Application of Machine Learning in Disease Prediction

Pahulpreet Singh Kohli
Dept. Of Electronics and Communication Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi, India
pahulpreet86@gmail.com

Shriya Arora
Dept. Of Electronics and Communication Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi, India
shriyaarora080696@gmail.com

***Abstract*—The application of machine learning in the field of medical diagnosis is increasing gradually. This can be contributed primarily to the improvement in the classification and recognition systems used in disease diagnosis which is able to provide data that aids medical experts in early detection of fatal diseases and therefore, increase the survival rate of patients significantly. In this paper, we apply different classification algorithms, each with its own advantage on three separate databases of disease (Heart, Breast cancer, Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. The results of the study strengthen the idea of the application of machine learning in early detection of diseases.**

Keywords— Machine Learning, Disease Prediction, Wisconsin Breast Cancer Dataset, Pima Indians Diabetes dataset, Heart Disease Dataset

I. INTRODUCTION

As per the Centers for Medicare and Medicaid services, 50% of Americans have multiple chronic diseases with a total US health care expenditure in 2016 to be about $3.3 trillion, which amounts to $10,348 per person in the US. The early detection of common diseases such as breast cancer, diabetes, coronary artery and tumor, could control and reduce the chance of these diseases to be fatal for the patient. With the advancement in machine learning and artificial intelligence, several classifiers and clustering algorithms are being used to achieve this.

Following the methodologies used in, this paper presents the use of machine learning algorithms for prediction of diseases including breast cancer, which is a very common disease among women, heart diseases, which are the leading cause of deaths in the US, and diabetes, in which blood glucose or blood sugar levels are too high.

The datasets used for the building the predictive models in this paper are available and can be downloaded from UCI machine learning library [1] . The data is imported in CSV format and cleaned for use. After data munging and attributes selection, machine learning algorithms including Logistic Regression, Decision Trees, Random Forest, Support Vector Machine(SVM) and Adaptive Boosting, are used for prediction of the above-mentioned diseases, and a comparison of their accuracy is done for selecting best model for that disease dataset. All the analysis and visualization are carried out in python 2.7.

The paper is presented as follows: Section 2 gives brief explanation about various machine learning algorithms used. Followed by Section 3 which describes the proposed method for building predictive model. Section 4 explains the experiment and results and Section 5 concludes and provide the future scope of the paper.

II. MACHINE LEARNING ALGORITHMS

A. Logistic Regression

Logistic Regression [2] is a classification algorithm for the probability of occurrence of an event, whether that event will occur or not. It is used to portray a binary or a categorical outcome with only 2 classes. It is similar to linear regression with the only difference being that the outcome of the variable is categorical instead of a continuous variable. It uses Logit Link function, in which the data values are fitted, for prediction. The mathematical interpretation defines Logit function as the natural log of the odds that Y equals one of the categories [3] . If p is the probability then, the logit function for p is defined as:

$$Logit(p)=ln(p/1-p) \quad (1)$$

B. Decision Tree

Similar to the tree analogy in real life, the Decision tree is a machine learning algorithm, used for both classification and regression analysis [4]. It is a tree-like graph beginning with a single node, and branching into its possible outcomes. Unlike the linear models, a decision tree is a supervised learning, that maps non-linear relationships as well. The data sample is divided into homogeneous subsets based on the most notable splitter in input attributes. The splitter is identified using various algorithms such as Gini Index, Chi-Square, Information Gain and Reduction in Variance . For example [5]

a dataset with boolean target variable, the entropy function for the dataset is given as:

$$Entropy = -p_{\oplus}log_2p_{\oplus}p_{\ominus}log_2p_{\ominus} \quad (2)$$

C. Random Forest

Random forest is an ensemble of various decision trees, trained with the bagging methodology [6] . Bagging is used for making the model more stable and accurate by approaching averaging model technique. The random forest classifier [7] is basically a collection of decision tree classifiers where each tree is constructed with a number of

random vectors and is able to vote for the most favored class for prediction. The injection of randomness in the model prevents it from over fitting and provide better result for classification analysis.

D. SupportVector Machine

Support Vector Machines [8] , also called Support Vector networks are supervised learning algorithms used for both classification and regression analysis. It classifies the data points plotted in a multidimensional space into categories by parallel lines called the hyperplane. The classification of data points involves the maximization of margin between the hyperplane. There are different kernels [9] available for mapping of linear or no linear data points in a multidimensional space for separation. For our analysis, we have used only the *Linear* and *Radial basis function* as kernel.

E. Adaptive Boosting

Adaptive Boosting (AdaBoost) formulated by Yoav Freund and Robert Schapire [10] is a machine learning algorithm used for classification as well as for regression analysis. It involves the conversion of a weak classifier into a strong one using the ensemble technique. For this purpose, the prediction of each weak classifier is merged using weighted average or by taking into account their prediction accuracy as a metrics. Initially, all the attributes are given equal weights, then the algorithm assigns a higher weightage to the inaccurate observation [11] . The error is then propagated with every prediction and multiple iterations are done to reduce it until the prediction become accurate.

III. Proposed Method

The proposed method for building the predictive model for the diseases proceeds as follows:

- *Exploration of dataset*: Dataset is explored in the python environment along with data dictionary of the attributes involved.
- *Data Munging:* It refers to the estimation of missing values in some variables and is necessary as most of the interpretations cannot be done with missing data. The missing values are then replaced with the mean value in case of a continuous variable or with the mode value in case of a categorical variable.
- *Feature Selection:* is crucial for any predictive modeling and is done to take care of multicollinearity, remove any redundant features that are highly correlated with each other, therefore, improving the model's performance. For elimination of the attributes that are not significant for the diagnosis of a disease, we have adopted backward selection method. In this, we begin with all the attributes of the model, followed by their elimination based on p-value. This helps to determine the significance of the results while performing null hypothesis test in statistics. The attributes with the p-value greater than 0.05 were deleted and the model was refitted with the remaining variables. This process was iterated multiple times until every existing variable for the model was at a significant level. The adjusted R square value was also observed after each iteration to measure the proportion of variation described by only those independent variables that really contribute to the prediction of the target variable.
- *Model fitting and Testing*: After feature selection, 5 classifications algorithms including Logistic Regression, Decision Trees, Random Forest, Support Vector Machine(SVM) and Adaptive Boosting were used with the selected feature and a comparison between their prediction accuracy was done using Train/Test split method. The test size for comparison was set to 0.1, that is 90% of the dataset for training of classifier and the rest 10% for testing. *Fig. 1* summarizes the steps for our proposed method.

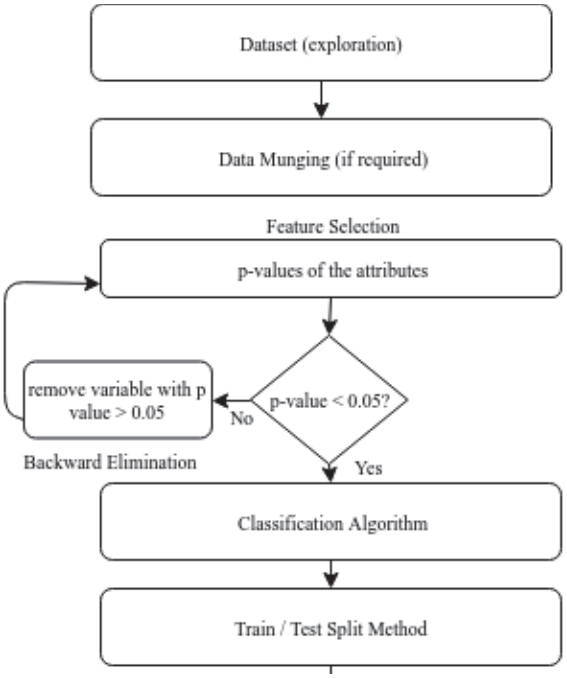

Fig. 1. *Proposed Method*

IV. Experiment and Results

The analysis and results of the different machine learning algorithm on the three datasets (Heart disease, Breast cancer, and diabetes) are summarized below.

A. Experiment with Breast Cancer Dataset

The Breast Cancer Wisconsin (Original) Dataset [12] consists of 9 numeric input attributes including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses and a target variable or the output in class attribute with only two values: 2 (non-cancerous) and 4 (cancerous).There are total 699 instances in the dataset, from them 16 instances have missing values for bare nuclei attribute. These missing values were filled with the mean value of bare nuclei of the dataset before modeling. The backward modeling was done using the p-value test, resulted in a total of 8 significant attributes for the predictive model including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli. *Fig. 2* compares the

prediction accuracy of the different machine learning algorithm for the breast cancer dataset .

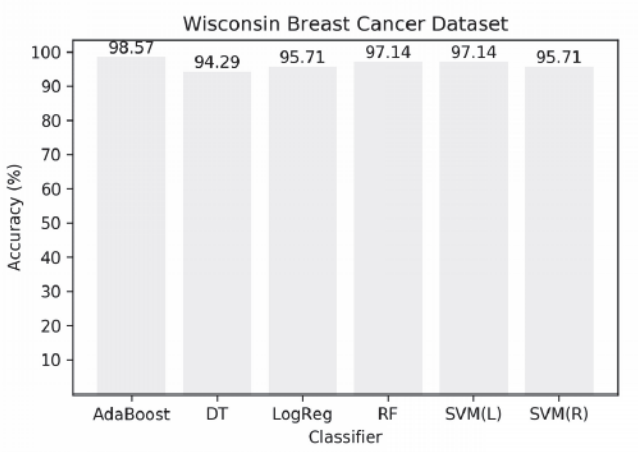

Fig. 2. *Comparison of different algorithm for Breast Cancer Dataset*

The accuracy of AdaBoost classifier was found to be highest among all machine learning algorithm. The AdaBoost model was then compared with the Multi-attributed lens model [13] , EM – PCA-CART fuzzy based model [14] using k fold cross-validation score.

TABLE I. COMPARISON

Classification Algorithm	**Cross Validation**	**Accuracy** (%)
Adaptive Boosting (proposed Method)	5 fold	95.29
	10 fold	95.86
Multi-attributed lens (Positive) [13]	5 fold	95.16
CART [14]	10 fold	94.56

B. Experiment with Diabetes Dataset

The Pima Indians Diabetes Dataset [15] contains a total of 768 instances, with 8 attributes including no of times pregnant, glucose concentration found in oral glucose tolerance test (glucose level),blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. The outcome attribute holds either 0 value (not diabetic) or value 1 (diabetic) for the instances. The missing values for the attributes glucose level,blood pressure, skin thickness, insulin, BMI found in the dataset were filled by their mean in the dataset respectively. The backward modeling of the data led to the elimination of 3 attributes, resulting in a total of 5 significant attributes comprising of no of times pregnant, glucose concentration found in oral glucose tolerance test (glucose level), BMI, diabetes pedigree function, and age. *Fig. 3* compares the prediction accuracy of the different machine learning algorithm for the diabetes dataset. Support Vector Machine (SVM) with linear kernel was found to perform superior than others.

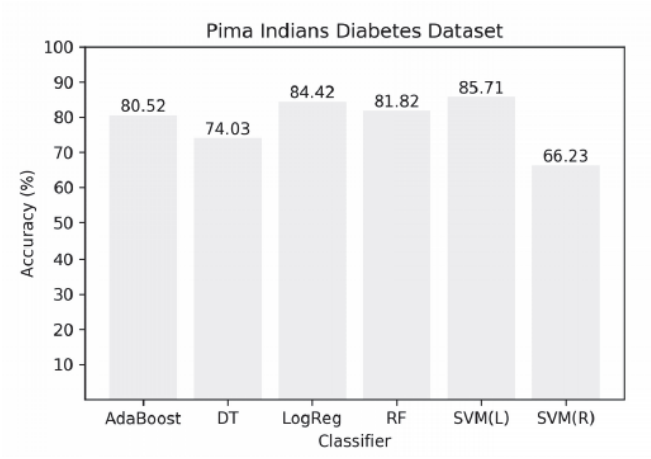

Fig. 3. *Comparison of different algorithm for Diabetes Dataset*

We then compared our model with Extreme Learning Machine (ELM) based model [16] and KNN based model with PSO optimization [17] on the metric defined in the paper respectively.

TABLE II. COMPARISON

Classification Algorithm	**Test Size**	**Accuracy** (%)
Support Vector Machine (Linear) (proposed Method)	0.40	77.92
	0.35	76.75
KNN based model with PSO optimization [17]	0.40	76.92
ELM based Model [16]	0.35	72.2

C. Experiment with Heart Disease Dataset

The Heart Disease Dataset [18] consists of 13 input attributes including age, gender, type of chest pain, blood pressure, cholesterol, blood sugar level, electrocardiograph result, maximum heart rate, exercise-induced angina, old peak, Slope, number of vessels colored, thal. The dataset contains 303 instances, from which 2 instances for a number of vessels colored attribute and 4 instances for thal attribute are missing, which are filled by their mean value for the dataset respectively. The prediction attribute consists of 5 classes ranging from integer value 0 - 4 where 0 indicate absence and the integer value from 1 - 4 indicate the presence of heart disease. We encoded the prediction attribute to class 0 and 1 to indicate absence or presence of heart disease respectively. The feature selection from 13 input parameter by backward elimination resulted in a total of 11 significant input parameters which include gender, type of chest pain, blood pressure, blood sugar level, electrocardiograph result, maximum heart rate, exercise-induced angina, old peak, Slope, number of vessels colored, thal. *Fig. 4* shows the comparison between the accuracy of the models obtained for heart dataset. Logistic Regression was found to have the highest accuracy among all.

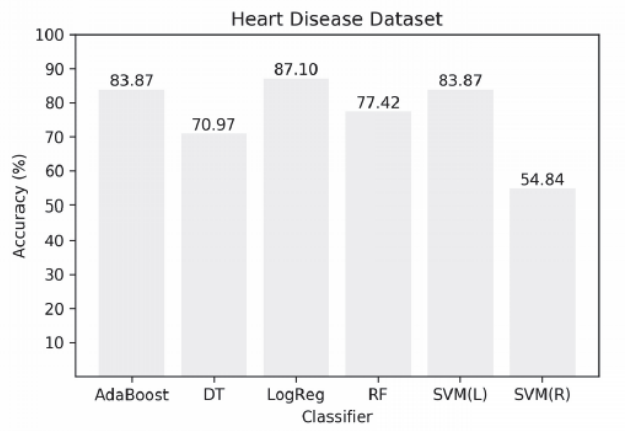

Fig. 4. *Comparison of different algorithm for Heart Disease Dataset*

On the same metric, we then compared the accuracy of our model with the SVM based model [19] , and Pruning Decision Tree method [20] and found it to perform better.

TABLE III. COMPARISON

Classification Algorithm	**Test Size**	**Accuracy** (%)
Logistic Regression (proposed Method)	0.34	80.77
	0.35	81.30
SVM based Model [19]	0.34	80.41
Pruning Decision Tree [20]	0.35	76.51

Table 4, present all the classification accuracies achieved by the algorithms following our proposed model.

TABLE IV. RESULTS

Classifier	**Wisconsin Breast Cancer Dataset** [12]	**Pima Indian Diabetes Dataset** [15]	**Heart Disease Dataset** [18]
Logistic Regression	95.71	84.42	87.1
Decision Tree	94.29	74.03	70.97
Random Forest	97.14	81.82	77.42
Support Vector Machine	97.14 (lin) 95.71 (rbf)	85.71 (lin) 66.23 (rbf)	83.87 (lin) 54.84 (rbf)
AdaBoost	98.57	80.52	83.87

V. CONCLUSION

The results of this study confirm the application of machine learning algorithms in prediction and early detection of diseases. To our best understanding, the model built according to the proposed method exhibits better accuracy than the existing ones [13,14,16,17,19,20]. The prediction accuracy of our proposed method reaches 87.1% in Heart Disease detection using Logistic Regression, 85.71% in Diabetes prediction using Support Vector Machine (linear kernel) and 98.57% using AdaBoost classifier for Breast Cancer detection. The future scope and improvement of the project involve automation of the steps such as data munging, feature selection and model fitting for best prediction accuracy. Use of pipeline structure for data preprocessing could further help in achieving improved results.

REFERENCES

[1] "UCI Machine Learning Repository." [Online]. Available: https://archive.ics.uci.edu/ml/index.php. [Accessed: 21-Apr-2018].

[2] W. Bergerud, "Introduction to logistic regression models with worked forestry examples: biometrics information handbook no. 7," no. 7, p. 147, 1996.

[3] S. Sperandei, "Lessons in biostatistics Understanding logistic regression analysis," *Biochem. Medica*, vol. 24, no. 1, pp. 12–18, 2014.

[4] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[5] T. M. Mitchell, "Decision Tree Learning," *Machine Learning*. pp. 52–80, 1997.

[6] L. Breiman, "Random Forest," pp. 1–33, 2001.

[7] M. Denil, D. Matheson, and N. De Freitas, "Narrowing the Gap: Random Forests In TheDenil, M., Matheson, D., & De Freitas, N. (2014). Narrowing the Gap: Random Forests In Theory and In Practice. Proceedings of The 31st International Conference on Machine Learning, (1998), 665–673. Retrieved from ht," *Proc. 31st Int. Conf. Mach. Learn.*, no. 1998, pp. 665–673, 2014.

[8] V. Jakkula, "Tutorial on Support Vector Machine (SVM)," *Sch. EECS, Washingt. State Univ.*, pp. 1–13, 2006.

[9] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," vol. 22, no. 2, pp. 103–104, 2000.

[10] Y. Freund and R. Schapire, "A Tutorial on Boosting," pp. 1–35, 2013.

[11] R. Rojas, "AdaBoost and the Super Bowl of Classifiers A Tutorial Introduction to Adaptive Boosting," *Writing*, pp. 1–6, 2009.

[12] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). [Accessed: 21-Apr-2018].

[13] C. Sirisomboonrat and K. Sinapiromsaran, "Breast Cancer Diagnosis Using Multi-Attributed Lens Recursive Partitioning Algorithm," *2012 Tenth Int. Conf. ICT Knowl. Eng.*, pp. 40–45, 2012.

[14] D. Lavanya and D. K. U. Rani, "Analysis of Feature Selection with Classification : Breast Cancer Datasets," Indian J. Comput. Sci.Eng., vol. 2, no. 5, pp. 756–763, 2011..

[15] "UCI Machine Learning Repository: Pima Indians Diabetes." [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/. [Accessed: 21-Apr-2018].

[16] R. Priyadarshini, N. Dash, and R. Mishra, "A Novel approach to Predict Diabetes Mellitus using Modified Extreme Learning Machine," *Int. Conf. Electron. Commun. Syst. (ICECS), IEEE*, pp. 1–5, 2014.

[17] "Diagnosis of Diabetes Mellitus using KNN and PSO Classifier," pp. 32–38, 2017.

[18] "UCI Machine Learning Repository: Heart Disease Data Set." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/heart+Disease. [Accessed: 21-Apr-2018].

[19] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and Artificial Neural Network," *2010 Int. Conf. Comput. Commun. Technol.*, pp. 741–745, 2010.

[20] A. M. Mahmood and M. R. Kuppa, "Early Detection of Clinical Parameters in Heart Disease by Improved Decision Tree Algorithm," *2010 Second Vaagdevi Int. Conf. Inf. Technol. Real World Probl.*, pp. 24–29, 2010