

Comparative analysis in the prediction of early-stage diabetes using multiple machine learning techniques

Leonard Flores^{1,3}, Rowell Marquez Hernandez^{2,3}, Lloyd H. Macatangay^{2,3}, Shiela Marie G. Garcia³,
Jonah R. Melo³

¹College of Arts, Sciences and Technology, Occidental Mindoro State College, San Jose, Philippines

²National Research Council of the Philippines, Division of Engineering and Industrial Research (VII), Batangas City, Philippines

³College of Informatics and Computing Sciences-Batangas State University, Batangas City, Philippines

Article Info

Article history:

Received Dec 18, 2021

Revised Jul 18, 2023

Accepted Jul 31, 2023

Keywords:

Comparative analysis

Diabetes prediction

Feature selection

Neural network

Random forest

Support vector machines

ABSTRACT

Diabetes is caused by high levels from blood glucose and it is characterized as a chronic disease, and also will disrupt fat and protein absorption. The levels rise from blood glucose because it cannot be burned in the cells from the pancreas because of the deficiency of insulin secretion or the insulin produce by the cell are insufficient. By means of early detection, it may decrease the hazards and frequency of diabetes. The application of technology has been an essential part of providing accurate and acceptable results in the prevention and early detection of the illness. This research provided the best machine learning used for predicting the early stage of diabetes. The methods involve the feature selection or dimension reduction using relief-based filter (reliefF), tenfold cross-validation for testing and training data, and different machine learning classifiers such as the random forest (RF), support vector machines (SVM), and neural network (NN) is used. In this research, RF recorded the highest precision point at 98.5%, which was able to provide a higher evaluation in terms of accuracy followed by SVM at 96.6% and NN at 96.2%. The results generated from this experiment are essential in contributing a new way that is highly accurate in determining diabetes among patients.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Rowell Marquez Hernandez

College of Informatics and Computing Sciences-Batangas State University

Batangas City, Philippines

Email: rowell.hernandez@g.batstate-u.edu.ph

1. INTRODUCTION

Diabetes is caused by high levels from blood glucose and it is characterized as a chronic disease, and also will disrupt fat and protein absorption. The levels rise from blood glucose because it cannot be burned in the cells from the pancreas because of the deficiency of insulin secretion or the insulin produce by the cell are insufficient [1]. Many people are not fully aware of having early-stage diabetes symptoms, it causes frequent urination, intensifies thirst and hunger. The factors in having a diabetes such as weight, genes, and insulin, but the most of time is sugar absorption, and for preventing diabetes the better way is to detect early as possible.

To determine the early stage of diabetes by providing correct classifier method and important significant attributes. Some proponents used machine learning classification to diagnose diseases [2]. The machine learning researchers used are random forest (RF), decision tree, naïve bayes, support vector machines (SVMs), and so on, as the result of their study the models or algorithms in machine learning are most efficient and effective in diagnosing diseases [3]–[5]. Machine learning and data mining handle large data from various sources and study context information are incorporated [6]–[8].

Objectives of the study:

- To train and test data in different machine learning models and compare their performance using the early-stage diabetes 2020 dataset in terms of accuracy, precision, confusion matrix, recall, and specificity, and to train and test the data using tenfold cross-validation to see the unseen data.
- To propose a fine-tuning method using relief-based filter (ReliefF) for feature selection to optimize accuracy performance.
- To analyze a model that can distinguish the likelihood of early-stage of diabetes in patients with the best result of accuracy.
- To address the importance of machine learning in health issues to decrease the hazardous and prevalence of diabetes.

To determine the prediction of diabetes of patients, machine learning models such as RF, neural network (NN), and SVM are used and tested on the early-stage diabetes 2020 dataset. On different scales, the experimental results of all three algorithms are evaluated based on their accuracy, precision, recall, and specificity [9]. The dataset is approved by a doctor from Sylhet Diabetes Hospital in Sylhet, Bangladesh.

2. RELATED LITERATURE

Diabetes, a chronic metabolic disorder characterized by abnormal blood glucose levels, has been the subject of extensive research aimed at understanding its causes, developing effective treatments, and improving patient outcomes. Related works in diabetes encompass a wide range of topics, including disease diagnosis, glycemic control, risk prediction, management of diabetic complications, and the exploration of digital health solutions. Researchers have focused on early diagnosis through the identification of novel biomarkers and genetic risk factors, enabling timely interventions and personalized treatment plans. Moreover, studies have explored various insulin therapies, including continuous glucose monitoring and artificial pancreas systems, to optimize glycemic control for patients. Predictive models have been developed to assess an individual's risk of developing diabetes, taking into account factors such as age, body mass index (BMI), family history, and blood glucose levels.

2.1. Related works in diabetes

Based on the study using the early-stage diabetes dataset the result of accuracy was achieved percentage split technique with 97.4%, by applying 10-fold-cross validation and RF algorithm without using feature selection [10]. Researchers conclude that the performance of all three algorithms is assessed using various metrics such as precision, accuracy, F-measure, and recall. In comparison to other algorithms, the results show that naive bayes has the highest accuracy of 76.30% [11]. Based in the findings of Alam *et al.* [12] the k-means, artificial neural networks (ANNs), and RF algorithms technique was used to predicted diabetes. The highest accuracy percentage is ANN with having a 75.7% and could be useful in assisting medical professionals with treatment decisions [12].

WEKA and LIBSVM were used for feature selection; in WEKA, the wrapper and ranker methods were used; accuracy of 71% was achieved using the wrapper method, and accuracy of 72% was obtained using the ranking method; and in LIBSVM, the fselect script was used for feature selection; accuracy of 63% was obtained using the LIBSVM Fselect script [13]. However, both the SVM and the Naive Bayes algorithms demonstrated high overall classifier performances of 95.52% and 94.52%, respectively [14]. The RF and decision tree have the highest specificity percentage of 98.00% and 98.20%, for the classification of diabetic data. Additionally, naïve bayesian has the highest percentage in terms of accuracy with 85.30%. To improve classification accuracy. The study generalizes features selection from a dataset [15].

Based on the study of Kandhasamy and Balamurali [16] RF and k-nearest neighbour (KNN) has the highest accuracy provided 100%. Researchers conclude that removing noisy data has a big effect in generating the best result. For future studies, the said proposed method will be able to use for any disease study a suitable dataset [16]. In addition, diabetes risk classification is based on patient symptom information using ANN. The SCG backpropagation or scaled conjugate gradient technique was used for NN [17].

2.2. Related works in random forest algorithm

The RF technique was established by Breiman [18], and it is a very effective standardized approach for classification and regression tools. The technique used in circumstances that the number of variables is significantly more than the number of observations, to perform well, the average prediction was combined with multiple randomized decision trees. The RF algorithm can efficiently classify large amounts of data [18]. It's a statistical learning theory-based strategy that uses produces several copies of sample sets from the original training datasets performed by bootstrap randomized re-sampling. The RF algorithm is linked to the following works.

Based on the study of Chen *et al.* [19], for risk analysis, using training model of RF and the optimization experiment are used, RF algorithm with a maximum of 0.86 reach the classification accuracy, in terms of large-scale activities it will good predictive analysis in the risk assessment [19]. However, two algorithms are used in their research. The NN with 72.33% indicates the average recognition accuracy in terms of backpropagation, However, the highest feasibility accuracy in this paper was reaching 98.49% by RF model [20].

The models' accuracy is determined, and evaluations are based on efficiency calculations. The accuracy of the naïve bayesian classification network for diabetes reaches 74.46%, coronary heart disease reaches 82.35%, and cancer data reach 63.74%, respectively [21]. However, classification using the RF model yields accuracy values of diabetes reaches 74.03%, coronary heart disease reaches 83.85%, and cancer data reaches 92.40%. The accuracy percentage of RF algorithm is higher than naïve bayes for three illnesses. Furthermore, with MRI structure, cognitive function, and patient data, the researchers' RF model demonstrated good classification accuracy for distinguishing AD healthy controls and moderate cognitive impairment (MCI) (HC vs. MCI 80.8 %; area under curve (AUC) 0.88, HC vs. AD 93.5%, AUC 0.99). Furthermore, researchers found that the used algorithm's segmentation processing time of five minutes was considerably quicker than Freesurfer's reaching to six to eight hours [22].

2.3. Literature in support vector machines algorithm

The SVM as frequently used in classification. The purpose of SVM is to find the gap in the hyperplane from the best highest-margin between two classes. The hyperplane of data points from the other class should not be closer [23]. The data points far from hyperplane in each category must be chosen. The support vectors are the points closest to the margin of the classifier [24]. The researcher developed a strategy for finding a suitable feature subset and SVM classifier using the genetic algorithm (GA) to increase classification accuracy [25]. The least square support vector machine (LS-SVM) and decision support system in medical based from GA for subset selection features was demonstrated for diabetes diagnosis [26]. One key application area where precision is critical is medical mining. K-means and GA were used to reduce dimensionality, while SVM was used to categorize the diabetes dataset [27]. The following are related works of literature for support vector machines.

Based on the conclusion of Vijayarani [28] these algorithms are evaluated based on the performance variables of classification accuracy and completion time. Based on the experimental findings, this paper shows that the SVM classifier is the best method due to its high classification accuracy. In comparison to other classifiers, the naïve bayes classifier requires the shortest execution time [28]. In addition, with excellent sensitivity and specificity, the SVM method supported a genetic foundation for diagnosis. Researchers demonstrated that gene expression profiling is beneficial in the categorization and diagnosis of soft tissue sarcoma, offering insights into pathophysiology and pointing to possible novel treatment targets [29]. Based on the study of Osisanwo *et al.*, the findings reveal that SVM is the method with the highest precision and accuracy. The most accurate algorithm classification are naïve bayes, SVM, and RF. Factors, on one hand, are the time spent building a model and precision, on the other hand, mean absolute error (MAE) and kappa statistics are also a factor based on their research. To have supervised predictive machine learning, ML algorithms must have precision, accuracy, and a low error rate [30]. However, naïve bayesian classification obtains the greatest accuracy and specificity among all classifiers. However, the SVM approach achieved the worst classification since it has the lowest sensitivity, accuracy, and specificity. Furthermore, when compared to their contemporaries, SVM obtains the greatest error rate with a considerable number of false positive (FP) and false negative (FN) situations [31].

2.4. Related works in neural network algorithm

It is possible to conclude that the ANN obtains improved classification performance and produces accurate results, and therefore it is regarded as the best classifier when compared to the SVM classifier method. SVM classifiers can potentially classify data in a short amount of time. In the future, the ANN algorithm will be improved to reduce execution time [30]. In addition, small blue round cell tumors (SBRCTs) are a diverse collection of tumors with overlapping morphologic, immunohistochemical, and clinical characteristics that make diagnosis challenging. CIC-DUX4-related fusions are seen in almost two-thirds of EWSR1-negative SBRCTs, with a small minority having BCOR-CCNB3 X-chromosomal paracentric inversion. The classification of the proposed study reached 100% accuracy; the said study was used the ANN model to predict SBRCT cancer [31].

The survey clearly illustrates the efficacy of NN methods in cancer diagnosis. The majority of NN produce excellent results when it comes to properly classify tumor cells. multi-layer perceptron (MLP) has a 97.1% accuracy, probabilistic neural network (PNN) has a 96% accuracy, perceptron has a 93% accuracy, and ART1 has a 92% accuracy. Findings improve when missing values are removed from the dataset. The results of NN structures can be improved by properly configuring NN parameters. Although NN approaches have a

high classification rate, their training period is quite long [32]. However, SVM with 82% gave correct predictions and in terms of Matthews cc reached 0.63 of total 525 descriptors, however, ANN got 80% gave correct predictions in terms of Matthews cc it reached 0.58. SVM overpower ANN classifiers in terms of FN, true positive (TP), true negatives (TN), and FP, the output of the two classifiers the SVM and ANN are not similar. Two classifiers training theory is briefly discussed [33].

3. MATERIALS AND METHOD

The combination of feature selection techniques and state-of-the-art machine learning algorithms allows researchers to create parsimonious and interpretable models with enhanced generalization capabilities. By reducing the number of input features and focusing on the most informative ones, the risk of overfitting is mitigated, resulting in more reliable and clinically applicable classification systems. The thorough evaluation using multiple statistical validation measurements provides a comprehensive assessment of model performance, enabling researchers to make data-driven decisions and select the most suitable algorithm for diabetes classification.

3.1. Model diagram

Figure 1 shows the workflow of the research used to create the model below is shown in the diagram. In the early stage of diabetes research, relevant patient data is collected to build a dataset. Pre-processing techniques handle missing values, outliers, and normalize data. Two applied ML algorithms, SVM and RF, are used for diabetes classification. Performance evaluation measures, including accuracy, precision, recall, F1-score, and AUC-ROC, are employed for comparative analysis. Results show high accuracy, indicating the potential of these models for early diabetes detection.

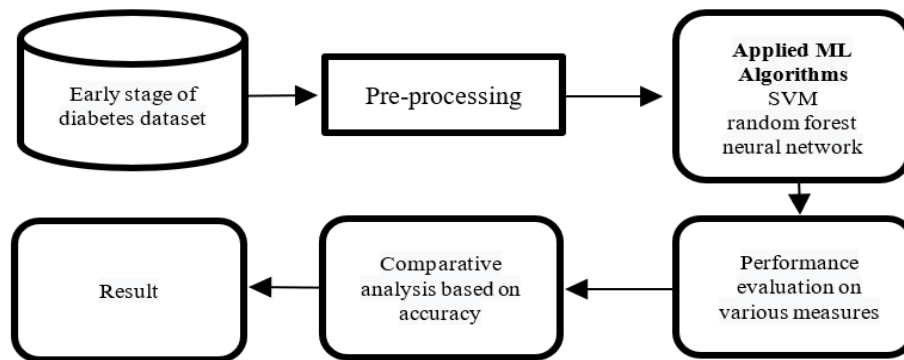


Figure 1. Model diagram

3.2. Dataset details

According to Shirol [34], the dataset came from Sylhet Diabetes Hospital in Sylhet, Bangladesh patients. The research was conducted using direct questionnaires from the patients. This dataset report contains 520 persons with diabetes-related symptoms. Including data that peoples have symptoms that may cause diabetes. Dataset has an attribute of 16 and 520 instances. The dataset was validated by a doctor from Sylhet Diabetes Hospital in Sylhet, and also the data set was recorded in the year 2020 [34].

Table 1 shows the dataset that consists of various attributes related to diabetes diagnosis. The “Class” attribute indicates whether the patient is positive (1) or negative (2) for diabetes. Other attributes include “Obesity,” “Alopecia,” “Muscle stiffness,” “Partial paresis,” “Delayed healing,” “Irritability,” “Itching,” “Visual blurring,” “Genital thrush,” “Polyphagia,” “Weakness,” “Sudden weight loss,” “Polydipsia,” and “Polyuria.” Each of these attributes can take the values of either 1 (yes) or 2 (no), representing the presence or absence of the corresponding symptom or condition. Additionally, the “Sex” attribute has two values, 1 for male and 2 for female. The “Age” attribute is divided into four categories: 1 for 20-35 years old, 2 for 36-45 years old, 3 for 46-55 years old, and 6 for above 65 years old. These attributes and values form the basis for building a classification model to predict diabetes diagnosis based on patient data.

Table 1. Description of attributes

Attributes	Values
Class	1. Positive, 2. Negative
Obesity	1. Yes, 2. No
Alopecia	1. Yes, 2. No
Muscle stiffness	1. Yes, 2. No
Partial paresis	1. Yes, 2. No
Delayed healing	1. Yes, 2. No
Irritability	1. Yes, 2. No
Itching	1. Yes, 2. No
Visual blurring	1. Yes, 2. No
Genital thrush	1. Yes, 2. No
Polyphagia	1. Yes, 2. No
Weakness	1. Yes, 2. No
Sudden weight loss	1. Yes, 2. No
Polydipsia	1. Yes, 2. No
Polyuria	1. Yes, 2. No
Sex	1. Male, 2. Female
Age	1.20-35, 2.36-45, 3.46-55, 6 above 65

3.3. Features selection

The nearest neighbors in the space are identified it's all characteristics via relief features selection model [35]. The (1) the distance between instances R_i and R_j is calculated in the space of all attributes $\alpha \in A$, generally using a Manhattan ($q=1$) metric, but it may alternatively be calculated using a Euclidean ($q=2$) metric: $D_{ij} = (\sum_{\alpha \in A} |diff(a(R_i, R_j))|^q)^{1/q}$, in which the usual "diff" shown in (2) formula for a real-valued attribute a between two instances R_i and R_j is: $diff(a(R_i, R_j)) = \frac{|value(a, R_i) - value(a, R_j)|}{max(a) - min(a)}$. This difference is suitable for gene expression and other genuine predictors. For genome-wide association study (GWAS) data with categorical characteristics, just modify the diff, but the algorithm remains unchanged. The diff function is needed by relief techniques to construct the distance matrix for locating nearest hit and miss neighbors, but it is also required to compute the relief significance scores [36].

3.4. 10-fold cross-validation

The given number of folds split from the data is called 10-fold cross-validation, which is usually 5, 10, or 20 folds. From one at a time, 10-fold cross-validation was tested by holding out examples; to classify the model from the held-out point the other folds and examples are induced. To test the model's simplification ability the 10-fold cross-validation was a highly recommended procedure to use where there is insufficient data for machine learning and data mining model development. Preventing the NN, SMV, and RF, and other machine learning algorithms from over-fitting during the training process, this preventing process was done by 10-fold cross-validation. To conduct a 10-fold cross-validation process, model development data were first separated into 10 nearly equally sized segments or folds. After the ten folds procedure was done validation and ten iterations of training are performed, with each iteration holding a separate fold of the data for validation, while the remaining nine folds are utilized for learning, and the learned models are then used to forecast the data in the validation fold. As a conclusion, a model is created and evaluated each time with an "unseen" dataset. The performance of each learning algorithm on each fold may be tracked using certain predefined performance functions. For particular, averaging may be used to create an overall evaluation from these samples, or these samples can be utilized in a statistical hypothesis test. Because this technique necessitates much more computing work than a typical trained-and-tested approach, it may perform reliable and impartial testing on small datasets.

3.5. Machine learning algorithms

3.5.1. Random forest

The classification algorithm that classifies data from uses numerous decision trees is called RF. It selects random samples B times with replacement of the training set for a training set of $X = x_1, \dots, x_n$ and $Y = y_1, \dots, y_n$ and fits trees to these samples. After training, it averaged all of the different regression trees' predictions for x' and making predictions for unseen samples x' and, like with classification trees, taking the majority result. $\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x')$.

3.5.2. Support vector machines

The SVM are machine learning algorithm that divides the attribute space with a hyperplane, optimizing the gap between occurrences of distinct classes or class value. The approach frequently produces

excellent predicted performance results. It is becoming increasingly used in the disciplines of pattern recognition and machine learning. In the input space, classification is achieved by creating a linear or non-linear separation surface. As demonstrated in (4), the separating function in support vector classification may be represented as a linear combination of kernels associated with the support vectors. $f(x) = \sum_{x_j \in S} a_j y_j K(x_j, x) + b$ where x_i denotes the training patterns, $y_i \in \{+1, -1\}$ denotes the corresponding class labels and S denotes the set of support vectors [37].

3.5.3. Neural network

An input layer of neurons or nodes, one, two, or even three hidden layers of neurons, and output neurons make up the last layer of the artificial NNs three layers. Simple architecture with lines connecting neurons is shown in Figure 2. A weight is assigned in each connection, which is a numeral value h_i is the output of neuron I in the hidden layer $h_i = \sigma(\sum_{j=1}^N V_{ij} x_j + T_i^{hid})$, the function of activation or transfer where $\sigma()$, the number of input neurons is N , V_{ij} are the weights, x_j are the inputs to the input neurons, and T_i^{hid} are the hidden neurons' threshold terms. In addition to providing nonlinearity into the NN, in order to prevent the NN from freezing due to split neurons, the activation function must remain [38]. The sigmoid or logistic function is a typical illustration of an activation function as $\sigma(u) = \frac{1}{1 + \exp(-u)}$.

	#	Relieff
C Polyuria	2	0.482
C Polydipsia	2	0.406
C Gender	2	0.298
C Alopecia	2	0.188
C visual blurring	2	0.178
C Itching	2	0.152
C partial paresis	2	0.134
C Irritability	2	0.120
C sudden weight loss	2	0.086
C Polyphagia	2	0.070
C Genital thrush	2	0.064
C muscle stiffness	2	0.042
C delayed healing	2	0.036
C Obesity	2	0.034
N Age		0.028
C weakness	2	0.022

Figure 2. Likelihood attributes with corresponding value

3.6. Confusion matrix

Table 2 shows the confusion matrix that describes the overall performance of the model and provides a matrix result. In this case, TP accurately anticipated the dataset's positive real values. Error rate of predicted values for incorrectly forecasted the positive real values, also known as FP. The FN, or error rate, refers to incorrectly projected numbers that are negative in reality. The dataset's negative real values were predicted using TN [39].

Table 3 shows the evaluation of the performance of each model using the elements of the confusion matrix. The classification models' accuracy, precision, recall, and specificity may all be calculated. Several measures for evaluating the performance of different categorization methods may be extracted from the confusion matrix. Consideration of classification accuracy or its balance error rate is a popular assessment technique. Several more metrics are calculated and compared based on the values in the confusion matrix. The algorithm's accuracy measures the likelihood of properly predicting positive and negative records. Precision is computed by dividing the number of correct positive outcomes by the number of positive outcomes predicted by the classifier. Precision is defined as the proportion of relevant results in the list of all returned search results. The probability of correctly predicting negative cases is referred to as specificity [40].

Table 2. Confusion matrix elements

		Predicted value	
		Negative	Positive
Actual value	Negative	TN	FN
	Positive	FP	TP

Table 3. Performance measure

Measure	Derivations
Accuracy	$TP+TN/TP+TN+FN+FP$
Precision	$TP/FP+TP$
Recall	$TP/FN+TP$
Specificity	$TN/TN+FP$

4. RESULTS AND DISCUSSION

By identifying relevant features, feature selection is essential for improving the performance of machine learning models. The model’s classification performance may be evaluated to see how well it performs the task at hand. Dealing with overfitting keeps the model from being too complicated and generalizing poorly, which assures impartial prediction outcomes.

4.1. Features selection

We used the relief features selection method for dimensional reduction to select the best attributes in our experimentation. Use a value as a proxy for the feature to rank all the characteristics, and then use an arbitrary cutoff to determine the feature subset. The predictive or subjective likelihood of relevance, or simply the number of characteristics in the target subset, can be used to calculate this cutoff. The following are the result of all attributes with define value. Based on the result, the ‘Weakness’ attribute had the lowest subjective likelihood significance having 0.022. Furthermore, the ‘Weakness’ attribute is no longer one of the attributes utilized in the dataset because of unlikelihood symptoms for the early stage of diabetes, additionally, it affects the result of accuracy in getting high accuracy percentage in the machine learning model. As a result, from the initial dataset of 16 attributes, a total of 15 attributes and 468 instances were utilized by applying features selection for dimensional reduction.

4.2. Machine learning models

4.2.1. Random forest

To provide accurate prediction without overfitting the data is used by RF and it is a new classification method that uses ensemble learning. All of the trees’ outputs are combined to form a single classification. The number of trees depends on the number of rows in the data set. The more rows in the data set, the more trees are needed. In this experiment, we are using 100 to 1,000 trees rather than 10 to 1,000 trees. It is because, when doing a hyper-parameters search, this saves a significant amount of processing time. The final prediction is the average of the 10 RFs trained with internal 10-fold cross-validation. In this approach, I iterate through all of the trees in the forest, computing predictions from single trees. The RF prediction is the average of all trees in the subset. In this experiment, we started with 100 trees and worked our way up to 1,000 trees. From 100 to 500 trees, an accuracy average percentage of 98.2% was obtained. However, we have a 98.5% accuracy percentage from 600 to 1,000 trees. The findings suggest that the more trees you utilize, the greater the accuracy percentage becomes. As a result, the highest accuracy percentage among the number of trees are 600, 700, 800, 900, and 1,000 having a 99.5% for the AUC, 98.5% for the CA, 98.5% for the precision, 98.5% for the recall, and 98.0% for the specificity, with having a parameter of do not split the subsets smaller than 2 that shown in Figure 3. Figure 4 shows the result of RFs performance in terms of proportion of predicted and error rate and brings the best-unbiased prediction and handled overfitting.

		Predicted		Σ
		Negative	Positive	
Actual	Negative	98.9 %	1.7 %	179
	Positive	1.1 %	98.3 %	289
Σ		176	292	468

Figure 3. Random forest performance

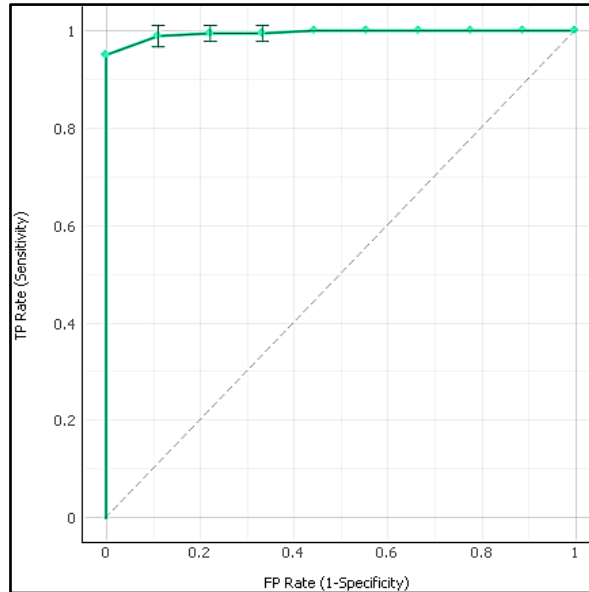


Figure 4. Classification of random forest

4.2.2. Neural network

In this experiment, we test the NN based on the perception model, the model used for the NN are the ReLu, tanh, logistic, and identity activation functions. Additionally, we included also the optimization solver such as Adam, Stochastic gradient descent (SGD), and L-BFGS-B for better accuracy results. To provide better results in an AUC, accuracy, precision, recall, and specificity. We provide all the results based on the perception model and an optimization solver was applied. The following are the classification results of AUC, accuracy, precision, recall, and specificity. It shows that Adam and ReLu got the highest AUC percentage of 99.3%. While Adam with tanh provided the 98.9% AUC percentage, and also L-BFGS-B with tanh got 98.3%, SGD with ReLu provide the least AUC percentage with 92.1%. Additionally, the result for the accuracy, SGD with ReLu got the highest percentage with 96.2%, the second was the L-BFGS-B with tanh having 95.7%. The third highest accuracy was Adam with ReLu with 95.5%. But the least accuracy percentage got 89.7% with SGD and logistic. With regards to the highest percentage for precision, L-BFGS-B with ReLu provide 96.2%. The second highest percentage was L-BFGS-B with the having 95.7%. However, the least precision is the SGD and logistic having an 89.7%. Furthermore, with regards to recall the highest percentage having a 96.2% is L-BFGS-B and ReLu. While the second highest is L-BFGS-B with tanh providing 95.7%. Moreover, the least recall got the percentage of 89.7% is SGD and logistic. The last performance statistic used in this experiment is specificity. Moreover, Figure 5 shows the highest specificity percentage is L-BFGS-B and tanh provides 95.4%. Secondly, L-BFGS-B with ReLu got a specificity percentage of 95.3%. However, SGD and logistic is the least specificity percentage among the activation and solver used, having an 88.5%.

To sum up the result, Adam and ReLu got the highest AUC percentage having a 99.3%. With regards to accuracy L-BFGS-B and ReLu provide 96.2%. And also, L-BFGS-B and ReLu got the highest percentage for precision with 96.2%. Moreover, L-BFGS-B and ReLu also have the highest recall percentage provide 96.2%. Lastly, L-BFGS-B and ReLu got the highest specificity percentage with 95.3%. The result for performance statistics was the majority to L-BFGS-B optimization solver and ReLu activation function. Figure 6 shows the performance of NN for the proportion of predicted and error rate, and handled overfitting.

		Predicted		Σ
		Negative	Positive	
Actual	Negative	96.0 %	3.1 %	179
	Positive	4.0 %	96.9 %	289
Σ		177	291	468

Figure 5. Neural network performance

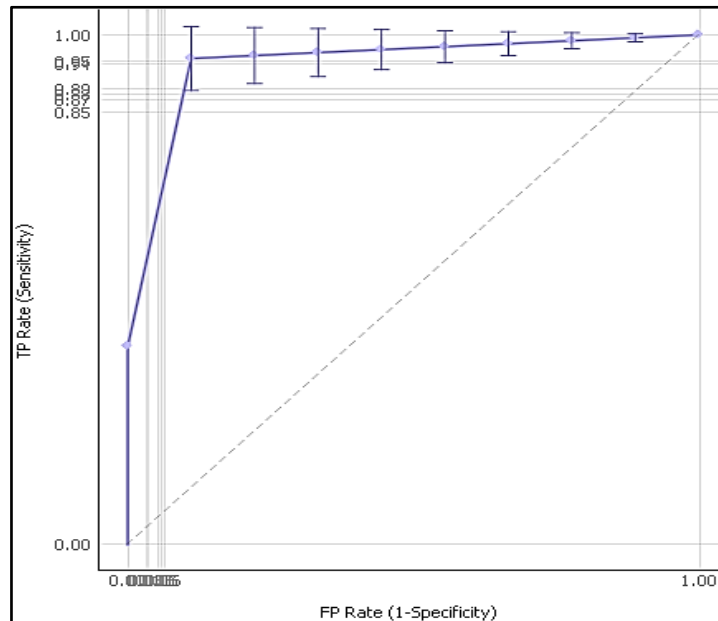


Figure 6. Classification of neural network

4.2.3. Support vector machines

The technique typically yields findings with strong predictive performance is SVM. A hyperplane is used by the SVM, a machine learning technique, to partition the attribute space and increase the distance between instances of different classes or class values. Orange includes a widely used SVM implementation from the LIBSVM package. Its graphical user interface is this widget. The following are the classification results of AUC, accuracy, precision, recall, and specificity.

Figure 7 shows the result of support vector machines, with regards to linear kernel with cost (c) parameter the highest accuracy percentage having 91.7% at cost parameter 0.1. The second is cost parameter 0.2 and 0.3 with 89.3%. The third lowest accuracy percentage is cost parameter 0.4 got 84.0%. Lastly, the least accuracy percentage for linear kernel with cost parameter of 0.5 with 84.0% and, 0.6, 0.7, 0.8, 0.9, and 1 they have all got 80.8%. Moreover, the polynomial kernel with cost (c) parameter 0.6 got the highest accuracy percentage with 96.4%. While the second-highest accuracy percentage is cost parameter 0.4 and 0.5 with 96.2%. The third is 0.7, 0.8, 0.9, and 1 cost parameter having a 95.9%. However, the least accuracy percentage for polynomial kernel with cost parameter 0.1 with 94.9%. Furthermore, 0.6 cost parameter got the highest accuracy percentage having 96.2% for RDF kernel, the second-highest is 0.5 and 0.7 having a 95.9%. Third highest accuracy percentage with 95.5% is cost parameter 0.4, 0.8, 0.9, and 1. Lastly, the least accuracy percentage is 0.1 having a 93.8%. With regards to sigmoid kernel, the highest accuracy percentage for cost parameter is 0.4 having an 84.6%. The second is 0.3 cost parameter with 84.2%, and the third-highest accuracy are 0.2, 0.6, and 0.7 cost parameter they are both got 84.0%. However, the least is cost parameter 0.9 with 82.1%. As a result, for the support vector machines, the polynomial kernel with cost parameter 0.6 got the highest accuracy percentage among the other kernel method used such as linear, RDF, and sigmoid kernel [37].

Figure 7 and Figure 8 shows the performance of SVM for the proportion of predicted and error rate, and handled overfitting is shown below. The distribution of the values for the confusion matrix corresponds to the performance of the classification algorithm. Moreso, it is evident that SVM provided an exemplary result.

		Predicted		Σ
		Negative	Positive	
Actual	Negative	72.2 %	10.8 %	179
	Positive	27.8 %	89.2 %	289
Σ		209	259	468

Figure 7. SVM performance

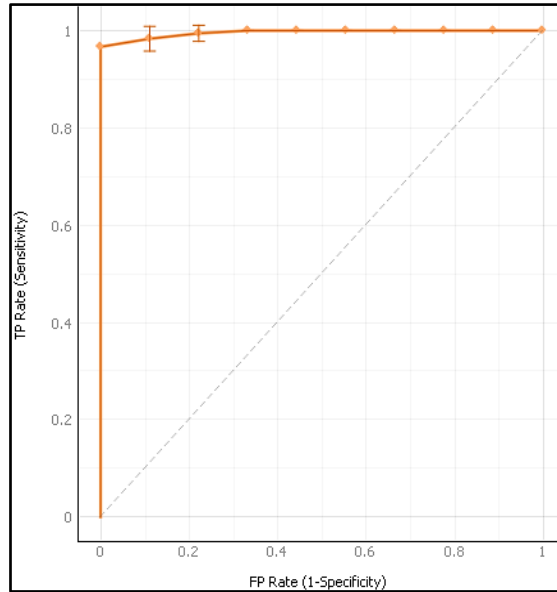


Figure 8. Classification for support vector machines

4.3. Comparative of model performance

Table 4 shows the result classification proposed model using 10-fold cross-validation with feature selection. It shows that RF having a 98.5% got the highest accuracy percentage, followed by SVM with 96.6% and lastly the NN got 96.2%. Furthermore, Figure 9 shows below the RF delivered unbiased prediction and handled overfitting among the three classifiers.

Table 4. Classification model results using 10-fold cross validation and features section

Proposed model	AUC	Accuracy	Precision	Recall	Specificity
RF	99.5%	98.5%	98.5%	98.5%	98.0%
NN	99.3%	96.2%	96.2%	96.2%	95.3%
SVM	99.1%	96.6%	96.6%	96.6%	96.0%

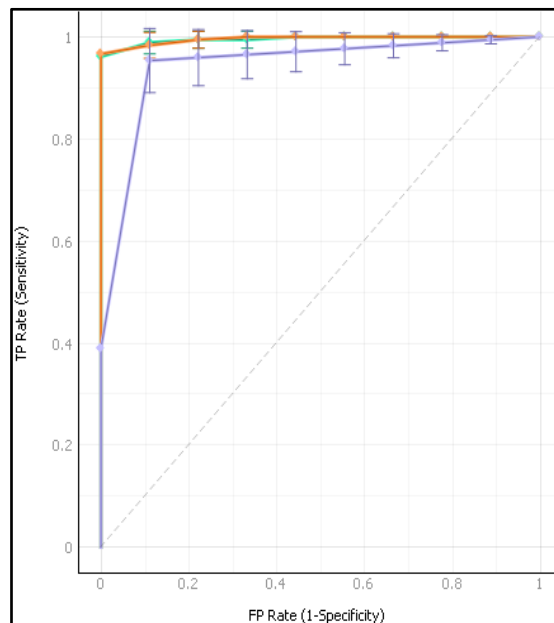


Figure 9. Classification comparative model

5. CONCLUSION AND RECOMMENDATION

The best accuracy classification result is important for the early stage of diabetes. It will prevent the patient from severe diabetic damage that causes damage to many parts of the body, including the eyes, heart, feet, nerves, and kidneys. In the worst scenario, it can cause fatality. However, experts can prevent this situation. The advent of new technology aligned with machine learning applications can be a great help in improving the early detection of diabetes. This study employed three machine learning classification models used are RF, SVM, and NN. It has been tested with the early-stage diabetes dataset. Each method has different parameters to determine the best classification results in terms of accuracy, AUC, recall, precision, and specificity. In this study, the outstanding result that attained the highest accuracy percentage of 98.5% was the RF model, with the use of 10-fold cross-validation for testing and training data and feature selection to optimize the accuracy result and obliterate the unlikelihood attribute. Additionally, the proposed study handled the overfitting and gave the best results for unbiased prediction using ROC analysis. Adding more instances and attributes will help the study improve its performance. Additionally, this can help the medical experts in providing accurate treatment to the patient that suffers from the early stage of diabetes. This study demonstrates an experiment with the application of machine learning models in early-stage diabetes classification. Hence, it can be great a help in decreasing the hazardous and prevalence of the early-stage of diabetes.




REFERENCES

- [1] G. Roglic, "WHO Global report on diabetes: a summary," *International Journal of Noncommunicable Diseases*, vol. 1, no. 1, p. 3, 2016, doi: 10.4103/2468-8827.184853.
- [2] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus-a machine learning approach," in *2015 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015*, Dec. 2016, pp. 122–127, doi: 10.1109/RAICS.2015.7488400.
- [3] H. Das, B. Naik, and H. S. Behera, "Classification of diabetes mellitus disease (DMD): a data mining (DM) approach," in *Advances in Intelligent Systems and Computing*, vol. 710, 2018, pp. 539–549.
- [4] B. D. Kanchan and M. M. Kishor, "Study of machine learning algorithms for special disease prediction using principal of component analysis," in *Proceedings - International Conference on Global Trends in Signal Processing, Information Computing and Communication, ICGTSPICC 2016*, Dec. 2017, pp. 5–10, doi: 10.1109/ICGTSPICC.2016.7955260.
- [5] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.
- [6] A. A. Aljumah, M. G. Ahamad, and M. K. Siddiqui, "Application of data mining: diabetes health care in young and old patients," *Journal of King Saud University - Computer and Information Sciences*, vol. 25, no. 2, pp. 127–136, Jul. 2013, doi: 10.1016/j.jksuci.2012.10.003.
- [7] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 09, no. 01, pp. 1–16, 2017, doi: 10.4236/jilsa.2017.91001.
- [8] L. Chaves and G. Marques, "Data mining techniques for early diagnosis of diabetes: a comparative study," *Applied Sciences (Switzerland)*, vol. 11, no. 5, pp. 1–12, Mar. 2021, doi: 10.3390/app11052218.
- [9] A. Iyer, S. Jeyalatha, and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 1, pp. 01–14, Jan. 2015, doi: 10.5121/ijdkp.2015.5101.
- [10] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Advances in Intelligent Systems and Computing*, vol. 992, 2020, pp. 113–125.
- [11] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.
- [12] T. M. Alam *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019, doi: 10.1016/j.imu.2019.100204.
- [13] A. Negi and V. Jaiswal, "A first attempt to develop a diabetes prediction method based on different global datasets," in *2016 4th International Conference on Parallel, Distributed and Grid Computing, PDGC 2016*, 2016, pp. 237–241, doi: 10.1109/PDGC.2016.7913152.
- [14] Z. Tafa, N. Pervetica, and B. Karahoda, "An intelligent system for diabetes prediction," in *Proceedings - 2015 4th Mediterranean Conference on Embedded Computing, MECO 2015 - Including ECyPS 2015, BioEMIS 2015, BioICT 2015, MECO-Student Challenge 2015*, Jun. 2015, pp. 378–382, doi: 10.1109/MECO.2015.7181948.
- [15] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data*, vol. 6, no. 1, p. 13, Dec. 2019, doi: 10.1186/s40537-019-0175-6.
- [16] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Computer Science*, vol. 47, no. C, pp. 45–51, 2015, doi: 10.1016/j.procs.2015.03.182.
- [17] R. P. C. Gamara, A. A. Bandala, P. J. M. Loresco, and R. R. P. Vicerra, "Early stage diabetes likelihood prediction using artificial neural networks," in *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2020*, 2020, pp. 1–5, doi: 10.1109/HNICEM51456.2020.9400075.
- [18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [19] Y. Chen, W. Zheng, W. Li, and Y. Huang, "Large group activity security risk assessment and risk early warning based on random forest algorithm," *Pattern Recognition Letters*, vol. 144, pp. 1–5, Apr. 2021, doi: 10.1016/j.patrec.2021.01.008.
- [20] T. Chen, X. Yin, L. Peng, J. Rong, J. Yang, and G. Cong, "Monitoring and recognizing enterprise public opinion from high-risk users based on user portrait and random forest algorithm," *Axioms*, vol. 10, no. 2, p. 106, May 2021, doi: 10.3390/axioms10020106.
- [21] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.
- [22] J. Y. Kim *et al.*, "Development of random forest algorithm based prediction model of alzheimer's disease using neurodegeneration pattern," *Psychiatry Investigation*, vol. 18, no. 1, pp. 69–79, Jan. 2021, doi: 10.30773/pi.2020.0304.




- [23] D. Sisodia, S. K. Shrivastava, and R. C. Jain, "ISVM for face recognition," in *Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010*, Nov. 2010, pp. 554–559, doi: 10.1109/CICN.2010.109.
- [24] D. Sisodia, L. Singh, and S. Sisodia, "Fast and accurate face recognition using SVM and DCT," in *Advances in Intelligent Systems and Computing*, vol. 236, 2014, pp. 1027–1038.
- [25] G. R. Kumar, G. A. Ramachandra, and K. Nagamani, "An efficient feature selection system to integrating SVM with genetic algorithm for large medical datasets," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 2, p. 2277, 2014, [Online]. Available: <https://www.researchgate.net/publication/280922177>.
- [26] S. Aishwarya and S. Anto, "A medical decision support system based on genetic algorithm and least square support vector machine for diabetes disease diagnosis," *International Journal of Engineering Sciences & Research Technology*, vol. 3, no. 4, 2014, [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.676.8650>.
- [27] T. Santhanam and M. S. Padmavathi, "Application of k-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," *Procedia Computer Science*, vol. 47, no. C, pp. 76–83, 2015, doi: 10.1016/j.procs.2015.03.185.
- [28] D. Vijayarani, "Liver disease prediction using SVM and naïve bayes algorithms," *International Journal of Science, Engineering and Technology Research*, vol. 4, no. 4, pp. 816–820, 2015.
- [29] N. H. Segal *et al.*, "Classification and subtype prediction of adult soft tissue sarcoma by functional genomics," *American Journal of Pathology*, vol. 163, no. 2, pp. 691–700, Aug. 2003, doi: 10.1016/S0002-9440(10)63696-6.
- [30] Osisanwo *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/ijett-v48p126.
- [31] R. Al-Massri, Y. Al-Astel, H. Ziadia, D. K. Mousa, and S. S. Abu-Naser, "Classification prediction of SBRCTs cancers using artificial neural network," *International Journal of Academic Engineering Research*, vol. 2, no. 11, pp. 1–7, 2018, [Online]. Available: www.ijeais.org/ijaer.
- [32] S. Agrawal and J. Agrawal, "Neural network techniques for cancer prediction: a survey," *Procedia Computer Science*, vol. 60, no. 1, pp. 769–774, 2015, doi: 10.1016/j.procs.2015.08.234.
- [33] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification.," *ChemInform*, vol. 35, no. 5, Feb. 2004, doi: 10.1002/chin.200405237.
- [34] S. Shirol, "Early-stage diabetes 2020," 2020. <https://www.kaggle.com/sujan97/early-stage-diabetes-2020>.
- [35] T. T. Le, R. J. Urbanowicz, J. H. Moore, and B. A. McKinney, "STatistical inference relief (STIR) feature selection," *Bioinformatics*, vol. 35, no. 8, pp. 1358–1365, Apr. 2019, doi: 10.1093/bioinformatics/bty788.
- [36] G. Singh and R. K. Panda, "Daily sediment yield modeling with artificial neural network using 10-fold cross validation Method: A small agricultural watershed, Kappari, India ," *International Journal of Earth Sciences and Engineering*, vol. 04, no. 6, pp. 443–450, 2011, [Online]. Available: <https://www.researchgate.net/publication/265988179>.
- [37] S. V. N. Vishwanathan and M. N. Murty, "SSVM: a simple SVM algorithm," in *Proceedings of the International Joint Conference on Neural Networks*, 2002, vol. 3, pp. 2393–2398, doi: 10.1109/ijcnn.2002.1007516.
- [38] S. C. Wang, "Artificial neural network," in *Interdisciplinary Computing in Java Programming*, Boston, MA: Springer US, 2003, pp. 81–100.
- [39] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.
- [40] T. K. Das, "A customer classification prediction model based on machine learning techniques," in *Proceedings of the 2015 International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2015*, Oct. 2016, pp. 321–326, doi: 10.1109/ICATCCT.2015.7456903.

BIOGRAPHIES OF AUTHORS






Leonard Flores    is a full-time instructor from the College of Arts, Sciences, and Technology of Occidental Mindoro State College-Mamburao Campus. He is a graduate of Bachelor of Science in Information Technology from the same school and currently pursuing his Master's Degree in Information Technology at Batangas State University Alangilan Campus. His research interest including machine learning and data mining. He can be contacted at email: floresleonard429@gmail.com.






Dr. Rowell Marquez Hernandez    is a graduate of Doctor of Information Technology from the Technological Institute of the Philippines-Manila. A permanent lecturer from the College of Informatics and Computing Science of Batangas State University-Alangilan. He was the Department chairman of Bachelor of Science and Computer, Master of Science in Compute Science and Master of Science in Data Science. He is currently the Head of External Affairs office of his campus. His research interest is data mining, data science and machine learning. He has several publications under the Scopus index publication and a reviewer of IEEE Conferences in Malaysia and India. He is a member of the National Research Council of the Philippines and faculty researcher under the Digital Transformation Center of STEER Hub of his university. Currently, he is the project leader of the two funded researches of his university. He can be contacted at email: rowell.hermamdez@g.batstate-u.edu.ph.






Lloyd H. Macatangay    is an instructor at College of Informatics and Computing Sciences, Batangas State University, Philippines. He Holds a Master degree in Computer Science and Master in Development Management. His research areas are deep learning, machine learning, and computer networks. He is currently the department chairman of BS Computer Science, Master of Science in Computer Science and Master of Science in data science. He is handling cisco networking academy courses. He can be contacted at email: lloyd.macatangay@g.batstate-u.edu.ph.



Shiela Marie G. Garcia    is a graduate of Master of Science in Computer Science at Batangas State University Main Campus. Her research interests are about data science and business analytics, as well as systems analysis and design and mobile applications. A permanent lecturer handling research, capstone project, and several programming languages courses, and currently the OIC-Dean of the College of Informatics and Sciences of Batangas State University-JPLPC Malvar. Currently, she is the project leader of one funded research of her university. She can be contacted at email: shielamarieggarcia@g.batstate-u.edu.ph.



Jonnah R. Melo    is an instructor at College of Informatics and Computing Sciences, Batangas State University, Philippines. She holds a master's degree in Computer Engineering. Her research interest areas in machine learning, image processing and deep learning. Currently she is the Head of the Library Services in BatStateU, Lipa City. She handles software engineering and networking subjects. She can be contacted at email: jonnah.melo@g.batstate-u.edu.ph.