ISSN 2515-8260 Volume 07, Issue 09, 2020

# SURVEY ON VARIOUS PREDICTION MODELS FOR SURVIVAL OF BREAST CANCER PATIENTS USING WARM BOOT RANDOM FOREST CLASSIFIER

[1]Vibinchandar .S, [2]Dr. Krishnapriya .V

*[1]Research Scholar, Dept. of Computer Science, Sri Ramakrishna College of Arts and Science, Coimbatore, India.*
*[2]Professor & Head, Dept. of Computer Science with Cognitive Systems, School of Computing, Sri Ramakrishna College of Arts and Science, Coimbatore, India.*

***ABSTRACT: - The rapid growth of genomics and proteomics in science has led to the exponential development of information that requires a complex computational analysis to find details. Review of statistical science or bioinformatics using knowledge mining centres using bioinformatics to resolve a range of certifiable problems in the field of medical services. Breast cancer malignant growth is the second most deadly form of disease that causes a woman to die. Numerous experts have led to the early detection, visualisation and improved management of malignancy in the breast cancer over the last 20 years, contributing to a reduction in the rate of death. However the problem of malignancy in the breast cancer remains concerning and requires further study in the territory of the development of locations and forecasts other than treatment methods. This article explore the present situation with the technique of estimating breast cancer disease status, which includes the study on breast cancer malignancy, breast cancer, the prediction of the risk of malignant growth, and the prediction of survival for breast cancer disease.***

***Keywords: Breast Cancer (BC), Breast Cancer Survival Prediction, Breast Cancer Risk Prediction, Random Forest, Decision Tree, AdaBoost.***

## 1. INTRODUCTION

Information mining is the way toward breaking down colossal measure of information so as to extricate patterns or examples from the information. These patterns or examples are shroudedin something else. They can be deciphered to find significant information or business knowledge (BI). The information mining is effectively utilized in the territory of bioinformatics. Information mining procedures are being created for taking care of natural issues. The organic information consists of enormous datasets that should be attentively prepared so as to derive valuable expertise. Information mining is the part of software engineering that can be utilized with bioinformatics for examining natural information, for example, protein structure, DNA arrangements, quality grouping, etc. The order, grouping and other information mining techniques can be rightly blended with bioinformatics to use comprehension of natural cycles. The exploration regions in bioinformatics where information mining can be utilized through incorporate arrangement, numerous succession sustenance, quality discovering, protein space examination, design ID, genomic investigation, and theme finding.

ISSN 2515-8260 Volume 07, Issue 09, 2020

As indicated by WHO, breast cancer disease is the second deadly malignancy which should be attendedclosely in reality. It has 16.1% of all disease cases on the planet. It is truly disturbing that ladies are gravely affected with breast cancer malignant growth around the globe. The creative therapy endurance pace diagnoses more breast cancer malignancy among women. Luckily, the mechanical and medical advancements mitigate the malignancy. In any case, the breast cancer malignant growth is an issue to be scrutinized effectively. In this exploration, the spotlight is to examine information digging strategies for bioinformatics to achieve productive visualization strategy with higher precision. This paper tosses light into the audit of information mining procedures utilized in breast cancer malignant growth forecast, hazard evaluation, and growth analysis and disease endurance expectation.

AI procedures are generally utilized in medication for diagnosing BC. A few AI strategies have been presented in different examinations. Montazeri, L., et al. have looked at exhibitions of Naive Bayes, Trees Random Forest, k-Nearest Neighbor, AdaBoost, Support Vector Machine, RBF Network and Multilayer Perceptron AI strategies. Chao et al have utilized help vector machine for calculated/Calculating relapse and a C5.0 choice tree model to anticipate BC endurance. The rightful blend of these strategies may possibly defeat issues like collinearity, heteroskadacity, complex communications among factors.The higher request associations between indicators will intensify the probable outcomes. While there are a few investigations demonstrating the lower blunder rates and the higher precision in grouping issues for information mining strategies and the customary techniques (LDA and LR), there can be discovered examinations that shows this greatness isn't the situation for all informational indexes. There is an irregularity over the consequences of different investigations with regard to the order exactness of information mining strategies. Yet, the contrasted results showed that the conventional techniques are less PC requesting.

## 2. LITERATURE REVIEW

Tapak, L, et al., (2019) compared the presentation of six AI procedures two conventional techniques for the forecast of BC endurance and metastasis. Authors utilized a dataset that incorporate the records of 550 breast cancer malignant growth patients. Guileless Bayes (NB), Random Forest (RF), AdaBoost, Support Vector Machine (SVM), Least-square SVM (LSSVM) and Adabag, Logistic Regression (LR) and Linear Discriminant Analysis were utilized for the forecast of breast cancer malignant growth endurance and metastasis. The exhibition of the pre-owned procedures was assessed with affectability, particularity, probability proportion and absolute exactness. Results show that out of 550 patients, 83.4% were alive and 85% didn't encounter metastasis. In expectation of endurance, the normal particularity of all methods was ≥94% and the SVM and LDA have more noteworthy affectability (73%) in contrast with different procedures. The more prominent all out precision (93%) had a place with the SVM and LDA. For metastasis expectation, the RF had the most elevated particularity (98%), the NB had most elevated affectability (36%) and the LR and LDA had the most elevated absolute precision (86%). Their discoveries demonstrated that the SVM outflanked other AI techniques in forecast of endurance of the patients regarding a few measures. In any case, the LDA strategy as a traditional technique demonstrated comparative presentation.

Cui, Z., et al., (2015) introduced a novel system to post-measure any ATM classifier to separate an ideal noteworthy arrangement that can change an offered contribution to an ideal class with a base expense. Specifically, they demonstrate the NP hardness of the ideal activity extraction issue for ATMs and plan this issue in a whole number direct programming definition which can be productively tackled by existing bundles. They likewise observationally show the

adequacy of the proposed structure by directing extensive analyses on testing genuine world datasets.

Sinha, N. K., et al., (2020) built up an electronic determination framework for which we have done the near investigation of the managed AI classifiers to become acquainted with which classifier is giving the best exactness. For that they have taken dataset from the Wisconsin breast cancer malignancy information base (WBCD) which is the benchmark data set for contrasting the outcomes through various calculations. In which they utilized after arrangement strategies of AI like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Adaboost Classifier and XGboost Classifier for the order of considerate and harmful tumor in which the machine is found out from the past information and can anticipate the classification of new information.

To store informational index that is enormous in scale, the word Big Data is utilized. Data is an undeniably expanding idea and wide records incorporates sorted out, unstructured and semester documents. Enormous information is generally used in the examination of maladies today in medical services. The kind of malignancy recognized in a lady is breast cancer disease and this comes in just short of the leader as the essential driver of a lady's end. On the off chance that distinguishes this illness in beginning stages there will be more possibilities for recuperating. Srilakshmi et al., (2020) utilized strategy 1 as CNNMDRP calculation for better expectation of the illness with precision and technique 2 as PC supported analysis (CAD) for hazard forecast. Both the strategies concoct information digging procedures for arrangement of informational collections.

Ganggayah, M. D., et al., (2019) examination utilized AI procedures to assemble models for recognizing and envisioning huge prognostic pointers of breast cancer malignancy endurance rate. An enormous clinic based breast cancer disease dataset recovered from the University Malaya Medical Center, Kuala Lumpur, Malaysia (n = 8066) with finding data somewhere in the range of 1993 and 2016 was utilized in this examination. The dataset contained 23 indicator factors and one ward variable, which alluded to the endurance status of the patients (alive or dead). In deciding the critical prognostic variables of breast cancer malignant growth endurance rate, expectation models were assembled utilizing choice tree, arbitrary woods, neural organizations, outrageous lift, strategic relapse, and backing vector machine. Next, the dataset was bunched dependent on the receptor status of breast cancer malignant growth patients recognized through immunohistochemistry to perform progressed displaying utilizing arbitrary woodland. Along these lines, the significant factors were positioned by means of variable determination techniques in irregular timberland. At last, choice trees were constructed and approval was performed utilizing endurance examination. Results regarding both model precision and alignment measure, all calculations created close results, with the most minimal got from choice tree (exactness = 79.8%) and the most noteworthy from arbitrary timberland (precision = 82.7%).

The significant factors recognized in this investigation were disease stage characterization, tumor size, number of all out axillary lymph hubs eliminated, number of positive lymph hubs, kinds of essential treatment, and techniques for analysis.

New strategies for time-to-occasion expectation are proposed by expanding the Cox relative perils model with neural organizations. Expanding on procedure from settled case-control examines, Kvamme, H., et al., (2019) proposed a misfortune work that scales well to enormous informational indexes and empowers fitting of both corresponding and non-relative augmentations of the Cox model. Through recreation contemplates, the proposed misfortune work is checked to be a decent estimation for the Cox incomplete log-probability. The proposed philosophy is contrasted with existing approaches on certifiable informational collections and is

discovered to be exceptionally serious, regularly yielding the best exhibition as far as Brier score and binomial log-probability.

Dhanya, R., et al., (2020) utilized the current troupe strategies alongside a mix of managed AI calculations to build up another model for breast cancer malignant growth expectation. They likewise utilized element choice strategies to upgrade the presentation of the group model. For this reason, AI calculations like Support Vector Machines, Naive Bayes, K-Nearest Neighbors, Logistics Regression and highlight determination procedures like Variance limit and f-test have been mulled over. To accomplish higher exactness for the troupe model, stowing, boosting and stacking strategies are utilized.

A few gathering malignant growth survivability prescient models are introduced and tried dependent on three variations of AdaBoost calculation. In the models Vincent, E., et al., (2017) utilized Random Forest, Radial Basis Function Network and Neural Network calculations as base students while AdaBoostM1, Real AdaBoost and MultiBoostAB were utilized as outfit procedures and ten different classifiers as independent models. There has been significant examination in gathering displaying in insights, medication, innovation and man-made brainpower over the most recent thirty years. This may be a direct result of the adequacy and dependability of the procedure in clinical finding and occurrence expectations contrast and the independent classifiers. Creators utilized Wisconsin breast cancer disease dataset in preparing and testing the models. The exhibitions of the outfit and independent models were assessed utilizing Accuracy, RMSE and disarray lattice prescient boundaries.

The outcome shows that notwithstanding the unpredictability of the outfit models and the necessary preparing time, the models didn't beat a large portion of the independent classifiers.

Table 1 shows the list of techniques used by various authors for their research work which includes classification and prediction.

**Table 1: Reviewed List of Techniques Used By Authors For Prediction And Analysis**

| S.NO | AUTHORS | USED TECHNIQUES |
|---|---|---|
| 1 | Sinha, N. K., et al., (2020) | SVM, KNN, RF AND ADABOOST |
| 2 | Srilakshmi et al., (2020) | CAD |
| 3 | Dhanya, R., et al., (2020) | SVM AND KNN |
| 4 | Tapak, L, et al., (2019) | SVM and LDA |
| 5 | Ganggayah, M. D., et al., (2019) | RF AND DECISION TREE |
| 6 | Vincent, E., et al., (2017) | ADABOOST |
| 7 | Cui, Z., et al., (2015) | SVM |

**Features Comparison**

| SNO | TITLE | AUTHOR | YEAR | ALGORITHM | FEATURE SELECTION METHODS | ENSEMBLE METHODS | RECORDS | FEATURES | ACCURACY |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Prediction of Breast Cancer Survivability using | Vincent F. Adegoke / Daqing Chen, | 2017 | AdaM1 + RF, MBAB + RF | - | Adaboost M1, Real Adaboost, Multiboost AB | 683 | 10 | 97% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Ensemble Algorithms | EbadBanissi / SafiaBarikzai | | | | | | | |
| 2 | Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm | Zeinab Arabasadi ,RoohallahAlizadehsani, Mohamad Roshanzamir, Hossein Moosaei , Ali Asghar Yarifard | 2017 | Genetic + NN | Gini Index, Weight by SVM, Information Gain, Principle Component Analysis | - | 303 | 54 | 93.85% |
| 3 | Comparison of skin disease prediction by feature selection using ensemble data mining techniques | Anurag Kumar Vermaa, Saurabh Palb, Surjeet Kumarb | 2019 | Radius Neighbour Classifier with Gradient Boosting | Feature Importance method | Bagging, Adaboost, Gradient Boosting | 366 | 15 | 99.68% |
| 4 | Developing A Web based System for Breast Cancer Prediction using XGboost Classifier | Nayan Kumar Sinha, Menuka Khulal, Manzil Gurung, Arvind Lal | 2020 | XGBoost | - | - | 569 | 30 | 98% |
| 5 | A Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm For Breast | Ahmed Abdullah Farid1, Gamal Ibrahim Selim, and Hatem A. Khater | 2020 | JRIP - SVM | Information Gain,Gain Ratio - Filter, Wrapper - Improved GO, Embedded C4.5 | Adaboost–Stacking | 569 | 10 | 98.25% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Cancer Detection | | | | | | | | |
| 6 | F-test feature selection in Stacking ensemble model for breast cancer prediction | Dhanya R, Irene Rose Paul, Sai Sindhu Akula, Madhumathi Sivakumar, Jyothisha J Nair | 2020 | F-Test - KNN, Variance Threshold - SVM, LR, KNN, DT | F-Test, Variance Thresholding | Bagging, Boosting, Stacking | (Wisconsin) 699 | 11 | F-Test - KNN - 97.86%, Variance Threshold - SVM, LR, KNN, DT 96.49% |
| | | | | F-Test - Naïve, SVM, Variance Threshold - LR, KNN, NB, DT, MLP | | | (WDBC) 569 | 32 | F-Test - Naïve, SVM, Variance Threshold - LR, KNN, NB, DT, MLP - 100% |
| | | | | F-Test - SVM, Variance Threshold - KNN | | | (Micro Array) 133 | - | F-Test - SVM - 97.14%, Variance Threshold - KNN - 85% |

## 3. DATA MINING TECHNIQUES USED FOR BREAST CANCER RESEARCH

Following are the various kinds of techniques utilized for characterization and forecast of information for different issues.

- Classification,
- K-Nearest Neighbor,
- Decision Tree (DT),
- Support Vector Machine (SVM),
- Neural Network (NN),
- Bayesian Methods,
- Regression,
- Clustering,
- Partitioned Clustering,
- Hierarchical Clustering,
- Density based Clustering,
- Association rule mining,
- Random Forest,
- AdaBoost,
- Adabag,
- RandomTree.

ISSN 2515-8260 Volume 07, Issue 09, 2020

A portion of the fundamental techniques which are utilized much of the time for forecast and investigation measure are clarified underneath.

- **Support vector machine (SVM)**

  SVM is an AI strategy that has been broadly utilized in relapse and order issues. In this strategy, the order condition for two gatherings (for endurance are dead and alive status and for metastasis are yes and no) in light of highlight space (age, Grade, stage, ER, PR, HER2, Pathological sort and surgical methodology).

- **Neural networks**

  This examination applied the multi-layer-perceptron based fake neural organizations (MLP-ANN); a feed-forward and managed learning method made out of information, covered up and yield layers. The information esteems (23 indicator factors) were introduced to the perceptron and if the anticipated yield was like the ideal yield, the exhibition was viewed as palatable and no weight was changed, depicting uncommon exactness. The neural organization was chosen in this investigation to perform model assessment as it functioned admirably when information instability was high. The feed forward neural organization was chosen to evade intricacies from criticism networks that present circle in the organization.

- **AdaBoost (AD)**

  AdaBoost has a place with the AI methods family and can be considered as a meta-calculation that improves the exhibition along with other learning strategies. In a grouping issue, AdaBoost centers on the successively applying powerless classifiers. Thusly, the calculation is over and over applied on the adjusted information. For instance, let Y { 1, 1} + be the yield variable with the −1 for death and +1 for alive statuses. Additionally, let X be a vector of potential danger factors (here age, grade, and so forth.). Along these lines, any classifier, state G X( ) , predicts the status of the patients in { 1, 1} + set and the blunder rate on the preparation set and the normal mistake rate on the test set.

- **Decision tree**

  This examination utilized the part bundle, which actualized the order and relapse tree (CART) capacity to construct DT for expectation and assessment of the 'all information'. This capacity prepared the information and yielded the model precision and an ideal tree as the final product. The DT contained a root hub at the head of the tree to mean the most significant variable, trailed by choice hubs and terminal hubs with rates of grouping.

- **Adabag**

  Stowing is an AI procedure that works dependent on consolidating bootstrapping and amassing. In this strategy, the quantity of B bootstrap tests is chosen from the preparing set, say Tb (b = 1, 2, . . . , B). By bootstrapping, the loud perceptions are diminished and even wiped out from some of Tbs. In this manner, these sets will furnish the classifiers with a superior conduct contrasted and the first set. This makes packing method a valuable device to manufacture a superior classifier at the presence of boisterous perceptions in the preparation set. At long last, better outcomes can be accomplished by the outfit of these B classifiers contrasted and the single classifiers.

- **Random forest (RF)**

  RF amasses order and relapse trees. The dataset is examined by substitution to frame the trees in RF. Irregular arrangements of indicators are chosen at the hubs which are made by the trees. It is conceivable to locate the most significant indicators utilizing

ISSN 2515-8260 Volume 07, Issue 09, 2020

mean abatement Gini and mean lessening exactness. The significant factors arrange the double result so the forecast is completed with the most elevated exactness.

- **Naïve Bayesian (NB) arrangement**

The Naïve Bayes grouping model works dependent on the celebrated Bayes' hypothesis following an unmistakable, basic, and quick classifier. Utilizing the Bayes rule, the earlier likelihood of having a place with each class can be learnt and assessed utilizing the preparation information with overlooking the minimal probabilities dependent on the contingent likelihood of every factor $X_j$ (age, Grade, stage, ER, PR, HER2, Pathological sort and Surgical methodology) given the class mark C (for endurance c is one of in any condition status and for metastasis c is yes or no).

## 4. PERFORMANCE ANALYSIS PARAMETERS

For analysis process Classification Accuracy, Classification Time, Precision, Recall and F-measures are used. For prediction process, to find out the things which are predicted or not, True Positive, False Positive, True Negative and False Negative of Confusion matrix are used. Figure 1, shows the prediction type of confusion matrix. For medical datasets sensitivity and specificity are need to be included in the analysis part. Accuracy will be calculated by using the below formula.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

| | Predicted **No** | Predicted **Yes** |
|---|---|---|
| Actual **No** | **TP** True Positive | **FN** False Negative |
| Actual **Yes** | **FP** False Positive | **TN** True Negative |

**Figure 1: - Confusion Matrix Prediction Type**

## 5. CONCLUSION

As breast cancer disease cases are as yet disturbing on the planet, this paper tosses light into the current scholarly considering breast cancer malignant growth forecast techniques and their points of interest and impediments. This study can help in making more effective and reliable breast cancer disease prediction and diagnostic system which will contribute towards developing better prediction model by reducing overall cost, time and mortality rate.

From this review it is been observed that the various classification techniques yields the highest classification accuracies when used with most predictive variables. It also greatly reduces the cost of treatment and improves the quality of life by predicting breast cancer at early stage of development.

It is important to move on to the next level that is to improve the performance of the prediction accuracy because to set up and train a classifier by reducing error rates, training time and increasing the accuracy of the status classification. The future research aiming to develop a new breast cancer prediction model using Levenberg-Marquardt algorithm in GRU for reduce the neuron tinning time with less Means Square rate. The future work will focus on exploring the

breast cancer malignant growth patients utilizing AdaBoost and Random Forest as prediction models.

## REFERENCES

[1] Cui, Z., Chen, W., He, Y., & Chen, Y. (2015). Optimal action extraction for random forests and boosted trees. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 179-188).

[2] Adegoke, V. F., Chen, D., Banissi, E., & Barikzai, S. (2017, October). Prediction of breast cancer survivability using ensemble algorithms. In 2017 International Conference on Smart Systems and Technologies (SST) (pp. 223-231). IEEE.

[3] Verma, A. K., Pal, S., & Kumar, S. (2019). Comparison of skin disease prediction by feature selection using ensemble data mining techniques. Informatics in Medicine Unlocked, 16, 100202.

[4] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access, 8, 150360-150376.

[5] Sinha, N. K., Khulal, M., Gurung, M., & Lal, A. Developing A Web based System for Breast Cancer Prediction using XGboost Classifier.

[6] Dhanya, R., Paul, I. R., Akula, S. S., Sivakumar, M., & Nair, J. J. (2020). F-test feature selection in Stacking ensemble model for breast cancer prediction. Procedia Computer Science, 171, 1561-1570.

[7] Thawkar, S., & Ingolikar, R. (2018). Classification of Masses in Digital Mammograms Using Firefly based Optimization. International Journal of Image, Graphics & Signal Processing, 10(2).

[8] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

[9] Ramani, R. (2016). Hybrid optimized framework for classification of breast cancer.

[10] Ezhilraman, S. V., Srinivasan, S., & Suseendran, G. Breast Cancer Detection using Gradient Boost Ensemble Decision Tree Classifier.

[11] U. Srilakshmi, V. Shravya, Amar, "Early Prediction of Breast Cancer Using Big Data Analytics and Data Mining Techniques", International Journal of Grid and Distributed Computing, Volume 13, Issue 1, (2020), pp. 786-801.

[12] Aavula, R., & Bhramaramba, R. (2018). A survey on latest academic thinking of breast cancer prognosis. Int J Appl Eng Res, 13, 5207-5215.

[13] Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2019). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clinical Epidemiology and Global Health, 7(3), 293-299.

[14] Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. BMC medical informatics and decision making, 19(1), 48.

[15] Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-event prediction with neural networks and Cox regression. Journal of machine learning research, 20(129), 1-30.

[16] Daoud, M., & Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. Artificial intelligence in medicine, 97, 204-214.

[17] Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., ... & Virtanen, T. (2017, November). DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*.

ISSN 2515-8260 Volume 07, Issue 09, 2020

[18] Es, H. A., Montazeri, L., Aref, A. R., Vosough, M., & Baharvand, H. (2018). Personalized cancer medicine: an organoid approach. *Trends in biotechnology*, *36*(4), 358-371.