Performance Analysis of Supervised Learning Algorithms based on Classification Approach

Fazeel Ahmed Khan
Department of Computer Science
International Islamic University Malaysia
Selangor, Malaysia
fazeelahmedkhan15@gmail.com

Adamu Abubakar Ibrahim
Department of Computer Science
International Islamic University Malaysia
Selangor, Malaysia
adamu@iium.edu.my

Mohammad Salman Rais
Department of Mechatronics Engineering
International Islamic University Malaysia
Selangor, Malaysia
salmanmech2018@gmail.com

Priyanka Rajpoot
School of Electrical and Electronics Engineering
Universiti Sains Malaysia
Penang, Malaysia
priyankarajpoot9760@gmail.com

Ambareen Khan
School of Aerospace Engineering
Universiti Sains Malaysia
Penang, Malaysia
khanambareen5@gmail.com

Mohammad Nishat Akhtar
School of Aerospace Engineering
Universiti Sains Malaysia
Penang, Malaysia
nishat.akhtar2000@gmail.com

***Abstract*— Machine learning algorithm integrates with various features and methods for learning. The performance metric has a major role in determining the capacity and ability of any machine learning algorithm. The supervised approach supports learning based on predefined parameters. The classification approach classifies the set of data with labels for training and evaluation purposes. This paper applies method of supervised learning algorithm based on classification approach to analyzed in terms of performance metrics. Some algorithms were applied in the experiment process to determine the clear outcome from them. The analysis gives an outcome based on certain parameters to obtain the results to understand the ability of these algorithm and their implementation in any potential use-cases.**

Keywords*— Machine Learning, Support Vector Machines, Backpropagation Neural Network, Decision Forest, Bayes Point Classifier, Logistic Regression, Evaluation Model.

I. INTRODUCTION

Machine learning (ML) has broad range of application in various domains. Each use-case has its unique features which mould it in comparison with other approaches. The ML research is rapidly developing new and robust algorithms to solve complex statistical and industrial problems. The beginning from Probably Approximately Correct (PAC) learning framework introduces the supervised learning concept in ML research. The Bayes' Theorem or Bayes' Law introduces the Bayesian statistical approach which describes probability of events and probabilistic interpretations in statistical inference. The approach as described in Bayes' theorem or Bayes' Law is referring to the Bayesian Inference approach. Similarly, the Linear Discriminant Analysis (LDA) introduces a generalization of linear discriminant applies for pattern recognition and determining features characterized on multiple events in a linear combination. The LDA proximate with Principal Component Analysis (PCA) and factor analysis for linear combination analysis to define best possibility in the data. Furthermore, the Support Vector Machines (SVMs) introduces concept of hyperplane in infinite-dimensional space for classification, regression and outlier's detection. The inspiration from biological neuron introduces the Artificial Neural Network (ANN) which have the ability to learn without been explicitly programmed with guided tasks. The nodes in ANN is similar with neuron in biological brain with a goal to solve complex problems in human-like approach [1]. The formatter will need to create these components, incorporating the applicable criteria that follow.

In the wake of assessing the applied model, we expect that getting the assessment based on confusion matrix comprises of four classifiers namely, true negative (TN), true positive (TP), false negative (FN) and false positive (FP). The classifiers that accurately predicts positive with same positive condition, then it is known as true positives (TP). If the prediction is negative with same negative condition, then it is known as true negatives (TN) respectively. Similarly, if the prediction is positive but the condition is negative, the is known as false positives (FP) while if the prediction is negative with same negative condition, then is known as false negatives (FN) respectively. Similarly, other parameters i.e. accuracy, precision, recall, sensitivity, specificity, f-measure, error-rate were also considered for analysis. Furthermore, the graph will be drawn on the structure of Area Under Curve Receiver Operating Characteristic (AUC-ROC) approach [2].

II. LITERATURE REVIEW

A. Support Vector Machine

Support Vector Machine (SVM) is an approach to implement algorithms based on hyperplane in N-dimensional space which classify the data points between two classes. Hyperplanes separate two classes with many possible outcomes to find a plan that has maximum distance between data points e.g. margins. They are decision boundaries to classify data points falling in between both ends based on respective dimensions. SVMs are relatively a type of classifier firstly proposed in [3]. It was later updated by researchers in ML research community to achieve robust performance curve specially with noisy and complex datasets. The SVMs with sets of binary labelled training data

978-1-7281-4082-7 /19/$31.00 ©2019 IEEE

separate the hyperplane when working with classification problem known as "Maximal margin hyperplane". Similarly, in a case where linear separation can be applied, they apply combination with technique that effectively detect feature space of a non-linear mapping known as "Kernels" [4].

Similarly, SVM algorithm focuses to maximize the margin between data points and hyperplane. The function applied to measure the limitation in the model with respect to its ability between x and y relationship is known as "Cost Function" states as below:

$$(\theta) = C \sum_{i=1}^{m} \left[b^{(i)} c_1 \left(\theta^j a^{(i)} \right) + \left(1 - b^{(i)} \right) c_0 \left(\theta^j a^{(i)} \right) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \tag{1}$$

The cost function (c) is applied to train the SVM by minimizing the θ (theta) value to ensure the accuracy of SVM. The optimization of SVM cost function applies to minimize the cost function using "Gradient Descent". It enables to learn gradient (direction) which a model should implies to reduce the occurrence of errors. Gradient descent work with two possibilities, when there is a misclassification or no misclassification. The misclassification model make mistake during prediction of class on the data point which gives loss to regularization parameter.

$$w = w - \alpha . (2\lambda w) \tag{2}$$

On the contrary, the model which predicts the class of data points can have the updates from regularization parameter. The gradient descent update for both cases is states as below:

$$w = w + \alpha . (y_i . x_i - 2\lambda w) \tag{3}$$

B. Decision Tree

A Decision Tree (DT) is a predictive model approach comprised of procedures to implement recursive separation of data into its sub-partitions based on branches in the tree. It is a multistage procedure for decision-making procedure based on decision theory applying conversion from decision table into optimal decision trees and sequential approaches [5]. In DT, each node shows an attributes (feature), branch (link) to show the decision rule and leaf to represent the result whether it is categorical or continuous. It comprises of root node, branches of interior nodes and terminal (leaf) branches. The data is classified with sequential order and each node has a single parent node along with two or more from its descendants. The framework defines a class for each node which fall under observation category. DTs are nonparametric by design which does not require any assumption on input data distribution. DT manages nonlinear values for features in relation with classes. DT allows for missing values and can easily manages numeric as well as categorical inputs with normal procedure [6].

The major approach for solving decision problem is by using ID3 (Iterative Dichotomiser 3) algorithm. The ID3 employs top-down greedy search approach on possible branches without any backtracking. It comprises of two sub-approaches including information gain and entropy to construct a decision tree.

$$I(x,y) = \frac{-x}{x+y} \log_2 \left(\frac{x}{x+y} \right) - \frac{y}{x+y} \log_2 \left(\frac{y}{x+y} \right)$$

$$-x_1 . \log x_1 - x_2 . \log x_2 - \cdots x_n . \log x_n \tag{1}$$

$$E(X) = \sum_{i=1}^{v} \frac{x_i + y_i}{x+y} (I(x,y)) \tag{2}$$

$$G = I(x,y) - E(X) = E(X) - \sum_i \frac{|S_i|}{|S|} . E(S_i) \tag{3}$$

The common approach in DT is the recursive binary splitting which integrate all the required features and various split points. They were tried and tested with utilizing the function known as "Cost Function". The best cost or worst cost is selected with splitting approach. The cost function tries to find homogenous branches having similar responses to make sure that data will follow the correct path within test domain.

a) Decision Forest (Random Forest)

Decision Forest or (Random Forest) - (RF) comprises of ensemble methods applied in classification and regression problems. The RF solves the problem of overfitting in DT during the course of training dataset. There were two approaches introduce in RF methods including random subspace and bagging methods respectively. The random subspace method utilizes subspace selection procedure in feature space. In a case of any feature space having *n* dimensions, there are 2*n* subspaces which can be utilized to create decision trees. The implementation of randomization for components selection in feature vector has large possibility in this regard. The random subspace implements the discriminant function to maintain the accuracy on the training set [7]. The discriminant function is stated as in which to decision approach is to assign *y* to class *x* with maximum f_x (y) as:

$$f_x(y) = \frac{1}{z} \sum_{i=1}^{z} \hat{\dot{R}}(x|c_i(x)) \tag{1}$$

Similarly, bagging is a predicting method which implements generation of multiple variants of predictors to get an aggregated predictor. The aggregate calculates the average of predictors versions in case of numerical predictions while in case of class it applies plurality voting procedure. The versions of predictors were generated by replicating datasets and utilize them to generated new learning datasets. The tests on bagging methods shows significant boost in accuracy while simulated with classification and regression dataset. The methods for prediction can cause some instability but it can be minimized by unsettling the learning dataset then bagging can significantly improve the accuracy [8]. The prediction from learning set by averaging all the predictors version is as:

$$\hat{f} = \frac{1}{E} \sum_{e=1}^{E} f_e(\acute{s}) \tag{2}$$

C. Backpropagation Neural Network

The backpropagation algorithm was introduced in late 1970s but that time the importance of this algorithm was negligible. But now it is one of the key algorithms in artificial neural network research and application. The main core of this algorithm is the proximity of relationship with partial derivatives of the cost function in contrast with the given weights inside the network. The expression elaborates the change between the cost with the change in the weights to avoid any biases. In a situation of complex expressions, the algorithm has the capability to classify the values in a more defined interpretation for better analysis on the results appearing from the calculations. The essence of this algorithm is to check on the insights of changing in the weights and relevant variables to analyse the overall behaviour of the network [9].

Furthermore, it is classified into three different layers of input and output. The initial layer is the input layer in which the initial weights along with relevant variables have been placed. The second layer comprises of a hidden layer in which the calculation was perform and the implementation of this layer is also hidden from outside intervention. The last layer is the output layer in which in the final output will be transferred and the final weights will also be adjusted respectively. It is an iterative process meaning thereby is after reaching to the final result if the values are not according to the desired output then the moving backward compatibility feature is available to return back to perform previous steps again and again until the final conclusion will match the goal-oriented criteria of the required analysis [10].

$$y_i(z) = \sum_{i=1}^{n} x_j\,(z)\,.\,w_{ij}\,(z) - \theta_j\;;\; S(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$y_k(z) = \sum_{j=1}^{m} x_{jk}\,(z)\,.\,w_{jk}\,(z) - \theta_k \tag{2}$$

$$\delta_k(z) = y_k(z).[1 - y_k(z)]\,.x_k(z);\; x_k(z) = y_{j,k}\,(z) - y_k(z) \tag{3}$$

$$\Delta x_{jk}\,(z) = \alpha\,.y_j\,(z).\delta_j\,(z);\; x_{jk}(z+1) = x_{jk}(z) + \Delta x_{jk}(z) \tag{4}$$

$$\delta_j(z) = y_j(z).\left[1 - y_j(z)\right].\sum_{k=1}^{l} \delta_k(z)\,x_{jk}(z) \tag{5}$$

$$\Delta x_{ij}\,(z) = \alpha\,.x_i(z)\,.\delta_j(z);\; x_{ij}\,(z+1) = x_{ij}\,(z)\,.\Delta x_{ij}\,(z) \tag{6}$$

D. Bayes Point Machine Classifier

Bayes Point Machines (BPM) algorithm is an approximation in linear classification for Bayesian inference. The weight vector consideration is nearly ||y|| = 1 due to spatial direction for weight vector for classification. To define the resulting ball-shape hypothesis space, the centre of mass is determine using zero training error region. To determine the centre of mass, the suggested method includes kernel billiard and the perceptron algorithm, to apply approximation on the transposition of training sample by using permutations methods. The latter approach proves to be more robust and efficient to train BPM applicable for large sets of data [11] [12].

To determine the Bayes classification procedure for determining the loss occurred on the hypothesis which is h ϵ H, applies to its probability on inputs and weights $P_R|Z^m = {}_z(r)$. To achieve the minimum predicted loss:

$$B_z(x) = arg\,min_{y\,\epsilon\,\boldsymbol{Y}}\; P_{R|Z^m} =_z [l(R(x), y)] \tag{1}$$

In case of given static condition in the hypothesis space $\boldsymbol{R} \subseteq \boldsymbol{X}^{\,y}$ and static loss l: Y x Y – □+. therefore, determining measures from any multiple variable $\mathbf{P_x}$ and $\mathbf{P_H}$, the BPM B_{bp} is given by,

$$B_{bp}(z) = arg\,min_{h\,\epsilon\,\boldsymbol{H}}\; \mathbf{P}_x[P_{R|Z^m} =_z \left[l\big(r(X), R(X)\big)\right] \tag{2}$$

In case of determining static mapping for given hypothesis α space is given,

$$R := \{\, x \mapsto \text{sign}(\langle \mathbf{x}, \mathbf{y}\rangle)\;\mathbf{y} \in Y\,, \quad \mathbf{Y} := \{\mathbf{y} \in \boldsymbol{K} \mid \|\mathbf{y}\| = 1\} \tag{3}$$

To assume the input distribution in feature space K inside spherical Gaussian for dimensionality d = dim (K),

$$f_y(y) = \frac{1}{\pi^{\frac{x}{z}}} \exp(-\|y\|^2) \tag{4}$$

Therefore, the center of mass will be,

$$y_{cm} = \frac{\mathbf{R}_{\mathbf{Y}|Z^m} =_z [\boldsymbol{Y}]}{\left\|\mathbf{R}_{\mathbf{Y}|Z^m} =_z [\boldsymbol{Y}]\right\|} \tag{5}$$

It shows an exceptional approximation towards Bayes Point classification B_{bp} and connect towards the posterior's confidence on $\mathbf{R_Y}|Z^m$ = z to become effectively increased [13].

E. Logistic Regression Classifier

Logistic regression (LR) algorithm applies on the observations on discrete classes in a classification approach. It is based on probability and applies the predictive analysis approach on given scenarios. The LR approach uses complex cost function known as "Sigmoid function" in comparison with linear function. The hypothesis limits the linear function between 0 and 1 which is not the case with logistic regression [14].

$$0 \le h_\theta(x) \le \tag{1}$$

The sigmoid function maps between any real values from 0 and 1 to predict the required probabilities.

$$f(x) = \frac{1}{1+e^{-(x)}} \tag{2}$$

The hypothesis will be expected between values 0 and 1 to comply with the expected assumption.

$$h\theta(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 \boldsymbol{X})}} \tag{3}$$

The classifier is expected to give values based on set of outputs on probability when applies via prediction function. This set the decision boundaries with a chosen threshold value. The cost function with minimum error optimization represent the accuracy of optimized model.

$$f\theta = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{4}$$

To reduce the cost values using gradient descent to minimize the cost on each parameter.

$$\theta_j := \theta_j - \alpha\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \tag{5}$$

III. Experiment Design

A. Model Development

In our model development, we apply supervised classification algorithms to evaluate and analyse the efficiency and performance of selected algorithms. The analysis includes Support Vector Machines, Decision Forest and Artificial Neural Network, Logistic Regression, Bayes Point classifiers. The applied algorithm creates classification models that is applied for predicting a dependent value in a labelled set of data. It is a classification based on supervised learning modelling techniques. It requires to label data with specific label known as "Class". The given labels tagged inside a set of data with classification into independent and dependent values. When the definition stage is successfully done then the training is provided to the tagged dataset and the define model is applied as an input to train final model. Therefore, the train model can then be able to predict values for new inputs.

In the model development, we apply data transformation method to create calculation applied to numeric input in the dataset. The dataset was check with missing values to avoid any noisy impact during training process. The data was screen through required categorical values for mapping them to ordinal features to ensure about implementation of algorithm to ordinal features effectively. Similarly, it also ensures to label the encoded nominal classes for creating unique value features in the nominal column. The crucial step of feature scaling is applied to make sure that algorithm perform effectively with features on same scale. The normalization feature is used to rescale the feature to a specific range [0,1] referring to the case of min-max scaling. The min-max scaling feature will scale the feature to each required column.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

The values and data types were checked with class label to ensure about its types i.e. numeric or Boolean. The process of converting values to categorical from its previous state i.e. Boolean or numeric is applied. It is applied for categorizing or prediction of values to make column as features. The model is also applied to check with optimization and generalization problems. Optimization is the process in which model is adjusted to give best performance results. Similarly, generalization refers to the performance of model on unseen data to obtain best generalization ability. During training process, the underfitting arises when there is low rate on loss of training data with loss in test data. After passing with couple of iterations, the generalization process is stop and the validation metrics start degrading referring to the model's overfitting status. To improve these problems, the regularization has ability for modulating the amount of information a model can hold or put restriction on its holding ability. It is applied with reducing the size of the model with learning parameter and adding weight regularization to force model to hold less weight.

$$C = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij} Z_j)^2 + \lambda\sum_{j=0}^{M}|Z_j| \tag{2}$$

$$C = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij} Z_j)^2 + \lambda\sum_{j=0}^{M} Z_j^2 \tag{3}$$

The model is split into training and test with fractions of row in the column to 0.7 randomized split.

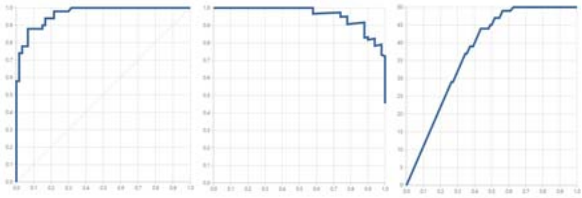

Figure 1. Support Vector Machine algorithm efficiency and performance evaluation. The left corner graph shows the AUC-ROC curve between TPR and FPR, the middle graph shows precision and recall. The right corner graph shows the lift proposition state of SVMs.

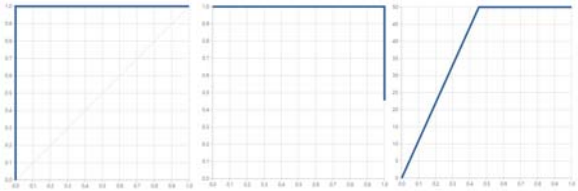

Figure 2. Decision Forest algorithm efficiency and performance evaluation. The left corner graph shows the AUC-ROC curve between TPR and FPR, the middle graph shows precision and recall. The right corner graph shows the lift proposition state of decision forest.

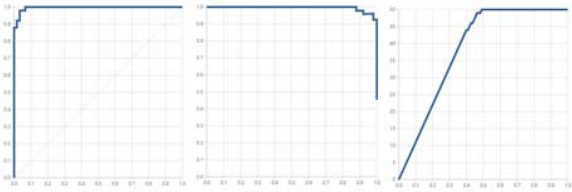

Figure 3. Bayes Point algorithm efficiency and performance evaluation. The left corner graph shows the AUC-ROC curve between TPR and FPR, the middle graph shows precision and recall. The right corner graph shows the lift proposition state of Bayes point.

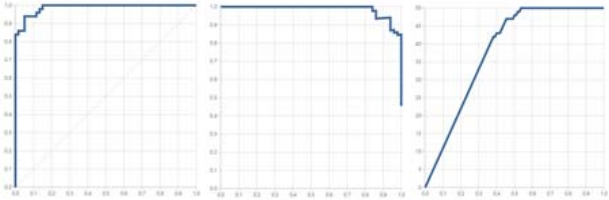

Figure 4. Logistic Regression algorithm efficiency and performance evaluation. The left corner graph shows the AUC-ROC curve between TPR and FPR, the middle graph shows precision and recall. The right corner graph shows the lift proposition state of logistic regression.

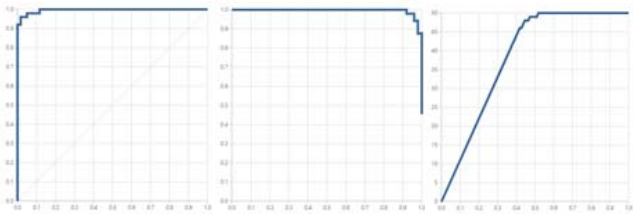

Figure 5. Backpropagation Neural Network algorithm efficiency and performance evaluation. The left corner graph shows the AUC-ROC curve between TPR and FPR, the middle graph shows precision and recall. The right corner graph shows the lift proposition state of neural network.

B. Evaluation Methods

The efficiency and performance are evaluated from outcome of the experiments. The model outputs visualize the results based on confusion matrix. There will be a certain prediction before and after the experiments. The predicting cases includes true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The result graph visualizes the performance based on Area Under Curve Receiver Operating Characteristic (AUC - ROC). It is an important evaluation metric in any classification evaluation at different threshold. It elaborates the capability of a model between two distinguish classes. The higher the AUC - ROC curve starting from 0 to nearly to 1, the better model shows its performance. It means that the model has a good separability measure between classes. In case of AUR - ROC nearly to 0.5, then it shows that the model has no separability of class. The ROC curve plots TP versus FP as a function to classify the positives while AUC calculates the comprehensive performance of classification based on ROC metrics. The model also comprises of analytical variables for better insights on the prediction. The classification model has accuracy, error rate, precision, recall, sensitivity, specificity, f-measure, lift metric for detailed evaluation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN} = 1 - Accuracy \quad (2)$$

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$F - measure = \frac{2\,.TP}{2\,.TP + FP + FN} \quad (8)$$

IV. Analysis Framework

A. Evaluating Model Performance Method – 1

The classification analysis framework methods - 1 checks the algorithms in terms of confusion matrix, accuracy, error rate and f-measure. The confusion matrix categorized prediction to compare it with actual values. The positive and negative class relationship depict 2 x 2 confusion matrix into four categories i.e. TP, TN, FP, FN respectively.

Table 1. Confusion Matrix

Algorithms	TP	TN	FP	FN
Support Vector Machine	44	52	8	6
Decision Forest	50	60	0	0
Backpropagation Neural Network	48	57	3	2
Logistic Regression	45	37	3	5
Bayes Point	47	58	2	3

The accuracy is a proportion of TP and TN divided by total number of predict conditions. The error rate is the proportion of incorrectly classified predictions. It is the difference of 1 from the accuracy. The performance analysis is used to examine the applied algorithms capacity and their ability to work in different circumstances.

Table 2. Model Score Evaluation based on Method – I

Algorithms	Accuracy	Error Rate
Support Vector Machine	0.873	0.127
Decision Forest	1.000	0
Backpropagation Neural Network	0.955	0.045
Logistic Regression	0.927	0.073
Bayes Point	0.955	0.045

B. Evaluating Model Performance Method – II

The classification analysis framework method – 2 checks the algorithms in terms of sensitivity, specificity, precision, recall and f-measure. The sensitivity is the proportion of correctly classified positives examples calculated by the fraction of TP upon the sum of TP and TN respectively. Similarly, the specificity is the proportion of correctly classified negative examples calculated by the fraction of TN upon the sum of TN and FP respectively. The precision is the positive predicted value calculates the real proportion of positive examples. It is the fraction of TP upon the sum of TP and FP respectively. The recall measures the overall completeness of the results calculated by the fraction of TP upon the sum of TP and FN respectively. The F-measure combines both precision and recall in single metric by using

harmonic mean (rates of change). It is the fraction of multiple by two for both precision and recall respectively.

Table 3. Model Score Evaluation based on Method – II (1)

Algorithms	Sensitivity	Specificity	Precision
Support Vector Machine	0.880	0.866	0.846
Decision Forest	1.000	1.000	1.000
Backpropagation Neural Network	0.960	0.950	0.941
Logistic Regression	0.900	0.950	0.938

Table 4. Model Score Evaluation based on Method – II (2)

Algorithms	Recall	F-Measure
Support Vector Machine	0.880	0.863
Decision Forest	1.000	1.000
Backpropagation Neural Network	0.960	0.950
Logistic Regression	0.900	0.918
Bayes Point	0.940	0.949

C. Performance Evaluation

The support vector algorithm has the lowest performance in all metrics while the decision forest stands higher in terms of performance rate. Similarly, backpropagation neural network and Bayes point algorithm have close difference in their respective values. The logistic regression has second lowest performance after support vector machine with less difference between their values.

V. Conclusion

In this analysis, we apply various algorithms and methods to examine the performance of five algorithm in the given set of data. The algorithms were analysed in the beginning section of this paper with their possible implementation methods. The performance analysis is based on certain metric and evaluation techniques which gives productive results based on expected assumptions. Overall, the analysis gives minimum to maximum outcomes which will be utilized to apply the given algorithm in different required use-cases according to the requirements.

VI. References

[1] S. H. Walker and D. B. Duncan, "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika,* vol. 54, no. 1, pp. 167-179, 1967.

[2] S. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics ,* vol. 21, no. 3, pp. 660 - 674, 1991.

[3] D. E. Rumelhart, R. J. Williams and G. E. Hinton, "Learning representations by back-propagating errors," *Cognitive modeling,* vol. 5, no. 3, p. 1, 1988.

[4] B. Potetz, "Estimating the Bayes Point Using Linear Knapsack Problems," *Proceedings of the 28th International Conference on Machine Learning (ICML-11),* pp. 257-264, 2011.

[5] C. Penaloza, Y. Mae, K. Ohara and T. Arai, "BMI-based learning system for appliance control automation.," *2013 IEEE International Conference on Robotics and Automation,* pp. 3396-3402, 2013.

[6] M. Mohri, A. Rostamizadeh and A. Talwalkar, Foundations of machine learning, 1 ed., MIT Press Ltd, 2012.

[7] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, p. 436, 2015.

[8] T. K. Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition,* 1995.

[9] R. Herbrich, T. Graepel and C. Campbell, "Bayes Point Machines," *Journal of Machine Learning Research,* vol. 1, pp. 245-279, 2001.

[10] D. F. Gordon and M. Desjardins, "Evaluation and selection of biases in machine learning," *Machine Learning,* vol. 20, no. 1-2, pp. 5-22, 1995.

[11] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics,* vol. 16, no. 10, pp. 906-914, 2000.

[12] U. M. Fayyad and K. B. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine Learning,* vol. 8, no. 1, pp. 87-102, 1992.

[13] S. F. Ahmed et al., “Robotic Exoskeleton Control for Lower Limb Rehabilitation of Knee Joint,” International Journal of Engineering & Technology, vol. 7, no. 2.34, p. 56-59, 2018.

[14] L. Breiman, "Bagging Predictors," Department of Statistics, University of California Berkeley, California, 1994.