

Unraveling the Multifaceted Nature of Breast Cancer: Harnessing the Power of Machine Learning and Statistical Models for In-depth Investigation, Prognosis, and Treatment Optimization

Fatma Dafallah Mohamed Alhasan¹, Nirmla Sharma^{2 *}, Olfa Ben Said³, Hanaa Abu-Zinadah⁴, Fahad Mohammed Alhomayani⁵, Saadia H. Abdalaha Hamza⁶ and Azhari A. Elhag⁷

1 Computer Science Department, King Khalid University Abha, Saudi Arabia, falhasan@kku.edu.sa,

2 Computer Science Department, King Khalid University Abha, Saudi Arabia, nprasad@kku.edu.sa

3 Computer Science Department, King Khalid University Abha, Saudi Arabia, omessai@kku.edu.sa

4 - Department of Statistics, College of Science, University of Jeddah, Jeddah 21959, Saudi Arabia; hhabuznadah@uj.edu.sa

5- College of Computers and Information Technology, Taif University, Saudi Arabia; fahad@tu.edu.sa

5- Applied College, Taif University, Saudi Arabia

6-Department of Computer Science College of Science and Humanities in Al-Sulail, Prince Sattam Bin Abdulaziz University, Saudi Arabia; s.hamza@psau.edu.sa

7-Department of Mathematics and Statistics, College of Science, Taif University, P. O Box 11099, Taif 21944, Saudi Arabia; a.alhag@tu.edu.sa

Abstract

The most prevalent form of cancer in all of human history. Women typically have a shorter lifespan than men. The death toll from breast cancer is higher than that of any other type of cancer. After reaching puberty, women all around the world are at risk for developing breast cancer, however, the disease is more prevalent in elderly individuals. The death rate from breast cancer did not significantly change from the 1930s through the 1970s. In the 1980s, countries that were healthcare conscious and had early detection systems, as well as a variety of treatment modalities to eliminate invasive diseases, began to see an increase in their overall survival rates. This article employed decision tree models and time series models to anticipate breast cancer using time series data from 1930 to 2019. The breast cancer cases that occurred in the United States of America served as the basis for this study. The most precise prediction models will be produced as a result of this data analysis. In order to evaluate how well the predictions were made, we determined how accurate they were by calculating the symmetric mean absolute percentage errors (sMAPE), measured mean absolute percentage errors (MASE), and mean absolute percentage errors (MAPE).

Keywords: Decision tree, Cluster analysis, Survival rates, Collections of patterns, Breast cancer, Healthcare.

1. Introduction

This study [1] gives a test process based on a machine learning (ML) system for the early diagnosis of breast cancer sickness. As a result, we finished presenting six algorithms that are inadequate for predicting cancer diseases and assessed their performance using the confusion matrix, a statistical method. The experimental findings show that the Nave Bayes and Random Forest classifiers perform better in cancer datasets than the other classifiers. In this study, six ML techniques were applied to forecast cancer illness based on a few characteristics. In the study [2] Metastatic breast cancer is the leading cause of death in women. Mostly as a result of a delayed diagnosis. According to our extensive data study, blood profile data can be utilized to detect early breast cancer metastasis without the usage of machine learning. The ensemble and logistic regression techniques were outperformed by the 83%-accurate Decision Tree (DT) classifier. Serum biomarkers, vitamin D3 and B12 levels, liver and kidney function tests, and other testing can improve the existing poor accuracy rate. If used in several medical specialties, our web interface would confirm our theory and generate a cancer-specific database with more attributes. With just one blood drop, we can create a precision medicine platform using statistical approaches and machine learning algorithms. This will increase cancer patient survival while decreasing healthcare costs. [3] compares J48, AD Tree, BF Tree, and regression trees (CART) classification accuracy using FP rate, TP rate, Recall, Precision, ROC Area, and F-measure. Rule extraction is easy using decision trees. The implementation only considers breast cancer data attribute numerical values. J48 has the highest accuracy (99%), followed by CART (96%), AD Tree (97%), and BF Tree (98%). J48 outperforms the other three algorithms for the supplied data set. This study [4] mined patient records using Bayesian Classification and the Nave Bayes Algorithm. Bayesian breast cancer classification predicted better. Prioritized EPR mining. We reviewed scientific and practical articles. Bayesian classification helps medicine. Sinology "experience base" classification and prediction models predicted breast cancer patients. These models simplify complex issues and provide accurate, easy-to-access information. 100 breast cancer cases confirmed them. Data and model validation are essential. Bayesian classification helped us classify the language using the Breast Imaging and Reporting Data System, a standard vocabulary. Our software analyzes huge breast cancer data to predict recurrences. To illustrate the Naive Bayes method's breast cancer screening and diagnosis classification/prediction capabilities. The Wisconsin Prognostic Breast Cancer dataset was used to evaluate Naive Bayesian Classification for computer-aided diagnosis. Finally, we tested the Weighted Naive Bayes Classifier for breast cancer classification/prediction. Medicine benefits from data mining. They help doctors diagnose, decide, screen, monitor, treat, and care for patients. Six distinct machine learning algorithms were studied in this research [5] for breast cancer diagnosis. When dealing with unbalanced data, the performance was competitive (98.1% accuracy). However, the dataset must be pre-processed before executing the algorithm, as it does not deal with missing values and performs better when learning from a dataset with discretized nominal values. In this study [6] We are studying the best methods for dealing with missing values in unbalanced data in order to increase the accuracy of our breast cancer classification. This will allow us to increase the number of women accurately diagnosed with the condition. The classifier algorithms J48, NB, and SMO were applied to two independent breast cancer datasets during the course of our investigation. The results indicate that using the resample filter during the preprocessing stage improves the classifier's performance. Different datasets and classifiers will be used in the following iterations of the project. During the course of this scholarly research [7], The effectiveness of multiple various deep learning-based techniques for the purpose of diagnosing breast cancer is specifically being investigated. After searching for and gathering information from various studies, 95 pieces of research were

determined to be eligible for analysis after being evaluated for quality. The Convolution Neural Network (CNN) is the most popular and accurate model for diagnosing breast cancer, and accuracy metrics are the most frequently utilized performance evaluation technique, according to the systematic review. The systematic review was conducted, which supports both of these findings. Researchers employed in this paper [8] Suction and three machine-learning algorithms distinguished malignant from benign tumors. Experiments reveal that machine learning classification algorithms affect classification performance. SVM-LDA and NN-LDA outperformed other ML classifier models in simulations. SVM-LDA is best since NN-LDA takes longer. A sophisticated breast cancer diagnosis method employs linear discriminant analysis and a support vector machine with an RBF kernel. It worked on the validation datasets. 98.82% classification accuracy, 98.41% sensitivity, 99.07% specificity, and 0.9994 area under the curve. This research [9] Breast cancer kills many women worldwide. This review explored the use of several machine learning (ML) algorithms and biosensors for early breast cancer diagnosis. Biosensors and ML identify tumors in microscopic pictures, requiring automation. ML helps computers learn. Instead of pre-programmed rules and models, it uses observed data to develop models to predict outcomes. We compared and analyzed Fuzzy support vector machine-radial basis function (ELM-RBF), For analyzing data from a variety of databases, including digitized images from fine needle aspirations, several machine learning algorithms were utilized. These algorithms included Support Vector Regression (SVR), Relevance Vector Machine (RVM), Naive Bayes, K-Nearest Neighbors (K-NN), Decision Tree (DT), Artificial Neural Network (ANN), Back-Propagation Neural Network (BPNN), and Random Forest. The research [10] found many systematic reviews of disease diagnosis and treatment trials using AI. AI has great sensitivity for diagnosing specific diseases, however SRs of AI applications in disease diagnosis and treatment, employing digestive system diseases as a test case, need to be improved. High-quality primary research may improve clinical AI deployment and promotion. This study [11] used breast cancer data to diagnose early-stage cancer. This study found that Association Rule Mining, Classification, Clustering, and Evolutionary Algorithms detect and categorize breast cancer data well. Identifying symptoms precisely also enhances detection. The classification technique improves detection accuracy, according to previous methods. This paper [12] compared supervised learning classifiers Nave Bayes, SVM RBF kernel, RBF neural networks, Decision trees J48, and Simple CART. This study seeks the best classifier. All classifiers classify WBC, WDBC, and Breast tissue. Experiments using WEKA. SVM RBF In terms of accuracy, sensitivity, specificity, and precision, the performance of Kernel is superior to that of other classifiers when applied to binary as well as multiclass datasets. Our investigation will center on both linear and nonlinear SVMs, with and without dimensionality reduction. Both approaches will be tested. This study [13] investigates breast thermogram interpretation and uses thermography to diagnose breast cancer. An artificial neural network (ANN) was suggested to classify and assess thermograms during screening.

2. Time Series

A technique for analyzing a set of data points gathered over time is called time series analysis. [14] In time series analysis, analysts are required to collect data points at regular intervals throughout a predetermined length of time. This is in contrast to the more common practice of collecting data points infrequently or randomly. [15]. Nevertheless, there is more to this kind of research than simply accumulating data over a period. Time series analysis, in contrast to other types of data analysis, is primarily concerned with presenting the changes that occur in variables over the course of time. It is distinguished from other methods of data analysis by virtue of this

one-of-a-kind approach. A more straightforward explanation would be that time is an essential component since it not only represents the results but also illustrates the development of data during the course of the research. Time series analysis, in contrast to other types of data analysis, is primarily concerned with presenting the changes that occur in variables over the course of time. It is distinguished from other methods of data analysis by virtue of this one-of-a-kind approach. A more straightforward explanation would be that time is an essential component since it not only represents the results but also illustrates the development of data during the course of the research. In addition to a predetermined hierarchy of data dependencies, it provides an extra source of data for use in the analysis. To guarantee reliability and consistency, time series analysis typically necessitates the collection of a substantial number of data points. The usage of a large data set ensures that your sample size will be representative, and that your analysis will be able to handle noisy data. Both of these things are very important. [16]. Additionally, it guarantees that any trends or patterns detected may consider seasonal variance and are not outliers. This is achieved by ensuring that any outliers are eliminated. Data from time series can be utilized to make educated guesses about the future by extrapolating from the present.[17]. There are many different approaches to researching data due to the fact that there are many different types and models.[18]In particular, the Box-Jenkins and Auto-regressive Integrated Moving Averages (ARIMA) models piqued our interest. These are univariate models that are employed in order to gain a deeper comprehension of a single time-dependent variable.[19]the practice of estimating future data points of variables, such as temperature over time. The functionality of these models is predicated on the assumption that the data is stationary. Analysts must consider and remove as much of the variability and seasonality as possible from earlier data points. Moving averages, seasonal difference operators, and auto-regressive terms are all factors that the ARIMA model takes into consideration. There is a wide range of ARIMA models to choose from. [20] It is usual practice to refer to the broad non-seasonal model as ARIMA (p, d, q): where q represents the number of terms in the moving average, p represents the number of terms in the auto-regressive model, and d represents the number of differences. A white noise model is referred to as ARIMA (0, 0, 0) because there is no AR component y_t because it does not depend on anything else y_{t-1} , so there is no difference, and there is also no MA component y_t because it does not depend on anything else e_{t-1} , so there is no difference. Consequently, there is no difference between the two. If y_t , for example, is not stationary, then the first difference of y_t that we take will cause Δy_t it to become stationary. $\Delta y_t = y_t - y_{t-1}$, (d = 1 implies one time difference).

$$y_t = c + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + \dots + \theta_p \Delta y_{t-p} + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \dots + \theta_q u_{t-q} + u_t \quad (1)$$

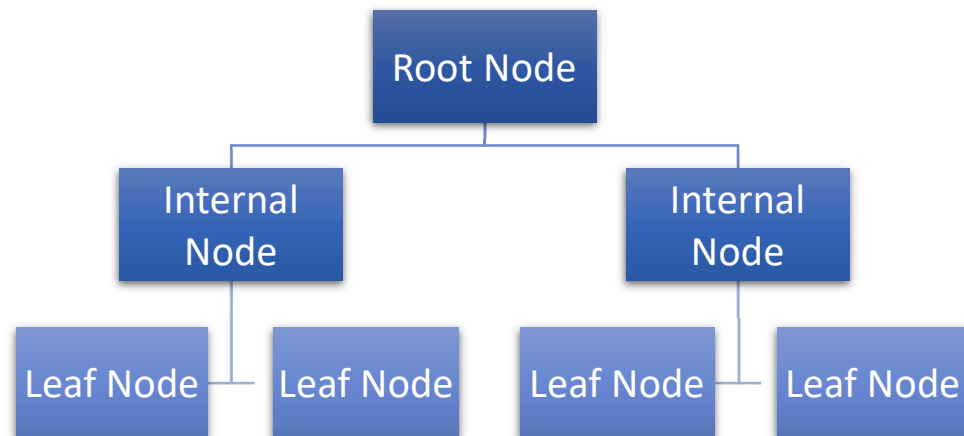
3. Decision Tree

Decision trees are a popular machine learning technique that may be applied to both classification and regression problems [21]. It is both classification and regression work can be accomplished with the help of a non-parametric supervised learning technique known as a decision tree. Decision trees can be used. It is organized in a hierarchical tree fashion, with a root node in the center and branches, internal nodes, and leaf nodes surrounding it [22].

Root Nodes are the nodes that are located at the beginning of a decision tree. It is from this node that the population begins to be segmented based on the numerous characteristics.

Internal Nodes are the nodes that are obtained once the root nodes are divided. These nodes are given the name "Decision Node."

Nodes on the Leaves The nodes on the leaves of a plant that cannot further divide are referred to as leaf nodes or terminal nodes.



4. The accuracy measurement

The basic idea behind accuracy evaluation is to use a set of criteria to compare the original objective to the anticipated one[23]

Metrics for model evaluation

$$\text{Root mean square error(RMSE)} = \sqrt{\text{MSE}} \quad (2)$$

$$\text{Mean square error (MSE)} = \frac{1}{N} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

$$\text{Root mean square error(MAPE)} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (4)$$

1. Data Analysis

Annual data on breast Cancer disease in the United State of America were obtained from 1930 to 2019. The data source is from the Breast- Cancer Statistics Center website [24]

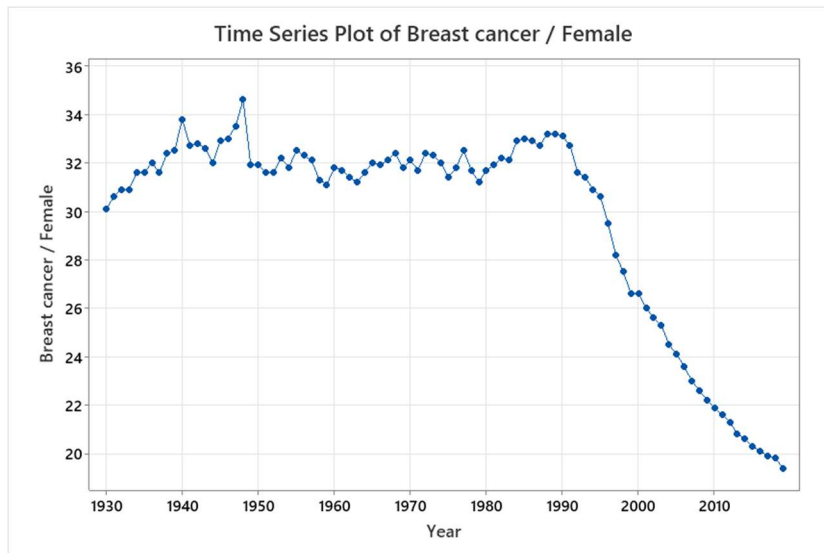


Figure.1 Breast Cancer Time Series Plot.

The graph depicts a chronological progression of breast cancer incidence from 1930 to 2019. Significantly, the data reveals a notable surge in the prevalence of breast cancer in the year 1948, with an incidence rate of 34.6 cases.

4.1 ARIMA Model for Breast Cancer

Table .1 Model Selection

Model (d = 0)	Log Likelihood	AICc	AIC	BIC
p = 1; q = 0*	-85.101	174.341	174.203	179.202
p = 1; q = 1	-84.381	175.040	174.761	182.261
p = 0; q = 2	-172.560	353.591	353.120	363.119
p = 0; q = 1	-204.865	416.009	415.730	423.230

The table displays the most effective model determined by minimizing the AICc criterion. The optimal model produces a result that is denoted as (1, 0.0).

Table 1.Final Estimates of Parameters

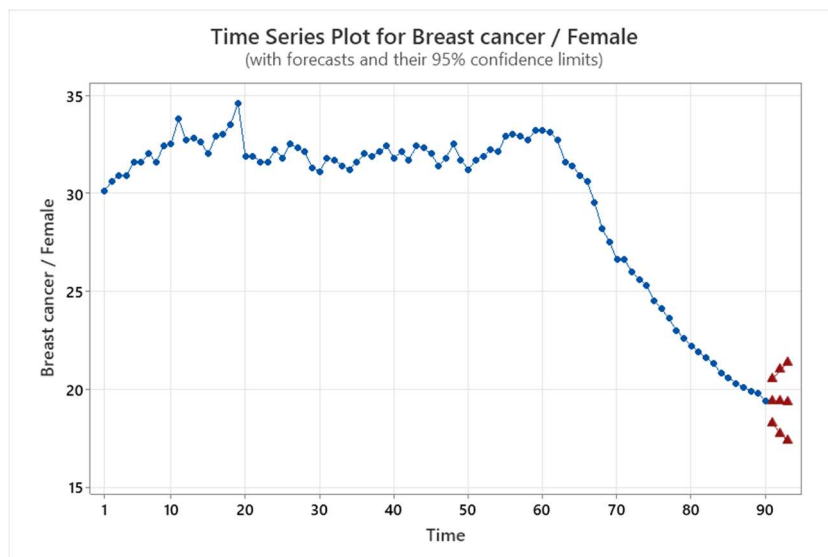
Type	Coef	SE Coef	T-Value	P-Value
AR 1	0.99904	0.00206	485.32	0.000

The p-value for the auto-regressive term is less than the significance level of 0.05, which shows statistical significance. Hence, it is recommended to keep the auto-regressive component in the model.

Table 2. Forecasts from Time Period 2019 with Best ARIMA Model for Breast Cancer / Female

Time Period	Forecast	SE Forecast	95% Limits		Actual
			Lower	Upper	
2020	19.3814	0.58837	18.2280	20.5348	
2021	19.3628	0.83168	17.7324	20.9932	
2022	19.3442	1.01810	17.3483	21.3401	

Breast Cancer in the USA decreases with time. The decrease may be caused reflect the use of a variety of treatment modalities to eliminate invasive diseases,

*Figure 1. Forecast with Best ARIMA Model for breast cancer*

4.3 Decision Tree (DT)

Table 3 Response Information

Mean	StDev	Minimum	Q1	Median	Q3	Maximum
29.7389	4.19930	19.4	28.025	31.7	32.325	34.6

As shown in Table 1. the mean of the cumulative averages is 29.7389

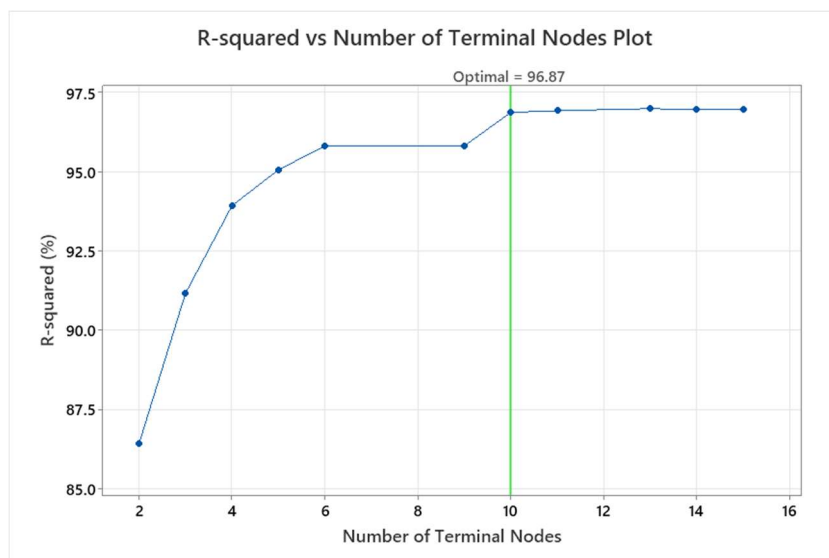


Figure3. According to the plot that illustrates the link between the R-squared value and the number of terminal nodes, the R-squared value reveals that it is 96.87% when the size of the tree is taken into consideration.

Optimal Tree Diagram

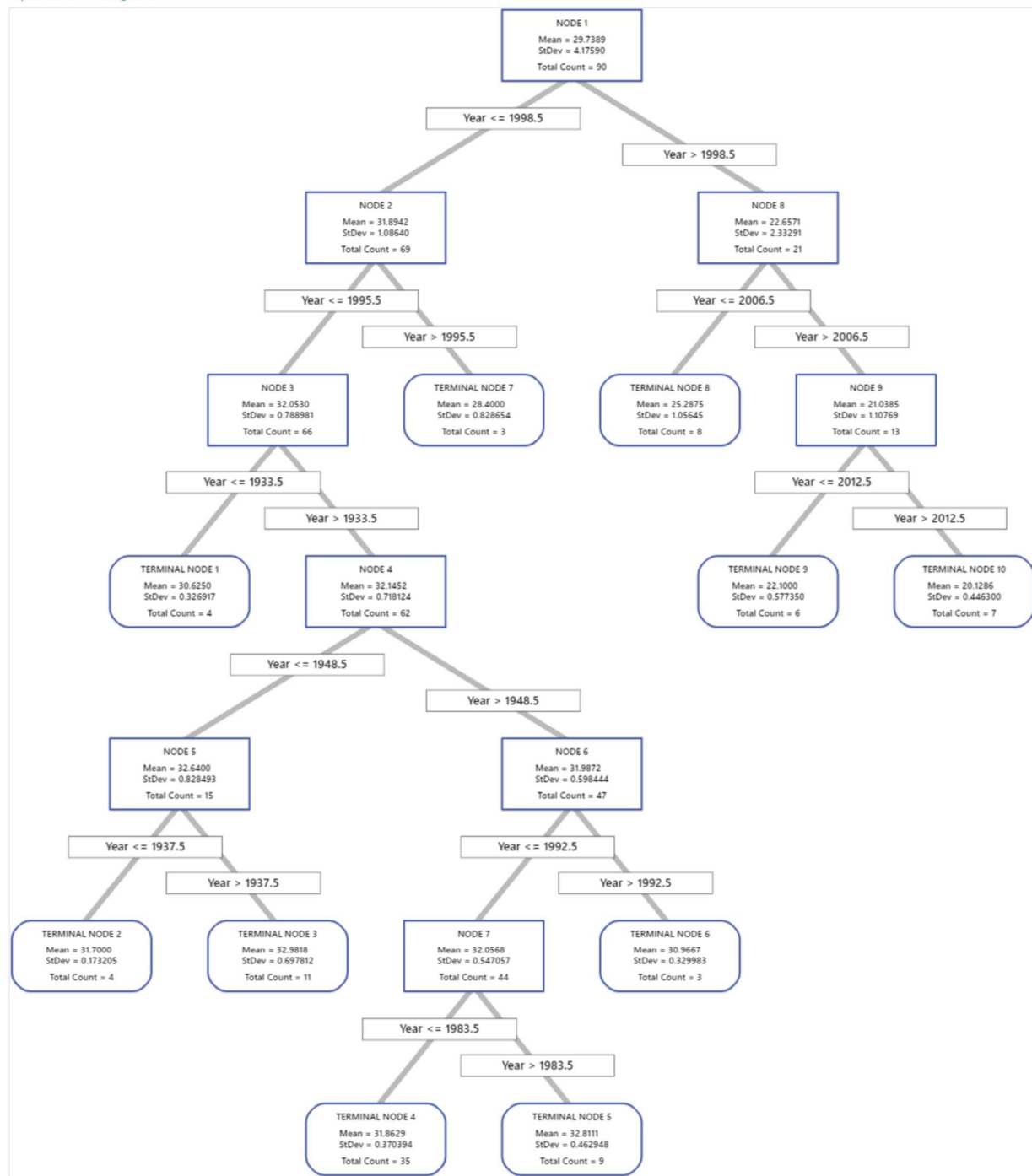


Figure 4. Optimal Tree Diagram

The subsets generated by splits are termed nodes, and the subsets that are not divided are called terminals, which equal three as illustrated in Fig.4.

Table 4. Model Summary

Statistics	Training	Test
R-squared	98.26%	96.87%
Root mean squared error (RMSE)	0.5512	0.7388
Mean squared error (MSE)	0.3038	0.5459
Mean absolute deviation (MAD)	0.4119	0.5404
Mean absolute percent error (MAPE)	0.0146	0.0193

The accuracy of each model using RMSE, MSE, MAD, MAPE, and R-square.

Table 5. Metrics calculation Results

Statistics	DT	TS
R-squared	97.57	96.90
Root mean squared error (RMSE)	0.645	0.738
Mean squared error (MSE)	0.42485	0.544644
Mean absolute percent error (MAPE)	0.01695	1.672

R-Squared, RMSE, MSE, and MAPE have been calculated using the original and forecasted data, as shown above.

6. Conclusion and Results

In this study, we used statistical and machine learning models to predict the occurrence of breast cancer in the United States of America (USA). In order to understand how well machine learning and statistical models predict the occurrence of breast cancer, our research compared the two types of models. We specifically compared the decision tree model against time series models based on statistical measurements in order to determine how well it predicted new cases of Breast Cancer disease. The outcomes, as shown in Table 5, indicate the decision tree model's better accuracy. Additionally, Figure 2 offers a visual depiction of the time series data, illuminating the current and anticipated patterns in the disease of breast cancer. This picture helps us better understand how well the used models anticipate the future. These results highlight the potential of machine learning techniques, notably the decision tree model, for accurately forecasting the disease of breast cancer. Policymakers and healthcare administrators can use these models to inform their decisions and carry out focused interventions to slow the spread of the disease. In conclusion, our study adds to the corpus of work that is being done to forecast and comprehend the dynamics of the breast cancer. Machine learning and statistical modeling are used to provide useful insights into the prognosis of breast cancer disease in the context of the USA. Compared to the research conducted by P. Hamsagayathri and P Sampath. An evaluation of the effectiveness of decision tree classifiers in classifying breast cancer. The author concludes that by a comparative analysis of classification outcomes, it is established that the Priority-based decision tree method outperforms alternative classification algorithms. This finding confirms the validates the findings presented in this study.

Such predictive algorithms can be extremely helpful in informing public health policies and supporting efficient decision-making as we continue to navigate this disease.

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP2/460/44.

REFERENCES

1. Shamrat, F.J.M., et al., An analysis on breast disease prediction using machine learning approaches. *International Journal of Scientific & Technology Research*, 2020. 9(02): p. 2450-2455.
2. Botlagunta, M., et al., Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Scientific Reports*, 2023. 13(1): p. 485.
3. Venkatesan, E.v. and T. Velmurugan, Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology*, 2015. 8(29): p. 1-8.
4. Demigha, S., Mining knowledge of the patient record: the Bayesian classification to predict and detect anomalies in breast cancer. *Electronic Journal of Knowledge Management*, 2016. 14(3): p. pp128-139-pp128-139.
5. Ak, M.F. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. in *healthcare*. 2020. MDPI.
6. Mohammed, S.A., et al. Analysis of breast cancer detection using different machine learning techniques. in *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5*. 2020. Springer.
7. Nasser, M. and U.K. Yusof, Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. *Diagnostics*, 2023. 13(1): p. 161.
8. Agarap, A.F.M. On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. in *Proceedings of the 2nd international conference on machine learning and soft computing*. 2018.
9. Amethiya, Y., et al., Comparative analysis of breast cancer detection using machine learning and biosensors. *Intelligent Medicine*, 2022. 2(2): p. 69-81.
10. Gang, Y., et al., *Intelligent Medicine. health care*. 6: p. 8.
11. Dubey, A.K., U. Gupta, and S. Jain, Breast cancer statistics and prediction methodology: a systematic review and analysis. *Asian Pacific journal of cancer prevention*, 2015. 16(10): p. 4237-4245.
12. Aruna, S., S. Rajagopalan, and L. Nandakishore, Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology*, 2011. 2(2011): p. 37-45.
13. YK Ng, L.U., FC Ng, LSJ Sim, E, Statistical analysis of healthy and malignant breast thermography. *Journal of medical engineering & technology*, 2001. 25(6): p. 253-263.
14. Xiao, H., et al., Using time series analysis to forecast the health-related quality of life of post-menopausal women with non-metastatic ER+ breast cancer: a tutorial and case study. *Research in Social and Administrative Pharmacy*, 2020. 16(8): p. 1095-1099.

15. Yasmeen, F., Predicting breast cancer incidence rates among white and black women in the United States: An application of Fts model. *Int J Statist Distrib Appl*, 2017. 3(103): p. 10.11648.
16. Chaolong, J., et al., Track irregularity time series analysis and trend forecasting. *Discrete Dynamics in Nature and Society*, 2012. 2012.
17. MARY, S., O. SO, and O. ES, Analysis of reported cases of breast cancer in Abuja, Nigeria: a time series approach. *IRE J*, 2020. 3.
18. Hyeda, A., É.S.M. da Costa, and S.C. Kowalski, The trend and direct costs of screening and chemotherapy treatment of breast cancer in the new coronavirus pandemic: total and interrupted time series study. *BMC Health Services Research*, 2022. 22(1): p. 1-8.
19. Tudor, C., A Novel Approach to Modeling and Forecasting Cancer Incidence and Mortality Rates through Web Queries and Automated Forecasting Algorithms: Evidence from Romania. *Biology*, 2022. 11(6): p. 857.
20. Wabomba, M.S., M. Mutwiri, and M. Fredrick, Modeling and forecasting Kenyan GDP using autoregressive integrated moving average (ARIMA) models. *Science Journal of Applied Mathematics and Statistics*, 2016. 4(2): p. 64-73.
21. Priyam, A., et al., Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, 2013. 3(2): p. 334-337.
22. Pandey, M. and V.K. Sharma, A decision tree algorithm pertaining to the student performance analysis and prediction. *International Journal of Computer Applications*, 2013. 61(13): p. 1-5.
23. Arroyo, J. and C. Maté, Introducing interval time series: Accuracy measures. *COMPSTAT 2006, proceedings in computational statistics*, 2006: p. 1139-1146.
24. <https://cancerstatisticscenter.cancer.org/>