How to Cite:
Kumar, H., Anuradha, A., Maini, T., Yadav, D. K., & Mohan, S. (2022). Advance cataloguing method for breast cancer detection. *International Journal of Health Sciences*, *6*(S1), 11208–11227. https://doi.org/10.53730/ijhs.v6nS1.7708

Advance cataloguing method for breast cancer detection

Harish Kumar*
Professor, Department of CSE-SET, Noida International University, Gr. Noida
*Corresponding author

Anuradha
Assistant Professor, Ajay Kumar Garg Engineering College, Ghaziabad

Tarun Maini
Associate Professor, Department of CSE-SET, Sharda University, Gr. Noida

Dileep Kumar Yadav
Professor, Department of CSE-SET, Galgotia University, Gr. Noida

Sumit Mohan
Assistant Professor, Sunder Deep Engineering College, Ghaziabad

***Abstract*---**This Data mining is a technique for extracting useful information from large amounts of data. In large databases, enormous patterns may be examined and evaluated utilizing statistics and artificial intelligence. Data mining can be used to anticipate future trends or uncover hidden patterns. Classification, clustering, association rules, regression, and outlier identification are examples of data mining techniques. The data mining technology is receiving a lot of traction in the healthcare industry. In the discipline of bioinformatics, several researchers are using data mining techniques. Bioinformatics is the science of storing, retrieving, organizing, interpreting, and exploiting data from biological sequences and molecules. A prediction is a statement regarding a future event based on the current condition. The major intend of this work is to predict the microarray cancer using machine learning (ML) algorithms. Different phases are comprised in the prediction of microarray cancer. This research makes the implementation of voting-based classification algorithm. The suggested algorithm assists in optimizing the performance up to 2% while predicting the microarray cancer.

***Keywords*---**Machine learning, Feature selection, Classification, Prediction, Breast cancer.

Introduction

Breast cancer seems to have a high incidence and fatality rate because it is the most common cancer in women. According to the most comprehensive cancer research, BC alone is expected to account for 25percent on average of all newly diagnosed cases and 15percent respectively of all cancer deaths among women worldwide [1]. Previous studies have proven about the dangers of BC since the beginning, and as a result, much early research has been done in the therapy of BC. The fatality rate has shown a constant and dropping pattern throughout the preceding many years, thanks to the efforts of specialists and early detection measures. According to the Cancer Research Institute (CRI) in the United Kingdom, the five-year survival rate for BC is over 100 % when diagnosed at the earliest stage, but can be as low as 15 % when diagnosed at the most advanced stage. The comprehensive overview of data mining breast disease prediction is presented in this chapter.

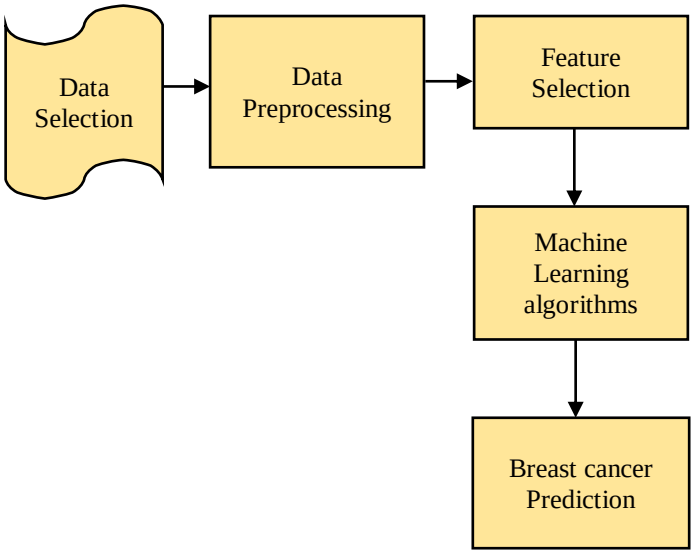

Figure 1: General process of Breast cancer prediction based on Machine Learning

Data Mining

As of late our capacities of both gathering and creating information have been expanding quickly. The far-reaching utilization of bar codes for most of business items, the computerization of numerous legislature and business exchanges, and the advances in information assortment devices have given us colossal measure of information. A huge number of databases have utilized in business handing, government organization, logical and designing information the executives, and numerous different applications. It has been observed that the number of such databases is rapidly rising as a consequence of the availability of fantastic and cost-effective data frameworks [2]. This enormous growth in data and datasets has necessitated the development of new strategies and equipment capable of quickly and organically converting prepared data and information into useful data and information. Thus, data mining has become an exploration zone with

expanding significance. Data mining, also known as KDD (Knowledge Discovery in Databases), is a cycle of nontrivial data extraction through databases of specific, already opaque and potentially useful data (for instance, informational rules, restrictions, and consistency).

With KDD, attractive information, possibility of establishing or significant level data can be extracted from connected datasets in databases and examined from multiple angles, and vast information bases along these lines serve as rich and solid hotspot for knowledge generation and verification. Many experts regard data and information mining from large databases as a key research topic in organized method and machine learning, and many mechanical companies regard it as a significant area with the potential for significant profits [2].

It is possible to apply the discovered knowledge or information to perform different tasks such as decision making, process control, data managing, and query processing etc. Scientists in a wide range of fields, such as database frameworks, knowledge base frameworks, computerized reasoning, AI, machine learning, knowledge gaining, spatial databases, and data visualization have indicated incredible interest in data mining. Figure 1.2 sketched below represents data mining as a stage in an iterative KDD (Knowledge Discovery in Databases) process.

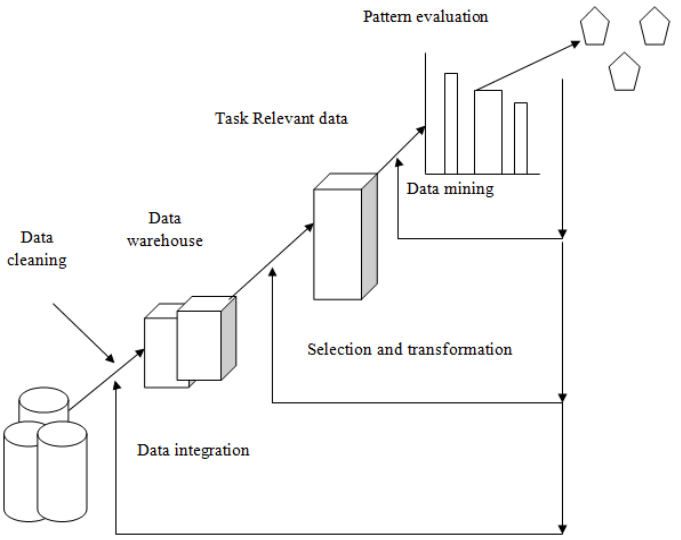

Figure 2: Data mining as a step in the process of knowledge discovery

The Knowledge Discovery in Databases measure entails a few phases that go from a collection of raw data to some type of new data. The repetitive cycle is made up of the following advancements:

- Data cleansing is also known as data purification. It is a phase in which noisy and unnecessary data are removed from the collection [3].
- During data integration, many information sources, which are frequently heterogeneous, may be combined into a single source.
- The data relevant to the inquiry is picked and retrieved from the gathered data during data selection.

- The stage of data transformation is also known as data consolidation. It is a stage in which the selected data is transformed into structures that are suitable for the mining procedure.
- Data mining is a critical advancement in which clever tactics are used to extract potentially useful patterns.
- Pattern evaluation distinguishes thoroughly intriguing situations when interpreting knowledge based on specified measures.
- The final stage of knowledge presentation is when the discovered knowledge is spoken to the client. This fundamental advancement employs perception tactics to aid consumers in understanding and deciphering data mining outcomes [4].

Data mining includes a wide range of machine learning algorithms to achieve various undertakings. These algorithms endeavor to fit a model to the information. The algorithms inspect the information and decide a model that is nearest to the attributes of the information being analyzed. The nature of a data mining model can be either predictive or descriptive. A predictive model makes a forecast about estimations of data utilizing realized outcomes found from various data sources. Predictive modelling might be made dependent on the utilization of other chronicle data. For instance, the data mining of electronic clinical records is performed using the headways in predictive analysis. Thusly, the danger of complexities which are expanding due toa chronic disorder can be anticipated. Data mining tasks using a predictive model include classification, regression, time series analysis, and prediction. Prediction may likewise be utilized to demonstrate a particular sort of data mining capacity.

A descriptive model recognizes examples or connections in data. In contrast to the predictive model, a descriptive model fills in as an approach to investigate the properties of the data inspected, not to foresee new features. Some popular descriptive data mining tasks are clustering, summarization, association rules, and sequence discovery [5].

Microarray cancer prediction

Breast cancer is one of the most well-known and a lethal tumour in women all over the world, and it is still the leading cause of cancer death among women in poor countries. Breast cancer has increased in frequency and mortality in recent years, putting women's lives and well-being at risk and causing enormous monetary, social, and family issues? Despite the fact that breast cancer is more common in younger women, postmenopausal women are also at risk of developing the disease. Investigating the characteristics of cancer aggressiveness in postmenopausal women, as well as discovering significant data from diverse sources in order to provide clinical analysis and therapy to rational dynamic and clinical investigation is crucial. Currently, certain Western countries have established a system for evaluating breast cancer risk assessment models for various explicit severity criteria. The risk variables contained in the risk of breast cancer forecasting model are primarily divided into two types of designs known as genetic and statistical models, as indicated by the risk variables incorporated in the model. Genotype models can compute the danger of breast malignant growth

at a particular age, and can likewise figure the likelihood of involving mutations through the incidence of breast cancer and ovarian malignancy in the family.

Artificial intelligence and, specifically, machine learning models have an obvious history in malignancy research and useful execution [11]. The great majority of these studies employ machine learning methodologies to demonstrate disease progression and identify useful characteristics that are then used in a classification strategy, with a focus on malignancy vulnerability, recurrence, and durability. The utilization of various ML models in malignancy research gives immense space to different applications. Artificial Neural Networks (ANNs) and Decision trees (DTs) have been utilized in disease prediction and detection for almost 30 years. Various models dependent on Support Vector Machine (SVM) applied to cancer prediction issues have been utilized for about a few decades. Different models for forecast of malignancy advancement and result have likewise been utilized for a few examinations.

Today, not exactly a portion of information science and bioinformatics techniques are utilized by ML-driven models with a wide scope of uses, from diagnostics to expectation and forecast in malignancy. All this examination contemplates are worried about utilizing ML strategies to recognize, classify, identify, or differentiate tumours and different malignancies, just as to foresee disease growth. Breast cancer prediction works dependent on machine learning models possess a considerable part of the contemporary examination in this domain. Machine learning algorithms are contributing significantly to diagnose and predict breast cancer by implementing classification schemes to recognize individuals with breast cancer, differentiate cancerous from non-cancer costumiers and to predict disease [12]. Precise clustering can additionally help clinicians to recommend the most proper treatment system. There are a few investigations considering the impact of an outfit of ML procedures to foresee the danger of breast cancer. These strategies provide more accurate predictive results on the breast cancer dataset in comparison to the outcomes of earlier approaches.

a. Data Selection: In this stage, a dataset is selected for extracting information. The datasets can be openly accessible (e.g., on the web) or they may result from a joint effort among foundations and exploration groups, not accessible for the overall population. Gathered data for the most part incorporates demographic features (age, stature, weight record), physiological richness factors (time of menarche, time of menopause, age of the first pregnancy, post-menopausal hormone discharge level), illness history and hereditary variables (history of non-cancerous breast cancer sickness, family background of breast cancer), social propensities (smoking, drinking) and so forth

b. Pre-processing: Pre-processing assignments are performed to diminish noise and increment the consistency of information [13]. The pre-processing steps generally tended to in various researches are data standardization/normalization, and missing data management. Two basic methods of information pre-processing are data cleaning, standardization (Min-Max change), and normalization (Z-Score conversion). All these tasks have been explained below:

- Data cleaning: Data cleanup schedules "clean" data by filling in incomplete data, smoothing noisy data, identifying and removing outliers, and settling irregularities. Clients who acknowledge that the data is dirty are unlikely to believe the results of any data analysis that has been done on it. Furthermore, soiled data can wreak havoc on the mining process, resulting in untrustworthy results. Although most mining programs include a few approaches for dealing with fragmentary or ambiguous information, they are usually ineffective. They may focus on avoiding generalization error the data to the capability being modelled, all things considered. As a result, passing acquired data through specific data cleaning algorithms is an important pre-processing step.
- Normalization: This refers to the scaling of a characteristic between its most basic and most extreme features, whereas standardization rescales the attributes in order to maintain a standard average distribution (zero mean and unitary standard deviation). The purpose of certification scheme is to create attributes with nearly comparable scales and scope of measurements (e.g., age, haemoglobin levels), such that none has a greater impact on the classification task than the others. Missing Data (MD) is a common test inside the healthcare industry that can occur from a wide range of functions.

c. Feature Selection: In data mining domain, a classification technique can profit altogether from utilizing just significant information regarding learning performance and learned outcomes, for example, improved conceivability [14]. Feature selection is a broadly applied procedure for discovering applicable information by eliminating irrelevant and superfluous information. A dataset comprises numerous features and so as to fabricate a precise normalization model of the association between breast cancer and its attributes, applicable attributes ought to be picked out to be utilized in the classification technique.
Feature selection plays an important role in accurate prediction. Data mining algorithms make use of feature selection strategies for choosing the ideal features from the dataset. These features ought to be stacked directly into the memory for pre-processing. Feature selection is a cycle wherein just the subset of the suitable features is chosen. This technique recognizes a small number most significant features and aids in result prediction. It is a type of dimensionality reduction utilized for pre-processing. The contrast between feature selection and dimensionality reduction is the primary strategy (Feature choice) which diminishes the features without changes the dataset. As the feature selection method is related to fewer parameters, it will reduce complexity. There are different strategies for feature selection algorithms implemented in classification. They are defined as:
   - Filter method: The filter methods select the features on the basis of scores in different statistical correlations.
   - Wrapper method: These methods perform feature selection using a greedy approach. These techniques evaluate all possible combination and generate the outcome for Machine learning.
   - Embedded Scheme: The embedded approach merges the benefits of both abovementioned methods. The inducer possesses its specific FSA (either explicit or implicit). The methods provide an example of this embedding by inducing logical conjunctions. Several classic machine learning algorithms,

such as decision trees or artificial neural networks are involved in this method.

d. Machine Leaning algorithms: Throughout the years, many Machine Learning (ML) algorithms have been utilized to predict the occurrence of breast cancer. A potential scientific categorization for the arrangement of these strategies comprises in partitioning them into "supervised, semi-supervised and unsupervised algorithms. Supervised learning is essentially an equivalent word for classification. The supervision in the learning originates from the labelled instances in the training dataset. Unsupervised learning, on the other hand, is basically an equivalent for clustering [15]. The learning cycle is unsupervised due to the non-labelling of the input instances. Ordinarily, classes within the data can be discovered with the help of clustering. Semi-supervised learning is a branch of machine learning algorithms that uses both labelled and unlabelled instances while learning a model. In one methodology, class models are learned using labelled instances while unlabelled instances refine the edges amongst classes. For a binary issue, a set of instances relevant to one class can be considered as the positive instances whereas those relating to the different class are regarded as the negative instance. Some prominent machine learning algorithms are:

K-Nearest Neighbors

The K Nearest Neighbor method is used in pattern recognition and grouping. It is commonly used in the analysis of predictions. The KNN algorithm recognizes available information examples that are nearest to new data when it appears. Features that differ to a substantial extent might have had a significant impact upon that interval between data occurrences. Number is defined as the most common K neighbor among the Training instances sample size in order to perform classification. The calculation in this example finds K adjacent neighbors of the original input pattern. When all of the data points are in metric space, measuring distance becomes a huge difficulty [16]. If N represents the number of neighbors in this procedure, then N observations are measured using different distance metric:
Minkowski Distance:

$Dist(x,y) = (\sum_{i=1}^{n}|x_{i-}y_i|^p)^{\frac{1}{p}}$.....(1)

In this expression if p=1, then it is known as Manhattan distance, p=2, represents Euclidean distance, and p=∞ denotes Chebyshev distance. Amidst various selections, Euclidean distance is universally implemented. Among these K neighbors, the calculation verifies the number of data relevant to every class, and then, it relegates the fresh data point to the classification which frames the more considerable share.

Decision Trees

DTs are forests in which examples are sorted by feature values and then categorized. Each node in a DT is used to demonstrate a feature for classification, and each branch is used to represent a value that the node can assume. In the beginning, instances are classified and sorted based on their attribute values at

the root node. C5.0, Id3, or CART are the DTs which are used to effectively handle real-world datasets. The C5.0 is a DT that was designed ID3 based on information gain owing to the selectivity of multi-valued characteristics ID3. C5.0 was created to address this issue, including the estimation of the data gain ration for each characteristic [17]. After that, the property with the highest Information Gain Ratio value is chosen as the training dataset's root node. The characteristic with the highest gain ratio is chosen for splitting in order to reduce the amount of information required for predicting a specific instance in the extracted features separation. The following is the Gain Ration evaluation for attribute A:

$$GainRatio(A) = \frac{Gain(A)}{Split\ Info(A)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

$Gain(A) = info(D) - Info_A(D)$.....(3)
Where D is the training dataset
$Info(D) = -\sum_{i=1}^{n} p(ci)\log_2 P(C_i)$.....(4)

$P(C_i) = |C_{i,D}|/|D|$ where $|C_{i,D}|$is the number of the tuples included in the class C_i used in the training dataset and $|D|$ is the number of the tuples of the training dataset, and n represents the number of the values of the class.

$$Info_A(D) = \sum_{i=1}^{n} (\frac{|a_{i,D}|}{|D|} \times (-\sum_{j=1}^{m} \frac{|C_{JD}|}{|a_{i,D}|}\log_2 \frac{|C_{J,D}|}{a_{i,D}}) \quad \ldots\ldots\ldots(5)$$

Where $|a_{i,D}|$is the number of the tuples of the value a_J of the attribute A in the training dataset and $|D|$ denotes the tuples of the training dataset and n is the number of the values of attribute A. $|C_{J,D}|$ corresponds to the number of the tuples of class $|C_J|$ associated with the value ai of the attribute A and m, the number of the classes of class C.

Support Vector Machine

The idea of SVM, which was proposed by Vapnik based on the statistical learning hypothesis, has become a fundamental part in ML strategies [18]. SVM was at first created for twofold classification, however it tends to be productively stretched out for multiclass issues with boundless use in fields of pattern recognition, handwriting recognition, text classification, and so forth. The vital component of a SVM classifier is to discover an improved decision boundary that speaks to the biggest partition (most extreme edge) amid the classes. The standard of SVM begins from providing a solution of linear separable issues, then it focusses to handle nonlinear issues. The method of solving nonlinear issues is mapping training instances from the new limited dimensional space to a higher dimensional space to acknowledge direct detachability. SVM is one of the most famous methodologies for predicting breast cancer forecast.

More precisely, support vector machine techniques generate a higher dimensional space or group of separating hyperplane in a high- or infinite-dimensional space, which can be used for regression, classification, and other tasks such as outlier detection. Intuitively, a good partition is achieved by the hyperplane with the

greatest separation from any class's closest preparation information purpose (the ostensibly utilitarian edge), which carries a lower classifier speculation blunder. A hyperplane with the greatest distance to the smallest training data point of any category (known as margin) offers a reasonable separation. This method also reduces the classification model's generalization error.

Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANN) can be characterized as a reasoning model having configuration similar to the human brain [19]. In the course of recent many years, ANNs have been utilized progressively by an ever-increasing number of analysts, and become a functioning exploration territory. ANNs have managed the cost of various victories with extraordinary advancement in breast cancer classification and diagnosis in the initial stages. Figure 1.3 represents a common ANN model consisting a chain of layers, i.e. input, hidden and output layers.

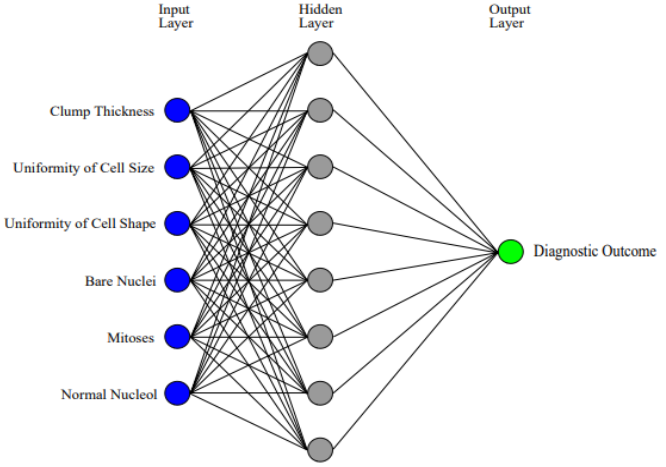

Figure 3: A typical ANN model for BC classification

As shown by the figure, layers are made out of interconnected neurons which contain an activation function for nonlinear change to fortify the nonlinear articulation capacity. The information layer gets the information and afterward sends the information to a hidden layer which is utilized for data processing and giving the training results to the output layer. The output layer generates the outcomes of classification [20]. In any case, contingent upon the issues, the way toward training an ANN may include long causal sequence of computational phases. Since 1986, a new effective gradient descent algorithmic approach known as backpropagation (BP) has been widely used, particularly for clinical data. It was made by generalizing the Widrow-Hoff learning strategy to numerous layer networks and nonlinear differentiable transfer functions. In spite of the fact that the BP approach is utilized, it actually has a few shortcomings when used with tremendous and convoluted data. The estimations of BP are broad and, therefore, training is moderate, subsequently an original BP calculation is hardly utilized in real-time. Specialists are attempting to improve the BP algorithm to increment computational effectiveness.

e. Prediction: Inside this step, selected characteristics are mapped to the training model, which is then used to classify the given features and forecast liver disease. A professional doctor labels the obtained breast cancer dataset in order to create predictions. The classification is constructed as a multi-class problem, with medical data being classified into many classifications. As a result, each stage corresponds to a certain type of breast cancer. Based on the identified essential features, this approach can determine the likelihood of a patient developing breast cancer.

Literature Review

Naveen, et.al discussed that the main concern of medical community was to predict the breast cancer [21]. This research focused on predicting the breast cancer accurately. The most effectual ensemble ML model was constructed using the breast cancer Coimbra dataset that was extracted from UCI. This model assisted in enhancing the system performance with un-biasing. There were 6 ML algorithm implemented namely DT, SVM, MLP, KNN, LR and RF and their predictive analysis was compared with ensemble as well as non-ensemble methods.

Anusha Bharat, et.al analysed that the breast cancer was a disease due to which a great number deaths were occurred in every year [22]. Various algorithms were available in order to classify and predict the breast cancer such as SVM, DT, NB and KNN.

Tanishk Thomas, et.al described that the major cause of breast cancer was the division of abnormal cell in the breast itself due to which benign or malignant cancer was developed [23]. Therefore, the earlier prediction of breast cancer was very essential. The life of numerous patients was saved if they got the proper treatment.

Madhuri Gupta, et.al suggested an ensemble model for predicting the breast cancer for which 4 ML schemes were employed namely SVM, LR, DT and KNN [24]. The outcomes demonstrated that the suggested model performed more accurately over the conventional single classification system. The weight was allocated to every classification technique applying Sequential Least Squares Programming technique.

Sidharth S. Prakash, et.al stated that the major intend was to develop a DNN algorithm for predicting the malignancy of the breast cancer [25]. The Wisconsin breast cancer data set that was taken from UCI had employed to acquire the data. The optimization of intended algorithm was obtained with the help of early stopping mechanism and dropout layers to deal with the over fitting. The obtained F1 score of this algorithm was computed 98. This CAD model was not the replacement of the expertise of professional doctors and medical practitioners. However, it assisted in performing the diagnosis procedure successfully. The future work would focus on expanding this approach further for predicting the malignancy of breast cancer image data with the utilization of CNNs.

Mamatha Sai Yarabarla, et.al investigated that the CAD systems were acted significantly while recognizing the breast cancer and they were employed to mitigate the death rate among women [26]. This approach aimed at the implementation of the current developments of CAD systems and its related methods. The RF algorithmic approach was utilized to detect and predict the breast cancer. The major intend of this project was that the person had to be predicted as normal or affected with BC. The machines were trained in the ML for learning and performing without any explicit program. This trained data was deployed to classify that the person had a BC or normal.

Tahreem Shouket, et.al (2019) discussed that the major purpose was that the OS and a number of years for the survival of patient who had not any sign of breast cancer had to be predicted [27]. The experiments were conducted considering the dataset of female patients of Pakistan who were diagnosed with breast cancer. The data that was collected from INMOL hospital of Pakistan had utilized for training the ML classification algorithms such as NB, DT, SVM, RF, JRip and AdaBoost. The outcomes of experiment exhibited that the JRip ML algorithm performed more efficiently for overall survival and disease-free survival in comparison with the others. The presented approach assisted the patients and doctors in predicting the upcoming situation.

Methodology

The smooth functioning of this part is very essential for the healthy lifestyle. If any kind of disease occurs in this organ, the other body parts are also disturbed. The researchers have often utilized the computer aided data which is extracted from the enormous databases. A number of businesses implement the DM systems and methods in extensive manner. The medical domain has made the deployment of these DM methods in order to predict the various diseases. Various risk factors may cause the breast cancer. There are different stages included while predicting the breast cancer. These stages are mentioned as:

A. Data Collection: This phase entails gathering information from many healthcare bodies in order to conduct the trials.
B. Data pre-processing: The whole process has been completed, and the data has been analyzed in order to deploy Machine Learning methods, as well as data pre-processing. To improve the efficacy of the training framework, superfluous attributes are eliminated from the data in order to transmit clean and de-noised data.
C. Feature selection: During this step, a subset with particularly unique qualities for identifying microarray cancer is deployed. The degree of sophistication of attributes is related to these chosen attributes. The introduced approach for selecting attributes uses the Random Forest (RF) algorithm. This technique builds the tree - like layout of the most significant bits using 100 as the estimated value. The RF algorithm is used to choose the acceptable or significant features for microarray cancer predictions.

   RF (Random Forest) is a probabilistic technique for building a DT (decision tree) ensemble model from random subsets of characteristics and bagged examples of training data. Even in the noisy environments in predictive qualities, the effectiveness of RF is found to be good. When the number of

observations is less than the number of features, this algorithm is useful. When Random Forest is combined with a randomizing system, the accuracy attained is proven to be effective on high dimensional data. This is owing to the bagged sample data's random sampling of a subspace of characteristics from hundreds of features in the creation of a tree. As a result, the tree derived from this type of packed subspace of attributes has a lesser accuracy in predicting the characteristic that has an impact on the Random Forest's final forecast.

A training dataset is presented as $\mathbb{L} = \left\{(X_i, Y_i)_{i=1}^{N} \middle| X_i \in RM, Y \in \{1,2,\dots,c\}\right\}$, Xi denotes the attributes, Y is used to show a class response feature, N represents the number of training samples and the number of attributes is denoted with M. A RF model is defined in Algorithm 1, let Yk be the prediction value of tree Tk given input X. The prediction of random forest with K trees is expressed as:

$\hat{Y} = majority\ vote\{\hat{Y}^k\}_1^K$ ..... (6)

Every tree is constructed from a packed sample set, which means which only three different of the observations in L, or in-bag samples, are used. Around one of the samples are not used, and these samples are referred to as out-of-bag (OOB) samples, which help estimate the prediction error.

The OOB predicted value is $\hat{Y}^{OOB} = (1/\|\mathcal{O}_{i'}\|)\sum_{k\in\mathcal{O}_{i'}} \hat{Y}^k$ in which $\mathcal{O}_{i'} = \mathbb{L}\backslash\mathcal{O}_i, i$ and i' denote in-bag and out-of-bag sampled indices, $\|\mathcal{O}_{i'}\|$ is the size of OOB sub dataset, and the OOB prediction error is

$\widehat{ERR}^{OOB} = \frac{1}{N_{OOB}}\sum_{i=1}^{N_{OOB}} \wp\left(Y, \hat{Y}^{OOB}\right)$.....(7)

In this $\wp(\cdot)$defines an error function and N_{OOB} is used to represent the size of OOB samples.

Using an RF to calculate the Feature Importance Score: Breiman introduces the out-of-bag importance score, a permutation method for assessing the value of characteristics in the prediction. The difference between the original average error and the randomised permuted ME may be determined in OOB samples, which is why such a score of characteristics is evaluated. This technique is used to efficiently reorder the values of the jth feature in OOB for each tree, allowing the construction of an RF model to forecast the permuted feature and derive the ME (mean error). This permutation was designed to examine the effect of eliminating the existing connection of the jth feature with Y values on the RF model further on. In order to dramatically reduce mean error, a characteristic must be represented in a strong connection.

During in the development of RF, the other type of correlation - based feature measure can be obtained. The separation is determined by the minimizing of node impurity R(t) at each node t in a decision tree. Node impurity R(t) determines the gini index. gini(t) is a comment thread in node t that contains samples from c classes and is described as:

$R(t) = 1 - \sum_{j=1}^{c} \hat{p}_j^2$ ..... (8)

In which $\hat{p}^2 j$ denotes the relative frequency of class j in t. Gini(t) is diminished, in case, the classes in t are skewed. When t is divided into two child nodes t_1 and t_2 with sample sizes $N_1(t)$ and $N_2(t)$, the gini index of the split data is expressed as:

$$Ginin_{split}(t) = \frac{N_1(t)}{N(t)} Gini(t_1) + \frac{N_2(t)}{N(t)} Gini(t_2) \quad \ldots\ldots\ldots \quad (9)$$

The feature providing smallest $Gini_{split}(t)$ is selected for splitting the node. The importance score of features X_j in a single decision tree T_k is defined as:

$$IS_k(X_j) = \sum_{t \in T_k} \Delta R(t) \ldots\ldots\ldots \quad (10)$$

and K trees are calculated in a random forest using it and this is defined as:

$$IS(X_j) = \frac{I}{K} \sum_{k=1}^{K} IS_k(X_j) \ldots\ldots\ldots \quad (11)$$

In bag samples, an RF (random forest) is used to generate an in-bag importance score, which is a type of important measure. This is the primary distinction between an in-bag assessment in terms and an out-of-bag measure, which is created through reducing prediction error. RF is used in OOB samples to attain this goal. In comparison to the out-of-bag measure, the in-bag significance score requires less computing time.

D. Classification: The training model uses the mapping of specified attributes to categorize the provided characteristics so that transcriptome cancer can be forecasted with ease. Each class represents a specific type of microarray cancer. This procedure is carried out using the linear regression (LR) technique. Linear regression is a widely used statistical technique for modeling the relationship between a set of "explanatory" variables and a real-valued outcome. The independent features are utilized in regression modeling to forecast a target class. As a consequence, this approach can be used to examine the link between independent and dependent variables, as well as for predicting. The most extensively used statistical method in machine learning for forecasting is linear regression, which is a type of regression modeling. Each observations is based on two values in linear regression: one is the regression coefficient, and the other is the independent component. Linear regression is used to find a linear relationship between different variables. Two aspects (x, y) are involved in linear regression analysis. The link between y and x is depicted in the equation below, which is known as regression.

$$y = \beta_0 + \beta_1 + \varepsilon \ldots\ldots\ldots \quad (12)$$

or equivalently

$$E(y) = \beta_0 + \beta_1 x \ldots\ldots\ldots \quad .(13)$$

Here, ε is the error term of linear regression. The error term here uses to account the variability between both x and y, β_0 represents y-intercept, β_1 represents slope.

To bring the principle of linear regression into a machine learning framework, x represents the input training dataset, and y denotes the class labels included in the input dataset. The machine learning algorithm's purpose is to identify the optimal values for (intercept) and (coefficient) in order to produce the best-fit regression model. To achieve the best fit, the gap between the actual and predicted values must be as small as possible, hence this reduction problem can be expressed as:

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \ldots\ldots\ldots \quad (14)$$

$$g = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \ldots\ldots\ldots \quad (15)$$

Here, g is called a cost function, which is the root mean square of the predicted value of $y(pred_i)$ and actual $y(y_i)$, n is the total number of data points. This algorithm uses the extracted characteristics as input. In this study, two classes are identified. Microarray cancer indicates that the person has a high risk of developing microarray cancer.
The normal is utilized for the person without any possibility of microarray cancer.

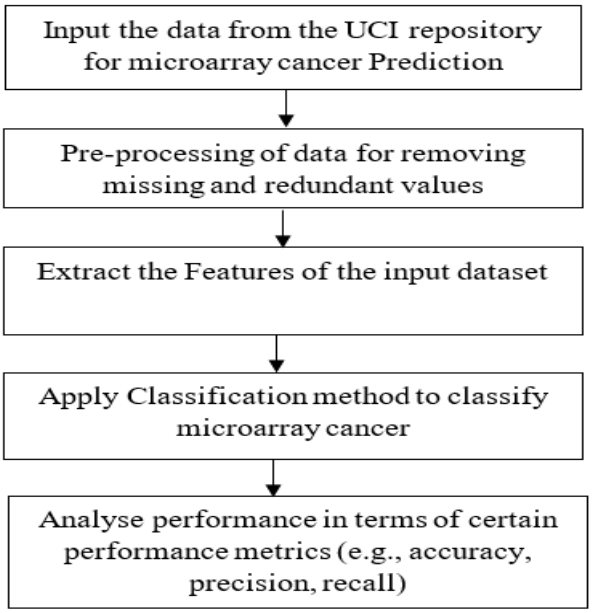

Figure 4: Proposed Methodology

Results and Discussions

In this work, the task of microarray cancer prediction has been accomplished by applying an openly available dataset called Cleveland. There are fourteen attributes included in this dataset. This work applies, and compares several classification models for predicting microarray cancer. Some of these classification models include DT, MLP (Multilayer perceptron), NB, an ensemble classifier combining random forest, and naïve bayes classifiers.

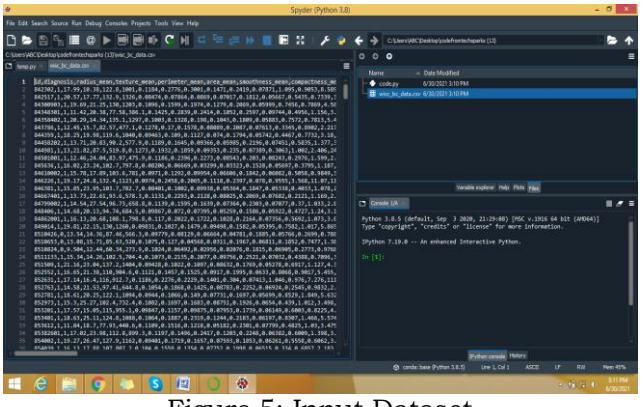

Figure 5: Input Dataset

As shown in figure 5, this research work is based on the microarray cancer prediction. The dataset is collected from UCI repository and it has various attributes which are used for the prediction analysis

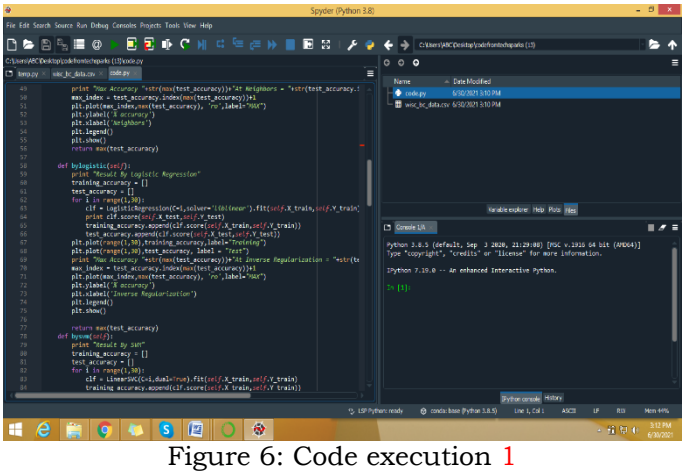

Figure 6: Code execution 1

As shown in figure 6, The datasets intake, feature extraction, and classifier phases of microarray cancer prediction are all important. Various machine learning techniques, such as decision tree, naive bayes, Multilayer, ensemble, and others, are used at this phase.

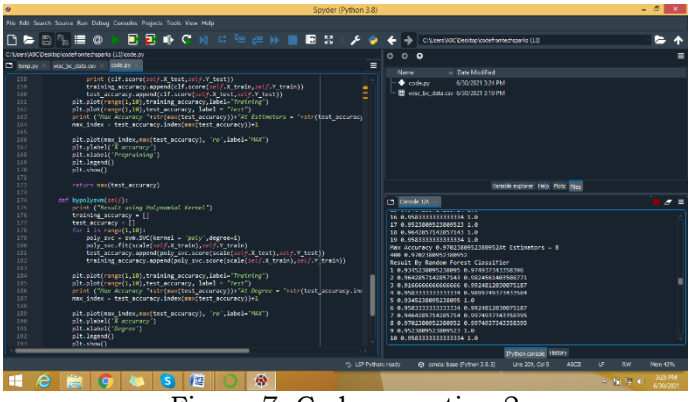

Figure 7: Code execution 2

Figure 7 shows the several phases of microarray cancer prediction, which include dataset input, feature extraction, and classification. Various machine learning techniques, such as decision tree, bayesian networks, Multi - layer, ensemble, and others, are used at this phase. For microarray cancer detection, this photo ensemble classification method is used.

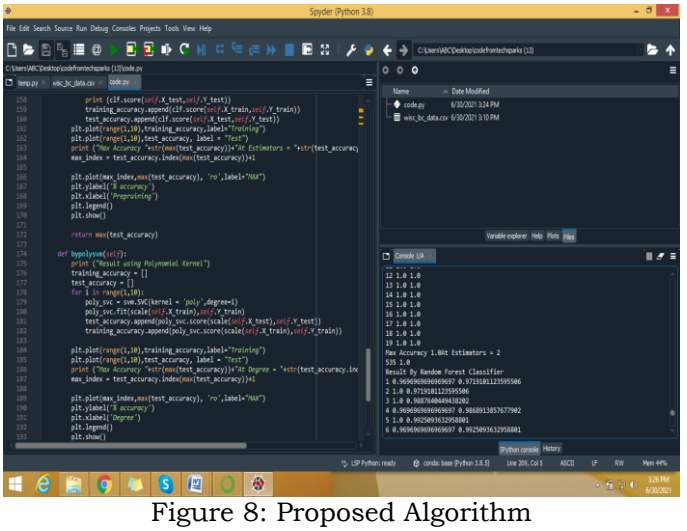

Figure 8: Proposed Algorithm

As shown in figure 8, the proposed algorithm is applied in this phase which is the combination of random forest and logistic regression for the microarray cancer prediction parameters.

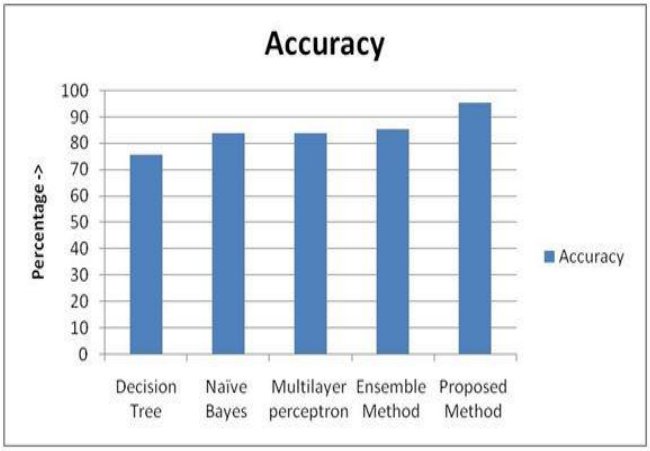

Figure 9: Accuracy Analysis

Figure 9 depicts the accuracy-based comparison of different classification models like DT, NB, multilayer perceptron, ensemble, and proposed models. The results of the analysis depict that the introduced model outperforms other models by obtaining an accuracy rate of 95%, and proves best.

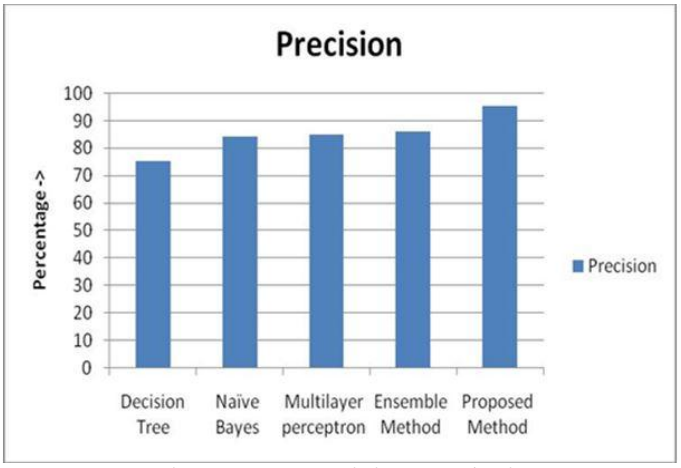

Figure 10: Precision analysis

Figure 10 depicts the precision-based comparison of different classification models like DT, NB, multilayer perceptron, ensemble, and proposed models. The results of the analysis depict that the introduced model outperforms other models by obtaining a precision rate of 95%, and proves best.

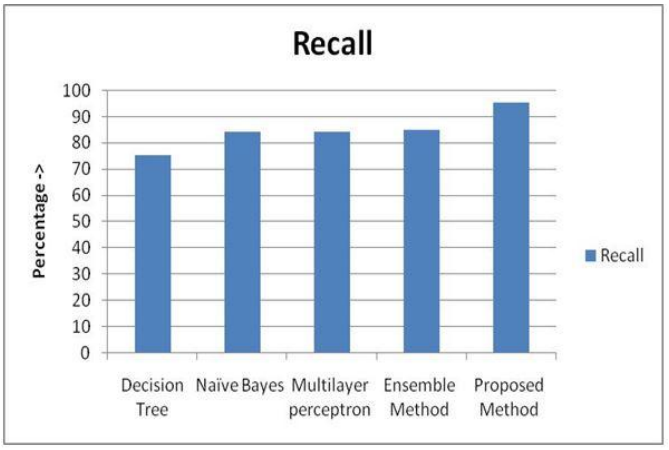

Figure 11: Recall Analysis

Figure 11 depicts the recall-based comparison of different classification models like DT, NB, multilayer perceptron, ensemble, and proposed models. The results of the analysis depict that the introduced model outperforms other models by obtaining a recall rate of 95%, and proves best.

Conclusion

The proposed method may evolve as a game changer in the field of medical diagnosis, especially in breast cancer prediction. The performance of the suggested technique is generally superior towards the other state-of-the-art methods, as shown in the preceding sections of results and discussion. The achieved accuracy reaches to more than 90%, which will be helpful for the medical community to diagnose such disease with more accuracy, by which patients will be benefited by having treatment in right time.

References

[1] Gopal K. Dhondalay, Dong L. Tong, Graham R. Ball, “Estrogen receptor status prediction for breast cancer using artificial neural network”, International Conference on Machine Learning and Cybernetics, Volume: 2, Issue: 24, PP: 256-264, 2011

[2] Ir Cath Tee, Ali H. Gazala, “A novel breast cancer prediction system”, International Symposium on Innovations in Intelligent Systems and Applications, Volume: 67, Issue: 9, PP: 984-992, 2011

[3] Chintan Shah, Anjali G. Jivani, “Comparison of data mining classification algorithms for breast cancer prediction”, Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Volume 14, Issue: 30, PP: 3980-3987, 2013

[4] Hon-Yi Shi, Jinn-Tsong Tsai, Wen-Hsien Ho, Ming-Feng Hou, “Application of artificial neural networks for the prediction of quality of life in breast cancer patients”, SICE Annual Conference, Volume: 53, Issue: 10, PP: 734-742, 2011

[5] Xiaoyi Xu, Ya Zhang, Liang Zou, Minghui Wang, Ao Li, “A gene signature for breast cancer prognosis using support vector machine”, 5th International

Conference on BioMedical Engineering and Informatics, Volume: 4, Issue: 28, PP: 2712-2720, 2012

[6] Glenn D Francis, Sandra R Stein, Glenn D Francis, “Prediction of histologic grade in breast cancer using an artificial neural network”, The 2012 International Joint Conference on Neural Networks (IJCNN), Volume: 12, Issue: 2, PP: 854-861, 2012

[7] Daphne Teck Ching Lai, Jonathan M. Garibaldi, “Improving semi-supervised fuzzy c-means classification of Breast Cancer data using feature selection”, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Volume: 29, Issue: 23, PP: 1092-1100, 2013

[8] Aida Ali, SitiManyamShamsuddin, Anca L. Ralescu, “Hybrid intelligent systems in survival prediction of breast cancer”, 12th International Conference on Hybrid Intelligent Systems (HIS), Volume: 80, Issue: 4, PP: 787-795, 2012

[9] Gul ShairaBanu, AmjathFareeth, NisarHundewale, “Prediction of breast cancer in mammagram image using support vector machine and fuzzy C-means”, Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, Volume: 27, Issue: 56, PP: 639-647, 2012

[10] Peter Adebayo Idowu, KehindeOladipo Williams, Jeremiah AdemolaBalogun and AdeniranIsholaOluwaranti, “Breast Cancer Risk Prediction Using Data Mining Classification Techniques”, Transactions on Networks and Communications, Volume: 3, Issue: 42, PP: 187-194, 2015

[11] R. Preetha, S. Vinila Jinny, “A Research on Breast Cancer Prediction using Data Mining Techniques”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume: 8, Issue:11, PP: 693-701, 2019

[12] Dr. C Nalini, D.Meera, “Breast cancer prediction system using Data mining methods”, International Journal of Pure and Applied Mathematics, Volume: 119, Issue: 12, PP: 10901-10911, 2018

[13] G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, “An Efficient Prediction of Breast Cancer Data using Data Mining Techniques”, International Journal of Innovations in Engineering and Technology (IJIET), Volume: 2, Issue:14, PP: 3677-3685, 2013

[14] Nitasha, “Review on Breast Cancer Prediction Using Data Mining Algorithms”,International Journal of Computer Science Trends and Technology (IJCST), Volume: 7, Issue: 25, PP: 732-740, 2019

[15] S. Yuvarani and Dr. C. JothiVenkateswaran, “Breast Cancer Detection In Data Mining: A Review”, Journal of Computer Science and Applications, Volume: 7, Issue: 1, PP:245-252, 2015

[16] A. Priyanga, Dr. S. Prakasam, “The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness”, International Journal of Computer Science and Engineering Communications, Volume: 10, Issue: 26, PP: 1740-1748, 2013

[17] Deneshkumar V, Manoprabha M, Senthamarai Kannan K, “Comparison of Datamining Techniques for Prediction of Breast Cancer”, International Journal of Scientific & Technology Research Volume: 8, Issue: 08, PP: 268-276, 2019

[18] K. Arutchelvan, Dr. R. Periyasamy, “Cancer Prediction System using Data Mining Techniques”, International Research Journal of Engineering and Technology (IRJET), Volume: 2 Issue: 68, PP: 797-804, 2015

[19] Shelly Gupta, Dharminder Kumar, Anand Sharma, “Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis”, Indian Journal of Computer Science and Engineering (IJCSE), Volume: 2, Issue: 25, PP: 3782-3790, 2011