IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SURVEY PAPER ON BREAST CANCER PREDICTION USING MACHINE LEARNING & CROSS-VALIDATION TECHNIQUES

Abhishek Powar, Aditi Batwal, Pranav Kale, Shubham Kamble

Department Of Computer Engineering, RMD Sinhgad School of Engineering, Savitribai Phule Pune University, Warje, Pune, India

Abstract: Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. We are using machine learning in our daily life even without knowing it such as Google Maps, Google Assistant, Alexa, etc. One of the major applications of machine learning is the cancer prediction system. Cancer is the most dangerous disease in the world. Amongst them, breast cancer is the most common type of cancer among women and its incidence is increasing day by day. Machine learning techniques can make a large contribution to the process of prediction and early diagnosis of breast cancer. Machine learning algorithms like decision trees, KNN, SVM, naïve Bayes, etc. give better performance in their own field. In this study, we applied five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree, and K-Nearest Neighbors (KNN) on the dataset, after obtaining the results, a performance evaluation and comparison are carried out between these different classifiers. The main objective of this research paper is to predict breast cancer, using machine-learning algorithms and apply cross-validation techniques on different machine-learning algorithms. This model can be used in medical departments to enhance Cancer Detection & result from accuracy. In the future, this model can be used to predict breast cancer using different datasets and parameters.

Index Terms **– KNN, SVM, Decision Tree, Random Forest, Breast Cancer, Cross Validation**

I. INTRODUCTION

In today's technologically advanced generation, artificial intelligence is something that is constantly changing. Artificial intelligence is a wide-ranging branch of the computer science field that builds smart machines that perform tasks that require human intelligence. Machine learning is a subfield of artificial intelligence. Machine learning designs and applies algorithms that learn things from past cases. Basically, machines learn from history to give appropriate results. Supervised machine learning is the most common method which uses labeled data for training models. In this method, the model gives results more accurately. In this learning, machines are trained with input and corresponding output and then this model is used to predict results for test data sets. Some examples of supervised machine learning are logistic regression, random forest, decision tree, support vector machine, etc.

Applications of Machine Learning: The real-world applications of machine learning that we are using daily are Google Maps, Google Assistant, Alexa, etc. The most trending real-world applications of machine learning are Image Recognition, Speech Recognition, Traffic prediction, Self-driving cars, etc.

One of the major applications of machine learning is the cancer prediction system.

Machine learning, as a modeling approach, represents the process of extracting knowledge from data and discovering hidden relationships, widely used in healthcare in recent years to predict different diseases.

Breast Cancer: Cancer is the most dangerous disease in the world. Amongst them, breast cancer is the most common type of cancer among women and its incidence is increasing day by day. According to WHO, 627,000 women died from breast cancer in 2018. Breast cancer is the main problem that spreads everywhere in the world but is mostly found in the United State of America. Breast cancer treatment can be highly effective, especially when the disease is identified early. Machine learning techniques can make a large contribution to the process of prediction and early diagnosis of breast cancer. Machine learning techniques can be used to identify the presence or absence of cancerous cells which helps doctors to start a proper treatment on patients. Machine learning algorithms like decision trees, KNN, SVM, naïve bayes, etc. give better performance in their own field. In this study, we studied some machine learning algorithms which are: Support Vector Machine (SVM), Random Forest, Decision tree, and K-Nearest Neighbors (KNN).

II. LITERATURE SURVEY

A. Ramik Rawal in the year 2020 has researched three domains. It comprises the first domain as a prediction of cancer before the diagnosis, the second domain is a prediction of diagnosis and treatment and the third domain focuses on outcome during treatment. Also, the paper gives a comparison between the performance of four classifiers: SVM, Logistic Regression, Random Forest, and kNN on the basis of their accuracy. A further 10-fold cross-validation method is used to evaluate data and analyze data in terms of effectiveness and efficiency.

B. Sarthak Vyas, Abhinav Chauhan, Deepak Rana, and Noman Ansari applied 3 main machine learning algorithms such as K-Nearest-Neighbor, Support Vector Machine (SVM), and Logistic Regression and presented a work that is centered around the advancement of predictive models to accomplish great precision in foreseeing legitimate disease results utilizing supervised machine learning techniques.

C. As there is a chance of fifty percent fatality in case one of two women diagnosed with breast cancer deaths in the cases of Indian women, Shubham Sharma, Archit Aggarwal, and Tanupriya Choudhury proposed a model which presents a comparison of the largely popular machine learning algorithms and techniques commonly used for breast cancer prediction, namely Random Forest, kNN (k-Nearest-Neighbor) and Naïve Bayes. The results obtained in the comparison can be used for breast cancer detection and treatment.

D. Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi applied the genetic programming technique to select the best features and perfect parameter values of the machine learning classifiers. The performance of the proposed method was based on sensitivity, specificity, precision, accuracy, and the roc curves. They have used the Wisconsin breast cancer dataset.

E. Reza Rabiei, Seyed Mohammad Ayyoubzadeh, Solmaz Sohrabei, Marzieh Esmaeili, and Alireza Atashi aimed to predict breast cancer using different machine-learning approaches by applying demographic, laboratory, and mammographic data. Dataset from Motamed cancer institute (ACECR), Tehran, Iran is used for training the machine learning model.

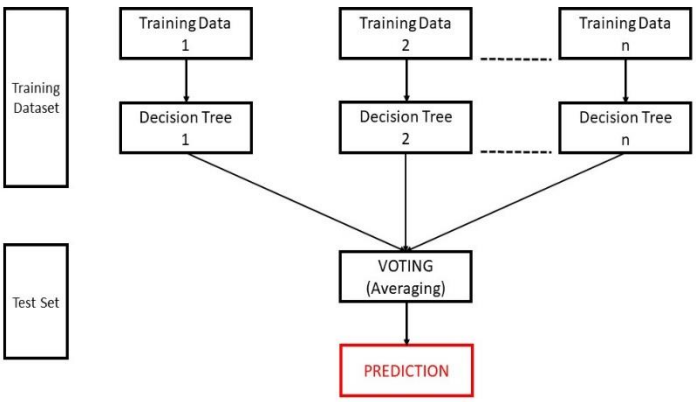
F. Manav Mangukiya, Anuj Vaghani, and Meet Savani reviewed various techniques of machine learning algorithms to detect breast cancer early, efficiently, and accurately. The research paper comprises Data Visualization and performance comparisons between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB), K Nearest Neighbors (k-NN), Adaboost, Xgboost, and Random Forest conducted on Wisconsin Breast cancer Dataset. They evaluated the accuracy, precision, sensitivity, and specificity of the model.

G. Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Yaman Afadar, and Omar Elgendy have reviewed previous work on detecting and treating breast cancer using genetic sequencing or histopathological imaging with the help of deep learning and machine learning.

The study's main aim was to detect and access the classification of breast cancer using a machine learning algorithm that provides high classification accuracy and effective diagnostic capabilities where classification is binary (benign cancer/malign cancer).

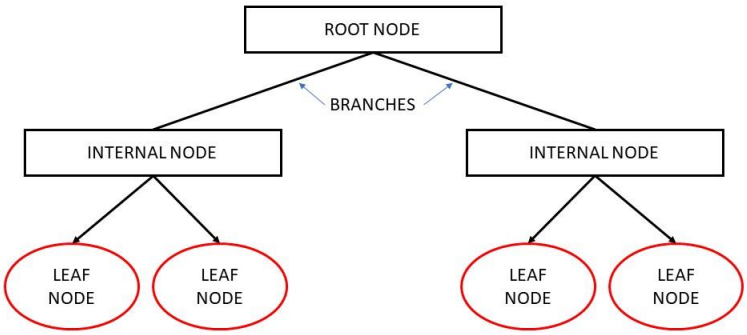
III. METHODS USED

- **Random Forest**-Random Forest is a classifier that consists of decision trees on various subsets of the given dataset and it takes the average of decision tree results to improve the predictive accuracy of that dataset. Greater the number of decision trees in the forest higher the accuracy of the model which will prevent the problem of overfitting.



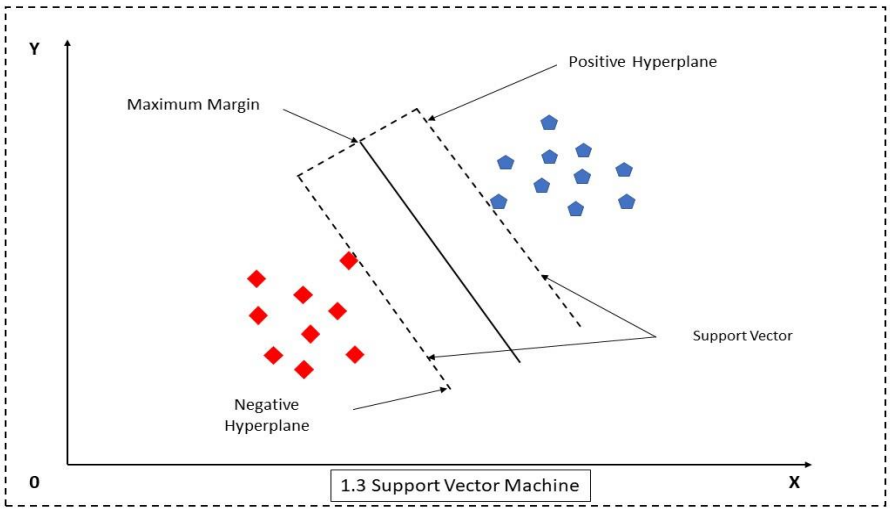
1.1 Random Forest

- **Decision Tree-** The decision tree shows the graphical representation of all the possible solutions to a decision based on given conditions. In this tree-structured classifier, internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome of that tree.



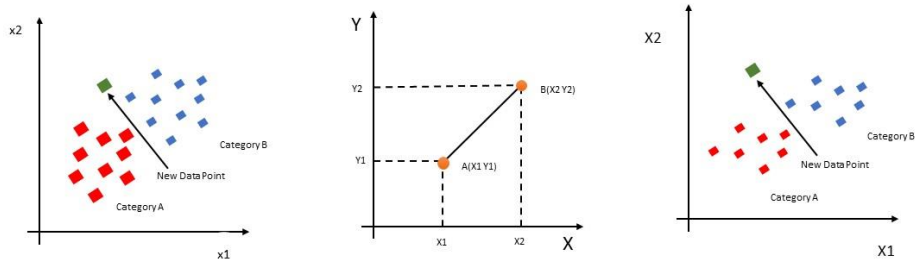
1.2 Decision Tree

- **Support Vector Machine**-The supervised machine learning algorithm which works on labeled data and is used for classification and regression problems. SVM creates a decision boundary which is called a hyperplane that segregates n-dimensional space into classes so that one can put the new data point in the correct category.[10]
  Here's how a support vector machine algorithm model works:
  - First, it finds lines or boundaries that correctly classify the training dataset.
  - Then, from those lines or boundaries, it picks the one that has the maximum distance from close data points.



1.3 Support Vector Machine

- **KNN- K-Nearest Neighbor**-KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into appropriate categories by using the KNN algorithm. KNN stores the dataset and it performs all actions on the dataset at the time of classification. Therefore, it is also called a lazy learner algorithm.[11]
  Here's how a kNN algorithm model works:
  - Input the dataset and split it into a training and testing set.
  - Pick an instance from the testing sets and calculate its distance with the training set.
  - List distances in ascending order.
  - The class of the instance is the most common class of the 3 first training instances (k=3).



1.4 k-nearest neighbors algorithm

Cross Validation- It is a technique for validating the model efficiency by training it on the subset of input data (i.e., training data) and testing on a previously unseen subset of the input data (test data). Cross Validation can be used to compare the performance of different predictive modeling methods. There are some techniques of cross-validation that are listed below:[12]

Holdout method: In this method, one subset of the training data is removed and used to get prediction results by training it on the rest of the dataset. Error obtained i.e., results of the model will tell about how well the given model will perform when given an unknown dataset.

Leave-P-out cross-validation: In this approach, the p datasets are left out of the training data. It means, if there are total n data points in the original input dataset, then n-p data points will be used as the training dataset and the p data points as the validation set. This complete process is repeated for all the samples, and the average error is calculated to know the effectiveness of the model.

Leave one out cross-validation: This method is similar to the leave-p-out cross-validation, but instead of p, we need to take 1 dataset out of training. It means, in this approach, for each learning set, only one data point is reserved, and the remaining dataset is used to train the model. This process repeats for each data point. Hence for n samples, we get n different training sets and n test sets.

K-fold cross-validation: K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called folds. For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test set.

Stratified k-fold cross-validation: This technique is similar to k-fold cross-validation with some little changes. This approach works on the stratification concept, it is a process of rearranging the data to ensure that each fold or group is a good representative of the complete dataset.

IV. CONCLUSION

We can conclude from this survey that using different machine learning model's prediction of breast cancer is more accurate and we can apply various cross-validation techniques to the model which will increase its accuracy on unseen datasets. The results of this can be useful for doctors to monitor these patients closely and start the treatment of the disease immediately. In the future, this model can be used to predict breast cancer using different datasets and parameters. Further research in this field should be carried out for the better performance of the classification techniques so that it can predict more variables.

V. REFERENCES

[1] Ramik Rawal. "BREAST CANCER PREDICTION USING MACHINE LEARNING." *Journal of Emerging Technologies and Innovative Research (JETIR)*, 2020, https://www.researchgate.net/publication/341508593

[2] Sarthak Vyas, Abhinav Chauhan, Deepak Rana, Noman Ansari. "Breast Cancer Detection Using Machine Learning Techniques." *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 10, no. V, 2021, https://doi.org/DOI: 10.22214/ijraset.2022.43055

[3] Shubham Sharma, Archit Aggarwal, and Tanupriya Choudhury . "Breast Cancer Detection Using Machine Learning Techniques." *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2021, https://doi.org/10.1109/CTEMS.2018.8769187

[4] Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi. "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms." *Hindawi Journal of Healthcare Engineering*, 2019, https://doi.org/10.1155/2019/4253641

[5] Reza Rabiei, Seyed Mohammad Ayyoubzadeh, Solmaz Sohrabei, Marzieh Esmaeili, and Alireza Atashi. "Prediction of Breast Cancer Using Machine Learning Approaches." *Journal of Biomedical Physics and Engineering*, 2021, https://doi.org/10.31661/jbpe.v0i0.2109-1403.

[6] Manav Mangukiya, Anuj Vaghani, Meet Savani. "Breast Cancer Detection with Machine Learning." *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 10, no. 2, 2022, https://doi.org/10.22214/ijraset.2022.40204.

[7] Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Yaman Afadar, and Omar Elgendy. "Breast Cancer Detection Using Artificial Intelligence Techniques: A Systematic Literature Review." *Artificial Intelligence in Medicine, Elsevier*, vol. 127, no. 2, 2022.

[8] "Random Forest Algorithm." www.javatpoint.com/machine-learning-random-forest-algorithm.

[9] "Decision Tree Classification Algorithm." www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[10] "Support Vector Machine Algorithm." www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[11] "K-Nearest Neighbor (KNN) Algorithm for Machine Learning." www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

[12] "Cross-Validation in Machine Learning." www.javatpoint.com/cross-validation-in-machine-learning.