International Journal of Reliable and Quality E-Healthcare
Volume 12 • Issue 1

Classifying Malignant and Benign Tumors of Breast Cancer:

A Comparative Investigation Using Machine Learning Techniques

Meshwa Rameshbhai Savalia, Institute of Technology, Nirma University, India
Jaiprakash Vinodkumar Verma, Institute of Technology, Nirma University, India*
https://orcid.org/0000-0001-6116-1383

ABSTRACT

Breast cancer is the second major cause of cancer deaths in women. Machine learning classification techniques can be used to increase the precision of diagnosis and bring it closer to 100%, thus saving the lives of many people. This paper proposed four different models, built using different combinations of selected features and applying five ML classification techniques to all the models to identify the best model with the highest accuracy. It analyzes five machine learning techniques, namely logistic regression (LR), support vector machines (SVM), naive bayes (NB), decision trees (DT), and k-nearest neighbor (KNN), for prediction of breast cancer using the Wisconsin Diagnostic Breast Cancer Dataset on these four models. The objective of the paper is to find the best ML algorithm that can most accurately predict breast cancer for a particular model. The outcome of this paper helps the doctors to improvise the diagnosis by knowing the effect of combinations of symptoms with the growth of breast cancer.

KEYWORDS

Breast Cancer Diagnosis, Classification, Decision Trees, K-Nearest Neighbor, Logistic Regression, Naive Bayes, Support Vector Machines, Wisconsin Diagnostic Breast Cancer Dataset

INTRODUCTION

Breast cancer is one of the major causes of death around the world. One in every ten women is affected by breast cancer (Ilbawi & Velazquez-Berumen, 2018). It is essential to diagnose and predict dreadful tumors as early as possible to save a woman's life. We need to improve efficiency and simplify the testing and treatment processes. Hence medical records in the form of images as well as numerical data are required for this purpose which is already stored digitally in repositories. These repositories are publicly available

DOI: 10.4018/IJRQEH.318483 *Corresponding Author

for research to improve the diagnosis process. As per WHO, there were 9.6 million deaths due to cancer in 2018, making it the second-largest cause of death in the world (Ilbawi & Velazquez-Berumen, 2018). Globally, about 1 in 6 deaths is due to cancer. As per the American Cancer Society, 1,762,450 new cancer cases and 606,880 cancer deaths are estimated to occur in the United States in 2019 (Siegel et al., 2019). According to (Bray et al., 2018), the risk of dying from cancer before the age of 75 years is 7.34% in males and 6.28% in females. Breast cancer is one of the most chronic and dreadful diseases and one of the most common types of cancer found in women in the world. It accounts for 14% of all cancers in women. Overall, 1 in 28 women is likely to develop breast cancer during their lifetime. There were about 2.09 million cases of breast cancer in 2018. Chances of survival can be improved by early detection. Chances of survival can be increased by 98% if the cancer is diagnosed early (Ilbawi & Velazquez-Berumen, 2018). The average accuracy of manually diagnosing breast cancer by a human being from Fine Needle aspiration cytology (FNAC) is only 90%. This percentage can be optimized by applying machine learning techniques on digitized images of breast cells. It is important to correctly detect and diagnose the patients as early as possible. AI can be used for better and accurate detection and diagnosis of breast cancer.

Machine learning employs a variety of statistical, probabilistic, and optimization techniques. It allows the machine to "learn" from past examples and detect hard-to-discern patterns from large, noisy, or complex datasets (Cruz & Wishart, 2007). It can be used in medical applications, especially those that depend on complex proteomic and genomic measurements. Recently, researchers have been using machine learning for cancer diagnosis as well as prognosis. There is also a growing trend of personalized predictive medicine by using artificial intelligence. Plenty of research has been done which implants Machine Learning Techniques on the medical diagnosis of breast cancer using the Wisconsin Breast Cancer Diagnosis Dataset (WDBC). (Meraliyev et al., 2017) applied K nearest neighbor (KNN), SVM, ANN, Logistic regression, and decision tree (DT) model to predict breast cancer from the WDBC dataset. It uses K-fold cross-validation techniques to find evaluation measures for the model such as accuracy, sensitivity, specificity, etc. It claims that ANN, DTC, and logistic regression give 98% accuracy whereas KNN gives 99% accuracy and finally SVM can give 100% accuracy. (Kathija & Nisha, 2016) applied SVM and Naïve Bayes techniques for breast cancer data classification. This paper finds the smallest subset of features from the Wisconsin Diagnosis Breast cancer (WDBC) dataset by applying a 5-fold cross-validation method and confusion matrix accuracy so that it can ensure a highly accurate ensemble classification of breast cancer. This paper suggests that the naive Bayes model gives the highest accuracy of 95.65%. (Borges, 2015) presents a detailed description of the WDBC dataset. In addition, he applies the NB algorithm and JV8 algorithm for classification which has 97.80% and 96.05% accuracy respectively. Pre-processing is done using tools available in Weka 3.6. This paper proposed a comparative analysis of five machine learning techniques namely Logistic Regression (LR), Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (DT), and K-Nearest Neighbor (KNN) for the prediction of breast cancer. We have used the Wisconsin Breast Cancer Diagnostic dataset (WDBC) (Dua & Graff, 2019) for the classification of benign and malignant tumors for breast cancer. This paper applies various machine learning classification techniques to the dataset to identify the best methodology for the classification task that gives the most accurate and reliable results.

The rest of the paper's organization is as follows: Section 2 shows a comparative study of the related literature on the different research done. Section 3 presents the proposed system for the proposed research work presented in this paper. Section 4 presents the methodology and concepts applied to achieve defined objectives. Section 5 and 6 presents performance analysis by describing the experiments and analysis of the experimental results. Section 7 concludes the paper and discusses future works.

RELATED WORK

This section describes the study of different Machine Learning (ML) approaches proposed or implemented by researchers in the area of cancer diagnosis. (Padma & Sowmiya, 2018) presents

a survey of various DM techniques used currently for breast cancer prediction. This research work concluded that the performance of C4.5 was better than ID3, K means, ANN, Naive Bayes, etc. (Chaurasia et al., 2018) applied Naive Bayes, RBF, and JV8 data mining algorithms to generate a prediction model. Naive Bayes predicted most accurately with 97.36% accuracy followed by RBF and JV8 with 96.77% and 93.41% accuracy respectively. For each of the three methodologies (NB, RBF, and JV8), this paper uses a 10-fold cross-validation procedure to calculate unbiased prediction accuracy and remove all 16 entries having missing attributes. (Kaushal et al., 2019) proposed four steps in CAD for cancer diagnosis. These steps are pre-processing, segmentation, feature extraction, and classification. Various pre-processing techniques discussed in the paper are H&E or IHC staining, Rotation, Cropping, Flipping, Noise expulsion, Morphological ðltering, ROI detection, Gaussian smoothing, etc. For classification, the pros and cons of neural networks, SVMs, and Decision tree classifiers were discussed.

(Hosni et al., 2019) reviews 193 papers related to cancer and aims at analyzing state of art in ensemble classification methods when applied to breast cancers. This paper says that the majority of researchers are interested in diagnosis tasks because if the disease is diagnosed correctly, further risks and costs can be reduced. (Yassin et al., 2018) also shows a survey of recent trends in CAD systems, different image modalities, and various ML classifiers used. Various image modalities used are digital mammography, ultrasound imaging, MRI, microscopic images, infrared thermography, etc. This paper suggests that out of various classification techniques such as SVM, ANN, KNN, DT, NB, Random Forest, Logistic regression, Deep Learning, etc. majority of papers used SVMs, followed by ANN. (Ahmad et al., 2013) experimented with C4.5, ANN, and SVM to predict breast cancer and achieved the accuracy of 93.6%, 94.7%, and 95.7% respectively. This research is performed on the Iranian Centre for Breast cancer program dataset. The paper (Westerdijk, 2018) uses predictive models such as Logistic regression, random forests, Support vector machines, Artificial neural networks, and ensembles to diagnose breast cancer. Various performance measures used in this study are accuracy, AUC from the ROC curve, sensitivity, and specificity. Four predictive models are then optimized and combined using ensemble techniques to achieve a better predictive model. The final accuracy result for the ensemble model is 98.23%.

This study suggests decreasing the number of false negatives for future research. In the study of (Huang et al., 2017) SVM and SVM ensembles are used to detect breast cancer over small- and large-scale breast cancer datasets. It achieves 98.28% classification accuracy for the WDBC dataset by using genetic algorithms for feature selection and RBF SVM ensembles for classification. It uses classification accuracy, ROC, F-Measure classifier training time as evaluation metrics. An automatic diagnosis system was proposed by Murat Karabatk which reported a classification rate of 95.6%. This system was based on association rules (AR) and neural networks (NN)(Karabatak & Ince, 2009). Ali Keles presented a decision support system based on neuro-fuzzy rules. This paper achieved a high positive predictive rate (96%) and specificity (97%) (Keleş et al., 2011). A least-square SVM classifier designed by Kemal Polat achieved a classification accuracy of 98.53% (Polat & Güneş, 2007). A new hybrid method was proposed by Sahan which was based on a fuzzy-artificial immune system and KNN algorithm. It achieved an accuracy of 99.14% (Übeyli, 2007). In addition, the above discussion Table -1 shows a comparative analysis of research work presented in various papers.

PROPOSED SYSTEM

In the proposed system, we have used the Wisconsin Breast Cancer Diagnostic dataset (WDBC) for the classification of benign and malignant tumors. The training and testing dataset is divided into 80% and 20% respectively. Figure 1 shows the process flow diagram, first of all, dataset selection is done. We have selected the WDBC dataset because it is widely utilized for this purpose as it has a large number of instances, is virtually noise-free, and has no missing values. After data selection, data preprocessing is done. Data preprocessing generally includes data cleaning, dealing with missing

Table 1. Comparative analysis of related work

Reference	Objective	ML Technique	Methodology	Evaluation Measures	Results	Findings
(Zafiropoulos et al., 2006)	SVM approach for prognosis and diagnosis of breast cancer	SVM	350 cases for training, the complete dataset for testing Kernels: guassian rbf, polynomial	Accuracy, Specificity, Sensitivity	Accuracy:96.91% Sensitivity:97.67% Specificity:97.84	SVM with Guassian rbf kernel and (sigma)=1
(Wang & Yoon, 2015)	Applying data mining techniques to predict breast cancer	1. SVM 2. ANN 3. NB 4. Adaboost Tree	PCA is used for feature space reduction	Accuracy, Model run time	SVM: 98.12% ANN: 99.63% NB: 93.32% Adaboost: 97.19%	ANN with PCA
(Nguyen et al., 2013)	Applying random forest classifier with feature selection for breast cancer diagnosis and prognostic	Random Forest	No. of tree in RF: 25 No. of features selected: 18.36	Accuracy, Sensitivity, Specificity, AUC	Accuracy: 99.82%, Sensitivity: 99.83%, Specificity: 99.72%, AUC: 99.78%	RF with feature selection
(Hazra et al., 2016)	Find the smallest subset of features that can ensure highly accurate classification of breast cancer	1.SVM 2.NB 3.Ensemble	Feature selection: 1. PCA 2. Pearson Correlation Coefficient	Accuracy, Execution Time	Accuracy: SVM:98.88% NB:97.39% Ensemble: 97.30	PCA+SVM(19 features) + without binning (98.88)
(Kathija & Nisha, 2016)	Find the smallest subset of features from the WDBC dataset that can ensure highly accurate ensemble classification	1.SVM 2. NB		Accuracy, sensitivity, specificity	Accuracy: NB: 95.65% SVM: 95.10%	NB
(Ali & Feng, 2016)	Breast Cancer Classification	1. SVM 2. NN	80% training, 20%test Kernel functions:mpl,rbf,guadratic,polynimial Training functions: trainbfg,train cgb,train gdx,train lm	Accuracy, Precision	SVM: Accuracy: 89%,Precision:88% NN:Accuracy: 92% Precision:88%	NN with traingdx training function
(Huang et al., 2017)	Assess the prediction performance of SVM and SVM ensembles over small and large-scale breast cancer datasets.	1. GA+SVM 2. GA+SVM ensembles(bagging/boosting)	90%training, 10%testing, 10 fold cross validation	Accuracy,ROC,F-measure,and computational times	Accuracy: 98.28% ROC: 98% F measure:98.8%	GA + linear SVM for classification accuracy(96.85%), GA + linear SVM for ROC (0.967), and GA + RBF SVM for F-measure (0.988)
(Sridevi & Anitha, 2018)	Prediction of Breast Cancer using Decision tree and random forest	1. DT 2. Random Forest	70%training, 30% testing	Accuracy, Specificity, Sensitivity	Accuracy: DT: 91.18% Random Forest: 95.72%	Random Forest
(Wang et al., 2018)	Breast Cancer diagnosis using SVM	1. SVM Ensemble	12 different SVMs are hybridized based on the proposed WAUCE approach	Error, Accuracy, Sensitivity, Specificity	Accuracy: 97.68%	SVM Ensemble
(Westerdijk, 2018)	Predicting malignant tumor cells in breasts	1. LR 2. RF 3. SVM 4. Neural Network 5. Ensemble	Feature selection is done	Accuracy, AUC, Specificity, Sensitivity	Accuracy: LR:97.35% RF: 97.35% SVM: 98.23% Neural Network: 97.35% Ensemble: 98.23	SVM and Ensembles
(Verma et al., 2019)	Breast cancer prediction using SVM	SVM	Missing values are substituted by the mean value	Accuracy	Accuracy: 96.09%	SVM

values, feature scaling, splitting the dataset for training and testing, etc. Here, in data-preprocessing, we normalize the values and scale them using the python library StandardScaler which standardizes a feature by subtracting the mean and then scaling to unit variance. Hence, it makes the mean of data equal to 0 and the standard deviation equal to 1. We also removed a few outliers from the dataset to

improve classification performance. In our dataset, there are no missing values hence we don't have to bother about it. After data preprocessing, data visualization is done to identify patterns and trends in the data and to gain insights from the data. Data visualization is followed by feature selection which is done using the 'feature_selection' module of sklearn library of python. This paper employs five main machine learning classification techniques: Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbour (KNN), Decision Trees (DT), and Naive Bayes (NB) for the classification task. Hence important features are identified for each algorithm which is shown in table 3. Four models are created with different criteria as mentioned in table 2. On each of these models, all five algorithms are implemented to compare the results. Various evaluation measures such as accuracy, precision, recall/sensitivity, and specificity are considered for evaluating the model. We have used 5-fold cross-validation to ensure that our model does not overfit. The five classification algorithms used in this paper are currently in demand and plenty of research work is going on to employ these machine learning techniques into breast cancer prediction for improved diagnosis.

MACHINE LEARNING CLASSIFICATION TECHNIQUES

This section describes the machine learning algorithms applied for classifying malignant and benign tumors.

Support Vector Machine (SVM): It is a supervised machine learning algorithm rooted in statistical learning theory which can classify both linear and non-linear data. It is a non-probabilistic binary classifier that supports both regression and classification tasks and can handle multiple continuous and categorical variables (Rajvanshi & Chowdhary, 2017). It constructs a hyperplane in a multidimensional

Figure 1. Process flow diagram of proposed approach

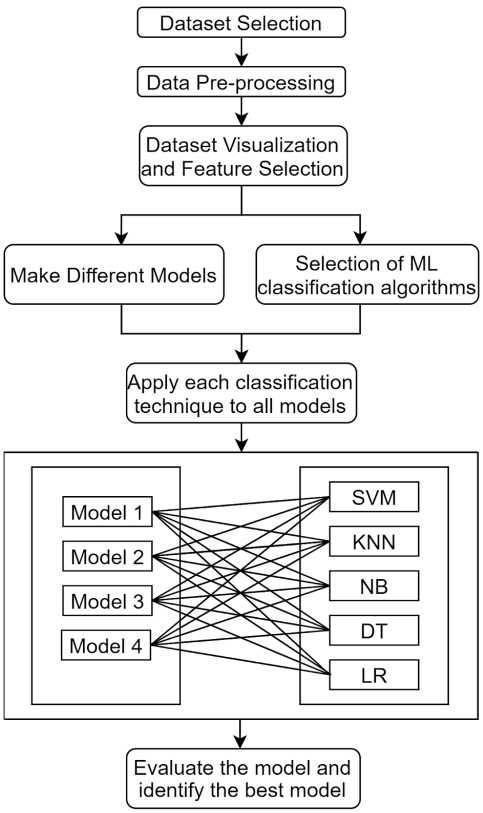


Table 2. Description of four models

Model Name	Description
Model 1	All features
Model 2	All features + Cross-Validation
Model 3	Feature Selection
Model 4	Feature Selection + Cross-Validation

Table 3. Selected features for each classification algorithms

Methodology	Selected Features
Logistic Regression (LR)	'mean radius', 'mean concavity', 'mean symmetry', 'texture error', 'worst texture', 'worst smoothness', 'worst compactness', 'worst concavity', 'worst concave points', 'worst symmetry'
Support Vector Machines (SVM)	All features
Decision Trees (DT)	'mean concave points', 'worst radius', 'worst texture', 'worst perimeter', 'worst concave points'
K-Nearest Neighbour (KNN)	All features
Naive Bayes (NB)	'worst concave points', 'worst area', 'area error', 'worst texture', 'mean texture', 'worst smoothness', 'mean smoothness', 'mean radius', 'mean symmetry'

space that classifies training data into two different classes ("Beginners Guide to Support Vector Machines," 2018). For our problem, a hyperplane is constructed such that all malignant tumors are on one side of the optimal hyperplane and benign tumors are on the other side. The data points touching the maximal margin hyperplanes are called the support vectors (Westerdijk, 2018). The distance between support vectors and the dividing line is called margin. Many hyperplanes are possible to separate data into two classes, but the optimal hyperplane has the maximum margin, i.e. maximum distance between data points of both classes as shown in Figure.2. Hence, we need to maximize the width of margin(w). Here, the two filled squares and one filled circle in Figure 2 are support vectors.

However, most advanced problems are not linearly separable. To classify the data which are not linearly separable, the original training data is transformed into a high dimensional feature space using a mapping function. (please refer to Figure 3) This transformation is done using kernels.

Figure 2. Possible hyperplanes and identification of an optimal hyperplane (Gandhi, 2018)

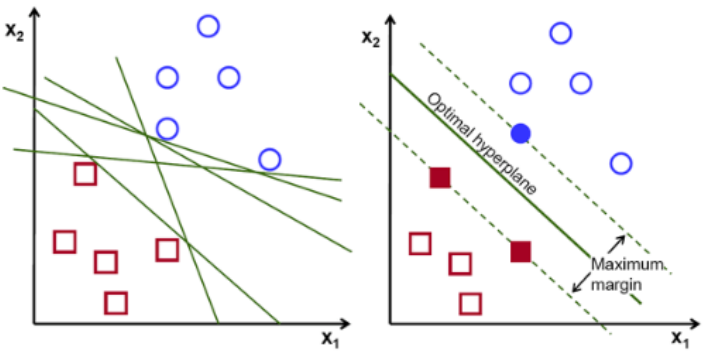
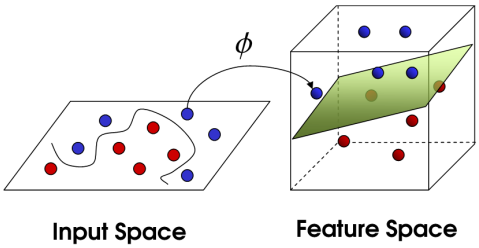


Figure 3. Transformation of data from input space to feature space using kernel function (Wilimitis, 2019)



Here in Figure 3, Φ is a transform function from 2-D to 3-D applied on x. As it can be seen in fig3, it seems impossible to find a single line to separate the two classes (green and blue) in the input space. But, after projecting the data into a higher dimension (i.e., feature space), it is possible to find the hyperplane which classifies the data. This transformation is done with the help of kernels. Kernel helps to find a hyperplane in the higher dimensional space without needing any extra memory and minimal extra computation time. For the given training set $\left\{\left(x_i, y_i\right) \vee x_i \epsilon R^N, y_i \epsilon \left\{-1,1\right\}, i = 1,...,n\right\}$ of a binary classification problem, each hyperplane should satisfy the following equation-1:

$$y_i\left(\left(w \cdot x_i\right) + b\right) \geq 1 - \xi_i \tag{1}$$

Where w is the corresponding weight, b is the intercept term, and $\xi_i \geq 0$ is a slack variable. The slack variable allows some instances to fall off the margin but penalizes them. In order to get the optimal hyperplane that will divide the input space into two classes, the following function (equation-2) should be minimized.

$$\frac{1}{2}\left\|w\right\|2 + C\sum\nolimits_{i=1}^{n} \xi_i \tag{2}$$

Where C>0 is a constant that determines the tradeoff between margin width and misclassifications. We have implemented SVM using the scikit-learn library of python. Various arguments such as regularization parameter, kernel type ('linear', 'poly', 'rbf', 'sigmoid', 'precomputed'), verbose, gamma value, etc should be set properly in order to get desired results.

Logistic Regression: It is one of the oldest algorithms used to solve classification problems. It is a predictive analysis algorithm and based on the concept of probability ("Regression Analysis in Machine learning," n.d.). It is popular because it is fast, it doesn't require scaling of input features, it doesn't require any tuning and it is easy to regularize. It predicts the probability of a categorical dependent variable. Unlike linear regression, which outputs continuous values, logistic regression outputs discrete values. In binary logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, positive, benign, etc.) or 0 (no, failure, negative, malignant, etc.). This outcome of either 0 or 1 is achieved using a sigmoid function. Hence the sigmoid function squeezes the output of the linear equation thus mapping real value into another value between 0 and 1. Hence, it is used to map predictions to probabilities. The formula of the sigmoid function is $f\left(x\right) = \frac{1}{1 + e^{-(x)}}$, where x is the linear equation used to get predictions. So, the prediction function outputs a probability value between 0 and 1. In order to find the class of a data point, a threshold

value is selected (say, 0.5). If the prediction is above the threshold, it will be classified into class 1 and if it is below the threshold, it will be classified into class 0. We use the Cross-Entropy cost function for Logistic regression which is as mentioned below equation -3:

$$J\left(\theta\right)=\frac{-1}{m}\Sigma\left[y^{(i)}log\left(h_{\theta}\left(x\left(i\right)\right)\right)+\left(1-y^{(i)}\right)log\left(1-h_{\theta}\left(x\left(i\right)\right)\right)\right] \tag{3}$$

Where $h_{\theta}\left(x\right)$ is the estimated probability that $Y=1$ (positive class) on input x and $0\leq h_{\theta}\leq 1$. The hypothesis of logistic regression is as mentioned below equation -4:

$$h_{\theta}\left(X\right)=\frac{1}{1+e^{-\left(\beta_0+\beta_1 X\right)}} \tag{4}$$

Naive Bayes: It is a probabilistic machine learning classification model based on Bayesian Theorem. The probability of an event occurring given the probability of another event that has already taken place can be found out using Bayes' Theorem. In this algorithm, it is assumed that each feature makes an independent and equal contribution to the outcome. Hence, it assumes that all features are not correlated to each other. It is fast, easy, and performs well with multiclass data. But in real life, it is nearly impossible to find a set of predictors which are independent of each other. There are mainly three types of Naive Bayes classifiers: Gaussian, Multinomial, and Bernoulli. In Gaussian Naive Bayes, it is assumed that continuous values associated with each feature that is used for prediction follow a normal distribution (Gaussian Distribution) ("Naive Bayes Classifiers," 2019). Hence, when they are plotted, a bell-shaped curve is formed which is symmetric about the mean of the feature as shown in equation 5.

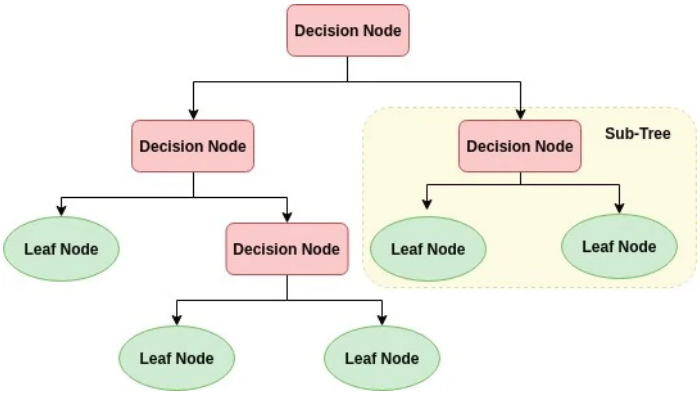
$$P(X\mid Y=c)=\frac{1}{\sqrt{2\pi\sigma_c^2}}e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}} \tag{5}$$

where mu and sigma are the mean and variance of the continuous X computed for a given class 'c' (of Y).

Decision Trees (DTs): It is a non-parametric supervised learning method that has influenced a wide area of machine learning by covering both classification and regression tasks. The "knowledge" learned by a decision tree through training is directly formulated into a tree-like model ("A Guide to Decision Trees for Machine Learning and Data Science," 2018). This structure holds and displays the knowledge in such a way that it is simple to understand and interpret. It is a recursive partitioning approach and similar to a flowchart where each internal node represents a *test* on a feature. The training data have some feature variables and a classification or regression output. A splitting attribute is selected from the dataset that "best" separates the data into individual classes. The splitting attributes can be continuous-valued or they can be restricted to binary trees. For continuous-valued attributes, a split point must be determined as part of the splitting criterion whereas for the binary trees a splitting subset must be determined. This splitting defines a node on the tree i.e each node is a splitting point based on a certain feature from our data which can be seen in Figure 4.

The most popular attribute selection measures are – Entropy (Information Gain), Gain Ratio, and Gini Index. There are various decision tree algorithms namely ID3 (Iterative Dichotomiser 3), C4.5, C5.0, CART (Classification and Regression Tree), CHAID (CHi- squared Automatic Interaction Detector), MARS, etc. We have implemented decision trees using scikit-learn which uses an optimized

Figure 4. Decision tree implementation (Navlani, 2018)



version of the CART algorithm. Hence, we will discuss only the CART algorithm. CART stands for Classification and Regression Trees. It supports both continuous and nominal attribute data and has an average speed of processing. It is constructed by the binary splitting of the attribute. The selection of splitting attributes is done by calculating the Gini Index. The Gini Index measures the impurity D as shown in equation 6:

$$Gini\left(D\right) = 1 - \sum_{i=1}^{m} p_i^2 \tag{6}$$

Where pi is the probability that a tuple in D belongs to class Ci and is estimated by |Ci,D|/|D| (Meghana & Deepika, 2017). The sum is computed over m classes. The attribute that has the minimum Gini index is selected as the splitting attribute.

K-Nearest Neighbors: It is one of the most basic supervised machine learning algorithms used for classification and regression. This non-parametric, simple, and efficient technique performs great in applications like pattern recognition, data mining, predictive analysis, statistical estimation, and intrusion detection. In this method, output interpretation is simple and calculation time is also less. Hence, it is one of the most popular classification techniques. In this technique, no explicit training step is required. Also, there is no segregation of training and testing datasets. Hence, whole data is used for predicting the class of the new data point. From the dataset, each data point is converted into a vector of multidimensional feature space, each with a class label. The first step simply includes storing these feature vectors along with their class labels. For a new data point, the distance between the arrived data point and each of the stored vectors is calculated, and hence 'k' instances of neighbors (feature vectors) of the new data point are identified which are nearest to the arrived data point. This is done by sorting all the data points in terms of the distance from the new data point. Here, 'k' is a user-defined positive integer. After identifying the group of neighbors, the class label possessed by the highest number of neighbors is assigned to the new data-point. It is important to note that the value of 'k' is usually kept odd to avoid ties if the number of classes is two. The distance between two points can be measured in various ways. Some distance metrics include Euclidean distance, Manhattan distance, Chebyshev distance, Cosine distance, Minkowski distance, etc.

We have used the Euclidean distance metric in our implementation as it is most popular and suggested by experts. Let the points P and Q be represented by feature vectors P = (x1, x2, …, xm) and Q = (y1, y2, …, ym), where m is the dimensionality of the feature space. The euclidean distance between P and Q is calculated by equation -7:

$$dist\left(P,Q\right)=\sqrt{\frac{\sum_{i=1}^{m}\left(x_i-y_i\right)^2}{m}} \tag{7}$$

It performs much better if all the data is of a similar scale. Hence, data normalization is recommended while using KNN. Also, feature selection will be beneficial as it will reduce the dimensionality of the feature vector.

EXECUTION AND IMPLEMENTATION

This section describes the performance analysis with the implementation details and results of the proposed approach.

Dataset Description

This paper uses Wisconsin Breast Cancer (Diagnostic) Dataset, created by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital in Madison, Wisconsin, USA (Yael, 2017). This dataset was created by a fine needle aspirate of patients with a solid breast mass and a computer program called Xcyt. Fine-needle aspiration (FNA) is a diagnostic procedure used to investigate lumps or masses. In this technique, a thin (23–25 gauge), hollow needle is inserted into the mass for a sampling of cells that, after being stained, will be examined under a microscope ("Fine-Needle Aspiration," 2020). Xcyt is an easy-to-use graphical computer program, which is capable of performing the analysis of cytological features based on a digital scan (Yael, 2017). This program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, then it calculates the mean value, extreme value, and standard error of each feature for the image, returning a 30 real-valued vector (Yael, 2017). Hence, Features that describe characteristics of the cell nuclei present in the image are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass (Dua & Graff, 2019).

Ten real-valued features are computed for each cell nucleus: 1) Radius: The average distance from the center of the nucleus to each of the points on the perimeter. 2) Texture: The standard deviation of gray-scale values. A gray-scale value represents the intensity of the shades of gray in each pixel of the image (Westerdijk, 2018). 3) Perimeter: The total distance of the boundary of the cell nucleus. 4) Area: The area of the cell nucleus. The number of pixels on the interior of the boundary and adding one-half of the pixels on the perimeter, to correct the error caused by digitization. 5) Smoothness: The difference between the length of a radius length and the mean length of the two radius lines surrounding it, hence the local variation in radius lengths (Westerdijk, 2018). 6) Compactness: The perimeter and area are combined using the formula $\frac{perimeter^2}{area}$ - 1 to obtain a measure of the compactness of the cell nuclei (Westerdijk, 2018). 7) Concavity: The severity of concave portions of the contour. A cell having many indentations in the boundary can be called a rough cell and thus has a high concavity value. Similarly, a smooth cell has a low concavity value. 8) Concave points: The number of concave portions of the contour of the cell nucleus. 9) Symmetry: The longest line from boundary point to boundary point through the center of the nucleus is found. Subsequently, the relative length difference between the lines perpendicular to the longest line to the boundary in both directions is measured. Attention should be given to nuclei where the longest line cuts through the boundary because of concavity (Westerdijk, 2018). 10) Fractal dimension: It is calculated by ("coastline approximation" - 1).

The dataset has 32 attributes and 569 instances. There are no missing values in the dataset. The first attribute is ID and the second attribute is Diagnosis. Diagnosis is a categorical variable. It has two values: M= Malignant (indicates the presence of cancer cells); B= Benign (indicates absence).

For each of the ten features mentioned above, the mean, standard error, and "worst" or largest (mean of the three largest values) values were computed for each image, resulting in 30 features (Dua & Graff, 2019). These 30 attributes, combined with ID and Diagnosis together make 32 attributes. There are 357 instances of benign tumors and 212 instances of malignant tumors. Hence, 62.7% of all observations indicate the absence of cancer cells and 37.3% of all observations show the presence of cancerous cells. The value of Radius and area conveys the size of the nucleus whereas perimeter conveys both shape and size of the nucleus. Moreover, features like smoothness, concavity, concave points, fractal dimension, compactness, and symmetry are responsible for expressing the shape of the nucleus. A higher value of shape feature corresponds to a higher probability of malignancy. It is important to note that all features are recorded with four significant digits.

Characteristics of Different Features of Dataset

We have plotted violin plots of each of the 30 features of our dataset to visualize the data and draw insights from it. Violin plots can be used to observe and make a comparison of distributions between multiple groups of numeric data. It is similar to a box plot, with the addition of a rotated kernel density plot on each side. Violin plots were implemented using python libraries such as seaborn, NumPy, pandas, and matplotlib. The thicker part means the values in that section of the violin have a higher frequency, and the thinner part implies a lower frequency. Unlike bar graphs with means and error bars, violin plots contain all data points. This makes them an excellent tool to visualize samples of small sizes. Violin plots are perfectly appropriate even if your data do not conform to normal distribution.

Figure 5 describes the violin plot. They are essential to determine interquartile range and outliers in the data. The following diagram depicts violin plots of all 30 features.

As seen in Figure 6, mean radius, mean perimeter, mean area, mean compactness, mean concavity, mean concave points are well separated between Malignant and Benign tumors, as the 75 percentile of Benign tumors is below the 25 percentile of Malignant tumors. Hence, these 6 features would be good candidates for the classifier. The mean fractal dimension has the same median for both tumor types, so it wouldn't be a good candidate for the classifier. We notice some similarities between 'worst radius' and 'worst perimeter' on one hand, and 'worst concavity' and 'worst concave points' on the other hand. If two violins look similar, it might indicate a correlation between the features, and if two

Figure 5. Common components of Box Plot and Violin Plot (Hintze & Nelson, 1998)

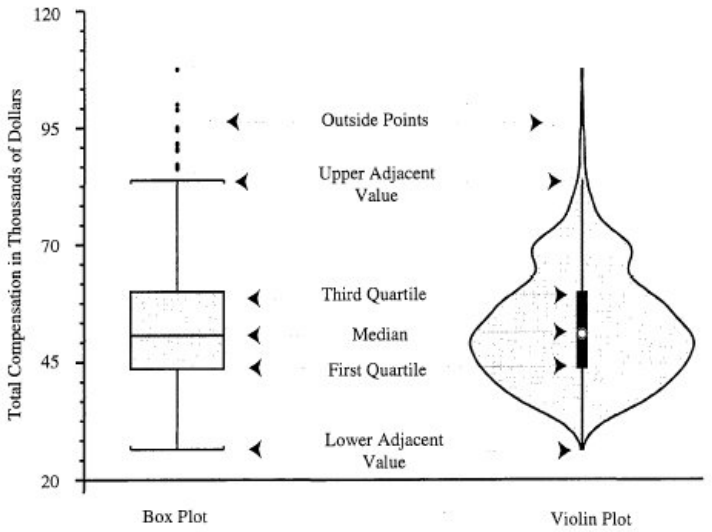
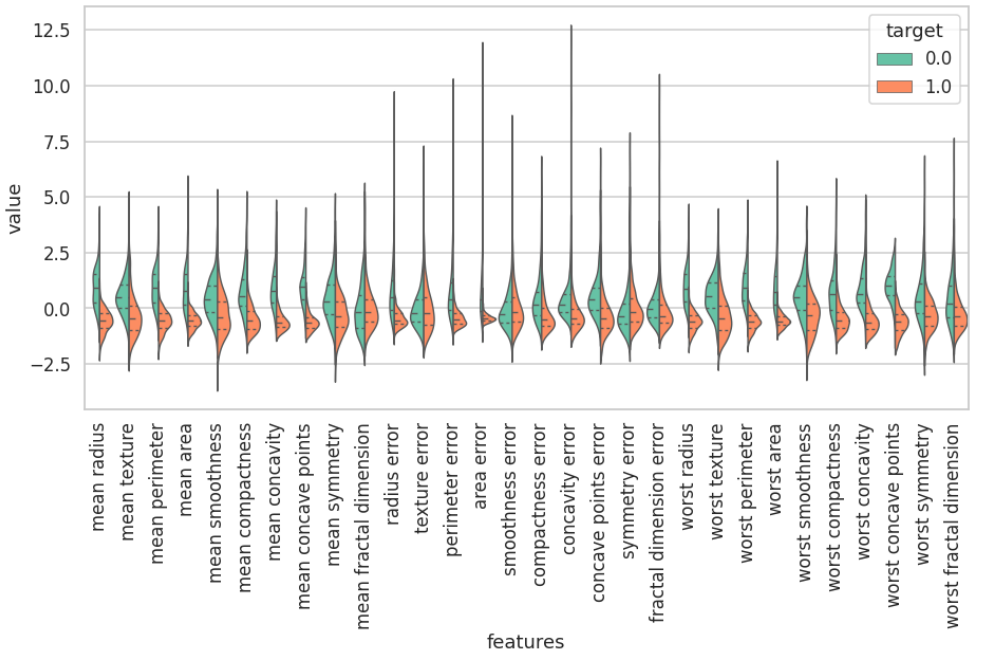


Figure 6. Violin plots of 30 features



features are correlated, one can ask if it's possible (or not) to drop one. These violin plots are useful to identify correlation and other trends between all the features of the dataset.

Dataset Pre-processing

Pre-processing of a dataset is essential before applying any algorithm. It reduces the computation time and increases the performance of the classifier. Data preprocessing generally includes data cleaning, dealing with missing values, feature scaling, splitting the dataset for training and testing, etc. First of all, we scale the values using the StandardScaler library of python which helps to normalize values within a particular range, and sometimes also helps in speeding up calculations. In our dataset, there are no missing values or duplicates hence we don't have to bother about it. We also removed a few outliers from the dataset to improve classification performance. Outliers are data instances with characteristics that are considerably different from the rest of the dataset. In Figure 7, we have plotted box plots of all 30 features to identify outliers. The box plots suggest that only 4 columns (12,13,16,19) contain few outliers. Hence, we remove these points before applying our classification algorithm to the dataset. Figure 8 shows the box plots after removing outliers.

Using the feature selection module of sklearn, we have identified the best features for every algorithm such that it gives maximum accuracy. We have implemented all algorithms using the sklearn package of python. For evaluating our model, various evaluation measures such as accuracy, precision, recall, specificity, etc are used. To ensure that our model does not overfit, we have done 5-fold cross-validation on the results. The training data and testing data are divided in a ratio of 8:2. That means 80% of the data is used for training the model.

Evaluation Measures

We have used four evaluation measures which are accuracy, precision, recall/sensitivity, and specificity to compare our models. As we are dealing with medical data, recall (proportion of people actually

Figure 7. Boxplot of data set before removing outliers

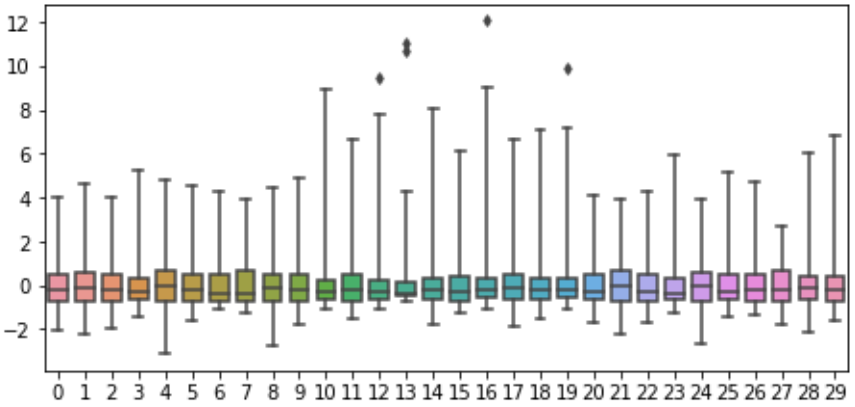
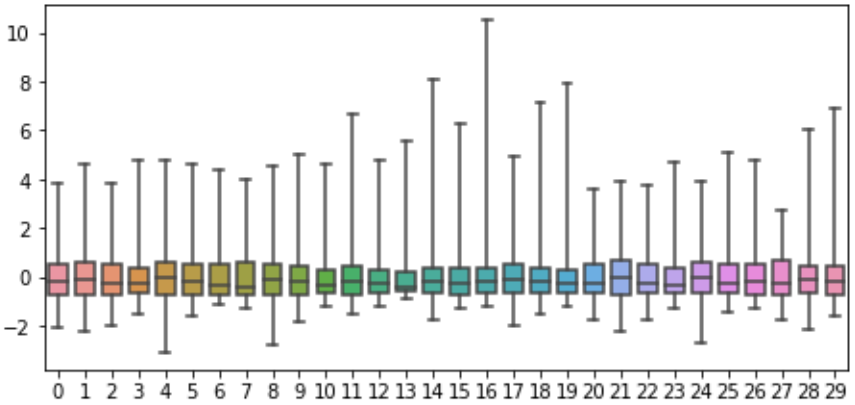


Figure 8. Box plot of the dataset after removing outliers



having cancer and identified correctly) is the most important evaluation measure. It is important that we don't miss any patient that has a malignant tumor. For instance, any data point which was originally malignant and was classified as benign is more harmful than a data point that was originally benign and was classified as malignant. Hence, the higher value of recall is more important than higher accuracy or precision.

RESULTS AND DISCUSSIONS

It is observed that for the model with all features and without cross-validation, KNN with 3 nearest neighbors outperforms all other classifiers by achieving 100% in all evaluation measures (As shown in table-4). Also, the SVM of model 1 achieves 99.12% accuracy and a perfect 100% recall score. After applying 5-fold cross-validation, we still received a pretty good accuracy of 98.83% in SVM and 97.17% in KNN. But the value of recall was higher in SVM compared to KNN. Logistic regression also showed good performance in both model 1 and model 2 by achieving an accuracy of 98.24% and 97.88% and recall of 100% and 96.19% respectively. Model 3 and Model 4 incorporate feature selection. It was observed that Naive Bayes showed an increased accuracy of 97.37% after feature

selection. For SVM and KNN, we have kept all features in feature selection models also as all features were important to achieve high accuracy in classification. Looking at recall values, it is observed that overall SVM and LR give higher recall values whereas SVM and KNN give higher precision values on an average. Although KNN gives 100% recall Model 1, the recall value drops to 94.28% in model 2 after cross-validation which suggests that KNN is not the best algorithm for this purpose. Hence, we suggest using SVMs with all features for the prediction of breast cancer as they give higher recall values and also an accuracy of 99.12% which is the highest among others after KNN.

Below table 5 below compares our model with existing models. It can be seen that our model outperforms all other models in terms of accuracy.

CONCLUSION

In this paper, we applied and analyzed various machine learning techniques to diagnose breast cancer. These techniques include SVM, KNN, NB, DT, and LR. We made four different models using feature selection and a 5-fold cross-validation technique and applied all five algorithms to each model. Models were evaluated based on accuracy, recall, precision, and specificity. Although KNN achieved the highest accuracy, it failed to achieve a higher recall value which is more important when we are working with medical data. Hence, we recommend SVM models trained using all features for diagnosing breast cancer which achieves an accuracy of 99.12% and recall of 100% before cross-validation and accuracy of 98.83% accuracy and 97.5% recall after cross-validation. Our model outperforms all

Table 4. Results of execution of various ML classification algorithms with different models

		Accuracy	Recall/Sensitivity	Precision	Specificity
Model 1	SVM	**99.12**	**100**	**97.56**	**98.65**
	NB	95.61	95	92.68	95.95
	LR	**98.24**	**100**	95.23	97.30
	DT	94.75	95	90.48	94.59
	KNN	**100**	**100**	**100**	**100**
Model 2	SVM	**98.83**	**97.50**	97.21	98.10
	NB	92.93	90	90.97	94.66
	LR	**97.88**	**96.19**	**98.08**	**98.88**
	DT	91.70	90.48	87.78	92.41
	KNN	**97.17**	**94.28**	**98.04**	**98.88**
Model 3	SVM	**99.12**	**100**	**97.56**	98.65
	NB	**97.37**	95	97.43	98.65
	LR	**98.24**	**95.24**	97.50	98.65
	DT	95.61	95	92.68	93.24
	KNN	**100**	**100**	**100**	**100**
Model 4	SVM	**98.83**	**97.50**	97.21	98.10
	NB	95.76	90.95	**97.51**	**98.65**
	LR	**97.00**	**95.71**	96.21	98.65
	DT	94.53	91.90	93.52	95.94
	KNN	**97.17**	94.28	**98.04**	**98.88**

Table 5. Comparison of our results with previous work

Method/Model	Accuracy	Reference
Supervised Fuzzy Clustering	95.57%	(Abonyi & Szeifert, 2003)
Fuzzy KNN	97.17%	(Ghazavi & Liao, 2008)
SVM with gaussian rbf kernel	96.91%	(Zafiropoulos et al., 2006)
Hybrid Neural Networks	94.37%	(Choi et al., 2009)
Rough set-based multiple criteria linear programming approach	89%	(Zhang et al., 2009)
Neural Network	92%	(Rani, 2010)
SVM - RBF kernel	98.06%	(Aruna et al., 2011)
CART algorithm	92.97%	(Lavanya & Usha, 2011)
SMO	97.71%	(Salama et al., 2010)
Neural Network with traingdx training function	92%	(Ahmad et al., 2013)
K-SVM	97.38%	(Zheng et al., 2014)
Weighted Vote Based Ensemble	95.09%	(Bashir et al., 2014)
Weighted area under the ROC curve ensemble(WAUCE)	97.68%	(Wang et al., 2018)
SVM	98.23%	(Westerdijk, 2018)
SVM	96.09%	(Verma et al., 2019)
Proposed Method - LR	98.24%	
Proposed Method - SVM	99.12%	

other models in literature. Despite the high performance of models, we suggest that they should not replace the doctors but only support their final decision because expertise and experience are pivotal elements in any decision-making process.

REFERENCES

A Guide to Decision Trees for Machine Learning and Data Science. (2018). https://www.kdnuggets.com/2018/12/guide-decision-trees-machine-learning-data-science.html

Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters*, *24*(14), 2195–2207. doi:10.1016/S0167-8655(03)00047-3

Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics*, *4*(2), 124–130.

Ali, E., & Feng, W. Z. (2016). Breast Cancer Classification using Support Vector Machine and Neural Network. *International Journal of Scientific Research*, *5*(3), 1–6.

Aruna, S., Rajagopalan, D., & Nandakishore, L. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science and Information Technology*, 2.

Bashir, S., Qamar, U., & Khan, F. H. (2014). Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. *Quality & Quantity*, *49*(5), 2061–2076. doi:10.1007/s11135-014-0090-z

Beginners Guide to Support Vector Machines. (2018, December 14). Retrieved June 13, 2020, from Ivy Professional School | Official Blog website: https://ivyproschool.com/blog/2018/12/14/beginners-guide-to-support-vector-machines/

Borges, L. R. (2015). Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection. *Proceedings of XI Workshop de Visão Computacional*.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a Cancer Journal for Clinicians*, *68*(6), 394–424. doi:10.3322/caac.21492 PMID:30207593

Broniee, J. (2016). *K-Nearest Neighbors for Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, *12*(2), 119–126. doi:10.1177/1748301818756225

Chen, H.-L., Yang, B., Liu, J., & Liu, D.-Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38*(7), 9014–9022. doi:10.1016/j.eswa.2011.01.120

Choi, J. P., Han, T. H., & Park, R. W. (2009). A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis. *Journal of Korean Society of Medical Informatics*, *15*(1), 49. doi:10.4258/jksmi.2009.15.1.49

Cruz, J. A., & Wishart, D. S. (2007). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*, *2*, 59–77. PMID:19458758

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, School of Information and Computer Science.

Fine-Needle Aspiration. (2020, March 28). Retrieved May 29, 2020, from Wikipedia website: https://en.wikipedia.org/wiki/Fine-needle_aspiration

Gandhi, R. (2018, June 7). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Retrieved from Towards Data Science website: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Ghazavi, S. N., & Liao, T. W. (2008). Medical data mining by fuzzy modeling with selected features. *Artificial Intelligence in Medicine*, *43*(3), 195–206. doi:10.1016/j.artmed.2008.04.004 PMID:18534831

Hazra, A., Kumar, S., & Gupta, A. (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. *International Journal of Computers and Applications*, *145*(2), 39–45. doi:10.5120/ijca2016910595

Hintze, J. L., & Nelson, R. D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, *52*(2), 181.

Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J., & Fernández Alemánb, J. (2019). Reviewing ensemble classification methods in breast cancer. *Computer Methods and Programs in Biomedicine*, *177*, 89–112. doi:10.1016/j.cmpb.2019.05.019 PMID:31319964

Huang, M.-W., Chen, C.-W., Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2017). SVM and SVM Ensembles in Breast Cancer Prediction. *PLoS One*, *12*(1), e0161501. doi:10.1371/journal.pone.0161501 PMID:28060807

Ilbawi, A. M., & Velazquez-Berumen, A. (2018). World health organization list of priority medical devices for cancer management to promote universal coverage. *Clinics in Laboratory Medicine*, *38*(1), 151–160. doi:10.1016/j.cll.2017.10.012 PMID:29412879

Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, *36*(2), 3465–3469. doi:10.1016/j.eswa.2008.02.064

Kathija, A., & Nisha, S. S. (2016). Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, *4*(12), 67–75.

Kaushal, C., Bhat, S., Koundal, D., & Singla, A. (2019). Recent Trends in Computer Assisted Diagnosis (CAD) System for Breast Cancer Diagnosis Using Histopathological Images. *Innovation and Research in BioMedical Engineering*, *40*(4), 211–227. doi:10.1016/j.irbm.2019.06.001

Keleş, A., Keleş, A., & Yavuz, U. (2011). Expert system based on neuro-fuzzy rules for diagnosis breast cancer. *Expert Systems with Applications*, *38*(5), 5719–5726. doi:10.1016/j.eswa.2010.10.061

Lavanya, D., & Usha, R. (2011). Analysis of Feature Selection with Classification: Breast Cancer Datasets. *Indian Journal of Computer Science and Engineering*, *2*(5), 756–763.

Logistic Regression — ML Glossary documentation. (2017). https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html

Meghana, L. & Deepika, N. (2017). A Survey on Different Classification Techniques In Data Mining. *International Journal of Science and Engineering Applications, 6*(1), 1–7.

Memon, M. H., Li, J. P., Haq, A. U., Memon, M. H., & Zhou, W. (2019). Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection. *Wireless Communications and Mobile Computing*, *2019*, 1–19. doi:10.1155/2019/5176705

Meraliyev, M., Zhaparov, M., & Artykbayev, K. (2017). Choosing best machine learning algorithm for breast cancer prediction. *International Journal of Advances in Science, Engineering and Technology*, *5*(3), 50–55.

Naive Bayes Classifiers - GeeksforGeeks. (2019, January 14). https://www.geeksforgeeks.org/naive-bayes-classifiers/

Navlani, A. (2018, December 28). *Review of Decision Tree Classification in Python*. Retrieved June 26, 2020, from datacamp.com website: https://www.datacamp.com/community/tutorials/decision-tree-classification-python

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, *06*(05), 551–560. doi:10.4236/jbise.2013.65070

Padma, P. S., & Sowmiya, P. (2018). Breast cancer prediction using data mining techniques. *International Journal of Advanced Research in Science and Engineering*, *7*(1), 41–44.

Pedregosa. (2011). Retrieved June 21, 2020:https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machines. *Digital Signal Processing*, *17*(4), 694–701. doi:10.1016/j.dsp.2006.10.008

Rajvanshi, N., & Chowdhary, K. R. (2017). Comparison of SVM and Naïve Bayes Text Classification Algorithms using WEKA. *International Journal of Engine Research*, *V6*(09).

Rani, K. U. (2010). Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique. *International Journal of Computers and Applications*, *10*(3), 1–5. doi:10.5120/1465-1980

Regression Analysis in Machine learning - Javatpoint. (n.d.). https://www.javatpoint.com/regression-analysis-in-machine-learning

Salama, G. I., Abdelhalim, M. B., & Zeid, M. A-E. (2012). *Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers.* Retrieved May 29, 2020: https://www.semanticscholar.org/paper/Breast-Cancer-Diagnosis-on-Three-Different-Datasets-Salama-Abdelhalim/ab6c4f08484db95f8950d26376dbd22c03b19b21

Seaborn.violinplot — seaborn 0.10.1 documentation. (n.d.). Retrieved May 29, 2020, from https://seaborn.pydata.org/generated/seaborn.violinplot.html

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics. *CA: a Cancer Journal for Clinicians*, *69*(1), 7–34. doi:10.3322/caac.21551 PMID:30620402

Sridevi, N., & Anitha, S. (2018). Prediction of Breast Cancer using Decision tree and Random Forest Algorithm. *International Journal on Computer Science and Engineering*, *6*(2), 226–229.

Support Vector Machine. (n.d.). Retrieved from https://www.saedsayad.com/support_vector_machine.htm

Tutorial4. (n.d.). Retrieved from http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial4/tutorial4.html

Tutorial4. (n.d.). Retrieved June 13, 2020, from http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial4/tutorial4.html

Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. *Expert Systems with Applications*, *33*(4), 1054–1062. doi:10.1016/j.eswa.2006.08.005

Verma, A., Kumar, A., & Kumar, S. (2019). Breast Cancer Prediction Using Support Vector Machine. *International Research Journal of Engineering and Technology*, *6*(4), 2640–2643.

Violin plot. (2020, May 5). Retrieved May 29, 2020, from Wikipedia website: https://en.wikipedia.org/wiki/Violin_plot

Wang, H., & Yoon, S. W. (2015). Breast Cancer Prediction Using Data Mining Method. *Proceedings of the 2015 Industrial and Systems Engineering Research Conference*.

Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*, *267*(2), 687–699. doi:10.1016/j.ejor.2017.12.001

Westerdijk, L. (2018). *Predicting malignant tumor cells in breasts*. Retrieved from https://www.math.vu.nl/~sbhulai/papers/paper-westerdijk.pdf

Wilimitis, D. (2019, February 21). *The Kernel Trick*. Retrieved June 9, 2020, from Medium website: https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

Yael, K. (2017). *Wisconsin Breast Cancer (Diagnostic)*. DataSet Analysis.

Yassin, I. R., Omran, S., El Houby, E. M. F., & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, *156*, 25–45. doi:10.1016/j.cmpb.2017.12.012 PMID:29428074

Zafiropoulos, E., Maglogiannis, I., & Anagnostopoulos, I. (2006). A Support Vector Machine Approach to Breast Cancer Diagnosis and Prognosis. In I. Maglogiannis, K. Karpouzis, & M. Bramer (Eds.), *Artificial Intelligence Applications and Innovations. AIAI 2006. IFIP International Federation for Information Processing, 204*. Springer. doi:10.1007/0-387-34224-9_58

Zhang, Z., Shi, Y., & Gao, G. (2009). A rough set-based multiple criteria linear programming approach for the medical diagnosis and prognosis. *Expert Systems with Applications*, *36*(5), 8932–8937. doi:10.1016/j.eswa.2008.11.007

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, *41*(4, Part 1), 1476–1482. doi:10.1016/j.eswa.2013.08.044

Meshwa Rameshbhai Savalia has completed her B. Tech. in Information Technology from Nirma University, Ahmedabad in 2020. Her current research interests include machine learning, deep learning, data science and artificial intelligence.

Jai Prakash Verma is working as an Associate Professor in Computer Science and Engineering Department. He has been associated with the department since July 2006. Dr Verma received his BSc (PCM) and MCA degree from University of Rajasthan, Jaipur and PhD degree from Charusat University, Changa in the area of Text Data Summarization and Analytics. His research interests include Data Mining, Big Data Analytics, Graph Data Analytics and Machine Learning. He has been contributing to the research in the area of said domain with several publications in international conferences and journals. He is actively involved in conducting various training programmes including customized training on Big Data Analytics to Naval officers at INS Valsura, Indian Navy and SAC-ISRO, Ahmedabad scientists.