A Comparative Study of Supervised Learning Models for Breast Cancer Tissue Classification

Rajendra Randa1[1], Sanjeev Gour[2]

[1]Research Scholar, Dept. of Computer Science, Medi-Caps University, Indore

[2]Research Supervisor and Asst. Professor, Dept. of Computer Science, Medi-Caps University, Indore

Abstract: breast cancer issues in women's life is increasing in the last few years leading to deaths. Machine learning technology helps to increase the chance to improve the quality of treatment and increase the process of making better plans to cure diseases via predicting complex clinical data. This paper presents early-stage breast cancer prediction using supervised learning from medical text data. Including real-time patient records to verify the prediction can improve the compatibility of models. This study found that SVM provides better accuracy than other supervised models such as LR, KNN, NB, XGB, SVM, RF, and GB. This accuracy can be improved using advanced and robust data analysis methods and feature engineering.

Keywords: machine learning, breast cancer, healthcare,

1. **Introduction:** Machine learning is an essential part of Artificial intelligence. In healthcare, AI/ ML can increase the treatment process, develop better treatment plans, analyse complex data, and find patterns faster than a human with minimal effort. deaths are increasing extremely because of cancer and all types of cancer are dangerous in healthcare areas. Still, breast cancer is extremely leading and increasing death rates in women's lives that need to be cured by better and faster treatment after predicting it from the early stage of breast cancer. This study aims to predict breast cancer more robustly and This study focuses on predicting breast cancer using supervised machine-learning algorithms such as logistic regression, decision tree, random forest, naïve bays, etc and provides a comparison between them to find compatibility to predict breast cancer effectively.

Literature review: Recent studies have explored machine learning (ML) methods to improve breast cancer detection. Random forest classification using urinary biomarkers showed high potential for early detection but was limited by sample size et.al. Alladio E. (1). Comparative analyses of ML algorithms, such as logistic regression, KNN, SVM, and random forest, highlighted the need for larger datasets and additional parameters to improve accuracy et.al. Naji M. (3). and Namade V. (4). Semi-supervised learning was found to perform comparably to supervised learning but required better validation strategies to address underfitting and overfitting et.al. Azzam N. (5).

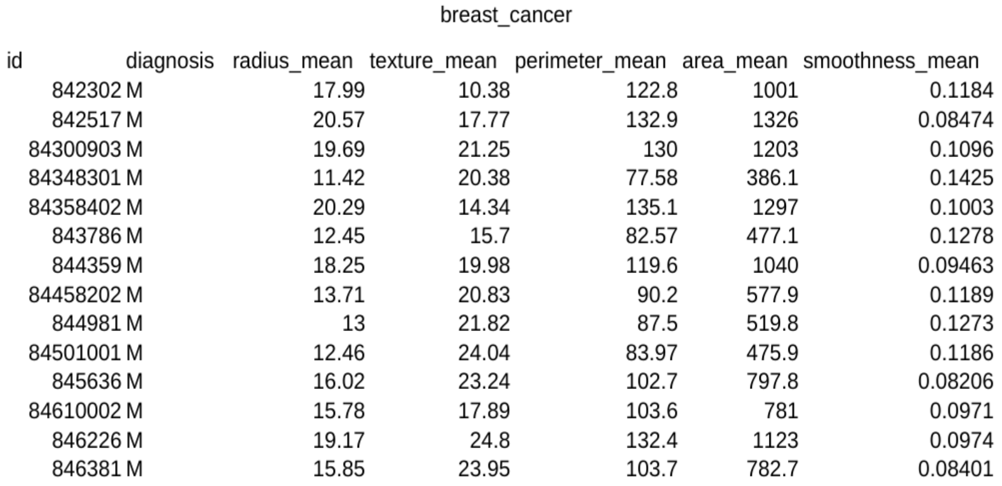
Meta-analysis integrated with ML techniques like SVM has been effective in identifying biomarkers, although advanced analytical methods are needed for further improvements et.al. Panahi R. (6). A multi-ensemble framework combining deep learning and random forest demonstrated promise for analyzing multi-omics data, though molecular validation of biomarkers is necessary et.al. Tembhare K.(7). Another study introduced a pan-cancer TEP stratification system using GMM and XGBoost, emphasizing the potential to link ML-identified clusters with cancer progression et.al. Chen X. (8).

Feature selection and hyperparameter tuning were key to improving model performance, as evidenced by studies evaluating KNN, XGBoost, and SVM et.al. khan Q.W. (9). Logistic regression was identified as a strong candidate for early detection but needs further comparison with other ML algorithms et.al. Anyachebelu T.K. (10). Combining advanced methods such as ANN and random neural networks has enhanced accuracy in breast cancer prediction et.al. Aamir S. (12). Similarly, the Neutrosophic Set and ML approach showed potential for handling complex biomedical data but could benefit from integration with deep learning et.al. Ashika T. (13).

Lastly, gradient boosting techniques, including LightGBM and CatBoost, emphasized the importance of diverse datasets for improving model generalizability and the development of user-friendly diagnostic tools et.al. Chibueze K.I. (19). These findings collectively underscore the potential of ML in breast cancer detection while highlighting key areas for further research.

2. **Methodology:** this research was done by following the phase of research from data source finding and then collecting or choosing the right dataset for prediction to apply multiple algorithms and compare them to check capabilities with different - different algorithms to predict more accurately. There are various phases to do this work, these are-

2.1. **Dataset:** This dataset, breast_cancer.csv, was taken from the Kaggle online public repository. It includes 568 patient records of breast cancer with 32 features, such as Id, diagnosis, radius_mean, texture_mean, concave, points_mean symmetry_mean fractal_dimension_symmetry_worst fractal_dimension_worst, etc. Our dataset has integer values, and for diagnosis, we have binary values 'M' and 'B'.



breast_cancer

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
842302	M	17.99	10.38	122.8	1001	0.1184
842517	M	20.57	17.77	132.9	1326	0.08474
84300903	M	19.69	21.25	130	1203	0.1096
84348301	M	11.42	20.38	77.58	386.1	0.1425
84358402	M	20.29	14.34	135.1	1297	0.1003
843786	M	12.45	15.7	82.57	477.1	0.1278
844359	M	18.25	19.98	119.6	1040	0.09463
84458202	M	13.71	20.83	90.2	577.9	0.1189
844981	M	13	21.82	87.5	519.8	0.1273
84501001	M	12.46	24.04	83.97	475.9	0.1186
845636	M	16.02	23.24	102.7	797.8	0.08206
84610002	M	15.78	17.89	103.6	781	0.0971
846226	M	19.17	24.8	132.4	1123	0.0974
846381	M	15.85	23.95	103.7	782.7	0.08401

Figure 1: sample of dataset

3.2 Data preprocessing and Feature Analysis: The following dataset has no null values and also doesn't have outliners. The dataset used has unique diagnosis features with binary values 'M' and 'B', which we are going to use to classify breast cancer recurrence. Here, 'M' stands for 'malignant' and 'B' for 'Benign'. We divided into two parts for make ready and validate the model using dataset into training and testing.

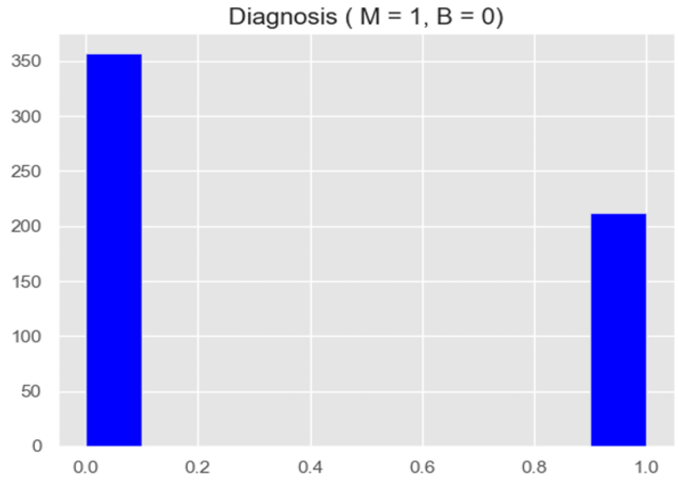


Figure 2: feature in dataset (M = 1, B = 0)

In dataset, we have 300 values for B and 270 values for M to predict the breast cancer to find the accuracy uniquely. There is graphical representation of relation between features with their densities.

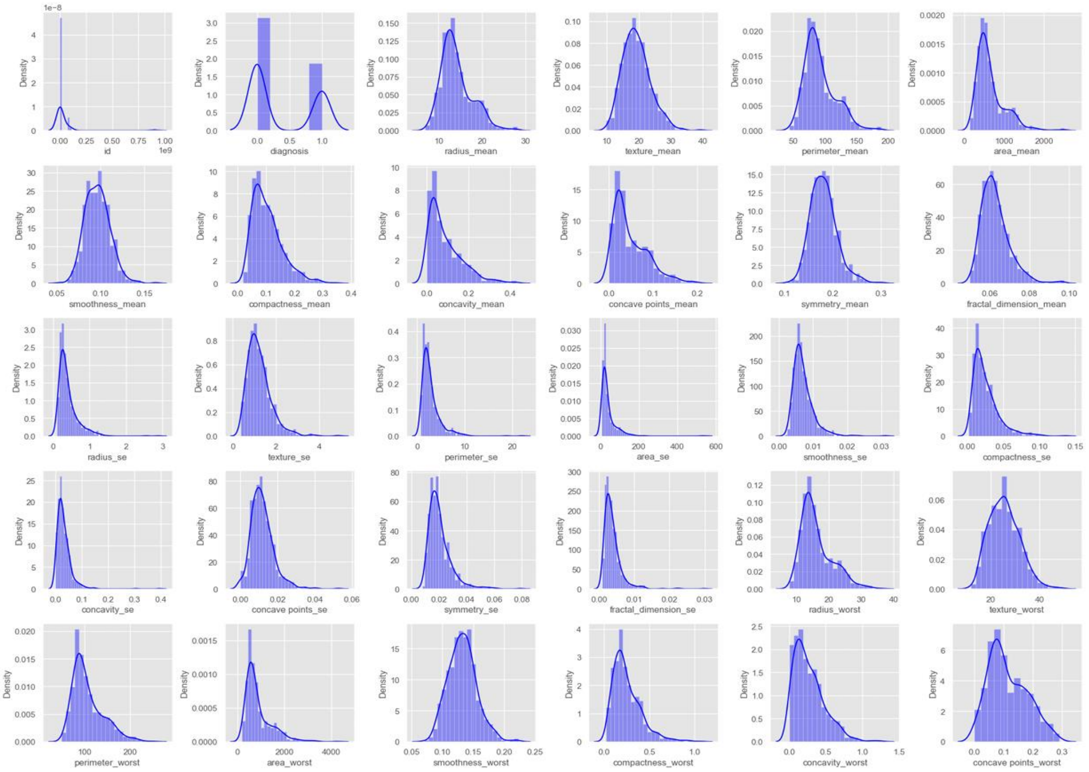


Figure 3: density-feature graph

3.3 Evolution Matrix: as the evolution matrix in this paper we used accuracy, precision, recall, and F1-score to measure the effectiveness of models in our dataset. Accuracy measures the correctness of our model, while precision and recall reflect the accuracy of true positives and model sensitivity to predict true positives. And F1-score measure the balance between these above terms, precision and recall. Additionally, the confusion matrix combines true positives, true negatives, false positives and false negatives. The AUC-ROC score is valuable in evaluating the capabilities of machine learning models to diagnose malignant and benign conditions of tumours.

3.4 Machine Learning Models Selection for Prediction: This paper uses supervised learning to represent breast cancer text data to predict the tumour condition. to choose the correct model is essential to get better accuracy. This study provides a comprehensive analysis of breast cancer prediction using supervised learning models and a comparative practical analysis of supervised learning algorithms such as logistic regression, KNN, SVM, DT, RF, gradient boosting, XG-boost, and NB. The following Models are used in this study:

3.4.1 Logistic Regression: logistic regression is used for classification using probability estimating that features belong to a particular class from given binary classes e.g., malignant and benign. This model is defined by the following sigmoid function:

$$P(y=1|x) = \sigma(w\ T\ x+b) = \frac{1}{1+e^{-(w\ T\ x+b)}}$$

Where **w** stands for the vector of weights, **b** for bias and **σ** represents a sigmoid function. The model minimizes the binary cross-entropy loss:

$$L=\frac{1}{N}\sum_{i-1}^{N}[y_i\log(P(y=1|x_i)) + (1-\ y_i)\log(1-P(y=1|x_i))]$$

This binary classification approach makes logistic regression compatible with breast cancer diagnosis.

3.4.2 K Nearest Neighbors: It is a distance-based supervised learning algorithm where points are classified based on the majority of their nearest neighbor. The distance between two data points is called Euclidean distance d. Consider two data points X1 and X2. The Euclidean distance will be measured by the following function.

$$d\left(x_1, x_2\right) = \sqrt{\sum_{m-1}^{n}\left(x_{im}-x_{jm}\right)^2}$$

For each test point, the algorithm calculates the distances to all training points, selects k closet ones and assigns the majority or nearest class. KNN allows it to adopt datasets without data distribution.

3.4.3 support vector machine: this algorithm aims to find the best optimal hyperplane that maximizes the margin between classes and SVM's decision boundary is defined by the following function:

$$f(x) = w^T x + b$$

SVM uses the following function used by the support vector to establish the optimal hyperplane for the current problem to differentiate m:

$$\min_{w,b} \frac{1}{2} \| w \|^2 \quad subject\ to\ y \quad y_i(w^T x_i + b) \geq 1, \forall i$$

Here, y_i is the class label of $x_{i,}$ enabling SVM to work on non-linear binary classification datasets. It also provides an optimal solution using a hyperparameter between malignant and benign classes.

3.4.4 Decision tree: The decision tree splits the dataset recursively based on provided classification feature values (e.g. Malignant and benign). Common measurement formulas for Gini index and entropy are:

$$Gini\ index: G = 1 - \sum_{k-1}^{K} p_k^2 \quad , entropy: H = -\sum_{k-1}^{K} \mathrm{p}_k \log\left(\mathrm{p}_k\right)$$

where P_k is the proportion of samples belonging to class k and with increasing impurity of values and split the data in different classes of M and B, this splitting creates a tree structure for binary classification.

3.4.5 Random Forest: RF builds randomly selected subsets or decision trees using bootstrap samples in each part to improve classification accuracy. By averaging the result from several trees or subsets. It reduces overfitting and improves robustness. In binary classification for breast cancer prediction with high accuracy and stability. We construct tree using the following formula,

$$\hat{y} = mode\left\{T_m(x) : m = 1, 2, 3, \ldots, M\right\}$$

where T_m is the m^{th} tree. This method increases stability and accuracy, making it robust for handling data.

3.4.6 gradient boosting:

it is a method where we make prediction using multiple decision trees with correct errors improvement sequentially from previous ones. If $F_m(x)$ is the updated prediction, η(etaη) is the learning rate, and $h_m(x)$ is the new model fitted to residuals. That formula will be:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

3.4.7 XG-boost: Extreme Gradient Boosting is an optimized boosting method that enhances accuracy using regularization and fast computation, it applies L1 and L2 regularization to reduce overfitting, and its objective term combines with loss function and a regularization term, such as:

$$obj = \sum_{i} L\left(y_{i}, \widehat{y_{i}} \right) + \sum_{k} \Omega\left(h_{k} \right)$$

Here L is the loss function, Ω (omega) is the regularization term, and h_k represents each tree for M and B classes to predict breast cancer recurrence robustly.

3.4.8 naïve bays: It is based on Bayas's theorem using the probabilistic classifier. It calculates the probability of each class given the features and makes classification based on the highest and posterior probability,

$$P\left(\frac{C}{X}\right) = \frac{P\left(\frac{X}{C}\right).P(C)}{P(X)}$$

Where P(C/X) is the posterior probability of class C on given data X. It works with large datasets and also works well with limited datasets.

These are various supervised machine learning algorithms to predict breast classification between malignant and benign classes for breast tissue recurrences in healthcare from clinical text dataset. After classification, each algorithm is compared based on its accuracy to find a better model to predict early-stage disease more robustly.

3.5 confusion matrix: The confusion matrix provides a summary of a machine learning model's accuracy during testing and tells us how better any model performed for that operation like how accurately they predict actual value or result from the dataset. It is a method to visualize the result of any classification algorithm.

3.6 Model evolution and comparison: after prediction using various supervised models. After training and testing the models, we compared them based on their accuracy and found best-fit models among base supervised learning. To compare various models, we are plotting bar charts and ROC curves to compare the thresholds easily.

5. **Experimental setup:** As the dataset, we are using 'breast_cancer.csv' was taken from the 'Kaggle' online public repository. To perform the prediction for breast cancer using supervised learning and for it, we installed numpy, pandas, matplotlib, seaborn, missingno, pickle, sk-learn and warning libraries. We used various libraries for different-different purposes such that,

5.1 Data handling: numpy and pandas are used for importing and handling the dataset and transforming the dataset into an organized way that can predict the diseases more accurately.

5.2 Data visualization: Seaborn and Matplotlib are used to visualize the information.

5.3 For Prediction: Sci-kit learn is used to import multiple machine learning models, evolution metrics and other features to make predictions.

6 Hyperparameter Tuning: Hyperparameter tuning is conducted to optimise machine learning models' performance using a grid search approach. For SVM, key hyperparameters, including the regularization parameter CC and kernel coefficient Gamma\text{Gamma}, were tuned over the ranges CC [0.01,0.05,0.1,1,10,15,20][0.01, 0.05, 0.1, 1, 10, 15, 20] and Gamma\text{Gamma} [0.0001,0.001,0.01,0.1][0.0001, 0.001, 0.01, 0.1]. The optimal combination C=10C = 10 and Gamma=0.01\text{Gamma} = 0.01 achieved a test accuracy of 99.12%. Similarly, the Gradient Boosting Classifier (GBC) underwent hyperparameter tuning for learning rate, number of estimators, and loss function. The grid search explored Learning Rate\text{Learning Rate} [0.001,0.01,0.1][0.001, 0.01, 0.1], Number of Estimators\text{Number of Estimators} [100,150,180][100, 150, 180], and Loss Function\text{Loss Function} [‘deviance’,‘exponential’][‘deviance’, ‘exponential’], identifying Learning Rate=0.1\text{Learning Rate} = 0.1, Number of Estimators=180\text{Number of Estimators} = 180, and Loss Function=’exponential’\text{Loss Function} = \text{'exponential'} as the optimal parameters, achieving a test accuracy of 96.49%—this systematic tuning process significantly improved model performance.

7 **Result and analysis:** This work represents the prediction of breast cancer recurrence using various supervised machine learning models, such as logistic regression, decision tree, support vector machine, KNN, random forest, XG boost, gradient boosting, and naïve bays. We can predict Breast cancer based on affected tissues. There are two main types of tissues: benign and malignant.

Following table provide complete analysis for supervised machine learning algoriths for breast cancer prediction used in this study with their accuracy.

Models	**Accuracy**
logistic regression	0.98
K-Nearest Neighbor	0.92
Support Vector Machine	0.99
Decision Tree	0.95
Random Forest	0.95
Gradient Boosting	0.96
Xtreme Gaussian Boosting	0.94
Naïve Bays	0.88

Table 1: accuracies for breast cancer using supervised learning models

The table shows that supervised learning performed well with binary classification. As per the medical dataset prediction, almost all algorithms performed well, but SVM provides better compatibility than other algorithms. During the training and test datasets, SVM provided nearly similar accuracy and predicted breast cancer with 99% accurate results. A bar plot compares various supervised learning models based on their accuracy during training and test duration. The ROC curve represents the compatibility of the breast cancer dataset to classify the diseases.

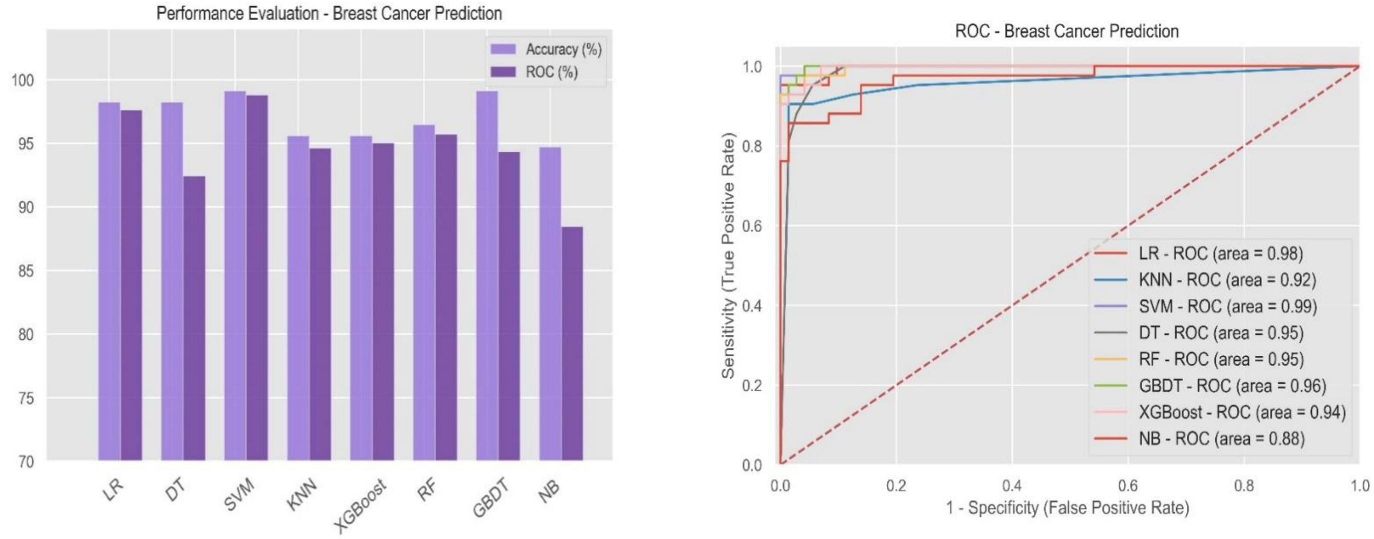


Figure 4: graphical comparison of models

Comparing supervised learning models, we found that a support vector machine provides more accuracy and compatibility in classifying breast cancer from complex clinical text data.

8. **Discussion:** This study evaluated several supervised machine learning models for breast cancer prediction, with Support Vector Machine (SVM) and XGBoost achieving the highest test accuracy of 99.12%. Logistic Regression and K-Nearest Neighbors (KNN) also performed well at 98.25%. Gradient Boosting, Decision Tree, and Random Forest followed with accuracy above 95%, while Naive Bayes showed the lowest accuracy (94.74%) due to its feature independence assumption.

SVM's success highlights its ability to balance model complexity and accuracy through hyperparameter tuning, while XGBoost's gradient boosting mechanism provided competitive performance. Although Gradient Boosting and Random Forest were slightly less accurate, their interpretability and scalability make them viable in clinical settings. Importantly, sensitivity to false negatives, which carry significant clinical implications, was emphasized.

While the dataset was sufficient for this comparison, its size (569 samples) may limit generalizability to larger populations. Future work should involve larger datasets and incorporate explainability techniques like SHAP to enhance clinical applicability. The findings demonstrate the potential of machine learning in breast cancer prediction and its importance in improving early diagnosis outcomes while addressing ethical and privacy considerations.

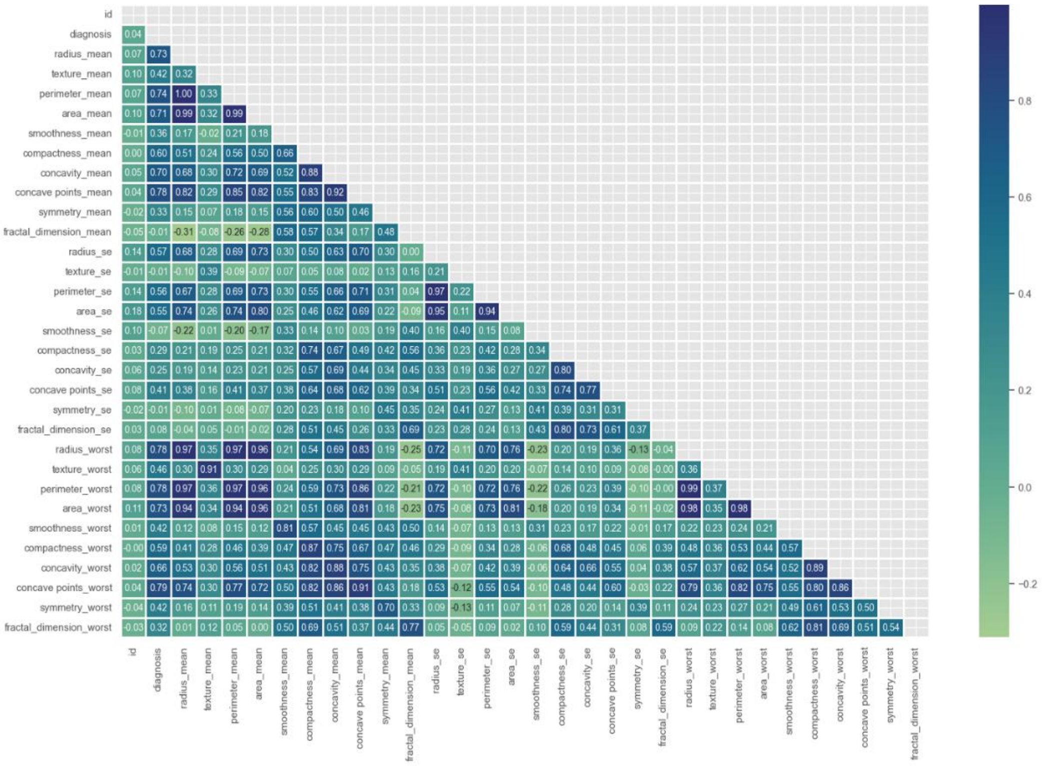


Figure 5: heatmap between breast cancer dataset features

9. Conclusion: This paper provides a practical analysis and comparative study for breast cancer using almost all supervised traditional learning algorithms and comparing them using clinical text data to classify breast cancer tissue conditions between malignant and Benign. We can use the more advanced ML algorithm to find more insights than only predicting the result from patient records.

References:

1. Alladio, E., Trapani, F., Castellino, L., Massano, M., Di Corcia, D., Salomone, A., Berrino, E., Ponzone, R., Marchiò, C., Sapino, A., & Vincenti, M. (2024). Enhancing breast cancer screening with urinary biomarkers and Random Forest supervised classification: A comprehensive investigation. *Journal of Pharmaceutical and Biomedical Analysis*, *244*, 116113. https://doi.org/10.1016/j.jpba.2024.116113
2. Pandey, S., Sharma, A., Siddiqui, M. K., Singla, D., & Vanderpuye-Orgle, J. (2020). AI3 PREDICTION OF BREAST CANCER USING K-NEAREST NEIGHBOUR: A SUPERVISED MACHINE LEARNING ALGORITHM. *Value in Health*, *23*, S1. https://doi.org/10.1016/j.jval.2020.04.006
3. Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Computer Science*, *191*, 487–492. https://doi.org/10.1016/j.procs.2021.07.062
4. Nemade, V., & Fegade, V. (2023). Machine Learning Techniques for Breast Cancer Prediction. *Procedia Computer Science*, *218*, 1314–1320. https://doi.org/10.1016/j.procs.2023.01.110

5. Al-Azzam, N., & Shatnawi, I. (2021). Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. *Annals of Medicine and Surgery*, *62*, 53–64. https://doi.org/10.1016/j.amsu.2020.12.043
6. Panahi, R., Ebrahimie, E., Niazi, A., & Afsharifar, A. (2021). Integration of meta-analysis and supervised machine learning for pattern recognition in breast cancer using epigenetic data. *Informatics in Medicine Unlocked*, *24*, 100629. https://doi.org/10.1016/j.imu.2021.100629
7. Tembhare, K., Sharma, T., Kasibhatla, S. M., Achalere, A., & Joshi, R. (2024). Multi-ensemble machine learning framework for omics data integration: A case study using breast cancer samples. *Informatics in Medicine Unlocked*, *47*, 101507. https://doi.org/10.1016/j.imu.2024.101507
8. Chen, X., Shen, R., Lv, L., Zhu, D., You, G., Tian, Z., Chen, J., Lin, S., Xu, J., Hong, G., Li, H., Luo, M., Cao, L., Wu, S., & Huang, K. (2023). Unsupervised and Supervised Machine Learning to Identify Variability of Tumor-Educated Platelets and Association with Pan-Cancer: A Cross-National Study. *Fundamental Research*, S2667325823002844. https://doi.org/10.1016/j.fmre.2023.09.004
9. Khan, Q. W. (2024) Advances in Breast Cancer Prediction: Evaluating KNN, XGB, and SVM Methods. Sci Set J Cancer Res 3(3), 01-03.

10. Anyachebelu, K. T., Hosea, S. H., Abdullahi, M. U., & Ibrahim, M. A. (2024). Comparative analysis of machine learning algorithms for breast cancer prediction. *Dutse Journal of Pure and Applied Sciences*, *9*(4b), 71–82. https://doi.org/10.4314/dujopas.v9i4b.8
11.
12. Aamir, S., Rahim, A., Aamir, Z., Abbasi, S. F., Khan, M. S., Alhaisoni, M., Khan, M. A., Khan, K., & Ahmad, J. (2022). Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques. *Computational and Mathematical Methods in Medicine*, *2022*, 1–13. https://doi.org/10.1155/2022/5869529
13. Grace, H., Martin, N., & Smarandache, F. (2024). *Enhanced Neutrosophic Set and Machine Learning Approach for Breast Cancer Prediction*. *73*.

14. Jb, A., T, D. R., & L, N. E. (2024, September 23). A Comprehensive Review of Breast Cancer Detection Using Machine Learning and Deep Learning Classifiers. *International Conference on Green Technology and Management for Environmental Sustainability*. International Conference on Green Technology and Management for Environmental Sustainability (ICGMES-2024). https://doi.org/10.59544/EXZW6527/ICGMES24P2

15. Tiwari, K., Pandey, S., Singh, V., Soni, G., & Sudarshan, D. B. G. (2024). *Breast cancer detection using novel ML algorithm*. *11*(2).

16. Thakur, R., Panse, P., & Bhanarkar, P. (2023). Machine learning and deep learning techniques. In Machine Learning and Metaheuristics: Methods and Analysis (pp. 235–253). Springer Nature Singapore. https://doi.org/10.1016/j.measen.2022.100437.

17. Potta, M., Narayanan, B., & Rani Balmuri, K. (2024). A Review on Breast Cancer Prediction Using Machine Learning and Deep Learning Techniques. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *15*(3), 1–24. https://doi.org/10.61841/turcomat.v15i3.14761.

18. Thakur, R., & Panse, P. (2022). Design of an ensemble deep learning model for improving satellite image classification efficiency via temporal analysis. Measurement: Sensors, 24, Article 100437. https://doi.org/10.1016/j.measen.2022.100437.

19. Chibueze, K. I., Ezigbo, L. I., & Kwubeghari, A. (n.d.). *BREAST CANCER PREDICTION WITH GRADIENT BOOSTING CLASSIFIERS.*