Classification of Benign or Malignant Tumor Using Machine Learning

MKalaiyarasi¹, RDhanasekar¹, S Sakthiya Ram¹, P.Vaishnavi¹
¹Assistant Professor, Department of Electronics and Instrumentation, Bannari Amman Institute of Technology, Sathyamangalam, Erode 638 401,
kalaiyarasime@gmail.com, dhanaekar@bitsathy.ac.in, sakthiyaram@bitsathy.ac.in, vaishnavi@bitsathy.ac.in

Abstract. The abnormal growth of the cell in the human body may lead to tumor. There are almost 100 types of cancer that affect the different parts of the human. The affected human may have symptoms like lump, abnormal bleeding, prolonged cough, weight loss etc... And depends on the part where the tumor formed. Breast cancer is one of the types among the 100 types of cancer and it is most commonly found on the female than the male. Breast cancer is a disease that caused due to the overgrowth of the cell in the breast area and they started to form lump over the breast. The people who are suffering from the breast cancer will have many emotional side effects and they are supposed to undergo a lot of pain in their day to day life. The most important problem for the healthcare people is earlier prediction. So, it can be rectified by employing machine learning algorithms in the prediction stage. In this paper the classification algorithms are used to classify whether the tumor is benign or malignant. The supervised learning algorithms of machine learning such as logistic regression, Support vector machine and K Nearest neighbour algorithm are usually used to analyse the tumour detection. Stacking ensemble method used in order to combine the entire three algorithms is proposed and the performance of algorithm is compared with the logistic regression, Support vector machine and K Nearest neighbour algorithm in order to get an efficient model for the classification.

Keywords: Tumor detection, Machine learning.
1. Introduction

For teaching machine to handle data more effectively, the machine learning comes to picture. Use machine learning technique, in-case after analysing data, if the pattern interpretation and information extraction are not possible (1). The machine learning worth increases as the dataset quantity gets more. Many applications like process industries, medicine, defence and banking uses machine learning for extracting significant data and information. Many algorithms are available to make the machine to learn themself (2).

Breast cancer is a dangerous disease that leads to death and this type of cancer mostly affect only women in the entire world (3). For reducing the death rate and control the spreading of cancer to the other parts of the body. Image processing technique is used to detect the cancer at starting stage (4). Breast cancer may spread over the other organs of the body through blood or lymph system and it may cause other tumors to grow. The type of treatment procedures that need to be followed will be varying according to the type of the cancer and its stage. For every individual patient the treatment procedure includes at least one of the following methods like surgery, chemotherapy, and radiation therapy. Breast cancer may occur in two different categories like malignant and benign tumors and it is the difficult task for the physicians to classify the type of tumor. The malignant and benign tumor is shown in the figure 1. So, the health care domain requires some reliable methodologies to predict the tumor cell and its type. The automation in the tumor identification can be achieved by employing machine learning algorithm for the detection purpose. The earlier detection of the disease will help the doctors to cure the disease so; it may reduce the cause of death and can save many patients life.

![Figure 1: The benign and malignant tumor](image_url)

Machine learning is a technique where, the model will label from the given dataset according to the type of learning preferred. It builds the mathematical model based on the provided sample data. It can be mainly used for prediction of the label, classify the label and form a cluster according to their input parameters. In order to classify the tumor from the malignancies, the supervise algorithm can be used to classify the required part that may be considered as tumor. The classification will come under the type of supervised learning. In this study, the dataset is taken from Kaggle website and the three different supervised algorithm is used to classify the dataset and the output of each algorithm is compared with the proposed method by the metrics value.

The main objective of machine learning is classification of data. The classification may be anything either binary or multiple. The application of binary classifications like Decision tree, Boosted Tree, KNN, SVM, Logistic Regression and Random Forest. Is needed in many fields. These popular ML algorithms have been applied to many fields, such as recognition of hand pose (5), face recognition (6), text separation (7), emotion identification (8), object detection at multi view and point (9), prediction of molecular hybridization (10), image and graphic (11), (12), machine vision (13), stability estimation of hard rock pillar (14), packet classification (15), data mining (16), Chess Games (17), remote control vehicles (18) and classification of MRI sequence images (19).
1.1 Data
The dataset for tumor detection is taken from Kaggle website. The dataset contains radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension of each cell nucleus. The mean, standard error and largest value of the above mentioned features were computed and it results in 30 features. The target contains tumor types like benign and malignant. The features are really very large, it leads to usage of more memory and execution time. So the important features were extracted based on the correlation. By using the heatmap plot the correlation between the features and target was found. The correlation plot is given in the figure 2. From the plot the less correlated features like symmetry standard error, smoothness standard error, texture standard error, fractional dimensional mean and standard error were found and dropped.

Figure 2. Correlation between the features and tumor

1.2 Logistic Regression
For classification, one of the famous algorithm is Logistic Regression. It is used for both binary classification like ‘0’ or ‘1’, ‘Yes’ or ‘No’ and multiple classifications. The base of this algorithm is linear regression [20]. The Sigmoidal function is used for modelling the data which eliminate the drawbacks of linear regression classification. The sigmoidal function is given in the equation 1.

\[ g(z) = \frac{1}{1 + e^{-z}} \]  

The threshold value is very important for classification. The setting of threshold value changes based on the application. Precision and recall plays a vital role to decide the threshold value.

The feature [input] matrix is denoted as

\[ X = \begin{bmatrix} 1 & \cdots & x_{ij} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{ij} \end{bmatrix} \]  

Where in \( x_{ip} \),
\( n \rightarrow \)ith observation
\( j \rightarrow \)jth feature

Recall Linear regression

\[ h(x_i) = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{np} \]  

\[ (x_i) = \beta^T x_i \]  

Where,
\( h(x_i) \rightarrow \) Classified target for nth observation
\( \beta \rightarrow \) Regression coefficient vector/matrix
\[
\beta = \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_p
\end{bmatrix} \quad (5)
\]

By applying sigmoidal function in linear regression
\[
h(x_i) = g(\beta^T x_i) \frac{1}{e^{-\beta^T x_i}} \quad (6)
\]

The sigmoidal graph is shown below

![Sigmoidal Graph](image)

From the graph it is infer that
\[
g(z) = 0 \text{ at } z \rightarrow -\infty \quad (7)
\]
\[
g(z) = 1 \text{ at } z \rightarrow \infty \quad (8)
\]

After describing conditional probabilities for 2 targets
\[
P(y_i = 1|x_i; \beta) = h(x_i) \quad (9)
\]
\[
P(y_i = 0|x_i; \beta) = h(x_i) \quad (10)
\]

The equation 1 and 2 can be combined as
\[
P(y_i|x_i; \beta) = (h(x_i))^{y_i}(1 - h(x_i))^{1-y_i} \quad (11)
\]

Likelihood parameters of \( \beta \)
\[
L(\beta) = \prod_{i=1}^{n} P(y_i|x_i; \beta) \quad (12)
\]

Apply log on both side
\[
\log(L(\beta)) = \log(\prod_{i=1}^{n} P(y_i|x_i; \beta)) \quad (13)
\]
\[
\log(L(\beta)) = \sum_{i=1}^{n} y_i \log(h(x_i)) + (1 - y_i) \log(1 - h(x_i)) \quad (14)
\]

Cost Function of logistic Regression
\[
J(\beta) = \sum_{i=1}^{n} -y_i \log(h(x_i)) - (1 - y_i) \log(1 - h(x_i)) \quad (15)
\]

Gradient descent for minimizing the cost function
\[
\frac{\partial J(\beta)}{\partial \beta_j} = (h(x) - y)x_j \quad (16)
\]

Where,
- \( y \rightarrow \) Actual response
- \( h(x) \rightarrow \) Predicted response

For getting minimum \( J(\beta) \)
\[
\beta_j = \beta_j - \alpha \sum_{i=1}^{n} (h(x_i) - y_i)x_{ij} \quad (17)
\]

Where,
- \( \alpha \rightarrow \) Learning rate
- The \( \beta \) value updated until the error minimized.

By using the Logistic Regression Algorithm, the classification of cancer will be done by using the dataset after data pre-processing techniques. Then the confusion matrix of Logistic Regression is given.
in the figure 4. This figure gives the details of True Positive, True Negative, False Positive and False Negative of cancer detection using Logistic Regression.

![Confusion Matrix](image)

**Figure 4.** Confusion matrix of Logistic Regression

### 1.3 Support Vector Machine

The support vector machine comes under supervised learning. It is suitable for both prediction and classification. The target data plotted in n-dimensional space, n is nothing but the number of features in the dataset. After plotting the hyper-plane has to draw. In 1992, the SVM which was first projected by Vapnik et al. It depends on VC dimension concept. Now, the SVM algorithm has certain constraints. The performance of SVM purely based on kernel selection. Still, the perfect selection of kernel for some problems are not found. The prevailing SVM algorithm focused on penalty coefficient [25]

**Rules for hyper-plane:**

1. Hyper-plane should segregate 2 classes very well.
2. The distance between the hyper-plane and the nearest data in the class is called margin. The hyper-plane should have maximum margin for 2 classes.
3. Different kernel tricks are available. Based on the non-linearity of data distribution, the kernels are selected.

The details of margin, hyper planes of Support Vector Machine classification is shown in the figure 5

![Support Vector Machine](image)

**Figure 5.** Support Vector machine working [24]

The hyperplane line is used to segregate the dataset into two classes. The hyper plane is a straight line. The equation of straight line is

\[ y = a \times x + b \]  

(18)

The equation can be reframed as

\[ a \times x + b - y = 0 \]  

(19)

The compact form of the equation

\[ W \cdot X + b = 0 \]  

(20)
Where
\[ X = (x, y) \]  \hspace{1cm} (21)
\[ W = (a, -1) \]  \hspace{1cm} (22)

The SVM classifies data by the given format
\[ y_i = 1, \text{If } w \cdot x_i + b \geq 1 \]  \hspace{1cm} (23)
\[ y_i = 0, \text{If } w \cdot x_i - b \leq 1 \]  \hspace{1cm} (24)

Margin is the distance between hyperplane and the support vector. The margin will be represented as
\[ M = \frac{1}{||w||} \]  \hspace{1cm} (25)

For optimal classification the margin should be maximum
\[ \max \frac{1}{||w||} \text{ or } \min ||w|| \]  \hspace{1cm} (26)

For better result L2 optimization is used
\[ \min \frac{||w||^2}{2} \]  \hspace{1cm} (27)

Sometimes very few class ‘A’ object falls in class ‘B’. This is called slack variable. If tries to make the slack variable zero the margin will reduce. So regularization parameter ‘C’ is introduced to maintain the trade-off between the margin width and slack variable.
\[ \min (w, b) \left\{ \frac{||w||^2}{2} + C \sum_i \xi_i \right\}, y_i (w \cdot x + b) \geq -1 - \xi_i \]  \hspace{1cm} (28)

For non-linear distribution of data, kernel tricks is used for classification. The kernel just map the non-linear data into high dimensional space for easier classification.

Different Kernels:
- Linear Kernel: \( K(x_i, x_j) = x_i x_j \)
- Polynomial Kernel: \( K(x_i, x_j) = (x_i \cdot x_j + c)^d \)
- RBF Kernel: \( K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \)

Cost Function of SVM
\[ J(\beta) = \sum_{i=1}^{n} y_i \max(0, 1 - \beta^T x) + (1 - y_i) \max(0, 1 - \beta^T x) \]  \hspace{1cm} (29)

By using the SVM Algorithm, the classification of cancer will be done by using the dataset after data pre-processing techniques. Then the confusion matrix of SVM is given in the Figure 6. The Figure 6 gives the details of True Positive, True Negative, False Positive and False Negative of cancer detection using SVM.

![Figure 6. Confusion matrix of SVM](image)

1.4K-Nearest Neighbor

The KNN algorithm is also a supervised algorithm. It is used for both classification and prediction problems. The KNN classify the data based on the majority of nearest data to the tested data [21]. The KNN was first proposed by Cover and Hart in 1968, it is a type of classification algorithm. KNN is usually called lazy learning method. The explicit learning phase and training process are not available in KNN [26]. The K value chosen in order to reduce training error and validation error. If K is small
the training error will be zero but the validation error will become high. If K is high both errors will be high. So the K has to be chosen in order to minimise the two errors. K-Nearest neighbor for classification is shown in figure 7. The K value is nothing but the number of neighbour data taken for classification. Always choose odd number as a K value.

![Figure 7. K-Nearest neighbor for classification [22]](image)

The nearest neighbour is calculated by using the Euclidean distance

\[ d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots (q_n - p_n)^2} \]  

(30)

The Euclidean distance is calculated for all target points in the dataset. Then based on the K values minimum distance of K data points is taken. Based on the target majority, the classification is done. The number of samples in the region is denoted by \( k(y) \)

The posterior probability \( p(y/x) \) is given

\[ P(y/x) = \frac{p(x/y)p(y)}{P(x)} = \frac{k[y]}{K} \]  

(32)

The classification is done by using the following equation

\[ g(x) = \begin{cases} 1, & k[y = 1] \geq k[y = -1], \\ -1, & k[y = -1] \geq k[y = 1] \end{cases} \]  

(33)

By using the KNN Algorithm, the classification of cancer will be done by using the dataset after data pre-processing techniques. Then the confusion matrix of KNN is given in the Figure 8. The Figure 8 gives the details of True Positive, True Negative, False Positive and False Negative of cancer detection using KNN.

![Figure 8. Confusion matrix of KNN](image)

2. Proposed Method

Ensemble stacking technique is used for creating new model. It associate the choices from multiple models and create new model from those model for improving the performance of the tumour detection process [23]. The step involved in proposed method is given.

1) The training dataset divided into small parts. For each training dataset the Logistic regression model is fitted to 9/10 part of the set and test take place in 1/10 part of the set.

2) Then fit the logistic regression model to all the training dataset
3) By using this model, the classification is done for all the testing dataset.
4) Repeat the steps 2 to 4 for SVM and KNN model, it give some other classification results for training and testing dataset
5) The classification results from the training dataset are used as an input for building a novel model.
6) By using this model the final classification is made on classified testing dataset

By using the proposed method, the classification of cancer will be done by following the above procedure. Then the confusion matrix of proposed method is given in the Figure 9. The Figure 9 gives the details of True Positive, True Negative, False Positive and False Negative of cancer detection using proposed method.

![Confusion Matrix](image)

**Figure 9.** Confusion matrix of proposed method

### 3. Results and Discussion

The breast cancer is classified by using 3 conventional machine learning algorithm and it is compared with the proposed stacking ensemble method by using the metrics. The metrics used in this research are accuracy, precision and recall. The formula used to calculate those metrics are given in the below equation

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{34}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{35}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{36}
\]

The bar-graph is used for the comparison of proposed method with the conventional methods. The figure 10 displays different algorithms and its metrics value in percentage. From the below graph it is clear that the proposed method give better accuracy and recall. It shows that the proposed method gives more true prediction and less false prediction values than the conventional methods like SVM, KNN and Logistic Regression.
Figure 10. Performance comparison of Algorithm

Table 1. Performance comparison in numeric value

| Metrics       | SVM  | KNN  | Logistic Regression | Proposed Method |
|---------------|------|------|---------------------|-----------------|
| **Accuracy**  | 96.49| 95.90| 97.07               | 98.24           |
| **Precision** | 98.30| 98.27| 96.77               | 96.87           |
| **Recall**    | 92.06| 90.47| 95.23               | 98.41           |

This table 1 evidently displays the value of three metrics like accuracy, precision and recall. From this table the conclusion is made that the proposed algorithm gives better result than the conventional algorithms.

4. Conclusion
The breast cancer dataset are taken from the kaggle website. The data are split into 7:3 ratio for training and testing. The more important features are taken based on the correlation matrix. The three conventional algorithms like SVM, KNN and logistic regression are applied to find the classification model. Then the proposed method also used to predict the classification model. After designing the model, based on the metrics the more efficient model for classification is found. The proposed method gives better satisfactory result for tumour classification whether the given data is benign or malignant. In future for getting more than 99% metrics, the optimization techniques are planned to use.

Reference
[1] M. Bowles, ‘Machine Learning in Python: Essential Techniques for Predictive Analytics’, John Wiley & Sons Inc., ISBN: 978-1-118- 96174-2
[2] W. Richert, L. P. Coelho, ‘Building Machine Learning Systems with Python’, Packt Publishing Ltd., ISBN 978-1-78216-140-0
[3] J. Ferlay, I. Soerjomataram, R. Dikshit et al., ‘Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012’, International Journal of Cancer,
Comparative performance of six supervised learning methods

J. G. Elmore, C. K. Wells, C. H. Lee, D. H. Howard, and A. R. Feinstein, ‘Variability in radiologists’ interpretations of mammograms’, New England Journal of Medicine, vol. 331, no. 22, pp. 1493–1499, 1994.

T. Parag and A. Elgammal, ‘Unsupervised learning of boosted tree classifier using graph cuts for hand pose recognition’, Proc. Brit. Mach. Vis. Conf., pp. 1259-1268, Sep. 2013.

X. Peng, S. Seltur, V. Govindaraju and S. Ramachandrula, ‘Using a boosted tree classifier for text segmentation in hand-annotated documents’, Pattern Recognit. Lett., vol. 33, no. 7, pp. 943-950, 2012.

Demirkir and B. Sankur, ‘Face detection using boosted tree classifier stages’, Proc. IEEE 12th Signal Process. Commun. Appl. Conf., pp. 575-578, Apr. 2004.

M. Day, ‘Emotion recognition with boosted tree classifiers’, Proc. 15th ACM Int. Conf. Multimodal Interact., pp. 531-534, 2013.

B. Wu and R. Nevatia, ‘Cluster boosted tree classifier for multi-view multi-pose object detection’, Proc. IEEE Int. Conf. Comput. Vis., pp. 1-8, Oct. 2007.

W. Zhu, Y. Han, H. Wu, Y. Liu, X. Nan and Q. Zhou, ‘Predicting the results of molecular specific hybridization using boosted tree algorithm’, Concurrency Comput. Pract. Exper.

C. Shen, Y. Chen, G. Yang and X. Guan, ‘Toward hand-dominated activity recognition systems with wristband-interaction behavior analysis’, IEEE Trans. Syst. Man Cybern. Syst..

Bera and D. Bhattacharjee, ‘Human identification using selected features from finger geometric profiles’, IEEE Trans. Syst. Man Cybern. Syst..

G. Yu, J. Yuan and Z. Liu, ‘Unsupervised random forest indexing for fast action search’, Proc. IEEE Comput. Vis. Pattern Recognit., pp. 865-872, Jun. 2011.

J. Zhou, X. Li and H. S. Mitri, ‘Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction’, Natural Hazards, vol. 79, no. 1, pp. 291-316, 2015.

J. R. Quinlan, ‘Learning efficient classification procedures and their application to chess end games’ in Mach. Learn, Berlin, Germany: Springer, pp. 462-482, 1983.

X. Dong, M. Qian and R. Jang, ‘Packet classification based on the decision tree with information entropy’, J. Supercomput., vol. 74, pp. 1-15, Jan. 2018.

M. Jaworski, P. Duda and L. Rutkowski, ‘New splitting criteria for decision trees in stationary data streams’, IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 6, pp. 2516-2529, Jun. 2018.

M. McClelland and M. Campbell, ‘Probabilistic modeling of anticipation in human controllers’, IEEE Trans. Syst. Man Cybern. Syst., vol. 43, no. 4, pp. 886-900, Jul. 2013.

Y. Zhang, D. Kwon and K. M. Pohl, ‘Computing group cardinality constraint solutions for logistic regression problems’, Med. Image Anal., vol. 35, pp. 58-69, Jan. 2017.

Lei Liu, ‘Research on Logistic Regression Algorithm of Breast cancer Diagnose data by Machine Learning’, International Conference on Robots & Intelligent System (ICRIS) - e-ISBN: 978-1-5386-6580-0,2018

Osisanwo F.Y, Akinsola J.E.T, ‘Supervised Machine Learning Algorithms: Classification and Comparison’, International Journal of Computer Trends and Technology (IJCTT) – e-ISSN: 2231-2803 Volume 48 Number ,2017

Ayon Dey, ‘Machine Learning Algorithms: A Review’, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016

Optiz, R. Maclin, ‘Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose’, Journal of Artificial Intelligence Research, 11, Pages 169- 198, 1999

D. Meyer, ‘Support Vector Machines - The Interface to libsvm in package e1071’, August 2015

X. Li, M. Fang, H. Wang and J.-J. Zhang, ‘Supervised transfer kernel sparse coding for image classification’, Pattern Recognit. Lett., vol. 68, no. 12, pp. 27-33, Dec. 2015.
[26] X. Shen, L. Niu, Z. Qi and Y. Tian, ‘Support vector machine classifier with truncated pinball loss’, Pattern Recognit., vol. 68, no. 8, pp. 199-210, Aug. 2017