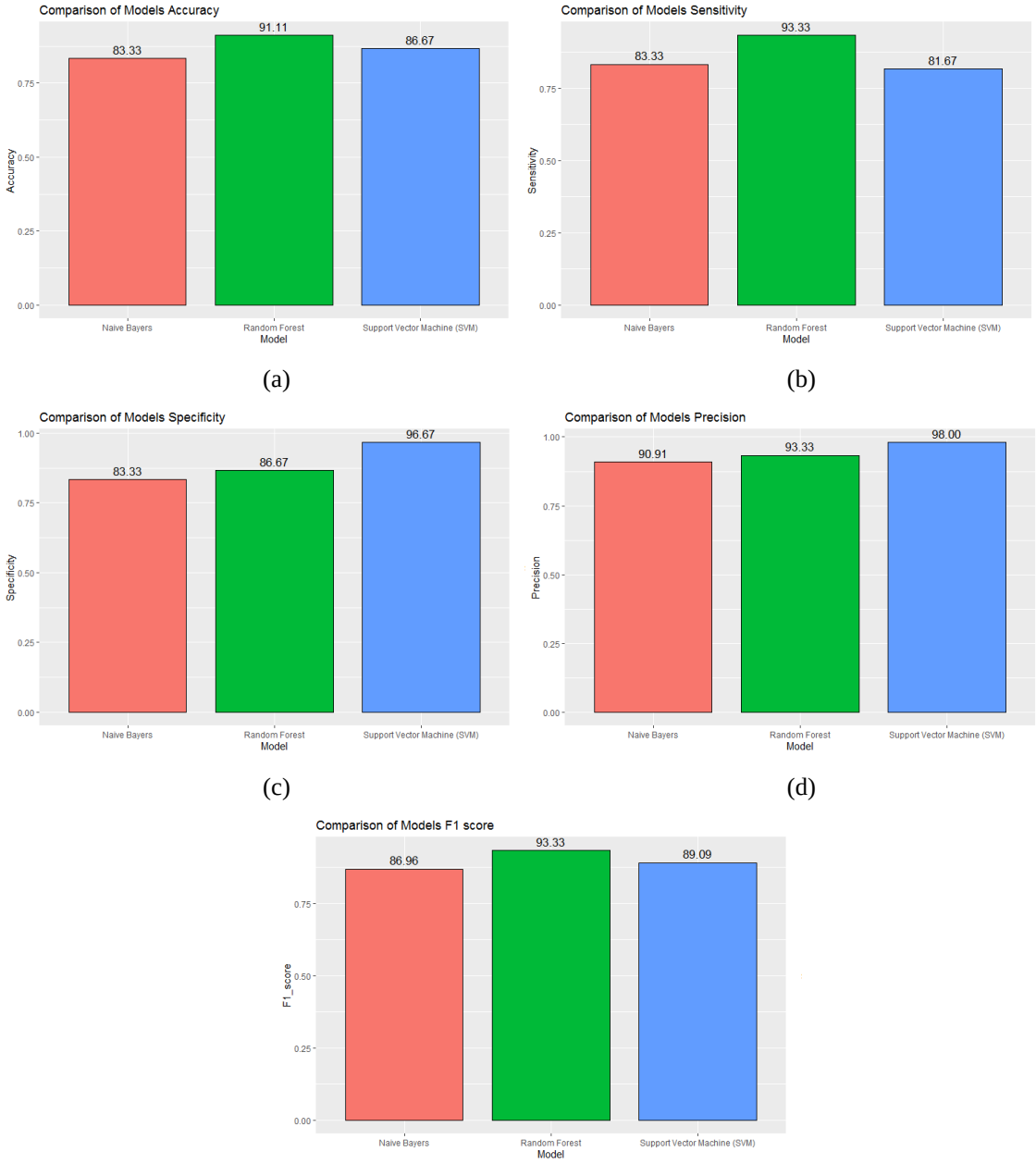

Fig. 4. Comparison of prediction performance between Naïve-Bayes, Random Forest and Support Vector Machine classifiers used in this study. Five main statistical indices are used to compare these performances; (a) Accuracy, (b) Sensitivity, (c) Specificity, (d) Precision and (e) F1-score

Similar to accuracy performance, the random forest model sensitivity value was the highest with 93% compared to SVM and Naïve Bayes models with 81% and 83%, respectively, as shown in table 2 and fig. 4(b). Sensitivity reflects the ratio of correctly classified diabetic individuals over the total number of predicted numbers in the diabetic group. High sensitivity indicates the low occurrence of falsely predicting diabetic patients as non-diabetic.

However, from the specificity result in table 2 and fig. 4(c), it can be seen that SVM classifier has the best specificity of 96%, compared to Naïve Bayes, with 83%, and random

forest model with specificity 80%. Specificity in this study refers to the proportion of patients correctly predicted as non-diabetic. Thus, the number of non-diabetic patient falsely predicted to be diabetic were the lowest from the SVM classifier.

Table 2 and fig. 4(d) also shows the precision performance between different classifier. The SVM classifier showed the best precision with 98% performance, followed by random forest at 93%. Lastly, Naïve Bayes showed the lowest precision value among the three with 89%. Precision here illustrates the proportion of diabetic patients that was predicted correctly. The low value of non-diabetic patients predicted to be diabetic determine the strength in this measurement, which can be seen from the ouput of SVM classifier. Lastly, table 2 and fig. 4(e) showed the comparison in F1 measurement among classifiers. The random forest classifier showed an F1 score of 93%, followed by the SVM classifier with 89%, and Naïve Bayes with 86%. The F1score weighs the performance of each classifier between their precision and sensitivity. This result suggests that the random forest classifier is balanced in terms of sensitivity and precision. Whereas the SVM classifier, although it is excellent in predicting non-diabetic patients, it has a slight tendency to falsely predict diabetic individuals as non-diabetic.

3.2. Roc Analysis

ROC is a plot of TPF (sensitivity) and FPF (1-specificity) with each point on the ROC curve representing a sensitivity/specificity pair corresponding to a particular probability threshold. The closer the curve to the upper left corner of ROC space, the higher the overall accuracy of the test. Subsequently, the closer the curve comes to the 45-degree diagonal (TPR = FPR) of the ROC space, the lesser the overall accuracy of the test [22]. In this study, the ROC curve was plotted from testing dataset with 90 samples consisting of diabetic and non-diabetic individuals. The purpose of the ROC was to provide further analysis on the performance of each classifier. From the plot in fig. 5, random forest curve located the closes to the upper left corner of the plot indicating higher accuracy in terms of correctly predicting diabetic and non-diabetic groups. The performance of each classifier can be better illustrated using the area under curve (AUC) metric. The area under curve (AUC) summarizes the entire ROC curve rather than at specific threshold. Thus, AUC measurement represents the overall accuracy of the models. In general, the closer AUC is to 1, the better the accuracy of the performing model. From the calculations, the random forest classifier showed the best accuracy among other classifier with an AUC value of 0.93, followed by Naïve Bayes classifier with a value of 0.84. Finally, the SVM classifier scored the lowest AUC value of 0.82. Standard error showed sampling fluctuations and as the data becomes more balance, standard error value decreases. Overall, the model from SVM outperforms the random forest classifier in terms precision and specificity measurement. But, in terms of accuracy, sensitivity, AUC and F1 score random forest classifier were the highest. Since diabetes classification rely on these measurements even more for predicting outcome, this study suggests that the random forest model is more suitable in predicting type 2 diabetes of high potential individual from local population (IIUMMC).

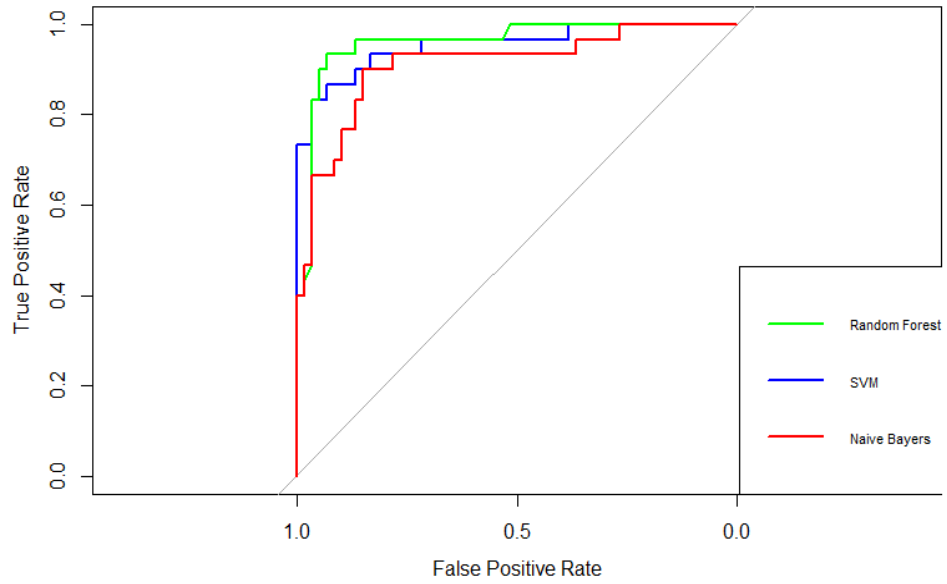

Fig. 5. Receiver Operating Characteristic (ROC) performance of the three classifiers; (red) Naïve-Bayes, (blue) Support Vector Machine (linear) and (green) Random Forest

Table 3: *The AUC and Standard Error of different models. (Delong et al, 1998)*.*

Models	**AUC (p=0.5)**	**Standard Error***
Random forest	0.93	0.0016
SVM	0.82	0.0011
Naïve Bayes	0.84	0.0012

4. CONCLUSION REMARKS

In this study, prediction model was constructed based on machine learning concept that focuses on Malaysian dataset. The proposed model is able to predict the likelihood of diabetes in high risk patients with high accuracy. Therefore, three machine learning classification algorithms namely Random Forest, SVM and Naïve Bayes were used in this study to detect type 2 diabetes at an early stage. The study was performed on IIUMMC dataset which is sourced from medical records department of IIUM Medical Centre for model evaluation. Before model fitting, several combinations of well-known pre-processing methods such as missing value imputation, data scaling, feature selection and normalization were implemented to the dataset. Following this, machine learning techniques were used on IIUMMC datasets to predict the outcome. The performances of all three algorithms were evaluated using various measurement such as Precision, Accuracy, F-Measure, and Recall. The results obtained suggest that the random forest classifier achieved the best overall performance with prediction accuracy of 91%, specificity of 86%, sensitivity of 93%, precision of 93%, and an F-measure of 93%, using IIUMMC dataset and 10 cross validation. The study also illustrates how each pre-processing step hugely affects the performance of the model such as in feature selection and data sampling. While this study focuses on a small

subset of a local population, it is possible to directly expand the methods to a wider subset of the Malaysian population.

ACKNOWLEDGEMENTS

The work in this paper has been funded by the International Islamic University Malaysia through grants from the IIUM Research Initiative Grant Scheme (RIGS17-011-0586). Data collection has been done at the IIUM Medical Centre, Kuantan. Consent for research conduct in this study has been approved by the IIUM Research Ethics Committee (IREC 2019-133).

REFERENCES

[1] Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J. E., & Makaroff, L. E. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, *128*:40–50.

[2] Institute for Public Health (2011). The Fourth National Health and Morbidity Survey Vol. II: Non-Communicable Diseases. *Ministry of Health Malaysia*, 188.

[3] Hussein, Z., Taher, S. W., Gilcharan Singh, H. K., & Chee Siew Swee, W. (2015). Diabetes Care in Malaysia: Problems, New Models, and Solutions. *Annals of Global Health*, *81*(6):851–862.

[4] Chan, Y. Y., Lim, K. K., Lim, K. H., Teh, C. H., Kee, C. C., Cheong, S. M., Khoo, Y. Y., Baharudin, A., Ling, M. Y., Omar, M. A., & Ahmad, N. A. (2017). Physical activity and overweight/obesity among Malaysian adults: findings from the 2015 National Health and morbidity survey (NHMS). *BMC Public Health*, *17(1),* 733.

[5] American Diabetes Association (2010). Diagnosis and classification of diabetes mellitus. *Diabetes care*, *33, Suppl 1:*S62–S69.

[6] Jothi, N., Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare – A Review. *Procedia Computer Science*, *72*:306 – 313.

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2016). Machine Learning and Data Mining. *Computational and Structural Biotechnology Journal*, *15*:104-116

[8] Fayyad U, Piatetsky-Shapiro G, Smyth P (1996). From data mining to knowledge discovery in databases. *AI Magazine*, fall:37–54.

[9] Awad, M., & Khanna, R. (2015). Efficient Learning Machines. Berkeley, Apress.

[10] Kaur, H., & Kumari, V. (in press). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*.

[11] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine, 17*, 195.

[12] Abdullah, N., Abdul Murad, N., Attia, J., Oldmeadow, C., Kamaruddin, M., Abd Jalal, N., Ismail, N., Jamal, R., Scott, R., & Holliday, E. (2018). Differing Contributions of Classical Risk Factors to Type 2 Diabetes in Multi-Ethnic Malaysian Populations. *International Journal of Environmental Research and Public Health*, *15*(12), 2813.

[13] Steyerberg, E. W., Uno, H., Ioannidis, J. P. A., van Calster, B., Ukaegbu, C., Dhingra, T., Syngal, S., & Kastrinos, F. (2018). Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of Clinical Epidemiology*, *98*:133–143.

[14] Famili, A., Shen, W., Weber, R., & Simodis, E. (1997). Data pre-processing and intelligent data analysis. *Intelligent Data Analysis*, *1*(1–4): 3–23.

[15] Bannon, W. (2015). Missing data within a quantitative research study: How to assess it, treat it, and why you should care. *Journal of the American Association of Nurse Practitioners*, *27*(4):230–232.

[16] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*:321–357.

[17] Gitman, I., & Ginsburg, B. (2017). Comparison of Batch Normalization and Weight Normalization Algorithms for the Large-scale Image Classification. *Computer Vision and Pattern Recognition*, *1709,* 8145.

[18] Novakovic, J., Strbac, P., & Bulatovic, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, *21*(1):119–135.

[19] Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*:1578–1585.

[20] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4):427–437.

[21] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontiers in Genetics*, *9*, 515.

[22] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8):861–874.