IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BREAST CANCER PREDICTION USING MACHINE LEARNING

Gunjan Dharsandiya[1], Shubh Gohil[2], Mrs. Ancy Almeida[3]

[1,2]*UG Student, Dept. of Computer Engineering, Universal College of Engineering, Mumbai, India*
[3]*Assistant Professor, Dept. Of Computer Engineering, Universal College of Engineering, Mumbai, India*

Abstract: One of the most dreadful disease is breast cancer and it has a potential cause for death in women. Every year, death rate increases drastically due to breast cancer. An effective way to classify data is through classification or data mining. This becomes very handy, especially in the medical field where diagnosis and analysis are done through these techniques. Wisconsin Breast cancer dataset is used to perform a comparison between SVM, Logistic Regression, KNN and Decision tree algorithm. Evaluating the correctness in classifying data based on accuracy and time consumption is used to determine the efficiency of the algorithms, which is the main objective. Based on the result of performed experiments, the support vector machine shows the highest accuracy (96.6%) with the least error rate. ANACONDA Data Science Platform is used to execute all the experiments in a simulated environment.

Keywords: SVM, KNN, LR, Decision tree algorithm, efficiency, mining.

I.INTRODUCTION

Being the most frequently occurring cancer in women, breast cancer affects around 10% of women at some point in their life. It is the second leading contributor to women's death after lung cancer. Out of 25% of cancers in women 12% are caused by breast cancer.

Big Data has seen a rise in value due to it being used in derivation of business intelligence, business analytics and data mining to obtain reports and result predictions. Topics like medial science rise rapidly when certain approaches like data mining is applied due to better possibility of prediction of diseases, reducing medicine costs, improving health of patient by revamping the quality of healthcare along with value by saving people's lives through real time decisions. The paper provides you with a analysis of performance and comparison of accuracy in classification between the algorithms such as: Logistic Regression, SVM, KNN and decision tree algorithm, being the major influential algorithms of data mining used in the research community.

II. LITERATURE SURVEY

In machine learning and data mining, classification should be a crucial task. Researchers have already done lot of researches by applying machine learning algorithm on medical dataset for classification and data mining algorithm to find a pattern in dataset for faster calculation and prediction. Many of the approaches provide good accuracy and result.

In [1], Author have implemented algorithms like C4.5, ANN, SVM to find classification accuracy in breast cancer dataset. Author's research shows SVM had produced higher accuracy in classification.

In [2], Author's research is about finding classification accuracy using machine learning algorithm known as k-Nearest Neighbour with different values of k. For each value of k they have received a different result.

In [3], Decision tree classifier was implemented in project to find sensitivity, time consumed and mean accuracy of two data set WBCPD and WBCDD.

In [4], KNN algorithm was implemented to test the classification accuracy of breast cancer dataset with specificity, sensitivity and mean accuracy.

In [5], two models namely Logistic Regression and ANN was implemented. They were used to compare prediction accuracy breast cancer in mammography. Author's study says logistic regression performed well in prediction.

The performance and efficiency of the algorithms such as SVM, Decision tree, Logistic Regression and KNN were compared to the similar works mentioned above. The goal is to achieve the lowest error rate and best accuracy in analysing data. The performance and efficiency of these approaches are compared using: accuracy and time to build model. SVM scores highest classification accuracy (96.6%) and least error rate. Unlike the other classifiers which we have chosen for this research has classification accuracy in the range of 94% and 99%.

III.SYSTEM ACHITECTURE

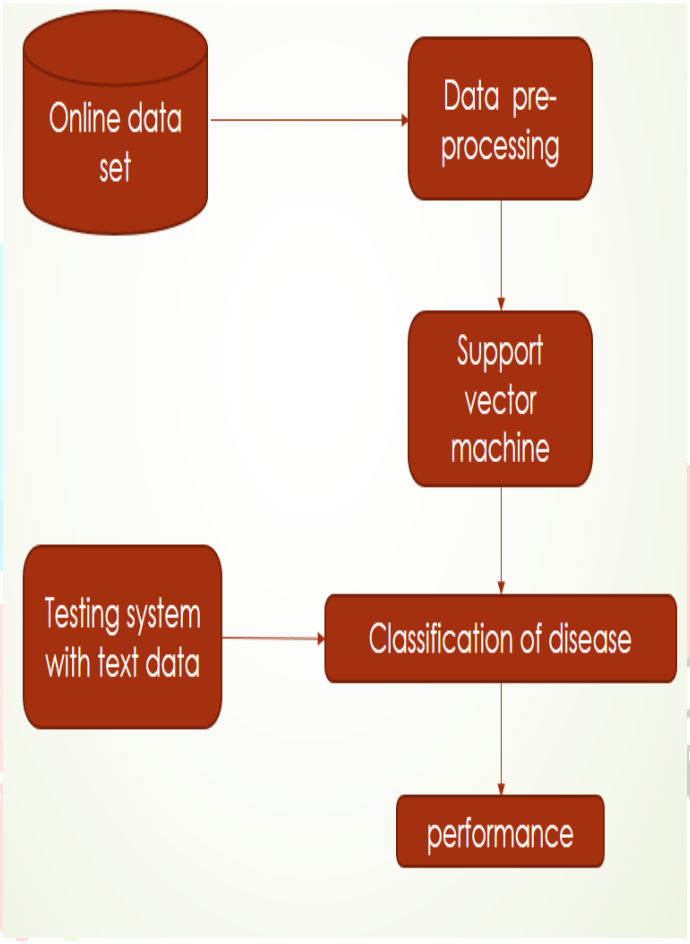

Figure 1. Architecture Diagram

As shown in architecture diagram we have download online dataset having 30 parameter and 570 real world cases. Also, we had transfer raw data into system understandable format through data pre-processing. Once data converted algorithm will start working and algorithm having high accuracy and good time efficiency will be selected. More we test and train the system it will provide higher accuracy. And there is higher chance of getting the perfect detection of disease.

IV.METHODOLOGY

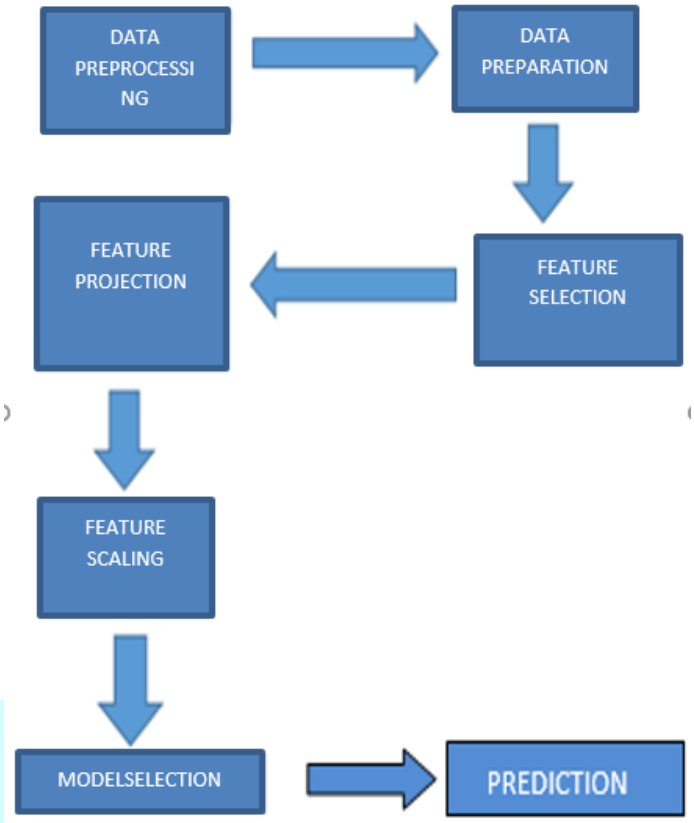

Figure 2. Methodology Diagram

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and lacking certain to contain many errors.

Data Preparation, where we load our data into a suitable place and prepare it for use in our machine learning training

In Feature selection we will select suitable subset of relevant features for model. Feature projection is used for transformation of high-dimensional space data to a lower dimensional space.

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations. We need to bring all features to the same level of magnitudes. This can be achieved by scaling.

In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B (Benign). So Classification algorithm of supervised learning is applied on it. We have chosen three different types of classification algorithms in Machine Learning. We can use a small linear model, which is a simple.

Prediction, or inference, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is real.

V. EXPERIMENTAL RESULTS

After creating predictive model, efficiency can be checked. For this, the models can be compared based on accuracy and time consumed. It was really hard to choose the algorithm which has higher performance, greater accuracy and efficiency, since all of them ended very close in accuracy. The time consumed and accuracy value of the algorithms from machine learning is shown in below table.

The accuracy of SVM is 96.6% which is higher than other models. We can say that the performance of SVM (96.6%) is better when it comes to classification with comparison to the other algorithm's accuracy obtained. Other algorithms have accuracy that varies between 94% and 99%.This means that the SVM portrays the highest correctly classified instance value and the lowest incorrectly classified instance value in comparison to the other classifiers.

Table 1. Experimental Result

Algorithm	Accuracy
LR	94.4%
SVM	96.6%
KNN	95.8%
Decision Tree	95.1%

VI. CONCLUSION & FUTURE WORK

In summary, SVM was able to show its efficiency on the basis of time and accuracy. SVM performs better when it comes to classification because First, it finds lines or boundaries that correctly classify the training dataset. Then, from those lines or boundaries, it picks the one that has the maximum distance from the closest data points.

Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables. We are intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy

REFERENCES

1. Fahad *et al.* ‘‘A survey of clustering algorithms for big data: Taxonomy and empirical analysis,’’ *IEEE Trans. Emerg. Topics Comput,* Sep. 2014.
2. M. I. Jordan and T. M. Mitchel ‘Machine learning: Trends, perspectives, and prospects,’’ *Science*, 2015.
3. L. G. Valiant, ‘‘A Bridging Model For Parallel Computation,’’ *Common. ACM*, vol. 33, no. 8, pp. 103–111, 1990.
4. Wang, D. Zhang and Y. H. Huang “Breast Cancer Prediction Using Machine Learning”(2018), vol. 66, No.7
5. Vikas Chaurasia and S. pal, ”Using Machine Learning Algorithm For Breast Cancer Risk Prediction And Diagnosis”(FAMS 2016) 83(2O16) 1064-1069
6. Joseph A. Cruz and David S. Wishart “Application Of Machine Learning In Cancer Prediction And Prognosis Cancer Informatics” 59-77. February 2007
7. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S “International Journal of Recent Technology and Engineering” (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019
8. DATASET:*http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29*