BREAST CANCER PREDICTION USING MACHINE LEARNING ALGORITHMS

1Murikipudi Nikhil’ 2 Thotakura Manish’ 3 P.Lakshmi Prasanna’ 4G S Lahari’ 5Yarkareddy Bhavana Reddy

1,2,3,4,5Department of Computer science and Engineering

1,2,3,4,5Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh.

Emails: 1[Nikhil.murikipudi@gmail.com](mailto:Nikhil.murikipudi@gmail.com) ,2[manishthotakura0216@gmail.com](mailto:manishthotakura0216@gmail.com) , 3[lakshmiprasannap87@gmail.com](mailto:lakshmiprasannap87@gmail.com) , 4[lahari.gandla@gmail.com](mailto:lahari.gandla@gmail.com) , 5[bhavana5161@gmail.com](mailto:bhavana5161@gmail.com)

Corresponding Author: 1Murikipudi Nikhil

***Abstract*—** Several of the important underlying factors hindering humans is leukaemia mortality in the emerging economies. Some cancer kinds still lack a cure, despite the fact that there are many ways to stop it from happening in the first place. One of the most common kinds of cancer is a carcinoma. Numerous research on predicting the kind of breast cancers are included in this project. The basis of this exertion is to pinpoint cancer cells applying machine learning concepts. The methodology of early diagnosis and prediction of carcinoma will benefit tremendously from the use of machine learning techniques, which have been proven to be effective and to be a hotspot for research. All experiments are carried out using the Jupyter platform and performed in a simulation environment. The proposed work may be used to forecast the results of many techniques, and based on the need, the appropriate approach can be applied. Towards better forecast accurateness, this inquiry is being undertaken.

Keywords— Risk Assessment, Early Diagnosis, Machine Learning, Cancer Prediction, Breast Cancer

# **1. Introduction :**

Breast Cancer is one weird sort of cancer which attacks breast. Carcinoma spreads as a side effect of escalating embryonic development. Breast cancer cells often develop a tumour, which is commonly noticeable on lump [1]. Mammary illness has the freedom to strengthen when leukemia cells touch the hepatic portal vein or lymph vessels system and are probably carried to other organs in the body. One of the underlying biggest reasons why carcinoma manifests is a Genetic amendment or mutation. Out of the countless variety of forms of chest chronic disease, D-C-I-S and invasive carcinoma are 2 of the most prominent.

Others are less frequent, such as phyllodes tumours and angiosarcoma. The study results of breast cancer are classified using a variety of formulas. ML has a tremendous favourable impact on the nomenclature of mammary melanoma. There are plenty varying screening methodologies, both of which are predominately reflected photographic images.

All such varieties of diagnostics digital knowledge are underlined using only machine learning. Overall efficacy of already-built paradigms is often being facilitated by webmasters incorporating ml classifiers.

Only when magnitude of knowledge is very little, machine learning typically exhibits handy insights; conversely, when the portion of info is humongous, machine learning is dysfunctional. Any of three predominant types of machine learning is said to supervise the design. In attempt to craft benefits, ml with watchful eye uses before existing statistics and every aid of the advisor. For meticulous judgement call, these schemes extricate utmost vital data from old expertise. AI technology is used in ML, which permits the systems to automatically adapt from their experiences and get better eventually without having to be manually programmed. Building algorithms that can take input data and apply numerical analysis to determine an output while modifying outcomes as new data arrives is the fundamental idea behind ML. Another very critical activity of learning is referencing evident data or acquiring data, such as when seeking patterns in statistics to obtain forthcoming judgments based on.

The immediate aim is to empower technology train purely on their own, without encouragement or meddling by inhabitants, and change state accordingly. The axiom of DL provides a way to understand text from images using algorithms such as the sliding window approach, single shot and region-based detectors, EAST (Efficient Accurate Scene Text Detector), and techniques for text recognition such as CRNN and Machine Learning OCR using Tesseract. Our experiment's mission is to more accurately classify the individuals' cancer types into normal and cancerous forms via classification models. The Kaggle website provided the information. A criterion for machine learning which we've incorporated is supervised learning, in which creating the system independent and dependent categories to learn from, and then when the learning activity has been completed, the system predicts the significance of the dependent for a specified input in independent variable format. The prediction models used to find the carcinoma include DTree, K-NN, Support VM, and NBayes. We tested above models in a Jupyter notebook and also included data visualisation while analysing data.

# **2. Literature Review:**

After gathering many more research articles which were previously published by some known authors, we have conducted some brainstorming session to find some gains and noted down all the main points which were mentioned below. This section goes into further detail on the supporting effort for the successfully completed research. In essence, we came to know there are 2 ways to identify mammary cancer. Following ml, and dl comes in second. Numerous studies that make use of machine learning are conducted. We came to know many problems were solved with machine learning techniques. This section provides details on machine and deep learning approaches conducted by some Authors.

Wang et al [2]. conducted data mining algorithms on several records in this experiment to discover the most efficient solution for breast cancer diagnoses. Individuals used the AdaBoost tree, the NBayes, SupportVM, and ArtificialNN. PCA, is made use with the intention of time decrease, was addressed after the feature space was decreased. They analysed the Wisconsin Cancer & Diagnostic datasets (1991,1995) to compare the effectiveness of the models [18, 30]. They offered a detailed review of the model's design and testing flaws.

According to Nithya et al. [3], the categorization of breast tumours is the primary problem regarding breast cancer. Breast cancer has been synthesised and characterised using computer-aided diagnosis (CAD). Their major goal was to apply data mining methodologies to enhance breast cancer prediction. The performance comparison of naive Bayes was boosted via bagging, multibooting, random subspace, supportVM-sequential minimum optimization ,and multiple layed perceptrons.

A whole other hybrid algorithm that utilizes machine learning was proposed by M. Tahmooresi [4] and others. We learnt that, SVM was an excellent classification algorithm that provided the high classification accuracy among all. SVM, K-Nearest Neighbour, artificial neural network, and decision tree each had been tested.

It was implemented towards the blood and picture datasets.

By using a wide selection of classifiers, Muhammet Fatih Aslan [5] and his companions developed a machine learning algorithm. As a classifier, the researchers utilised the Extreme Learning Machine, K-NN, SupportVM , ArtificialNN. The classifier was slightly modified to get enhanced performance. Nevertheless, the Extreme Learning Machine generated the most impressive results.

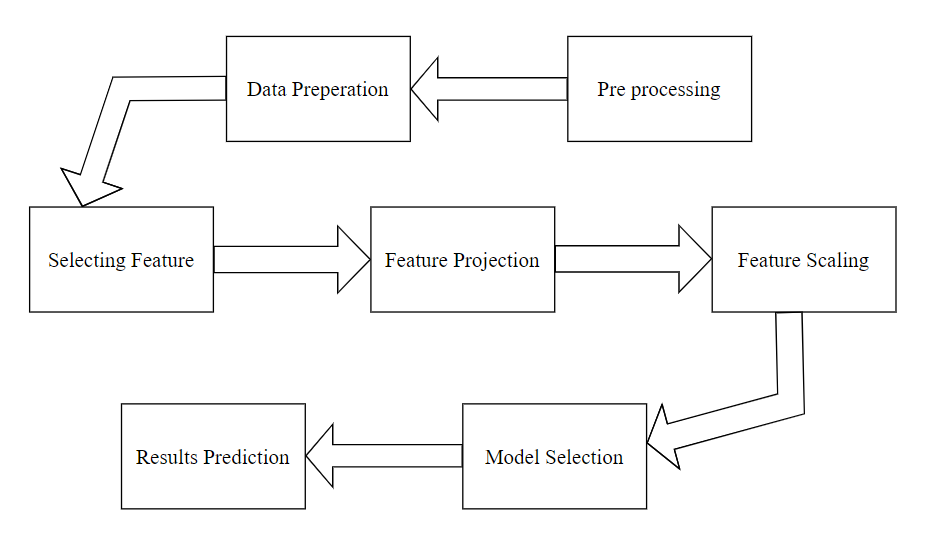
A machine learning-based model has been proposed by Anusha Bharat [6] and teammates. It included DTree, K-NNeigbor, Support VM, & Nave Baye as four distinct tacts. The article implies that KNN produced the maximum accuracy. We saw support vector machine has always had its drawbacks, even though support vector machine presented the amazing results for binary variables. Multi-SVM was adopted for this purpose.

Various Machine learning methodologies were investigated by Ebru Ayndindag Bayrak et al [7]. The WEKA & Wisconsin data were applied in their comparison. Their evidence suggests that support vector machine showed excellent performance matrix results. DL routines are invented to fix the challenges concerning ML.

Convolution neural network that relied on deep learning was the model that Shewtha K et al [8]. developed. There were several other models implemented with CNN, however they primarily adopted Mobile Network & Inception. The columnist had weighed the two routines and identified that the second tact furnished the greater guess. Regrettably, there is still potential to diagnose breast cancer using ML.

The supervised machine learning model was put forth by Ch. Shravya et al [9]. This study used classifiers like K-NNeighbor, Support VM, & LR. The datum is obtained through the UCI rep & efficiency statistics are obtained. As seen by its 92.7% accuracy in the Python environment, SVM was an excellent classifier.

# **3. Methodology:**



**Fig. 1:** Step to Step process of prediction of malignant cancer

**3.1 PRE-PROCESSING:**

A data mining approach entitled data pre-processing entails putting original data together into comprehensible format. Actual data is commonly insufficient, unreliable, and prone to various inconsistencies. A tried-and-true methodology for overcoming such conflicts is data initialisation. original data is prepared for subsequent distilling by data initialisation. The UCI information has been pre-processed by using standardized approach. This stage is crucial because the calibre and amount of data individuals acquire will greatly impact how effective your prediction model will be.

**3.2 DATA PREPARATION:**

Data loading and processing is the process by which we store our info into an appropriate database and make it prepared to be included in machine learning testing. All of our information will be loaded initially, and the arrangement will be randomly done after that.

**3.3 SELECTING FEATURES:**

Feature selecting, often referred to as variable choosing and attribute selection, is the way of choosing a subcategory of pertinent characteristics to be used in model building in statistical modelling and machine learning. For the choice of the features, we utilised the wrapper method.

**3.4 PROJECTING FEATURES:**

Data from a relatively high region is transmogrified into a way lower region using feature projection. Depending on the nature of correlations between the available features, both sequential and nonlinear mitigating measures can be applied.

**3.5 FEATURES SCALING:**

A good set of data will often include characteristics with greater magnitude, units, and span variation. However, because Euclidian distance among two variables has become the most common method used by machine learning techniques, we must adjust the amplitude of each characteristic. Scaling is one possible way to do this.

**3.6 SELECTION OF APPROPRIATE MODEL:**

Model selection is an important phase in this research because, without choosing an appropriate model, we cannot get results. While there are many models to select from, we choose a model based on our requirements and availability.

**3.7 OUTCOME PREDICTION:**

Metadata is used by machine learning to offer responses to queries. This is the culmination of all this effort, and this is where machine learning truly shines.

# **4. Implementation:**

Since breast cancer-related symptoms might not always manifest, adopting the standard precautions can assist. Therefore, the only strategy for reducing cancer fatalities is by early diagnosis and proper diagnosis, which is a challenging task in the medical world. The underlying concerns include incompetency in acquiring textual information as well as poor retrieval performance driven by poor differentiation of selection capabilities. Early detection, diagnosis, and screening are attained through mammography. Evaluation and processing are key components in this case for improved prognostic outcomes. Here, photo segmentation is carried out via the FCM concept [10]. Additionally, several characteristics are retrieved using these segmented images. Every data point with in FCM method, sometimes referred to as Fuzzy-C-means [11], is a member of several clusters that differ in membership level and are dependent on objective function. In order to look into the segmented region, the multileveled Discrete Wavelet Transformation is now being implemented. Matrix-based databases are used to store textual information like cancer cells and pixel values.

This proposed approach aims to classify the images as Initial Stage, Harmful, or Normal after extraction of features and training. It does this by applying the KNN algorithm [12], which labels it in accordance with the structure of cancer cells. Our system includes the detected boundaries along with the tumor cell after conducting a particular morphological operation and provides clear region attributes, such as Area, Euler Number, and so on. For the extraction of textural characteristics, a wide range of techniques are performed, including PCA, GLCM, and Multi leveled Discrete Wavelet Transform [13]. It is correctly noted and displayed to the doctor where the tumor's area lies.

# **5. Machine learning algorithms:**

Five different approaches, which are listed below, can be used to complete this project. We discovered which approach produces the most accurate results after testing each one.

**5.1 Logistic Regression:**

Early in the 20th century, the concept of logistic regression was ultimately used throughout the biological sciences. It has gained more importance in social studies as well. One of the popular methods of analysis for prediction is logistic regression. If you have a single binary dependent variable but several independent variables, logistic regression is the best possible technique to use. The dependent variable differs between linear and logistic regressions. For continuously varying variables, linear regression is indeed an effective method.

**5.2 K-Nearest Neighbour (KNN):**

KNN uses supervised learning, which implies, prior to generating projections, the labelling of the data is set. Regression as well as clustering are additional advantages for it. K denotes an arithmetical number for the neighbours which are adjacent to each other. A training program phase is absent from the KNN algorithm. The k-nearest neighbour metric is used to formulate predictions.

**5.3 Support Vector Machine:**

The svm is a ML paradigm which also is incredibly well known. This optimizer 's goal is to discover an N-dimensional hyperplane that categorises the data. Our main intention of svm approach is to locate plane with maximum margin. Feature values determine how the N-dimensional diversity changes. It is possible to compare two important characteristics easily. Greater forecast accuracy is attained by increasing the margin.

**5.4 Random and Rotation Forest:**

A form of ensemble learning called random forest may be applied to both classification and prediction. Undoubtedly, a random forest is made up over several decision trees. Consequently, using a random forest instead of a decision tree in any circumstances is more beneficial. Creating a classification model using attribute extraction is the rotation forest algorithm's main goal. K distinct subgroups are picked at random from the set of attributes. It strives to generate classifiers that are immense and accurate.

**5.5 Decision Tree:**

By depicting a tree structure, it attempts to provide solutions to issues. Decision nodes allude to features, meanwhile leaf nodes correspond to results. According to the decision tree mechanism, feature values are qualitative. The best possible attribute must always be handpicked at the

commencement of this task, and it must be mounted at the top of the tree representation before the tree is partitioned.

After examining the aforementioned approaches, we got to the conclusion that using the Support Vector Machine tactic yields a more reliable findings, assisting in the early detection of cancer.

A picture containing diagram

Description automatically generated

**Fig. 2:** Advantages of incorporating data science to this approach

# **6. Results and discussion:**

Once the Wisconsin Diagnostic set for breast cancer has been exposed to machine learning algorithms. Confusion Matrix, Accuracy, Precision, Recall, and F1 Score, were considered as performance indices to evaluate and contrast the models and choose the strongest algorithm for screening breast cancer.

**6.1 Confusion matrix:**

A Confusion Matrix is generated for grading classification tasks when the output may use two or more different types of classes. A confusion matrix is a table with the entries "Actual" and "Predicted"

Table

Description automatically generated

**Fig. 3:** Confusion matrix

*Formulas:*

A picture containing text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

After reading dataset, visualized using bar graph

Chart, bar chart

Description automatically generated

**Fig. 4:** WISCONSIN DATA

So, there are 357 benign cases and 212 malignant cases based on taken data.

We considered 11 values:

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

**Fig. 5:** Support Vector classifier

Table

Description automatically generated

**Fig. 6:** Support Vector classifier Stats

**Table

Description automatically generated**

**Fig. 7:** Accuracy %

**Table

Description automatically generated**

**Fig. 8:** ROC Curve Area

**Chart, line chart, scatter chart

Description automatically generated**

**Fig. 9:** Roc Curve

**Chart, bar chart

Description automatically generated**

**Fig. 10:** AccuracyBar plot

**Chart, line chart

Description automatically generated**

**Fig. 11:** AccuracyLine plot

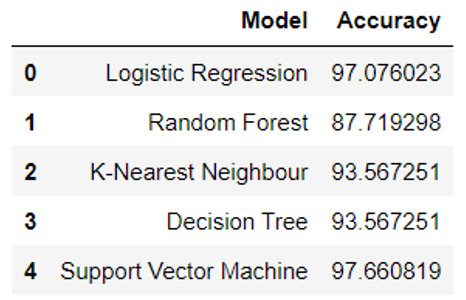
**Chart, box and whisker chart

Description automatically generated**

**Fig. 12:** AccuracyBox plot

There is a collection of five alternatives to carry out this project, which are described above. We encountered which approach produces the most accurate results after testing each.These respective algorithms' effectiveness rankings were created in this proposal: LR, K-N-N, SVM, random forest, DT, and NB classification. Every model was applied to three differentiated datasets with unique feature sets.

The very first dataset had all self-reliant features; the second encapsulated features that were intimately connected; and the third comprised features which were little associated.To appraise the exactness and reliability of each machine learning methodology, accuracy findings from three separate datasets were utilised differently for each approach. Compared to the other algorithms, SVM delivered outcomes with higher accuracy. SVM's biggest benefit is how fastly it can be trained. The SVM model assists in complicated algorithms and tends to produce more accurate results. When dealing with unbalanced data, a respectable achievement was shown (97.66% accuracy).



**Fig. 13:** Statistics report

Chart, treemap chart

Description automatically generated

**Fig. 14:** Heat Map

# **7. Conclusion and future work:**

With the help of this Python project, we understood how to design a breast cancer malignancy predictor that use the Wisconsin datasets and managed to produce visualizations and outcomes for the same. A large dataset has been determined to offer improvement in accuracy. Prediction systems are produced by determining the right algorithms and a solid training dataset. When a woman has been confirmed to have breast cancer, these algorithms can assist with determining the best treatment strategy. Regardless of the stage of a victim's breast cancer, there are a wide range of therapies available; data extraction and machine learning is indeed a big help in selecting the course of therapy to be taken by extracting useful information from such appropriate databases. There is a collection of five alternatives to carry out this project, SVM delivered outcomes with higher accuracy. SVM's biggest benefit is how fastly it can be trained. The SVM model assists in complicated algorithms and tends to produce more accurate results. When dealing with unbalanced data, a respectable achievement was shown (97.66% accuracy).

Future updates to the system might include more strengths including the ability to propose particular medications or therapies based on the degree of the cancer. Doctors can more accurately identify and handle the condition with the aid of this prediction and suggestion system.

# **8. References:**

[1] DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. CA Cancer J Clin. learning methods.” 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (2018):1-3.2011;61(6):409-418. doi:10.3322/caac.20134.

[2] Wang H., Yoon W.S. Breast cancer prediction using data mining method; Proceedings of the 2015 Industrial and Systems Engineering Research Conference; Nashville, TN, USA. 30 May–2 June 2015.

[3] Nithya R., Santhi B. Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *Int. J. Comput. Appl.*2011;28:0975–8887. doi: 10.5120/3391-4707.

[4] M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath, M. A. Bamiah, “Early detection of breast cancer using machine learning techniques,” Journal of Telecommunication, Electronic and Computer Engineering, vol. 10, no. 3-2, pp. 21-27, 2018.

[5] Muhammet Fatih Aslam, YunusCelik, KadirSabanci, AkifDurdu, “Breast cancer diagnosis by different machine learning method using blood analysis data,” International Journal of Intelligent System and Applications in Engineering, vol. 6, no. 4, pp. 289-293, 2018

[6] Anusha bharat, Pooja N, R Anishka Reddy, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” IEEE 3rd International Conference on Circuits, Control, Communication and Computing, pp. 1-4, 2018.

[7] Ebru Aydindag Bayrak, Pinar Kirci, TolgaEnsari,“Comparison of machine learning methods for breast cancer diagnosis.2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), pp. 1-3,2019.

[8] Shwetha K, Spoorthi M, Sindhu S S, Chaithra D, “Breast cancer detection using deep learning technique,” International Journal of Engineering Research & Technology, vol. 6, no. 13, pp. 1-4, 2018.

[9] Ch. Shravya, K. Pravalika, ShaikSubhani, “Prediction of breast cancer using supervised machine learning techniques,” International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 6, pp. 1106-1110, 2019.

[10] Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu, ‘‘A review of breast cancer detection in medical images,’’ in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2018, pp. 1–4.

[11] Reddy, V. Anji, and Badal Soni. "Breast Cancer Identification and Diagnosis Techniques." Machine Learning for Intelligent Decision Science. Springer, Singapore, 2020. 49-70.

[12] S. Singh, D. Srivastava and S. Agarwal, "GLCM and its application in pattern recognition," 2017 5th International Symposium on Computational and Business Intelligence (ISCBI), Dubai, 2017, pp. 20-25, doi: 10.1109/ISCBI.2017.8053537.

[13] C. Yanyun, Q. Jianlin, G. Xiang, C. Jianping, J. Dan and C. Li, "Advances in Research of Fuzzy C-Means Clustering Algorithm," 2011 International Conference on Network Computing and Information Security, Guilin, 2011, pp. 28-31, doi: 10.1109/NCIS.2011.104.