Using Machine Learning algorithms for breast cancer classification and risk prediction

Vanshika Tyagi
Dept of Information Technology
Galgotia’s College of Engineering and Technology
Uttar Pradesh, India
vanshikatyagi989@gmail.com

Tushar Mehrotra
Dept of Information Technology
Galgotia’s College of Engineering and Technology
Uttar Pradesh, India
tusharmehrotra86@gmail.com

Pranjal Singh
Dept of Information Technology
Galgotia’s College of Engineering and Technology
Uttar Pradesh, India
Pranjalsingh.singh24@gmail.com

Ranjit Kumar
Assistant Professor,
Dept of Information Technology
Galgotia’s College of Engineering and Technology
Uttar Pradesh, India
ranj.kumar@galgotiacollege.edu

Abstract—Breast Cancer is the most often identified cancer among women and major reason for increasing mortality rate among women. As the diagnosis of this disease manually takes long hours and the lesser availability of systems, there is a need to develop the automatic diagnosis system for early detection of cancer. Data mining techniques contribute a lot in the development of such system. We have used classification techniques of machine learning in which the machine learns from the past data and can predict the category of new input. The cancerous cells are classified as Benign (B) or Malignant (M). There are many algorithms for classification and prediction of breast cancer: Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes (NB), Logistic Regression, Random Forest Classifier and k Nearest Neighbours (kNN). In this project, Support Vector Machine (SVM) on the Breast Cancer data is used. The dataset is also trained with the other algorithms: KNN, Naives Bayes, Random Forest Classifier, Logistic Regression and CART and the accuracy of prediction for each algorithm is compared.

Keywords—Breast Cancer, knn, naives bayes, RFC, logistic regression, CART, SVM

I. INTRODUCTION
Breast Cancer is the prime reason for demise of women. It is the second dangerous cancer after lung cancer. In the year 2018 according to the statistics provided by World Cancer Research Fund it is estimated that over 2 million new cases were recorded out of which 626,679 deaths were approximated. Of all the cancers, breast cancer constitutes of 11.6% in new cancer cases and come up with 24.2% of cancers among women. The prior diagnosis of Breast Cancer can enhance the prediction and survival rate notably [1]. Cancerous cells are detected by performing various tests like MRI, mammogram, ultrasound and biopsy. The dataset used in this project contains features that are computed from a digitized image of a fine needle aspiration (FNA) biopsy of a breast mass. They describe characteristics of the cell nuclei present in the image. Diagnosis of breast cancer is done by classifying the tumour. Tumours can be either benign or malignant. Malignant tumours are more harmful than the benign. Unfortunately, not all physicians are expert in distinguishing between the benign and malignant tumours and the classification of tumour cells may take up to 2 days. Machine learning algorithms are used to predict the type of cancerous cells efficiently and accurately. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The different algorithms used are: Support Vector Machine (SVM), Logistic Regression, Decision Tree (CART), Naive Bayes (NB), Random Forest Classifier and k Nearest Neighbours (k-NN) each algorithm performs in a different way depending on the dataset and the parameter selection[5].

A. Logistic Regression
Logistic regression uses Sigmoid function to transform linear regression into the logit function. Logit is nothing but the log of Odds. And then using the log of Odds it calculates the required probability. So, let’s understand first what the Sigmoid function is and what is the log of Odds. A Sigmoid function is a mathematical function having a characteristic
“S”-shaped curve or Sigmoid curve. Often, Sigmoid function refers to the special case of the logistic function shown in the first figure and defined by the formula (source: Wikipedia):

\[ S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}. \]

So Sigmoid function gives us the probability of being into class 1 or class 0. So generally we take the threshold as .5 and say that if \( p \) > .5 then it belongs to class 1 and if \( p \) < .5 then it belongs to class 0. However, this is not a fixed threshold. This varies based on the business problem. And what threshold value should be, we can decide it with the help of AIC and ROC curves.

B. K – Nearest Neighbour (KNN)

KNN makes predictions using the training dataset directly. Predictions are made for a new instance \((x)\) by searching through the entire training set for the \( K \) most similar instances (the neighbours) and summarizing the output variable for those \( K \) instances. For regression this might be the mean output variable, in classification this \( \text{XXX-X-XXXX-XXXX-X/XXX/XXX-00} \) ©20XX IEEE might be the mode (or most common) class value. To determine which of the \( K \) instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance. Euclidean distance is calculated as the square root of the sum of the squared differences between a new point \((x)\) and an existing point \((x_i)\) across all input attributes \(j\).

Euclidean Distance \((x, x_i) = \sqrt{\sum (x_j - x_{ij})^2} \)

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, \( k \) is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the \( k \) training samples nearest to that query point.

C. Naives Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Bayes’ Theorem is stated as:

\[ P(h|d) = \frac{P(d|h) * P(h)}{P(d)} \]

Naive Bayes is a classification algorithm for binary (twoclass) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. Assume that we have a dataset with two classes of data inside. We have an equation for the probability of a piece of data belonging to Class 1:\( p(h|d) \), We have an equation for the class belonging to Class 2:\( p(h|d) \). To classify a new measurement with features \((h,d)\), we use the following rules: If \( p(h|d) > p(h|d) \), then the class is 1. If \( p(h|d) > p(h|d) \), then the class is 2.

D. Decision Tree (CART)

Decision Trees (DTs) are probably one of the most useful supervised learning algorithms out there. As opposed to unsupervised learning (where there is no output variable to guide the learning process and data is explored by algorithms to find patterns), in supervised learning your existing data is already labelled and you know which behavior you want to predict in the new data you obtain. This is the type of algorithms that autonomous cars use to recognize pedestrians and objects, or organizations exploit to estimate customers lifetime value and their churn rates. DTs are ML algorithms that progressively divide data sets into smaller data groups based on a descriptive feature, until they reach sets that are small enough to be described by some label. They require that you have data that is labelled (tagged with one or more labels, like the plant name in pictures of plants), so they try to label new data based on that knowledge.

DTs algorithms are perfect to solve classification (where machines sort data into classes, like whether an email is spam or not) and regression (where machines predict values, like a property price) problems. Regression Trees are used when the dependent variable is continuous or quantitative (e.g. if we want to estimate the probability that a customer will default on a loan), and Classification Trees are used when the dependent variable is categorical or qualitative (e.g. if we want to estimate the blood type of a person). The importance of DTs relies on the fact that they have lots of applications in the real world. Being one of the mostly used algorithms in ML, they are applied to different functionalities in several industries.

E. Random Forest Classifier

The random forest approach is a bagging method where deep trees, fitted on bootstrap samples, are combined to produce an output with lower variance. However, random forests also use another trick to make the multiple fitted trees a bit less correlated with each other: when growing each tree, instead of only sampling over the observations in the dataset to generate a bootstrap sample, we also sample over features and keep only a random subset of them to build the tree. Sampling over features has indeed the effect that all trees do not look at the exact same information to make their decisions and, so, it reduces the correlation between the different returned outputs. Another advantage of sampling over the features is that it makes the decision making process more robust to missing data: observations (from the training dataset or not) with missing data can still be regressed or classified based on the trees that take into account only features where data are not missing. Thus, random forest algorithm combines the concepts of bagging and random feature subspace selection to create more robust models.

F. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where \( n \) is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. Support Vectors are simply the coordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line). Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane
that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples. It is worth noting that probabilistic outputs can also be obtained for SVMs figure below illustrates how an SVM might work in order to classify tumours among benign and malignant based on their size and patients' age. The identified hyperplane can be thought as a decision boundary between the two clusters. Obviously, the existence of a decision boundary allows for the detection of any misclassification produced by the method.

II. LITERATURE SURVEY

From [2] the Breast Cancer dataset was obtained. In [3] Authors proposed a Support Vector Machines (SVMs) based classifier in comparison with Bayesian classifiers and Artificial Neural Networks for the prognosis and diagnosis of breast cancer disease. The paper provides the implementation details along with the corresponding results for all the assessed classifiers. A SVM model is implemented for the breast cancer diagnosis and prognosis problem using the Breast Cancer data datasets. The optimized SVM algorithm performed excellently, exhibiting high values of accuracy (up to 95.91%), specificity (up 95.67%) and sensitivity (up to 95.84%).

[4] states that artificial neural network have been the most widely used predictive technique in medical prediction, though its structure is difficult to understand. The paper lists out the benefits and limitations among various machine learning techniques such as Decision trees, Naïve Bayes, neural network and SVM. In [5], each algorithm performs in a different way depending on the dataset and the parameter selection. For overall methodology, KNN technique has given the best results. Naïve Bayes and logistic regression have also performed well in diagnosis of breast cancer. But SVM is the most suited technique for recurrence/non-recurrence prediction of breast cancer.

III. MATERIALS AND METHODOLOGY

Materials that we have used include: Python software for coding and breast cancer data. Our methodology involves use of machine learning techniques such as: SVM, KNN, decision trees, Logistic Regression, Random Forest Classifier and Naïve bayes.

A. Dataset

The Breast Cancer data dataset was obtained from the UCI machine learning depository (available at: http://archive.ics.uci.edu/ml). The dataset contains 357 cases of benign breast cancer and 212 cases of malignant breast cancer. The dataset contains 32 columns, with the first column being the ID number, the second column being the diagnosis result (benign or malignant), followed by the mean, standard deviation and the mean of the worst measurements of ten features. There were no missing values. The features are obtained from a digitized image of a fine needle aspiration biopsy of the tumour. These features describe the nuclei of the cell. The different features are as shown:

<table>
<thead>
<tr>
<th>TABLE I. FEATURES USED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Radius</strong></td>
</tr>
<tr>
<td><strong>Texture</strong></td>
</tr>
<tr>
<td><strong>Perimeter</strong></td>
</tr>
<tr>
<td><strong>Area</strong></td>
</tr>
<tr>
<td><strong>Smoothness</strong></td>
</tr>
<tr>
<td><strong>Compactness</strong></td>
</tr>
<tr>
<td><strong>Concavity</strong></td>
</tr>
<tr>
<td><strong>Concave points</strong></td>
</tr>
<tr>
<td><strong>Symmetry</strong></td>
</tr>
<tr>
<td><strong>Fractal dimension</strong></td>
</tr>
</tbody>
</table>

B. Methodology (Proposed Method)

The dataset is divided into training set and testing set. 80% of the data is used to train the system and the remaining 20% is used for testing. From the dataset, we analyse and build a model to predict if a given set of symptoms lead to breast cancer. The machine learning algorithms are trained on the training data, and tested on the untrained data. If the model is excessively complex, such as having too many parameters, it is likely to lead to the problem of overfitting. Likewise, if the model is excessively simple that cannot capture the underlying trend of the data, underfitting occurs. Both overfitting and underfitting lead to poor predictive performance. There are several techniques to overcome overfitting, such as cross validation, regularization and drop out. One of the most commonly used methods is k-fold cross validation, where the original data is randomly partitioned into k equal sized subsamples. Out of the k subsamples, one subsample is used to testing the model, and the remaining k-1 subsamples are used to train the model. The k results are then averaged to generate one single estimation. One advantage of k-fold cross validation is each testing subsample is used exactly once. Support vector machine (SVM), a binary
classifier, searches the hyperplane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of each class from the hyperplane. SVMs are a more recent approach of ML methods applied in the field of cancer prediction/prognosis. Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples.

IV. FLOWCHART

The data obtained from the patients is used to form a dataset. The dataset is divided into training and testing data, and it is ensured that the dataset has no incomplete values. The six machine learning algorithms are applied and the accuracy of prediction of each algorithm is compared. The algorithm with best accuracy is used for prediction.

A. Data Exploration

The distributions of the mean, standard error and worst average of the 10 features extracted from the fine needle aspiration slides show that compactness, concavity, fractal dimension, smoothness and symmetry each have relatively small values for the measurement. Perimeter, radius and texture each have relatively large values for the measurement, with areas that show the largest measurement value and amount of variation for all three measurements. From the distribution visualization, we can see overall the malignant diagnosis class has relatively higher mean for all the attributes.

B. Attributes Correlation

Among the mean measurement of the 10 attributes, we can see several of them are highly correlated between each other. The red around the diagonal suggests that attributes are correlated with each other. The yellow and green patches suggest some moderate correlation and the blue boxes show negative correlations.

C. Count of Benign(B) and Malignant(M):

From the dataset we can observe that the number of BENIGN patients are more than the number of MALIGNANT patients and this is proved by the bar graph below obtained in our jupyter notebook.

Malignant: 290
Benign: 357

D. Performance Comparison
From the initial run, it looks like Gaussian NB, KNN and CART performed the best given the dataset (all above 92% mean accuracy). Support Vector Machine has a surprisingly bad performance here. However, if we standardise the input dataset, its performance will improve.

After standardising, the data set, accuracy of SVM drastically improves as shown below.

![Performance Comparison Graph](image)

**E. Calculation of Accuracy:**

When we calculate accuracy we observe the output to be as shown below:

Accuracy score 0.951228

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95</td>
<td>75</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>0.93</td>
<td>0.95</td>
<td>39</td>
</tr>
<tr>
<td>Average / Total</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>114</td>
</tr>
</tbody>
</table>

**F. Confusion Matrix:**

```
M B
M [[74 1]
B [0 39]]
```

We can see that we achieve an accuracy of 94.7% on the held-out test dataset. From the confusion matrix, there is only 1 case of mis-classification. The performance of this algorithm is expected to be high given the symptoms for breast cancer should exhibit certain clear patterns.

**VI. CONCLUSION AND FUTURE SCOPE**

Our work mainly focused in the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. The analysis of the results signifies that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain. 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.

As our dataset contains 32 attributes dimensionality reduction contributes a lot in decreasing the multidimensional data to a few dimensions. Of all the three applied algorithms Support Vector Machine, k Nearest Neighbor and Logistic Regression, SVM gives the highest accuracy of 94.7% when compared to other algorithms. So, we propose that SVM is the best suited algorithm for the prediction of Breast Cancer Occurrence with complex datasets. Further research in this domain such as the creation of SVM classes like LIBSVM has taken place. Fine tuning of parameters used in algorithms can result in better accuracy. Furthermore, this can also be implemented on a cloud platform for ease of usage.

**REFERENCES**

1. Yi-Sheng Sun, Zhao Zhao, Han-Ping-Zhu, “Risk factors and Preventions of Breast Cancer” International Journal of Biological Sciences