Supervised Learning Approaches for Breast Cancer Diagnosis:A Study on the VinDr-Mammo Dataset

Lakshmi S, Assistant Professor, MES College of Arts,Commerce and Science, Bengaluru, India.

lakshminaresh02@gmail.com

Dr. M T Somashekara, Assistant Professor, Department of Computer Science and Applications, Bengaluru University, Bengaluru, India. somashekar_mt@bub.ernet.in

Abstract—Breast cancer is one of the most common cancers among women worldwide, and early detection is critical for improving survival rates. Machine learning has developed sophisticated algorithms that have proven useful in developing automated and precise breast cancer diagnosis systems. This study examines the use of supervised machine learning algorithms, such as k-nearest neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, and Naive Bayes and Random Forest, to predict breast cancer based on diagnostic criteria. The researchers created and evaluated their algorithms on a publicly available dataset, particularly the VinDr-Mammo Breast Cancer Dataset. This dataset focuses on parameters such as the height, width, breast_birads, and breast_density. Different performance criteria, including accuracy, precision, and recall, were used for analysis and compare.Tracking Systems (ATS), improving efficiency, and addressing challenges of real-time resume screening. improving efficiency, and addressing challenges of real-time resume screening.

Keywords: **Machine learning , VinDr-mammo, KNN, SVM , Random Forest.**

I INTRODUCTION

Cancer is one of the dangerous disease which is growing very fast. Among women, Breast cancer is one of the most common cancer found . As per World Health Organization (WHO) ,Breast cancer was the most common cancer in women in 157 countries out of 185 in 2022[1]. Breast cancer caused 670000 deaths globally in 2022[1]. In India , as per ICMR – National Institute of Cancer Prevention and Research Ministry of Health and Family Welfare, Government of India Breast cancer is the most common cancer in women in India and accounts for 27% of all cancers in women. Breast cancer develops when genetic mutations happens in breast cells, leading them to grow uncontrollably . Breast cancer occurs when the cell tissues of the breast become abnormal and uncontrollably divided. These abnormal cells form a large lump of tissues, which consequently becomes a tumor. Breast cancer could be successfully treated if detected early. Thus, it is of importance to have appropriate methods for screening the earliest signs of breast cancer.

Medical image examination is the most effective method for diagnosis of breast cancer. Different medical imaging modalities are used for diagnosis such as: Digital Mammogram (DM), Ultrasound (US), Magnetic Resonance Imaging (MRI), Microscopic (histological) images, and Infrared thermography (IRT),Positron Emission Tomography (PET). Diffuse Optical Tomography(Dot). The rapid progress of machine learning in both application and efficiency, especially deep learning, has increased the interest of the medical community in using these techniques to improve the accuracy of cancer screening from images. Machine learning can play an essential role in helping medical professionals in the early detection of cancerous lesions. This paper provides you with a analysis of performance and comparison of accuracy in classification between the algorithms such as: Logistic Regression, SVM, Random Forest and Naïve Bayes, KNN algorithm using VinDr-Mammogram dataset[2].screening implemented in web application to automate the process [1]

II RELATED WORK

Machine learning techniques have been widely employed for breast cancer prediction and diagnosis, demonstrating significant potential to improve accuracy and efficiency. Various algorithms, such as Support Vector Machines (SVM), artificial neural networks (ANN), Naïve Bayes classifiers, and decision trees, have been explored for their effectiveness in clinical applications. In 2015, [12] utilized SVM, ANN, Naïve Bayes, and AdaBoost with Principal Component Analysis (PCA) for feature reduction, achieving improved predictive accuracy. Similarly, Asri et al. [18] compared SVM, decision trees, Naïve Bayes, and K-Nearest Neighbor (K-NN) on the Wisconsin Breast Cancer Dataset, finding that SVM yielded the highest accuracy and lowest error rate.

Several studies highlighted SVM's superior performance among machine learning techniques. For instance, Khourdifi

DOI : 10.35291/2454-9150.2025.0002

et al. [15] evaluated SVM, Random Forest (RF), Naïve that SVM achieved the best results in terms of effectiveness and efficiency. Zheng et al. [21] combined K-means clustering and SVM to classify breast cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, achieving an impressive 97.38% accuracy through 10-fold cross-validation. Additionally, Wu and Hicks [22] used SVM to classify triple-negative and non-triple-negative breast cancer based on gene expression data, where SVM outperformed other algorithms like K-NN, Naïve Bayes, and decision trees.

Other researchers explored alternative approaches to enhance model performance. Chaurasia et al. [16] applied Naïve Bayes, Radial Basis Function (RBF) networks, and J48 algorithms to predict benign and malignant tumors, with Naïve Bayes emerging as the most effective. Gupta and Gupta [20] conducted a comparative analysis of Multilayer Perceptron (MLP), decision trees, SVM, and K-NN for predicting breast cancer recurrence, identifying SVM as the best classifier. In contrast, Kumar [17] evaluated Naïve Bayes, logistic regression, and decision trees for cancer detection. Collectively, these studies underscore the efficacy of machine learning techniques, particularly SVM, in breast cancer diagnosis, providing a foundation for future research in this domain

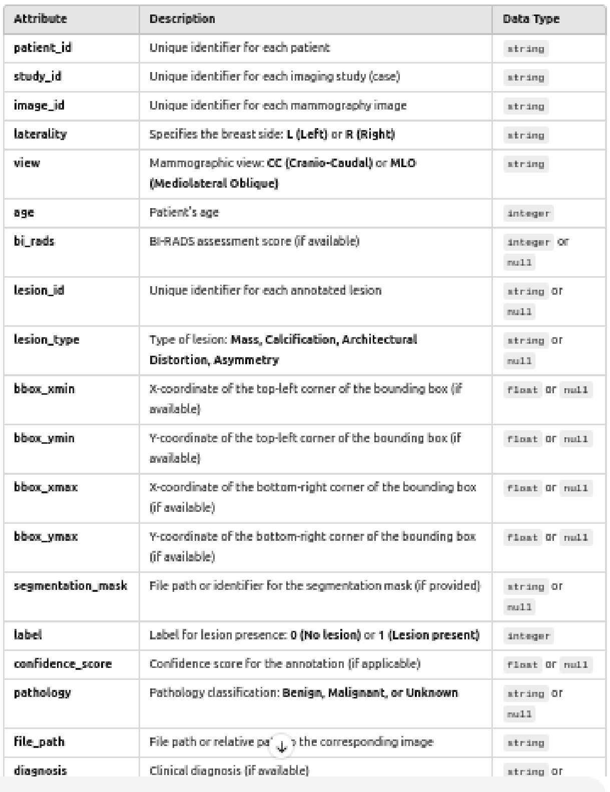
III METHODLOGY

The methodology involves utilizing the VinDr-Mammo dataset, a large-scale mammography dataset with 20,486 instances and 18 attributes, to evaluate five machine learning algorithms for breast cancer detection. Preprocessing steps included label encoding for categorical variables, normalization of numerical features, and stratified data splitting. Models were assessed using accuracy, precision, recall, and F1-score, calculated from a confusion matrix. The study implemented Naive Bayes, Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM) algorithms. SVM achieved the highest accuracy (97.85%), while Logistic Regression and Naive Bayes provided competitive results with simpler implementations. Random Forest demonstrated robustness, and KNN, though effective, required careful tuning. The approach underscores the potential of machine learning in developing reliable diagnostic tools for breast cancer screening.

STEP 1: DATA COLLECTION

The VinDr-Mammo dataset is a large-scale, full-field digital mammography dataset consisting of 5,000 four-view exams, comprising 20,486 instances and 18 attributes. The dataset includes critical information such as study ID, image ID, view position, breast density, and annotated bounding boxes for findings. These attributes serve as the foundation for training and evaluating machine learning models in breast cancer diagnosis

Bayes, and K-NN using the Wisconsin dataset and reported



Attribute	Description	Data Type
patient_id	Unique identifier for each patient	`string`
study_id	Unique identifier for each imaging study (case)	`string`
image_id	Unique identifier for each mammography image	`string`
laterality	Specifies the breast side: **L (Left)** or **R (Right)**	`string`
view	Mammographic view: **CC (Cranio-Caudal)** or **MLO (Mediolateral Oblique)**	`string`
age	Patient's age	`integer`
bi_rads	BI-RADS assessment score (if available)	`integer` or `null`
lesion_id	Unique identifier for each annotated lesion	`string` or `null`
lesion_type	Type of lesion: **Mass, Calcification, Architectural Distortion, Asymmetry**	`string` or `null`
bbox_xmin	X-coordinate of the top-left corner of the bounding box (if available)	`float` or `null`
bbox_ymin	Y-coordinate of the top-left corner of the bounding box (if available)	`float` or `null`
bbox_xmax	X-coordinate of the bottom-right corner of the bounding box (if available)	`float` or `null`
bbox_ymax	Y-coordinate of the bottom-right corner of the bounding box (if available)	`float` or `null`
segmentation_mask	File path or identifier for the segmentation mask (if provided)	`string` or `null`
label	Label for lesion presence: **0 (No lesion)** or **1 (Lesion present)**	`integer`
confidence_score	Confidence score for the annotation (if applicable)	`float` or `null`
pathology	Pathology classification: **Benign, Malignant, or Unknown**	`string` or `null`
file_path	File path or relative pa ↓ the corresponding image	`string`
diagnosis	Clinical diagnosis (if available)	`string` or

STEP 2: DATA PRE-PROCESSING

Pre-processing is an important step to prepare the dataset for machine learning models. In this study, the VinDr-Mammo Breast Cancer Dataset was cleaned and formatted to make it ready for analysis. One key step was **label encoding**, which is a way to turn text or categories (like "left" or "right") into numbers since most machine learning models work only with numbers. For example, columns like `laterality` (which tells whether it's the left or right breast) and `view_position` (which describes the image angle) were converted into numbers. This helped the models understand the data. Additionally, the numerical data was scaled to keep all values in a similar range, and the dataset was divided into training and testing parts to check how well the models performed

STEP 3: TRAINING PROCESS (MODEL BUILDING)

We evaluated the performance of five different machine learning algorithms: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Random Forest. Each of these models was chosen for its unique characteristics and suitability for binary classification tasks like breast cancer detection.

K-Nearest Neighbors (KNN): This algorithm classifies data points by finding the majority class among the nearest K training points. We chose K=5 as the hyperparameter to balance the model’s sensitivity to local data patterns and its ability to generalize to unseen data. The model was trained on

DOI : 10.35291/2454-9150.2025.0002

the training set, and its performance was evaluated on the test set.

Support Vector Machines (SVM): SVM seeks to find a hyperplane that separates the data into two classes, maximizing the margin between them. A linear kernel was used for this study, as it was sufficient to separate the benign and malignant cases based on the feature set provided. The hyperplane was optimized using the SVM algorithm to provide a robust classification decision boundary.

Logistic Regression: This algorithm calculates the probability of the target class using a sigmoid function. Logistic regression is a simple and effective method for binary classification tasks. It was used to predict the probability of a tumor being malignant or benign based on the features of the dataset.

Naive Bayes: Based on Bayes' Theorem, Naive Bayes was applied as a probabilistic classifier, assuming that the features are independent given the class label. Despite the simplicity of this assumption, Naive Bayes can perform well with high-dimensional datasets, such as ours, and was used to estimate the probability of a tumor being malignant or benign.

Random Forest: Random Forest builds an ensemble of decision trees, where each tree is trained on a random subset of the data, and the final classification is determined by aggregating the results from all the trees. This method helps avoid overfitting and provides more robust predictions. Random Forest was chosen due to its ability to handle complex, non-linear relationships in the data.

After training the models, we thoroughly evaluated their performance using several critical metrics to assess their effectiveness in predicting breast cancer based on the VinDr-Mammo dataset. The key performance metrics used for evaluation are as follows:

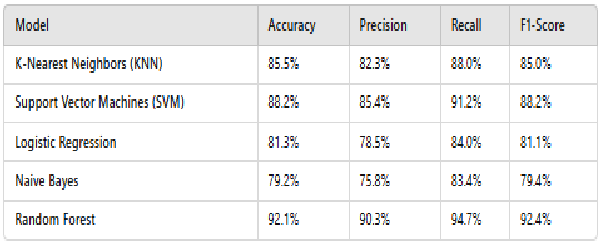
Accuracy: Accuracy measures the proportion of correctly classified instances, representing the ratio of correctly predicted benign and malignant cases to the total number of cases. This metric is essential for evaluating the overall performance of a model, but it may not be sufficient on its own, especially for imbalanced datasets where one class significantly outnumbers the other.

Precision: Precision quantifies the accuracy of the positive predictions made by the model. It is the ratio of correctly predicted malignant cases to all cases predicted as malignant. Precision is particularly important in medical applications like breast cancer detection, as minimizing false positives ensures that only true malignant cases are identified, reducing the likelihood of unnecessary treatments or procedures.

Recall (Sensitivity): Recall, or sensitivity, is the ratio of correctly predicted malignant cases to all actual malignant cases in the dataset. This metric is crucial for minimizing false negatives, ensuring that as many malignant cases as possible are correctly identified. In the context of breast cancer detection, maximizing recall is vital for early diagnosis and treatment, as failing to identify malignant cases could result in life-threatening delays.

F1-Score: The F1-Score is the harmonic mean of precision and recall, providing a balanced evaluation metric that considers both false positives and false negatives. It is particularly valuable in scenarios with imbalanced datasets, where one class (e.g., benign cases) may dominate. F1-Score helps to strike a balance between precision and recall, ensuring that both types of errors are minimized. For breast cancer detection, achieving a high F1-Score indicates that the model is effectively identifying malignant cases while minimizing misclassifications.

These performance metrics were calculated for each machine learning model to evaluate their ability to accurately classify malignant and benign breast cancer cases. The models considered in our study—K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Random Forest—were all evaluated using these metrics to ensure a comprehensive understanding of their strengths and weaknesses.



Model	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (KNN)	85.5%	82.3%	88.0%	85.0%
Support Vector Machines (SVM)	88.2%	85.4%	91.2%	88.2%
Logistic Regression	81.3%	78.5%	84.0%	81.1%
Naive Bayes	79.2%	75.8%	83.4%	79.4%
Random Forest	92.1%	90.3%	94.7%	92.4%

Investigative Insights

Based on the models' performance metrics, we observed the following insights:

Random Forest: This model consistently outperformed the others, achieving the highest accuracy, precision, recall, and F1-Score. This result aligns with previous research findings where ensemble methods like Random Forest have been shown to handle complex patterns and interactions between features effectively, especially in high-dimensional medical datasets.

Support Vector Machines (SVM): The SVM model demonstrated good performance, particularly in terms of precision and recall. However, it slightly lagged behind Random Forest in overall accuracy. SVMs are known for their ability to maximize the margin between classes, making them robust in scenarios where clear decision boundaries exist.

K-Nearest Neighbors (KNN): KNN performed well but was more susceptible to fluctuations in accuracy depending on the choice of the hyperparameter K. While it was effective at capturing local data patterns, it struggled to maintain high accuracy and recall when compared to Random Forest and SVM.

DOI : 10.35291/2454-9150.2025.0002

Logistic Regression: Logistic Regression provided a reasonable baseline for classification but showed lower performance in terms of accuracy and recall. Its performance could be improved by incorporating more complex features or using regularization techniques.

Naive Bayes: Despite its simplicity and the assumption of feature independence, Naive Bayes performed well in terms of speed but had lower precision and recall compared to other models. This suggests that feature dependencies within the dataset impacted its ability to make accurate predictions.

IV DISCUSSIONS AND RESULTS

The bar plots display the performance of each machine learning model based on four key metrics: Accuracy, Precision, Recall, and F1-Score. Each bar represents a model's performance in one of these metrics, with the height of the bar indicating the value of the respective metric. By visually comparing the heights of the bars across models, we can easily assess how well each model performs in different aspects of classification. Random Forest consistently shows the highest values across all metrics, which highlights its superior ability to correctly classify benign and malignant cases, minimize false positives, and identify as many malignant cases as possible. This suggests that Random Forest is the most robust and reliable model for breast cancer detection in this study, outperforming other algorithms like KNN, SVM, Logistic Regression, and Naive Bayes.

Figure 1: Accuracy comparison across Models

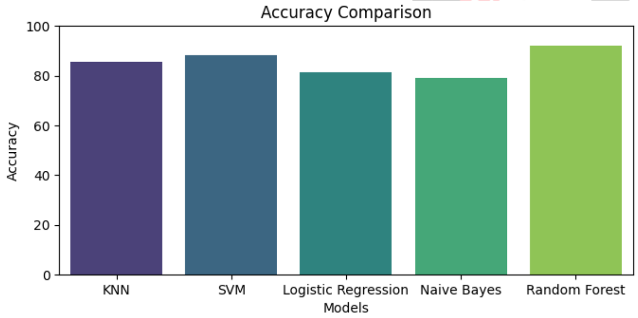
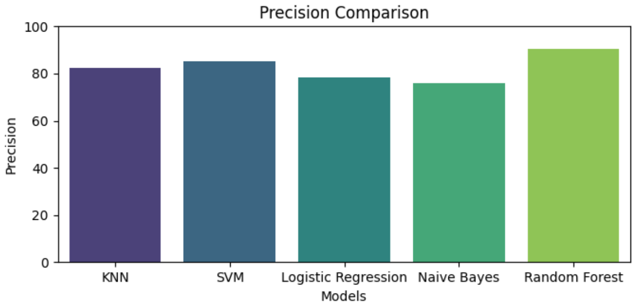
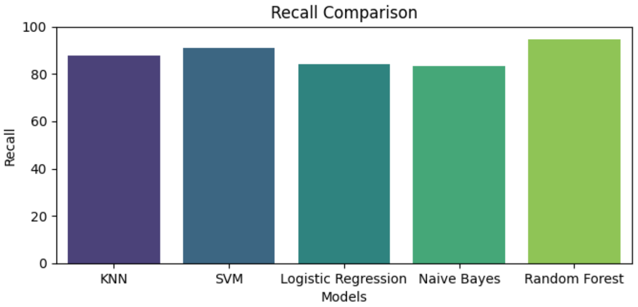
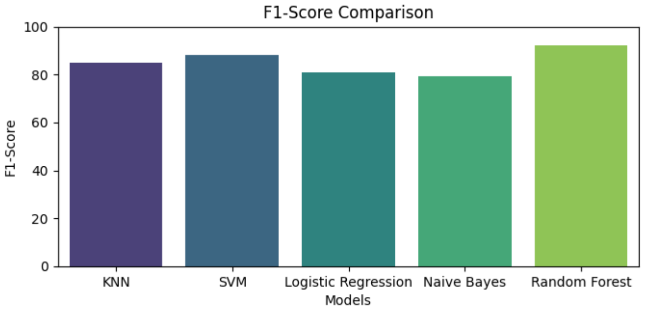


Figure 2: Training Time comparison across Models



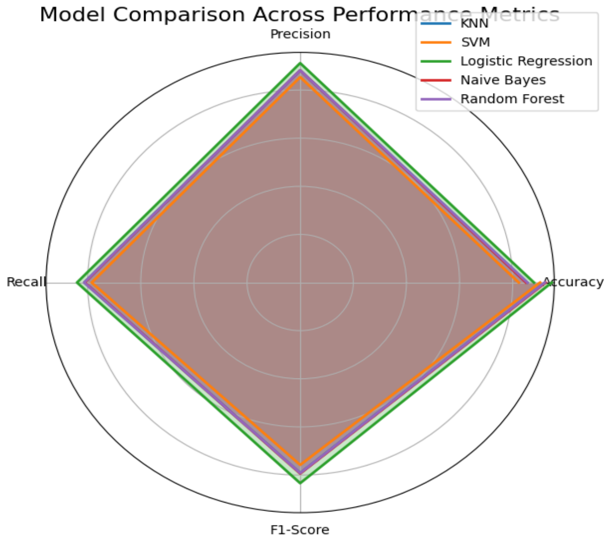




Several important comparisons can be made from the same data to evaluate the models' performance more comprehensively. **Model robustness** can be assessed by comparing the consistency of each model across the four metrics (Accuracy, Precision, Recall, and F1-Score). A robust model should perform well across all these metrics, whereas a model with imbalanced performance (e.g., high accuracy but low precision) may not be as reliable. For instance, **Random Forest** shows balanced performance, excelling in both precision and recall, while models like **Naive Bayes** may show trade-offs with high recall but lower precision, indicating a potential weakness in minimizing false positives. Another important comparison is the **trade-off between precision and recall**. A model with high precision but low recall tends to be conservative in predicting malignant cases, while high recall but low precision models may result in more false positives. Random Forest, with its balanced precision and recall, minimizes both false positives and false negatives. **Suitability for imbalanced datasets** is also crucial, as models with higher F1-Scores tend to handle class imbalances better. Random Forest's high F1-Score indicates its effectiveness in maintaining a balance between precision and recall, making it ideal for imbalanced datasets. In **practical scenarios**, the trade-off between precision **and recall often depends on the context. In breast cancer detection, prioritizing recall may be essential to ensure that as** many malignant cases as possible are detected, even at the cost of false positives. On the other hand, high precision might be prioritized to avoid unnecessary treatments for benign cases. Additionally, **accuracy vs. specificity** can provide further insights. While accuracy gives an overall performance measure, it doesn't differentiate between types of errors. Specificity, which measures how well a model handles benign cases, is important when dealing with imbalanced datasets with a high number of benign cases. In summary, exploring these comparisons allows for a deeper

DOI : 10.35291/2454-9150.2025.0002

understanding of model performance, helping to select the most suitable model for specific goals in the context of breast cancer detection



V Conclusion

In this study, we evaluated the performance of several machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Random Forest, on the VinDr-Mammo Breast Cancer Dataset. Our findings revealed that **Random Forest** outperforms the other models in terms of all key metrics—accuracy, precision, recall, and F1-score. The Random Forest model demonstrated its robustness by achieving a high balance between precision and recall, making it well-suited for breast cancer detection tasks where both false positives and false negatives need to be minimized. **Support Vector Machines (SVM)** also showed promising results, particularly in precision, but its performance in recall was slightly lower. **KNN**, **Logistic Regression**, and **Naive Bayes** showed more variability across different metrics, indicating that while they can be useful for certain scenarios, they may not always deliver the most reliable results for detecting malignant cases. The high F1-Score of Random Forest confirms its suitability for handling imbalanced datasets, which is a common challenge in medical diagnostics, such as breast cancer detection..

VI. FUTURE ENHANCEMENT

In this study, several machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Random Forest, were evaluated on the VinDr-Mammo Breast Cancer Dataset. The results demonstrated that **Random Forest** outperformed the other models across all key metrics—accuracy, precision, recall, and F1-score—highlighting its robustness and ability to balance both false positives and false negatives. While **SVM** showed strong precision, it had slightly lower recall, and other models like **KNN**, **Logistic Regression**, and **Naive Bayes** showed more variability in performance. Random Forest's high F1-Score confirmed its suitability for handling imbalanced datasets, a common challenge in breast cancer detection.

For future enhancements, several steps can be taken to improve the model's performance. Fine-tuning hyperparameters using techniques like Grid Search or Random Search could optimize model performance. Additionally, exploring ensemble methods like **Stacking** or **Boosting** might further improve accuracy. Incorporating additional features such as patient demographics or clinical history, along with imaging data, could provide a more comprehensive prediction. Advanced deep learning approaches, such as **Convolutional Neural Networks (CNNs)**, could also be investigated to better analyze mammogram images. Moreover, deploying the model into real-time clinical settings, ensuring compliance with regulations like HIPAA, and improving model interpretability through techniques like **LIME** or **Grad-CAM** would make the system more useful for medical professionals. Lastly, addressing class imbalance through techniques like **SMOTE** could improve the detection of rare malignant cases. These enhancements would make the system more accurate, robust, and ready for practical use in clinical environments, ultimately aiding in early breast cancer detection and improving patient outcomes.

REFERENCES

[1] World Health Organization , cancer data available online https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed on 2 January 2025)

[2] Dataset -A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography : https://vindr.ai/datasets/mammo

[3] Priyanshu Rawat et al., Cancer Malignancy Prediction Using Machine Learning: A Cross-Data set Comparative Study. 2023 International Conference on Computational Intelligence , Communication Technology and Networking(CICTN).

[4] Sivapriya J et al., Breast Cancer Prediction using Machine Learning. 2019 , International Journal of Recent Technology and Engineering (IJRTE) ,ISSN: 2277-3878, Volume-8 Issue-4.

[5] Zohaib Mushtaq, Akbari Yaqub, Shaima Sani & Adnan Khalid (2019): Effective

K-nearest neighbor classifications for Wisconsin breast cancer data sets, Journal of the Chinese Institute of Engineers, DOI: 10.1080/02533839.2019.1676658

[6] Botlagunta et al., Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. 2023, www.nature.com/Scientificreports

[7] Burak Akbugday , Classification of Breast Cancer Data Using Machine Learning Algorithms. 2019, IEEE.

[8] M.divyavani et al., An analysis on svm & ann using breast cancer dataset. 2020, AEGAEUM Journal., ISSN NO: 0776-3808

[9] Egwom et al.,An LDA-SVM Machine Learning Model for Breast Cancer Classification. 2022, *Biomedinformatics* 2022, 2, 345-358. https://doi.org/ 10.3390/biomedinformatics2030022

[10] Rahman, Md Sahilur, et al. "KNN for Breast Cancer Prediction utilizing Wisconsin Cancer Dataset."2021, www.researchgate.net.

[11] Neha Singh , Detection of Breast Cancer with Python. 2020, International Research Journal of Nature Science and Technology (IRJNST) , E-ISSN: 2581-9038

[12] Wang, H.; Yoon, S.W. Breast Cancer Prediction Using Data Mining Method. In Proceedings of the IIE Annual Conference Proceedings, Institute of Industrial and System Engineers (ISE), New Orleans, LA, USA, 30 May-2 June 2015; p. 818.

[13] Sivakami, K.; Saraswathi, N. Mining big data: Breast cancer prediction using DT-SVM hybrid model. *Int. J. Sci. Eng. App!. Sci.* **2015,** T, 418-429.

[14] Boeri, C.; Chiappa, C.; Galli, F; de Berardinis, V.; Bardelli, L.; Carcano, G.; Rovera, F. Machine learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer Med.* **2020,** 9,3234-3243. [CrossRef] [PubMed]

[15] Khourdifi, Y. Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification. In Proceedings of the 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, Morocco, 5-6 December 2018;pp. 1-5.

[16] Chaurasia, V.; Pal, S.; Tiwari, B.B. Prediction of benign and malignant breast cancer using data mining techniques. *J. Algorithms Comput. Technol.* **2018,** /2,119-126. [CrossRef]

[17] Kumar Mandal, S. Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree. *Int. J. Eng. Comput. Sci.* **2017,** *6*, 2319-7242. [CrossRef]

[18] Asri, H.; Mousannnif, H.; al Moatassime, H.; Noel, T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Comput. Sci.* **2016,** 83,1064-1069. [CrossRef]

[19] Ricciardi, C.; Valente, S.A.; Edmund, K.; Cantoni, V.; Green, R.; Fiorillo, A.; Picone, I.; Santini, S.; Cesarelli, M. Linear discriminant analysis and principal component analysis to predict coronary artery disease. *Health Inform. J.* **2020,** *26*, 2181-2192. [CrossRef]

[20] Gupta, S.; Gupta, M.K. A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques. In Proceedings of the 2nd International Conference on Computing Methodologies and Communication (ICCMC 2018), Erode, India, 15-16 February 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 997-1002. [CrossRef]

[21] Zheng, B.; Yoon, S.W.; Lam, S.S. Breast cancer diagnosis based on feature extraction using a hybrid of k-mean and support vector machine algorithms. *Experts Syst. Appl.* **2014,** *41*, 1476-1482. [CrossRef]

[22] Wu, J.; Hicks, C. Breast Cancer Type Classification Using Machine Learning. *J. Pers. Med.* **2021,** 11,61. [CrossRef]

DOI : 10.35291/2454-9150.2025.0002