Machine Learning Applications in Cancer Diagnosis

Seminar Member (s)

Sl. No.	Reg. No.	Student Name
1	18ETCS002102	S Sadhana
2	18ETCS002104	Sahil Salim

Supervisors: Mr. Nithin Rao R
Dec – 2021

B. Tech. in Computer Science and Engineering

FACULTY OF ENGINEERING AND TECHNOLOGY

M. S. RAMAIAH UNIVERSITY OF APPLIED SCIENCES

Bengaluru -560 054

FACULTY OF ENGINEERING AND TECHNOLOGY

Certificate

This is to certify that the Seminar titled "Machine Learning in Cancer Diagnosis" is a bonafide work carried out in the Department of Computer Science and Engineering by Ms. S Sadhana bearing Reg. No. 18ETCS002102 and Mr. Sahil Salim bearing Reg. No. 18ETCS002104 in partial fulfilment of requirements of the Course curriculum of 7th Sem Computer Science and Engineering of Ramaiah University of Applied Sciences.

Dec – 2021

(Name of Mentor) : Mr. Nithin Rao R
Designation : Asst. Professor

Acknowledgements

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my seminar.

First and foremost, I would like to offer my sincere gratitude to my seminar mentor,

Mr. Nithin Rao R , Assistant Professor, Department of Computer Science Engineering, MSRUAS for his valuable suggestions, guidance, and encouragement which he has provided throughout the duration of this project. The experience, which I have gained by working under him, is an invaluable possession.

I would like to express my sincere thanks to Dr. H.M. Rajashekara Swamy, Dean, Faculty of Engineering and Technology, MSRUAS for providing necessary support for this project.

I am also very thankful to Dr. Pushphavathi T P, Head of Department, Computer Science Engineering, MSRUAS for providing all the help and facilities to carry out the research work.

Summary

Cancer is characterized as a heterogenous disease and consisting of various different subtypes. The early diagnosis and detection of cancer has become a necessity in cancer research as it facilitates the subsequent clinical management of patients. The importance of classifying cancer patients into high risk or low risk has led to extensive research into the application of machine learning (ML) methods. The ability of ML tools to detect key features from complex datasets highlights their importance.

A variety of these techniques, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Even though it is evident that the use of ML methods can improve our understanding of cancer detection , an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice.

This seminar focuses on common ML approaches employed in the modelling of cancer progression. The most recent publications that employ these techniques as an aim to model cancer risk or patient outcomes are presented in this seminar .

The different models can be analysed using different evaluation metrics such as accuracy, sensitivity, specificity, etc. Finally, challenges are also highlighted for possible future work to be conducted in this domain,

Table of Contents

List of Tables

List of Figures

1. Introduction

The utilization of data science and machine learning approaches in medical fields proves to be prolific as such approaches may be considered of great assistance in the decision making process of medical practitioners. With an unfortunate increasing trend of cancer cases , comes also a big deal of data which is of significant use in furthering clinical and medical research, and much more to the application of data science and machine learning in the aforementioned domain.

Cancer is the leading cause of deaths worldwide . Both researchers and doctors are facing the challenges of fighting cancer . According to the American cancer society, 96,480 deaths are expected due to skin cancer, 142,670 from lung cancer, 42,260 from breast cancer, 31,620 from prostate cancer, and 17,760 deaths from brain cancer in 2019 (American Cancer Society, new cancer release report 2019. Early detection of cancer is the top priority for saving the lives of many. Typically, visual examination and manual techniques are used for these types of a cancer diagnosis. This manual interpretation of medical images demands high time consumption and is highly prone to mistakes.

With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type.

This study compares some machine learning techniques to detect the disease from the input features. Five supervised machine learning approaches have been used to diagnose the disease with proper outcome. The remaining part of the study is organized as follows. The next section outlines the current review of the state of the art in this field followed by which the methods and materials used for the study are illustrated. The theoretical concept of each machine learning technique is illustrated in the subsequent section. Then the performance measurement parameters are described. The final section draws a conclusion and inferences for future improvements.

2.Background Theory

In machine learning and data mining, classification should be a crucial task. Researchers have already done lot of research by applying machine learning algorithm on medical dataset for classification and data mining algorithm to find a pattern in dataset for faster calculation and prediction. Many of the approaches provide good accuracy and result.

With the evolution of medical research, numerous new systems have been developed for the detection of breast cancer. The research associated with this area is outlined in brief as follows

[1] This paper has implemented algorithms like C4.5, ANN, SVM to find classification accuracy in breast cancer dataset. Their research shows SVM had produced higher accuracy in classification.

[2] Their research is about finding classification accuracy using machine learning algorithm known as k-Nearest Neighbor with different values of k. Each value of k produces a different result .

[3] Their paper is about using powerful machine learning classification algorithm Naïve Bayes, C4.5 which is usually used in data mining and ANN a neural network algorithm for the tumour classification of breast cancer in dataset. Their work shows C4.5 did a better job in classification.

[4] Random forest classifier was implemented in their project to find sensitivity, time consumed and mean accuracy of two data set WBCPD and WBCDD.

[5]Performance criterion of classifiers is compared by Vikas Chaurasia and Saurabh Pal for SVM with the RBF kernel, naïve bayes, rbf kernel in neural networks, simple cart and algorithm in decision trees in breast cancer dataset to find the best classifier. Their experimental results say, SVM-RBF kernel produces an accuracy of 96.84% which is higher than other classifiers.

The performance and efficiency of the algorithms such as SVM, Random Forest, Logistic Regression and Naïve bayes were compared to the similar works mentioned above. The

goal is to achieve the lowest error rate and best accuracy in analysing data. The performance and efficiency of these approaches are compared using: accuracy and time to build model.

3. Aim and Objectives

- **Title**
  - ❖ Machine Learning Applications in Cancer Diagnosis
- **Aim**
  - ❖ To perform a comparative analysis study on cancer detection models that applies machine learning to address the topic of automated diagnosis.
- **Objectives**
  - ❖ To highlight, inform, and discuss a few selected applications on cancer predictive machine learning methods
  - ❖ To identify benefits and challenges in cancer detection models.

- **Methods and Methodology/Approach to attain each objective**

Objective No.	Statement of the Objective	Method/ Methodology	Resources Utilised
1	To highlight, inform, and discuss a few selected applications on cancer predictive machine learning methods	Literature review was carried out by referring reputed journals, books, manuals and related documents .	Reputed journals and books on cancer predictive machine learning models.
2	To identify benefits and challenges in cancer detection models.	Examining the drawbacks of existing models and providing recommendations for improvement of the same.	Reputed journals

Table 1 : Defining all objective and resources utilized for each objective

4 Discussion and Results

Typical Framework for Cancer Diagnosis

- **Pre-processing -** Raw images contain noise in it so the first step in detection procedure is pre-processing, i.e., improving the quality of an image to be used further by the removal of unwanted image information, which is referred to as the image noises. Several inaccuracies may occur in the classification if this issue is not entertained properly. In addition to inaccuracies, the requirement of performing this pre-processing is because of low contrast among skin lesion and surrounding healthy skin, irregular border and the skin artifacts, which are hairs, skin lines, and black frames. Many filters can be applied for removal of Gaussian noise, speckle noise, Poisson noise, and salt and pepper noise, including median filter, mean filter, adaptive median filter, Gaussian filter, and adaptive wiener filter.
  The right combination of pre-processing tasks gives more accuracy. Some of the preprocessing techniques are black frame removal techniques, automatic color equalization, hair removal technique, dull Razor, Karhunen–Loe've transform , Gaussian filter, pseudo-random filter, non-skin masking, color space transform, and contrast enhancement.

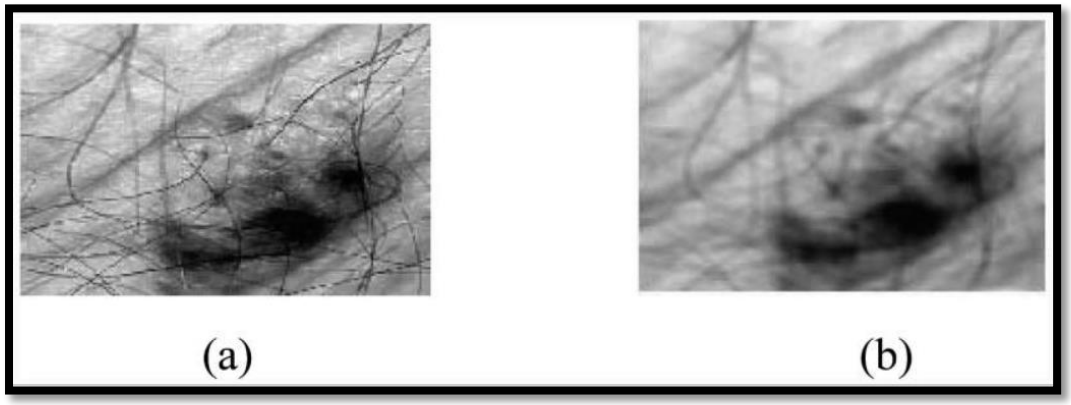

Fig1.1 Median filtering: (a) original image, (b) result after median filtering.

The algorithm of the median filter, which is a nonlinear filter, replaces each pixel by the median value of the neighbouring pixels. This filter is used on skin tumour images. It is noticed that noise is not completely eliminated and some residual noise depicting some hair traces remained.

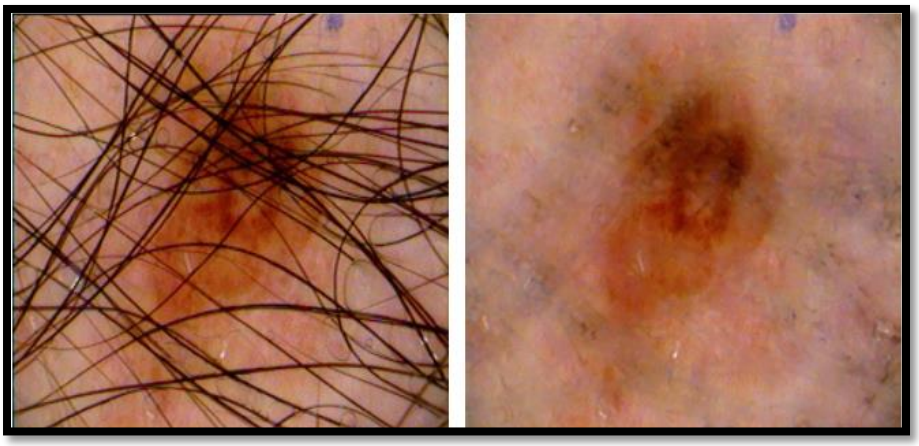

Fig1.2 Hair removal by the DullRazor technique: (a) original image, (b) image after removal of hair.

The DullRazor algorithm is as follows:

1. It identifies the dark hair locations by a generalized grayscale morphological closing operation,
2. It verifies the shape of the hair pixels as thin and long structure, and replace the verified pixels by a bilinear interpolation, and
3. It smooths the replaced hair pixels with an adaptive median filter.

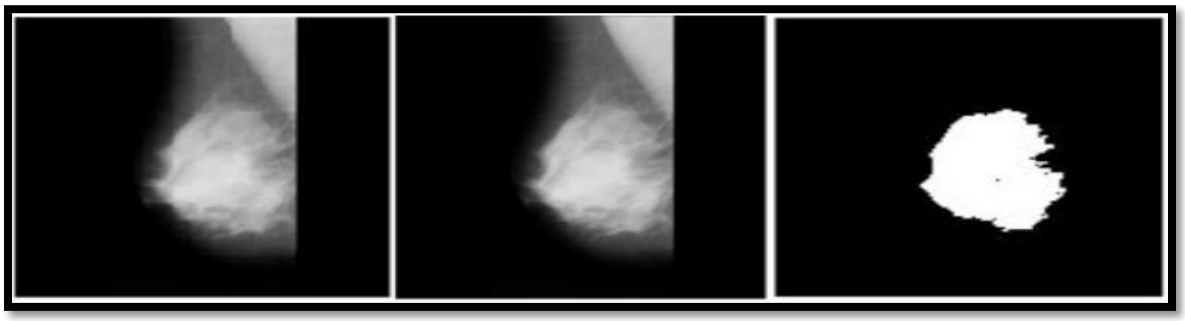

Fig.1.3 Segmentation (a) Original Image, (b) Preprocessed Image, (c) Segmented Image.

- Image Segmentation:

  Division of the input image into regions where the necessary information for further processing can be extracted is known as segmentation. Segmentation is basically the separation of a region of interest (ROI) from the background of the image. ROI is the part of the image that we want to use. In the case of cancerous images, we need the lesion part to extract the features from the diseased part. Segmentation can be divided into four main classes: (i) threshold-based segmentation; (ii) region-based segmentation; (iii) pixel-based segmentation; and (iv) model-based segmentation. Threshold-based segmentation includes Ostu's method, maximum entropy, local and global thresholding, and histogram-based thresholding.

- Post-Processing After passing through the stages of preprocessing and image segmentation, there awaits post-processing where the task is to grab features. To accomplish this, the most common post-processing methods are opening and closing operations, island removal, region merging, border expansion, and smoothing.

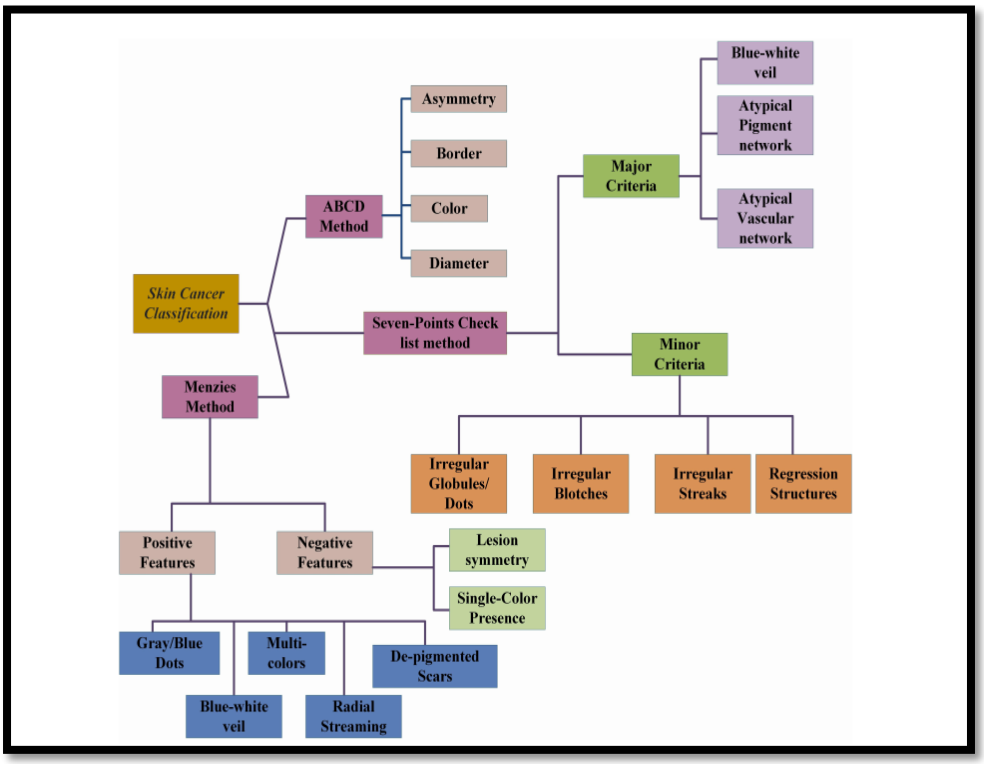

Fig 1.4 : Typically used classification methods

- ABCD-Rule:

ABCD-rule analysis refers to asymmetry (A), border (B), color (C) and diameter (D) of the lesion image). (A) Asymmetry: The input image is divided into a perpendicular axis in such a way that it gives the lowest possible value of asymmetry score. The score will be 2 if the asymmetry is with respect to the axes. If it is asymmetric on one axis, then its score will be 1. No asymmetry gives 0 scores. (B) Border: The image is divided into eight and checked for sharp and abrupt changes. Then, the score is checked, where a sharp cut off scores 1 and gradually scores 0. (C) Color: There are shades of colors for cancer detection: black and brown, but also sometimes white, red or pink. Colors are counted. (D) Diameter: The diameter of the lesion is carefully checked. If it is larger than 6 mm in diameter, then it is melanoma. Figure 1.4 shows the block diagram of all the methods described in this section.

- Seven-Point Checklist Method There are two types of criteria based on which classification is done. These are major and minor criteria. The major criteria have three points, and each point has a score value of 2, whereas minor criteria have four points each with a score value of 1. If the score value is at least 3, the classification result would be malignant melanoma
- Menzies Method There are a few positive features (Pos.F) and negative features (Neg.F). The presence of any negatives declares melanoma to be malignant. It would be benign if both negatives are absent and one or more positives are true.

Machine Learning

Machine Learning (ML), is a subfield of Artificial Intelligence (AI) that allows machines to learn without explicit programming by exposing them to sets of data allowing them to learn a specific task through experience . Over the last few decades, ML methods have been widespread the development of predictive models in order to support effective decision-making. In cancer research, these techniques could be used to identify different patterns in a data set and consequently predict whether a cancer is malignant or benign. The performance of such techniques can be evaluated based on the accuracy of the classification, recall, precision, and the area under the ROC.

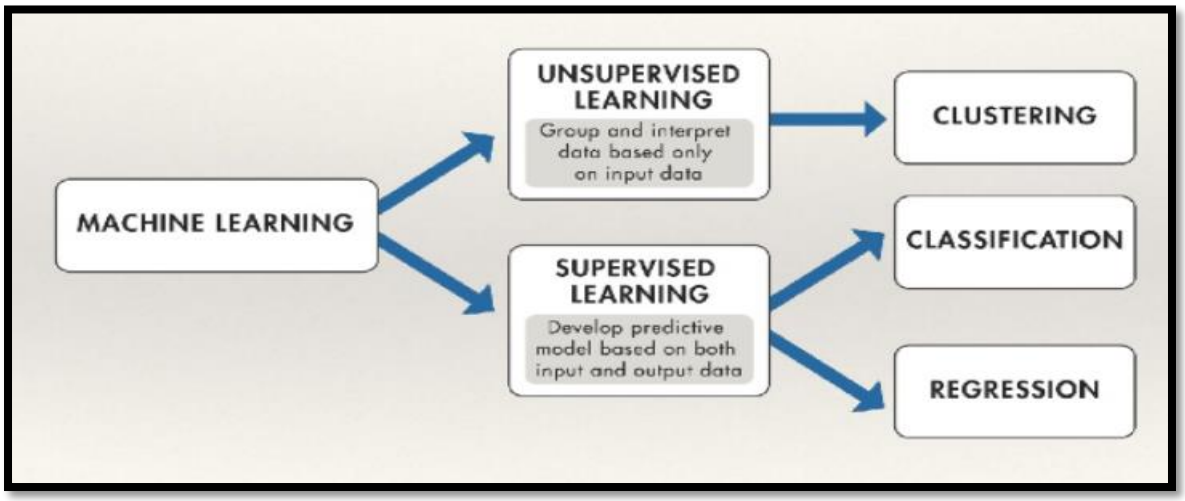

Fig.1.5 Supervised vs Unsupervised Learning

The learning process in ML techniques can be divided into two main categories, supervised and unsupervised learning. In supervised learning, a set of data instances are used to train the machine and are labelled to give the correct result. However, in unsupervised learning, there are no pre-determined data sets and no notion of the expected outcome, which means that the goal is harder to achieve.

The training phase extracts the features from the dataset and the testing phase is used to determine how the appropriate model behaves for prediction. The dataset is divided into

wo sections. These are the training and testing phase. K fold cross-validation depicts that a single fold is utilized for testing and k−1 folds are being used training circularly. Cross-validation is used for the avoidance of overfitting. In our study, a ten-fold cross-validation technique is used to partition data in which nine-fold are used for training and the remaining one-fold for testing in each iteration

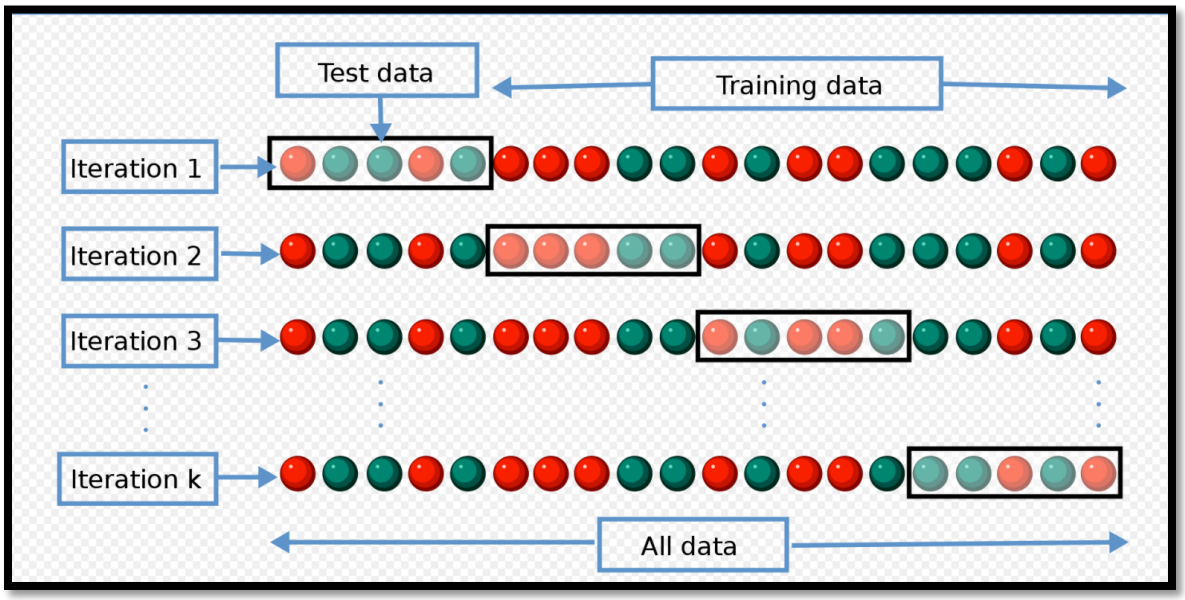

Fig 1.6 K fold cross-validation

Some of the common machine learning techniques employed for cancer prediction is highlighted in the section below :

- Support Vector Machine (SVM) is one of the supervised ML classification techniques that is widely applied in the field of cancer diagnosis and prognosis. SVM functions by selecting critical samples from all classes known as support vectors and separating the classes by generating a linear function that divides them as broadly as possible using these support vectors. Therefore, it can be said that a mapping between an input vector to a high dimensionality space is made using SVM that aims to find the most suitable hyperplane that divides the data set into classes . This linear classifier aims to maximize the distance between the decision hyperplane and the nearest data point, which is called the marginal distance, by finding the best suited hyperplane.
  The dataset categories cannot be divided using hyperplane, so feature space has to be enlarged using Gaussian radial basis function (RBF) or sigmoid function, cubic, quadratic or even higher order polynomial function.

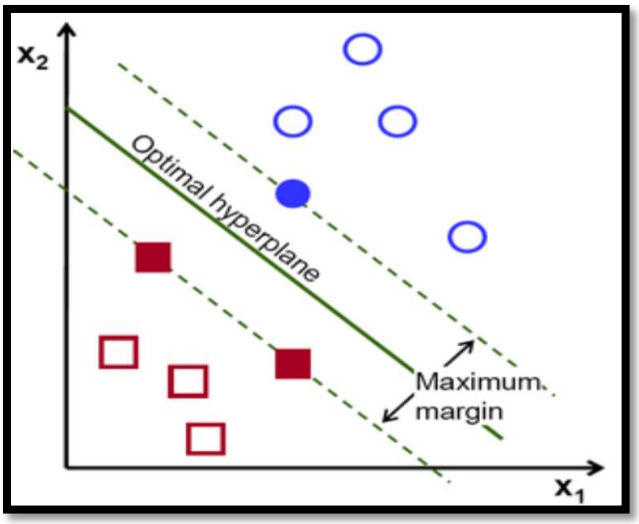

Fig.1.7 Hyperplane

The hyperplane that is used in p-dimensions is as follows:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots \beta_p X_p = 0$$

where $X1, X2, \ldots, and\ Xp$ are the data points in the sample space of p-dimension and $\beta 0, \beta 1, \beta 2, \ldots, and\ \beta p$ are the hypothetical values.

- K-Nearest Neighbours:

  K-nearest neighbour algorithm is utilized for grouping and used in pattern recognition. It is widely used in predictive analysis. On the arrival of new data, the K-NN algorithm identifies existing data points that are nearest to it. Any attributes that can differ on a large scale may have sufficient influence on the interval between data points . The feature vectors, as well as class labels, are stored in the training phase. K-NNs assume that the data samples are represented in a metric space. In the classification phase, first, the quantity is characterized by neighbours of K that is the K most regular among the K training sample. At that point, the calculation will discover K adjacent neighbours of the new data sample. As all the data points are in metric space, a significant concern is how the distance will be calculated.

  If the number of neighbors is denoted by N in K-NNs, then N samples are considered using:

  Minkowski Distance: Dist(x,y)= $(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$

  where if p = 1, then it is Manhattan distance, if p = 2, then it is Euclidean distance and if p = ∞, then it is Chebyshev distance.

- Random Forest

  Random forest classifier is a powerful supervised classification tool. The RF classification is an ensemble method that can be studied as a form of the nearest neighbour predictor. Ensemble learning is the method by which statistical methods like classifiers or experts are strategically developed and incorporated to solve a specific problem of computational intelligence. RF generates a forest of classification trees from a given data-set, rather than a single classification tree.

  The workflow of random forest is given below.

  i. From the training set, picked K data points randomly.
  ii. From these K data points, generate the decision trees.
  iii. From generated trees, choose the number of N-tree and repeat steps (i) and (ii).
  iv. Form the N-tree that predicts the category to which the data points relate for a new data point, and assign the new data point via the category with the highest probability.

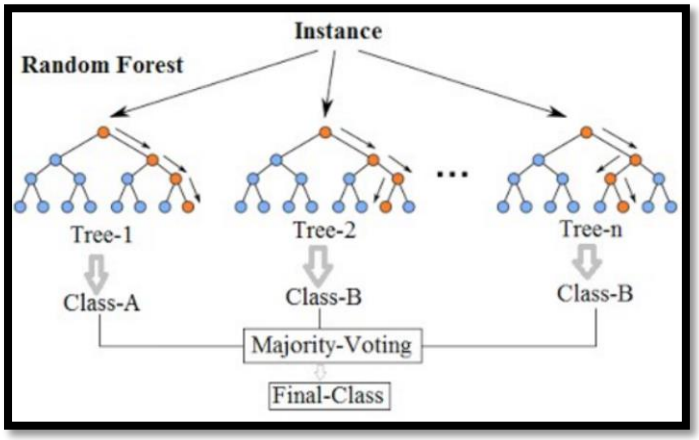

Fig.1.8 Random Forest

- Artificial Neural Networks

  Artificial neural network algorithm is slightly inspired by biological neuron and work by following the workflow of biological neurons dendrite, soma, and axon. The internal structure of every ANN is an artificial neuron and a simple mathematical function . The basic architecture of an artificial neural network is a set of interconnected neurons located in three different layers named input, hidden, and output layers. This type of network generally learns to per-form tasks by considering a sufficient number of examples. The neural network can be applied both for classification and regression problems.The representation for forward propagation and prediction of a single neuron:

$$\text{Output} = b_i + \sum_{j=1}^{n_x} w_{ij} x$$

where w_{ij} is the weight from input to output layer, $b_i$bias value, and x_i is the input value. An activation function is applied to the output value after the calculation of it.

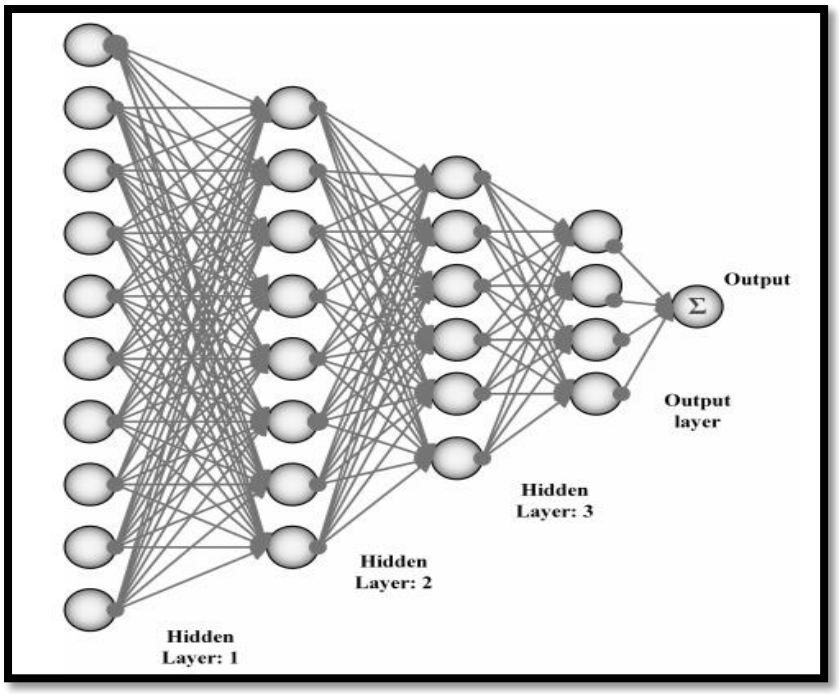

Fig .1.9 Artificial Neural Network

- **Logistic Regression**

  Logistic Regression is an analytical modelling technique where the likelihood of a level is associated with a set of explicative variables. It is used for analysing a dataset in which there are one or more independent variables that decide a result. The result is measured with a binary variable (in which there are only two possible results). It is applied to predict a binary result (True/False, 1/0, Yes/No) given a set of independent variables. The following equations are the representation of the LR model:

$$x = c_0 + \sum_{i=1}^{n} c_i x_i \qquad P(x) = {e^x}/{1 + e^x}$$

  where x is a quantity of the participation of the illustrative variables xi (i = 1,…,n), ci is the regression coefficient that is achieved by the highest probability in association with its usual errors. Δci and P(x) are the certain acknowledgments of variables that describe the likelihood of an excitement.

Performance Measure Parameters :

An evaluation metric quantifies the performance of a predictive model. For classification problems, metrics involve comparing the expected class label to the predicted class label or interpreting the predicted probabilities for the class labels for the problem.

- True positive (TP) is the correct classification of the positive class
- True negative (TN) is the correct classification of the negative class
- False positive (FP) is the incorrect prediction of the positives
- False negative (FN) is the incorrect prediction of the negatives

The classifier's performance is evaluated by the following formulas:

- Accuracy is the fraction of predictions our model got right:

$$Accuracy\ (Acc)\ =\ \frac{(TP\ +TN)}{(TP\ +TN\ +FP\ +FN)}$$

- Sensitivity is the amount of positive items correctly identified:

$$Sensitivity(Sens)\ =\ \frac{TP}{TN\ +\ FN}$$

- Specificity (Spec) Specificity means the relationship of observed negative examples with all negative examples, says the rate of predicted presence including entire examples by the presence of cancer

$$Specificity\ (Spec)\ =\frac{\text{TN}}{\text{TN +FP}}$$

- Precision (Prec) Precision is named the division of the examples which are actually positive among all the examples that we predicted positive:

$$Precision(Prec)\ =\ \frac{TP}{TP\ +\ FP}$$

- Matthews correlation coefficient (MCC) For binary classification, MCC is used. Here the range is +1 to −1. When the value is +1, the best performance is shown and when the value is –1, the worst performance is shown. It is rep-resented as:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- Receiver operating characteristic curve (ROC) curve, the true-positive rate (sensitivity) is plotted against the false-positive rate (1−specificity) at various

threshold settings. ROC curve expresses a relation between true-positive rate vs. false-positive rate

- Area under the ROC Curve (AUC) AUC provides the area under the ROC-curve integrated from (0, 0) to (1, 1). It gives the aggregate measure of all possible classification thresholds. AUC has a range from 0 to 1. A 100% correct classified version will have the AUC value 1.0 and it will be 0.0 if there is a 100% wrong classification. It is attractive for two reasons: first, it is scale-invariant, which means it checks how well the model is predicted rather than checking the absolute values; and, second, it is classification threshold invariant as it will check the model's performance irrespective of the threshold being chosen.

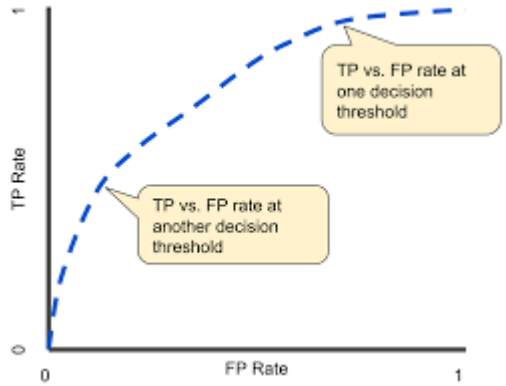

Fig 1.10 ROC curve

Case Study

Evaluation of the classifiers' performance based on the metrics.We have chosen a Cancer Database named WBC to train our Machine Learning Model.

The Wisconsin-Breast Cancer (Diagnostics) dataset (WBC) was taken from UCI machine learning repository.It is a classification dataset, which records the measurements for breast cancer cases. There are two classes, benign and malignant.

Breast cancer prediction

Wisconsin Breast Cancer Database (WBCD) data from the UCI Repository

	Accuracy	Sensitivity	Specificity
Forest Tree	0.926	0.935	0.907
ANN	0.955	0.953	0.947
SVM	0.934	0.948	0.926

Fig 1.11 Comparative study

5. Conclusions and Suggestions for Future Work

This study presented a comparative study of five machine learning techniques for the prediction of cancer, namely support vector machine, K-nearest neighbours, random forests, artificial neural networks, and logistic regression. The basic features and working principle of each of the five machine learning techniques were illustrated.

The system proposed that machine learning technique can be acted as a clinical assistant for the diagnosis of cancer and will be very helpful for new doctors or physicians in case of misdiagnosis. The developed model by ANNs is more consistent than any other technique stated, and it may be able to bring changes in the field of prediction of breast cancer. From the study, we can conclude that machine learning techniques are able to detect the disease automatically with high accuracy.

When implementing machine learning for cancer diagnosis, one of the major challenges becomes a lack of availability of datasets. Every learning algorithm requires a large amount of training for performance measure. However, efforts have been made to make medical images archives containing confidential information of many patients by picture archiving and communication society (PACS). To deal with the issue of limited dataset, a scheme of data augmentation was proposed. Many researchers use data augmentation, which includes techniques such as rotation, cropping and filtering to increase the number of available data.

Another issue that was observed is the inequality of training data distribution. If the positive data are made larger than the negative data, then the system will be automatically biased and will majorly give positive results while the same happens if the data have more negative than positive cases. Thus, equality of training data is very important, which was ignored by few researchers.

References

[1] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. Journal of Health & Medical Informatics

[2] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Applications (0975 - 8887)

[3] Abdelghani Bellaachia, Erhan Guven Predicting Breast Cancer Survivability Using Data Mining Techniques. 2006 SIAM Conference on Data Mining

[4] Cuong Nguyen, Yong Wang, Ha Nam Nguyen Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomedical Science and Engineering, 2013, 6, 551-560

[5] Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013

[6] Munir, Khushboo; Elahi, Hassan; Ayub, Afsheen; Frezza, Fabrizio; Rizzi, Antonello. 2019. "Cancer Diagnosis Using Deep Learning: A Bibliographic Review" Cancers 11, no. 9: 1235.

[7] Islam, Md & Haque, Md & Iqbal, Hasib & Hasan, Md & Hasan, Mahmudul & Kabir, Muhammad Nomani. (2020). Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN Computer Science. 1. 290. 10.1007/s42979-020-00305-w