International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Triple-Technique Diagnosis Using Machine Learned Classifiers

Siddhant Gada[1], Het Sheth[2], Meet Chheda[3], Ashwini Swain[4], Kriti Srivastava[5]

[1, 2, 3, 4] Student, IT Department, D. J. Sanghvi College of Engineering, Vile Parle, Mumbai, India

[5]Assistant Professor, Computer Department, D. J. Sanghvi College of Engineering, Vile Parle, Mumbai, India

Abstract: ***In present scenario, breast cancer has become most common disease among women. Despite the fact, not all public hospitals have the facilities to diagnose breast cancer in India through mammograms. Delaying the diagnosing may increase the chances of cancer to spread throughout the body. Machine learning techniques have been benevolent in the detection and diagnosis of various diseases due to their accurate prediction performance. Various classifiers may provide differently desired accuracies and it is, therefore, exigent to use the most fitting classifier which provides the best accuracy. This paper documents a study of four machine learned classifiers, namely, Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), Logistic Regression (LR) and K-Nearest neighbour (KNN) on the Wisconsin Diagnosis Breast Cancer (WDBC) dataset. The performance of these algorithms has been analysed using classification accuracy and a confusion matrix. We have also introduced an ensemble of the above mentioned classifiers. The results show that the performance of Multi-Layer Perceptron is far superior to other classifiers.***

Keywords: Machine Learning, Support Vector Machine, Artificial Neural Network,k Nearest Neighbour, Logistic Regression

1. Introduction

1.1 Introduction to Breast Cancer

Breast cancer has become one of the most common diseases among women which account for a huge amount of deaths. Breast cancer is second leading cause of death among women after lung cancer [1]. Breast cancer may be hereditary, arising in women with germline mutations in tumor suppressor genes, or sporadic. Approximately 12% of breast cancer occurs due to the inheritance of an identifiable susceptibility gene or genes. In India, there has been an increase in the number of patients diagnosed with breast cancer in the younger age groups and more than 60% of the women are diagnosed with breast cancer at stage III or IV, affecting the survival rate and treatment pattern.

1.2 Machine Learning in Healthcare

Disease diagnosis is in the vanguard of Machine Learning (ML) research in medicine. In 2015, a report on cancer stated that over 800 medicines and vaccines were in a trial to treat cancer [10]. Jeff Tyner, in an interview, said that though being interesting, it presents a challenge of finding ways to work with the resulting data [15]. The idea of biologist/doctors working together with information scientist and data scientist has become the prime factor for the diagnosis of diseases.

Although doctors will never be completely replaced by robots and computers, machine learning is improving outcomes and transforming the healthcare industry through improved diagnosis, prediction and a more personalized care. It puts another arrow in the quiver of healthcare decision making. The success of machine learning comes with an increase in the amount of data and healthcare sector has a rich profusion of data available. The main issue of machine learning in healthcare is finding out how to effectively collect and use data for better results and treatment of a disease. In this paper we will be concentrating on using data for correct prediction. The proliferating applications of machine learning in healthcare are indicators of a potential future. Some of the leading tech giants such as Google, IBM, and Philips are using machine learning and artificial intelligence to predict diagnose and treat diseases such as cancer.

1.3 Breast Cancer Diagnosis using Machine Learning

There are two different types of breast cancer tumors i.e. malignant and benign. Malignant tumors are the cancerous tumors which can invade nearby tissues. If cells are non-cancerous then it is classified as a benign tumor. Since the cells are non-cancerous, they won't invade nearby tissues or spread to other parts of the body. The emergence of new technologies in the field of healthcare has led to the extraction of large amounts of data relating to breast cancer and has been made available to the medical research community. Hence, many researchers have employed various classification techniques to classify a tumor.

We intend to do the same using various classifiers to classify a breast cancer tumoras either malignant or benign. This paper shows a comparison of four classifiers, namely, SVM, KNN, MLP, and LRto predict the correct class, determine its accuracy, display it in a tabular form and find out the best classifier out of the four. It even shows that combining two or more classifiers does not lead to an increase in accuracy. Section II discusses the present literature work in prediction of breast cancer, Section III explains the proposed model, Section IV compares various classifier results and Section V concludes the paper.

2. Literature Review

Researchers have come up with various machine learning models. Use of KNN, SVM, Neural networks, Naïve Bayes, RVM and Decision Tree for the diagnosis of breast cancer

using the WDBC dataset is very common. Each classifier gives a different accuracy based on the parameters used to classify. In 2004, Tuba kiyan et. al. [6] used four different neural networks (RBF, PNN, GRNN, and MLP). The classification accuracies obtained by them are 96.18%, 97%, 98.8% and 95.74% respectively. In 2006, Elias Zafiropouloset. al. [7] made use of SVM to predict breast cancer. They got an accuracy of approximately 90% while sensitivity and specificity indices were also satisfactory.

In 2010, Leena Vig [13] presented an analysis of ANN, NB, SVM and Random Forest. The analysis showed that ANN, Random Forest and SVM had better accuracies, sensitivity and specificity than NB. In 2012, Gouda Magdy et. al. [2] used the WEKA data mining tool to compare the accuracies of different classifiers individually and even the accuracies an ensemble of classifiers. They used Naïve Bayes (NB), MLP, J48 (decision tree), SMO (SVM) and IBK (KNN). They achieved the highest accuracy amongst all classifiers using Sequential Minimal Optimization (SMO) i.e. 97.7153. Hence, they combined all other classifiers with SMO and saw that the fusion between SMO and MLP and the fusion between SMO and IBK gives the same highest accuracy as of SMO alone. In 2012, Ali Raad et. al. [5] used a MLP classifier in which the input layer had 9 neurons and one neuron in output layer. They achieved an accuracy of 94%.

In 2012, D. Lavanya and Dr. K. Usha Rani [9] used a hybrid approach which involved a CART classifier with feature selection, and bagging technique to evaluate the performance. Using CART they got accuracy of 92.97%. With CART and feature selection they got accuracy of 94.72% and using hybrid approach they got accuracy of 95.96%.In 2013, Ahmad LG et. al. [4] used three machine learning techniques to diagnose breast cancer and used the Iranian centre for Breast Cancer dataset. They used Decision Tree (C4.5), MLP and SVM and found out that SVM has the best accuracy of 95.7%. In 2014,BichenZhenget. al. [8] used a hybrid of k-means and SVM. K-SVM reduces feature space dimensions significantly. They got an accuracy of 97.38% using K-SVM. In 2016, Kathija et. al. [3] used ensemble classification technique in Naïve Bayes and SVM and performed 10 fold cross validation on both the classifiers. The result showed that performance of Naïve Bayes (95.65%) was greater than SVM (95.1%). In 2016, Animesh Hazraet. al. [14] showed that NB classifier gave maximum accuracy with only 5 dominant features compared to the other two classifiers.

3. Proposed Model

3.1 Architecture

There are several classifications techniques which can and have been used to classify the WDBC dataset. The method of combining classifiers in order to improve accuracy is an interesting approach. We aim at using four machine learning classification techniques, namely, MLP, KNN, SVM and LR in order to predict the correct class for the given set of parameters. Then we will combine the various algorithms and perform an analysis of our results.

Multi-Layer perceptron (MLP) is a supervised learning algorithm which is class of feed forward artificial neural network (ANN). An MLP consists of three or more layers, i.e.

- 1 Input layer
- 1 Output layer
- 1 or more Hidden layer

Except for input nodes, each layer is a neuron that uses some activation function like relu, adam, sgd. It learns a function by training on a dataset. Figure 1shows one hidden layer MLP.

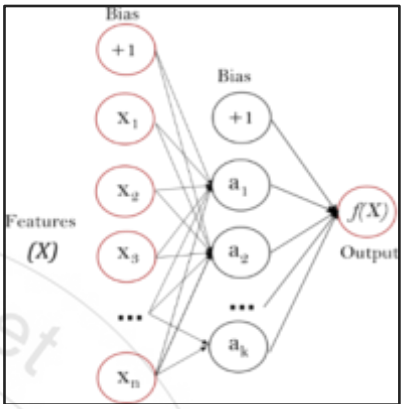

Figure 1: One hidden layer Multi-Layer Perceptron

The used parameters are:

- 31 inputs in the input layer.
- 500 neurons in our hidden layer.
- lbfgs optimizer which is an optimizer in the family of quansi-Newton methods.
- L2 penalty parameter (alpha) as 5.

KNN stands for K- Nearest Neighbours which is a classification algorithm that can work well with multi-class classification. It works by calculating the distance of a new input point, from the other training set points. We can set the number of neighbours to be checked for each class. So if an input point A is closer to 3 points of class B but only 2 points nearer to class C, then the new point belongs to class B. The used parameters are:

- n_neighbours(n_neighbours= 9): Specifies the number of neighbours.

A Support Vector Machine (SVM) is a supervised machine learning algorithm. It can be used for both classification and regression purposes. SVMs are more commonly used in classification problems and as such, this is what we will focus on in this paper. SVM's are based on the idea of finding a hyperplane that best divides a dataset into two classes. A hyperplane is a line that separates the data into classes. The used parameters are:

- Kernel: Specifies the model for SVM. The values can be linear, polynomial, and radial.
- C(C=0.1): penalty parameter for error.
- Gamma (gamma=0.1): kernel coefficient for models like polynomial, radial.
- Probability (probability=true): specifies whether we should enable probability estimates.

Logistic Regression is a classification algorithm. It makes predictions using probability and is best when it is used in a

binary classification. Logistic Regression can also solve multi-classification problems like whether it belongs to category A, B, C or D. For example, a credit card company typically receives thousands of applications for new card. The application contains several inputs such as gender, attributes, annual salary, past debit, etc. We need to categorize the people in to two types; good credit people and bad credit people. The parameters used are:

- Penalty (penalty='l1'): Used to specify the norm used in the penalization.
- C(C=0.5): Inverse of regularization strength.

3.2 Performance Evaluation Criteria

A confusion matrix is a visualization table that describes the classifier's accuracy in classification [11]. Each row depicts instances in a predicated class and each column depicts the instance in an actual class [12].

The four outcomes of the confusion matrix have the following meaning:

- True Negative (TN) is the class non-members which are classified as class non-members.
- False Positive (TP) is the class non-members which are classified as class members.
- False Negative (FN) is the class members which are classified as class non-members.
- True Positive (TP) is the class members which are classified as class members.

Table 1 shows the structure of a confusion matrix.

Table 1: Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

Accuracy of a classifier can be calculated from a confusion matrix using the formula given below:

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FN + FP}$$

4. Performance Analysis

The dataset that we have used has been taken from UCI's repository and is titled 'Wisconsin Diagnosis Breast Cancer' (WDBC) [16]. The WDBC dataset is a multivariate dataset with 569 instances, 32 attributes and no missing values. The first two attributes of the 32 attributes are ID number and Diagnosis (M= Malignant, B= Benign). The next 30 attributes consist of ten real-valued features used for each cell nucleus. They are: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension. Three different values have been recorded and documented for these ten real-valued features. They are: The mean, standard error and the worst (the mean of the largest three values). All the above feature values have been recorded with four significant digits.

We then ran our four classification algorithms on the WDBC dataset and visualised our observations in the form of a confusion matrix. The confusion matrices of the singular classifiers are given below followed by a bar graph that compares the accuracies of these four algorithms.

A) SVM

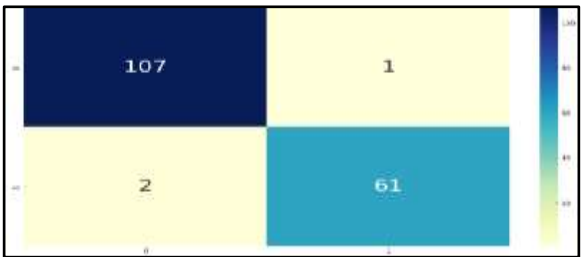

Figure 2: Confusion matrix using Simple Vector Machine

Figure 2 shows the confusion matrix of SVM whose values are 107, 61, 2, and 1 for TP, TN, FP, and FN.

B) KNN

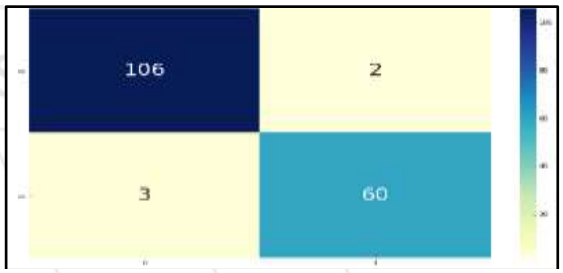

Figure 3: Confusion matrix using K-nearest neighbours

Figure 3 shows the confusion matrix of KNN whose values are 106, 60, 3, and 2 for TP, TN, FP, and FN.

C) LR

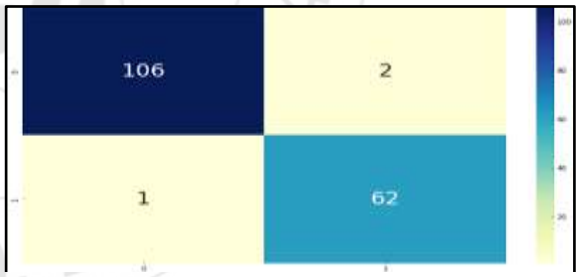

Figure 4: Confusion matrix using Logistic Regression

Figure 4 shows the confusion matrix of LR whose values are 106, 62, 1, and 2 for TP, TN, FP, and FN.

D) MLP

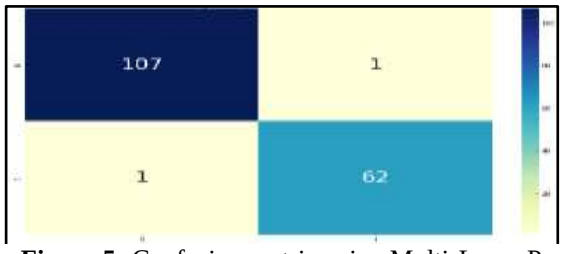

Figure 5: Confusion matrix using Multi-Layer Perceptron

Figure 5 shows the confusion matrix of MLP whose values are 107, 62, 1, and 1 for TP, TN, FP, and FN.

Table 2shows the accuracy of all the four classifiers individually.

Table 2: Accuracy of Individual Classifiers

Sr.No	*Model*	*Accuracy (%)*
1	SVM	98.25
2	LR	98.25
3	KNN	97.08
4	MLP	98.83

On checking the accuracy of each algorithm individually, we conclude that the accuracy for MLP in highest while it is lowest for KNN.

Figure 6 depicts the same accuracies with the help of a bar graph.

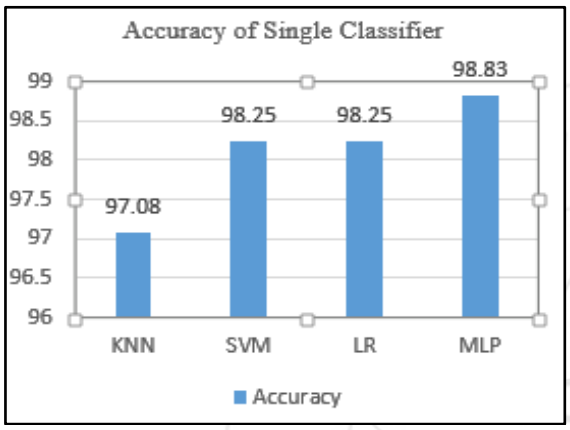

Figure 6: Graph for single classifier

We will use an ensemble of two or more classifiers with one classifier fixed as MLP. The accuracy of the new model does not show an increase in accuracy when compared to the individual classifier's highest accuracy.

Figure 7 visualises the confusion matrices of LR-MLP, MLP-SVM and MLP-KNN with the help of the bar graph.

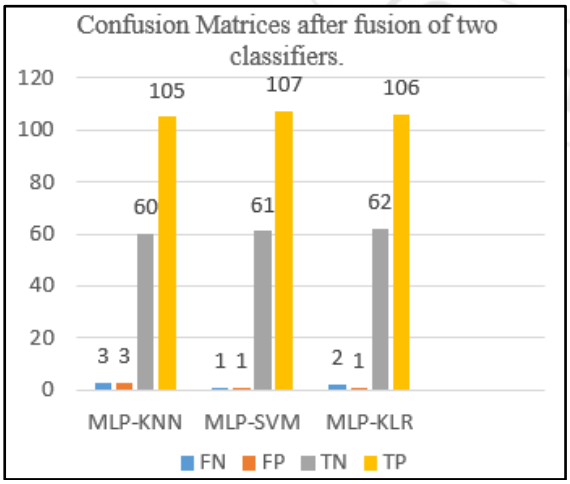

Figure 7: Confusion matrix of LR with other 3 classifiers

Figure 8 depicts bar graph of accuracies of the above fusion from its confusion matrices.

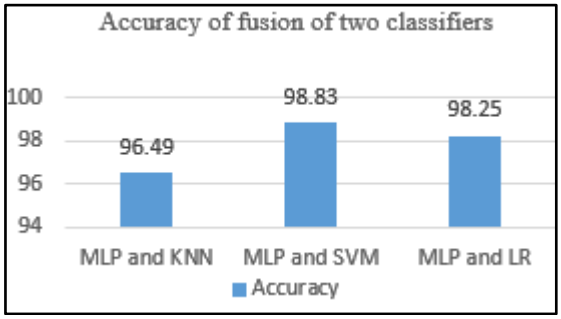

Figure 8: Accuracy after fusion of LR with other 3 classifiers

Figure 9visualises the confusion matrices of LR-MLP-SVM, LR-SVM-KNN and LR-KNN-MLP with the help of bar graph.

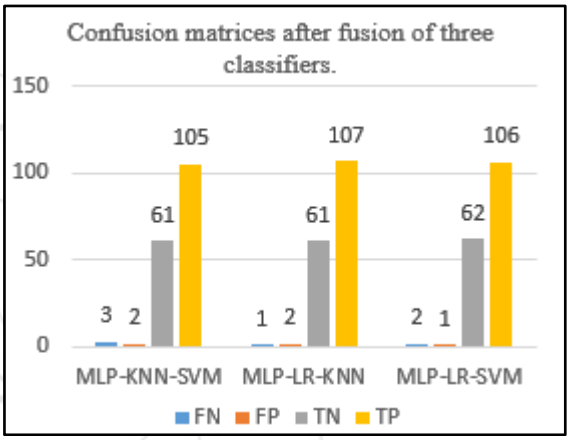

Figure 9: Confusion matrix after fusion of LR-KNN-SVM, LR-MLP-KNN, and LR-SVM-MLP

Figure 10 depicts bar graph of accuracies of above fusion from confusion matrices.

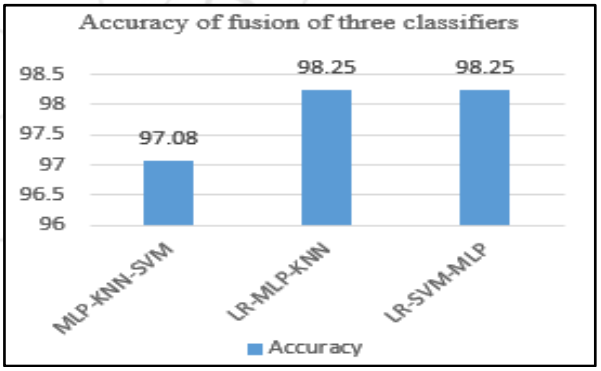

Figure 10: Accuracy after fusion of LR-KNN-SVM, LR-MLP-KNN, and LR-SVM-MLP

Table 3 shows accuracies of both two and three classifier fusion.

Table 3: Accuracies after combining 2 and 3 classifiers

S.No	Model	Accuracy (%)
1	MLP-KNN	96.49
2	MLP-SVM	98.83
3	MLP-LR	98.25
4	MLP-KNN-SVM	97.08
5	MLP-LR-KNN	98.25
6	MLP-SVM-LR	98.25

Hence, after evaluating the accuracies of both individual classifiers and the ensemble classifiers, we get MLP as the best classifier giving a testing accuracy of 98.83% and a fusion of two or more classifiers did not cause an increase in accuracy.

5. Conclusion

Being the most common disease among women, it accounts to accurately diagnosing breast cancer. Since mammograms and FNAC aren't available in all public hospitals, machine learning becomes an important aspect to diagnose breast cancer. It can add valuable minutes in a patient's treatment after getting diagnosed. In this paper, we evaluated four classifiers i.e. SVM, KNN, MLP and LR. We even evaluated an ensemble of these classifiers to check whether we get an increase in accuracy. The results showed that MLP has the best accuracy among all classifiers and an ensemble of two or more classifiers does not lead to an increase in accuracy. Multi-Layer Perceptron, known to have an edge over the other classifiers for binary classification (in this case Malignant and Benign), gave an accuracy of 98.83% and outdistanced all the other classifiers.

References

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report.
[2] Gouda I. Salama, M.B.Abdelhalim, and MagdyAbd-elghanyZeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers," International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012.
[3] Kathija, ShajunNisha, "Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques", International Journal of Innovative Research in Computer and Communication Engineering Vol.4, Issue12, December 2016.
[4] Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," J Health Med Inform 4: 124. doi:10.4172/2157-7420.1000124.
[5] Ali Raad, Ali Kalakech, and Mohammad Ayache," Breast cancer classification using neural network approach: mlp and rbf," The 13th International Arab Conference on Information Technology ACIT'2012 Dec. 10-13.
[6] Tuba kiyan, TulayYildirim "Breast cancer diagnosis using statistical neural networks", Journal of Electrical and Electronic Engineering, Year 2004, Volume 4, Number 2.
[7] Elias Zafiropoulos, IliasMaglogiannis, LoannisAnagnostopoulos, "A Support Vector Machine Approach to Breast Cancer Diagnosis and Prognosis," IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, Pages 500-507.
[8] BichenZheng, Sang Won Yoon, and Sarah S.Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-Means and Support Vector Machine," Expert System with applications Vol.41, Issue 4, Part 1, March 2014, Pages 1476-1482.
[9] D.Lavanya, and Dr.K.Usha Rani, "Ensemble decision tree classifier for breast cancer data" International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012.
[10] American Association for Cancer Research "Medicines in Development for Cancer" (Report 2015).
[11] J. Han and M. Kamber,"Data Mining Concepts and Techniques", MorganKuffman Publishers, 2000.
[12] David M W Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness& Correlation," Technical Report SIE-07-001, December 2007.
[13] LeenaVig, "Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset,"OALibJ, Vol.1 No.6, September 2014, page 1-7.
[14] AnimeshHazra, Subrata Kumar Mandal, and Amit Gupta, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", International Journal of Computer Applications, Volume 145, No 2, July 2016.
[15] Interview:http://www.livemint.com/Companies/uriDWin9uFqGSAzunT76bI/Microsoft-develops-AI-to-help-cancer-doctors-find-the-right.html. (available)
[16] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.