Revolutionizing Healthcare: An Intuitive Interface for Swift Breast Cancer Detection with XGBoost Models

[1] Aniket Dhoke
Computer Engineering Department
Vishwakarma Institute of Information Technology
Pune, India
aniket.22110835@viit.ac.in

[2] Ketan Choudhari
Computer Engineering Department
Vishwakarma Institute of Information Technology
Pune, India
ketan.22110060@viit.ac.in

[3] Ahilya Mote
Computer Engineering Department
Vishwakarma Institute of Information Technology
Pune, India
ahilya.22110787@viit.ac.in

[4] Pooja Vishwakarma
Computer Engineering Department
Vishwakarma Institute of Information Technology
Pune, India
pooja.22110402@viit.ac.in

[5] Guided by: Nitin Sakhare
Department: *Computer Engineering*
Name of Organization:
Vishwakarma Institute of Information Technology
Pune, India
Email; nitin.sakhare@viit.ac.in

[6] Guided by: DR. Shubham Dodia
Department: *Computer Engineering*
Name of Organization:
Vishwakarma Institute of Information Technology
Pune, India
Email: shubham.dodia@viit.ac.in

Abstract*—Breast cancer poses a global health threat, prompting exploration into AI's role in early detection. This study evaluates K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and decision tree (DT) for discerning benign from malignant tumour. SVM, particularly on larger datasets, demonstrates superior accuracy. Integrating ensemble voting and adaptive algorithms enhances early-stage identification, promising better prognoses. Beyond breast cancer, this research suggests robust diagnostic tools for early cancer detection in various medical fields. The synergy of artificial intelligence and healthcare offers proactive disease management. Emphasizing SVM's excellence and innovative artificial intelligence integration, this study showcases potential in improving patient outcomes through early intervention.

Keywords – Breast cancer; malignant tumours; benign tumours; early intervention; KNN; SVM; DT.

1.INTRODUCTION

Breast cancer, a common and potentially fatal disease, remains a major global health concern, especially among women. Early detection is critical for improving prognosis and treatment outcomes for people with breast cancer. In recent years, there has been a growing interest in using Machine Learning (ML) techniques to improve breast cancer detection and diagnosis. Traditional methods of breast cancer detection, such as mammography, biopsy, and ultrasound, have proven effective but have limitations, such as false positives and the need for invasive procedures. ML, a subset of AI, introduces a paradigm shift by allowing computer systems to learn patterns and make predictions from data, in this case, medical imaging and patient information.ML algorithms can analyse massive datasets of mammograms, histopathological images, genetic markers, and clinical records to identify subtle patterns indicative of breast cancer. These algorithms have the potential to improve accuracy, reduce false positives and negatives, and assist healthcare professionals in making more informed decisions.

Male breast cancer accounts for only 1% of all breast cancer cases. The prevalence of breast cancer in transgender people, as well as the impact of gender-affirming hormonal treatment (GAHT) on breast cancer risk, are largely unknown. However, it is unclear what risk breast cancer poses to transgender people and how, if at all, physicians should screen these patients. The medical literature contains reports of transgender men developing breast cancer. The incidence of breast cancer in trans women receiving GAHT is unknown. In 2018, two population-based studies assessed the risk of breast cancer associated with GAHT. Both studies had limitations due to the small number of breast cancer cases and the lack of genetic risk stratification.

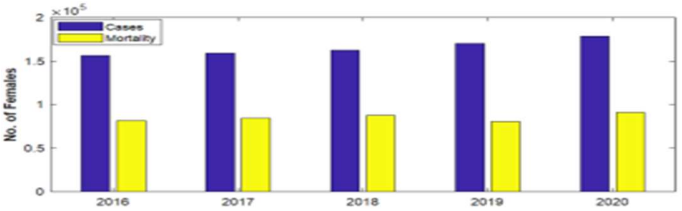

Fig. 1 Breast cancer cases and mortality in females in India [15]

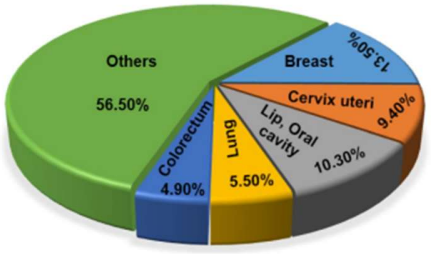

Fig. 2 Number of new cases of cancer found in India during 2020

2. LITERATURE REVIEW

Medical imaging, particularly mammography, is critical in the early detection of breast tumours. The need for skilled radiologists, as well as the trade-off between double reading for accuracy and cost/time constraints, are all challenges. CAD (Computer-Aided Diagnosis) systems provide radiologists with a second opinion by

assisting in the detection and differentiation of normal and abnormal tissues [1, 2]. The focus of this review is on recent advances in breast cancer detection using traditional machine learning (ML) and deep learning (DL) techniques. The study focuses on Multiview digital mammograms (DMs) acquired at the CC and MLO views. The review is divided into two sections: the first introduces traditional ML methods (enhancement, feature extraction, segmentation, and classification), and the second delves into DL techniques. The DL section focuses on Multiview mammographic data, which includes breast density discrimination, lesion detection, and classification [3]. Saliha Zahoor, focuses on improving model performance through feature extraction with fine-tuned MobileNetV2 and Nasset Mobile [1]. CNN architectures are used in studies such as BCDCNN and Guan and Loew, with accuracy rates of 82.71% and 90.5%, respectively. MLP architectures show accuracy rates on cross-validation datasets (CV) ranging from 72.5% to 100% in studies by Meha Desai and Manan Shah [5].

The [6] uses comparison of five nonlinear ML-based classification algorithms is presented in the paper: Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Gaussian Naive Bayes (NB), and Support Vector Machines (SVM). Table III of [6] shows the predicted accuracy of the five algorithms on testing data. With a prediction accuracy of 99.12%, the Multilayer Perceptron (MLP) stands out. Various machine learning algorithms were used in [12], including logistic regression (LR), multi-layer perceptron (MLP), decision tree (DT), K-nearest neighbour (KNN), random forest (RF), nave Bayes (NB), eXtreme Gradient Boosting (XGBoost), and adaptive boosting (AdaBoost).When compared to other models, the proposed model combining SVM and extra-trees demonstrated higher accuracy, precision, specificity, sensitivity, and AUC rates, with improvements of up to 80.23%, 82.71%, 78.57%, 78.57%, and 0.78, respectively. The Jin Huang [13], used a WSI dataset from the ICIAR2018 Grand Challenge on Breast Cancer Histology Images (BACH) competition, which included four classes: normal, benign, in situ carcinoma, and invasive carcinoma. The proposed BM-Net architecture combines a bilinear structure with the MobileNet-V3 network Data augmentation techniques were used to increase diversity, and focal loss was used to address class imbalance. The most recent approach [14], a transfer learning framework is introduced, and pre-trained models are fine-tuned for the task, specifically ResNet50 and Nasnet-Mobile. To increase the number of mammographic images, augmentation strategies such as rotation, scaling, and shifting are used. The proposed system is tested using the dataset from the Mammographic Image Analysis Society (MIAS). ResNet50 achieved an accuracy of 89.5%. Using Nasnet-Mobile, we achieved a 70% accuracy rate. By utilizing transfer learning and pre-trained models, the research contributes to the advancement of breast cancer detection. The achieved accuracies suggest that these methods have the potential to improve the efficiency of medical imaging, particularly in scenarios with limited training data. Roslidar, Mohd Syaryadhi, Khairun Saddami [15], introduce BreaCNet, a mobile neural network model for early breast cancer detection, is introduced in the study. It demonstrates the power of combining a segmentation algorithm with a modified ShuffleNet classifier to improve accuracy and diagnostic performance. In real-world implementations, the emphasis on on-device inference addresses privacy concerns as well as network reliability issues. Edge detection and second-order polynomial curve fitting techniques are used. Captures the region of interest (ROI) in breast thermograms effectively. Based on ShuffleNet with the addition of a convolutional layer with 1028 filters. Modified ShuffleNet is 22 MB in size and has 6.1 million parameters. Modified ShuffleNet achieved a 72% accuracy rate on its own. When combined with the proposed segmentation algorithm, performance improved to 100% accuracy. BreaCNet increased sensitivity from 43% to 100% and achieved 100% specificity. The Imaging Modalities for BC are provided by Nusrat Mohi ud din [9], Diagnosis: Mammography is commonly used to detect early-stage breast cancer; Ultrasound: Identifies cysts from solid masses in dense breasts. Highest sensitivity, but time-consuming and costly; Histopathology is regarded as the gold standard, but it has limitations in multi-classifying BC; CT: Used for preoperative staging, particularly in patients with BC who have pulmonary symptoms, Non-invasive thermography detects abnormalities based on heat patterns. The paper used a three-phase strategy (identification, evaluation, and selection) to examine research studies published between 2015 and 2021 on BC diagnosis using AI-based techniques. The Global Cancer Observatory provided mortality rate statistics. Science Direct, PubMed, Web of Science, ArXiv, SPIE, IEEE Xplore Digital Library, and Google Scholar were among the databases searched.

Key performance in [11], such as accuracy (Acc), precision (Pres), sensitivity (Sens), specificity (Spec), area under the curve (AUC), and F1-score are used in the validation tests. The proposed system performs well, with an Acc of 86.21%, a Pres of 85.90%, a Sens of 85.71%, a Spec of 84.51%, an F1-score of 88%, and an AUC of 0.89. These findings point to the possibility of reducing pathologist errors and efforts during the clinical process. When compared to other models, such as CNN-BiLSTM, CNN-LSTM, and existing ML/DL methods, CNN-GRU outperforms them by 4 to 5% in accuracy, as well as better Pres, Sens, Spec, AUC, F1-score, and lower time complexity (ms)

3. METHODOLOGY

3.1. Dataset Description:

Breast Cancer Wisconsin (Diagnostic) Dataset. Here, 569 patients' data were analyzed, with each instance having 32 Attributes with Diagnosis and Features. Each instance has a parameter of cancerous and non-cancerous cells, and we will predict cancer based solely on feature input. The feature values are in Numeric Format. The 'Target' is the patient who has either 'Benign' or 'Malignant' cancer. The term benign refers to the absence of cancer, whereas malignant refers to the presence of cancer.

Table I

Type of Patients	
Patient Type	Target
Benign	1
Malignant	0

3.2. Data Visualization:

A pair plot is a powerful data visualization tool that allows us to investigate relationships between multiple variables at the same time. A pair plot of features such as mean radius, mean texture, mean perimeter, mean area, and mean smoothness provides valuable insights into the distribution and potential correlations among these critical characteristics in the context of breast cancer detection using machine learning

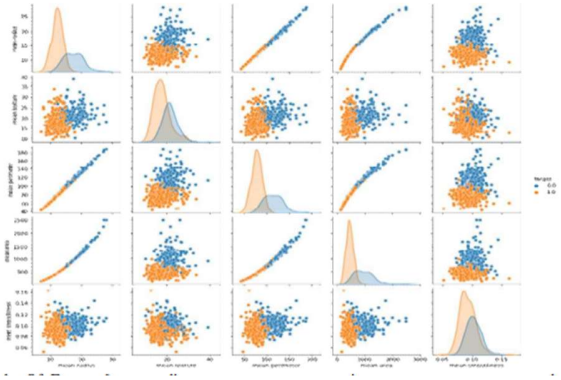

Fig. 3 Pair plot of the Features [mean radius, mean texture, mean perimeter, mean area, mean smoothness]

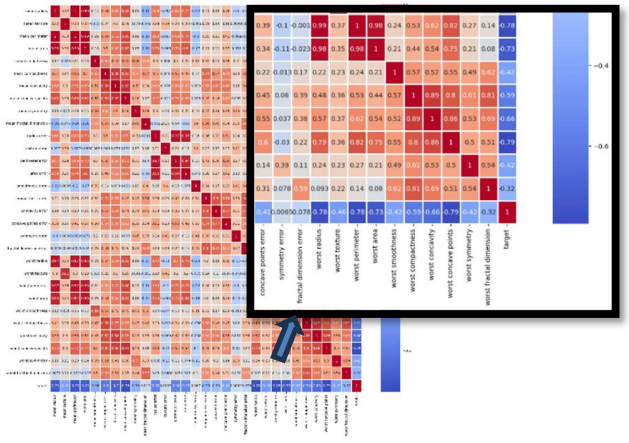

fig. 4 Dataframe heatmap of breast cancer correlation matrix

3.3. Data splitting and scaling:

The dataset under investigation is meticulously partitioned into two fundamental components in this research endeavor—input variables (X) and target variables (y). This preliminary stratification lays the groundwork for subsequent analyses and model development. To assess the robustness and predictive capability of the models, the dataset is split into training and testing sets using the widely used train, test, split methodology, ensuring an 80-20 distribution, respectively. This partitioning method allows for a thorough evaluation of the model's performance on both known and unknown data. Recognizing the significance of feature standardization for improving model interpretability and convergence, the features in the dataset are transformed using the Standard Scaler. This critical preprocessing step ensures that the features are rescaled to a common scale, avoiding potential biases introduced by varying units and magnitudes. This standardization process lays the groundwork for developing robust and generalizable machine learning models. The prudent combination of feature engineering, dataset partitioning, and standardization forms a cohesive methodology poised to yield valuable insights and advancements in the field under investigation.

3.4. Machine Learning Model Construction:

This study investigates the use of various classifiers to detect breast cancer using a large ensemble of machine learning models. Each classifier contributes to the overall goal of accurate and reliable detection in a unique way.

3.4.1. The Support Vector Classifier (SVC): is a powerful and versatile algorithm that is used to separate malignant and benign instances in a dataset. SVC improves its discriminatory capabilities by transforming the data into a higher-dimensional space.

3.4.2. K-Nearest Neighbors (KNN): A non-parametric and instance-based algorithm that makes decisions based on the majority class within the k-nearest data points. KNN uses similarities among feature vectors to classify instances in breast cancer detection, providing flexibility and simplicity in model interpretation.

3.4.3. Naive Bayes: Naive Bayes: which is based on Bayesian probability theory, assumes independence among features, which simplifies the modeling process. It is used in breast cancer detection to calculate the probability of malignancy based on observed feature distributions, providing quick and accurate predictions.

3.4.4. Decision Tree: Classifier employs a tree-like model, recursively partitioning the dataset based on feature values. In the context of breast cancer detection, decision trees reveal feature importance and capture intricate decision boundaries, allowing for intuitive interpretation.

3.4.5. Random Forest: A collection of decision trees that combine their outputs to improve accuracy and reduce overfitting. This ensemble method excels at capturing complex relationships within data and providing robust predictions when applied to breast cancer detection.

3.4.6. AdaBoost: AdaBoost focuses on weak learners, assigning

higher weights to misclassified instances iteratively. When applied to breast cancer detection, this ensemble technique combines multiple weak classifiers to create a strong, accurate model.

3.4.7. XGBoost: XGBoost, an optimized gradient boosting algorithm, builds decision trees sequentially to correct errors in previous models. XGBoost, known for its speed and predictive accuracy, has been fine-tuned to improve its breast cancer detection capabilities through Randomized Search and Grid Search.

This study evaluates each classifier's performance on both raw and scaled data, taking into account the impact of feature standardization on model efficacy. The XGBoost classifier is meticulously tuned to extract optimal hyperparameters using Randomized Search and Grid Search, ensuring the model's peak predictive performance in breast cancer detection. The combination of these classifiers creates a strong framework that has the potential to significantly advance breast cancer diagnosis and treatment.

Table II

Techniques	Accuracy Without Standard scale	Accuracy With Standard scale
SVM	57.89%	96.49%
KNN	93.85%	57.89%
Random Forest	97.36%	75.43%
Decision Tree	94.73%	75.43%
Naïve Bayes	94.73%	93.85%
Adaboost	94.73%	94.73%
XGboost	98.24%	98.24%

3.5. Model Evaluation:

A thorough assessment of breast cancer detection models is performed using key performance metrics such as accuracy, confusion matrix, and classification report. These metrics are critical indicators of a model's ability to make accurate predictions, particularly in a sensitive domain like medical diagnosis. A fundamental metric is accuracy, which represents the ratio of correctly predicted instances to total instances. In medical applications, however, it is critical to supplement accuracy with more detailed insights provided by the confusion matrix. The confusion matrix goes into detail about true positive, true negative, false positive, and false negative predictions. Each quadrant of the matrix provides important information about the model's performance, particularly in distinguishing between malignant and benign tumors. The confusion matrix heatmap, in the context of the tuned XGBoost model, provides a visual representation of the model's performance after fine-tuning. This detailed evaluation not only quantifies prediction accuracy but also identifies potential areas for improvement. The incorporation of these metrics and visualizations contributes to a comprehensive and in-depth evaluation of breast cancer detection models, assisting practitioners and researchers in making informed decisions about model deployment and refinement.

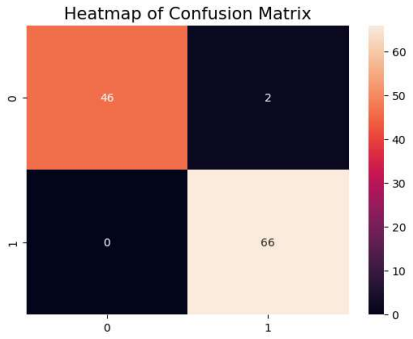

fig. 5 Heatmap of confusion matrix

3.6. Model cross-validation:

Cross-validation emerges as a crucial technique in the pursuit of evaluating the generalization performance of the breast cancer detection model. Cross-validation validates the model's robustness and reliability across different subsets of the dataset, providing a more complete understanding of its generalizability. The identification of overfitting, as mentioned in the context of our model, emphasizes the significance of cross-validation. Overfitting occurs when a model learns the training data too well, capturing noise and fluctuations that may or may not generalize to unobserved data. By systematically partitioning the dataset into multiple folds, training the model on different subsets, and validating its performance on distinct, non-overlapping portions, cross-validation protects against overfitting. Despite the dataset's small size, cross-validation becomes especially useful. The limited data range increases the risk of overfitting, making it critical to evaluate the model's performance across multiple scenarios. Cross-validation provides a more stable and reliable estimate of the model's true generalization capability by averaging its performance over multiple folds. This methodical approach to model evaluation aligns with machine learning best practices, promoting transparency and accountability in the evaluation of predictive models. It enables researchers and practitioners to determine whether the model's performance is truly indicative of its ability to generalize to new, previously unseen data or if changes are required to improve its reliability across a broader range of scenarios.

3.6 Model saving:

The preservation and reuse of machine learning models are critical components of developing a dependable and efficient predictive system. In the context of breast cancer detection, where the model is meticulously tuned and assessed, saving the final XGBoost classifier model with Pickle is critical. The act of saving the model with Pickle not only ensures the model's predictive capabilities are preserved, but it also facilitates its deployment in various contexts. This serialized version of the model encapsulates the complexities of the tuned XGBoost classifier, allowing researchers, clinicians, and stakeholders to seamlessly integrate it into a variety of applications ranging from

healthcare systems to research platforms. After saving the model, the next step is to load the serialized model to make predictions on new, unseen data. This procedure provides a streamlined and efficient method of utilizing the model without the need for extensive retraining. Because breast cancer detection is an evolving field, the ability to load a pre-tuned model improves adaptability by allowing for rapid deployment as new data becomes available or the model is applied to different datasets. The next step is to load the serialized model and make predictions on new, previously unseen data. This procedure provides a streamlined and efficient method of utilizing the model without the need for extensive retraining. Because breast cancer detection is a rapidly evolving field, the ability to load a pre-tuned model improves adaptability.

3.6.1 Pickle:

A Python library, is a versatile tool for serializing and deserializing Python objects, allowing the storage of complex data structures, such as machine learning models, in a compact and portable format. By saving the tuned XGBoost classifier model to a file with Pickle, we encapsulate the model's learned patterns, configurations, and hyperparameters, resulting in a portable artifact that can be seamlessly loaded into future Python environments.

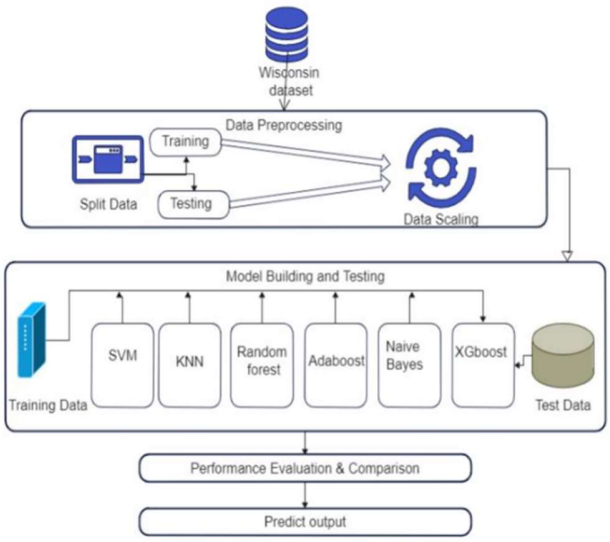

Fig.6 System architecture

3.7. Breast Cancer Detection web application:

The creation of a web application for breast cancer detection represents a significant step forward in the field of medical diagnostics, utilizing machine learning techniques to provide accurate and timely assessments. The application includes a user-friendly interface that collects patient data via essential features such as mean radius, mean texture, mean perimeter, mean area, mean smoothness, and mean compactness. These characteristics serve as critical input parameters for the model, allowing it to analyze and predict the likelihood of cancer, assisting healthcare practitioners in making informed decisions. Following input submission, the application generates comprehensive outputs, including cancer prediction, cancer probability, and tumor classification as benign or malignant. This multifaceted approach ensures a nuanced understanding of the diagnostic results, providing medical professionals with detailed information about the nature and severity of the detected abnormalities. Furthermore, the application goes beyond binary predictions by providing additional information such as the dataset's nearest data points, allowing clinicians to contextualize the patient's data within the larger dataset landscape. In addition to predictive outputs, the web application provides critical performance metrics such as accuracy, confusion matrix, and cancer threshold values. These metrics are invaluable for assessing the model's dependability and efficacy. The accuracy metric quantifies the overall correctness of predictions, whereas the confusion matrix provides a detailed breakdown of true positive, true negative, false positive, and false negative results. The addition of cancer threshold values adds another layer of interpretability, allowing healthcare professionals to fine-tune the model's sensitivity and specificity based on the specific diagnostic requirements. This comprehensive approach to breast cancer detection, encapsulated within a user-friendly web application, not only streamlines the diagnostic process but also improves the interpretability and accessibility of machine learning-driven medical insights.

4. Conclusion:

To summarize, the development of a web application for breast cancer detection that incorporates machine learning methodologies represents a significant step forward in medical diagnostics. The application's use of critical patient features allows for precise predictions, assisting healthcare practitioners in making informed decisions. The comprehensive output, which includes cancer prediction, probability, and tumor classification, provides a detailed understanding of detected abnormalities. Furthermore, the application's ability to provide nearest data points and critical performance metrics, such as accuracy, confusion matrix, and cancer threshold values, enhances its utility in clinical settings. The accuracy metric quantifies the model's overall correctness, whereas the confusion matrix provides detailed diagnostic breakdowns. The incorporation of cancer threshold values provides flexibility in balancing sensitivity and specificity. This comprehensive approach not only improves the interpretability of machine learning results, but it also aligns with the evolving landscape of personalized and data-driven medicine. The application's user-friendly interface ensures accessibility, making it a valuable tool for medical professionals seeking robust diagnostic support. As the field of medical diagnostics embraces technological advancements, this web application exemplifies a forward-thinking solution poised to significantly contribute to breast cancer diagnosis and treatment strategies.

5. References:

[1] Breast Cancer Mammograms Classification Using Deep Neural Network and Entropy-Controlled Whale Optimization Algorithm Saliha Zahoor,* Umar Shoaib,

[2] A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection P. Esther Jebarani; N. Umadevi; Hien Dang; Marc Pomplun

[3] Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review Syed Jamal Safdar Gardezi; Baiying Lei ; Ahmed Elazab [1];

[4] Breast Cancer Detection using Machine Learning Approaches Sri Hari Nallamala, Pragnyaban Mishra, Suvarna Vani Koneru

[5] An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)Author links open overlay panel Meha Desai, Manan Shah

[6] A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection; Ali Al Bataineh

[7] Machine Learning Classification Techniques for Breast Cancer Diagnosis David A. Omondiagbe[1], Shanmugam Veeramani[1] and Amandeep S. Sidhu[2]Published under license by IOP Publishing Ltd

[8] Breast cancer detection using artificial intelligence techniques: A systematic literature review Author links open overlay panel Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Yaman Afadar, Omar Elgendy

[9] Breast cancer detection using deep learning: Datasets, methods, and challenges ahead Author links open overlay panel Nusrat Mohi ud din [a], Rayees Ahmad Dar, Muzafar Rasool, Assif Assad

[10] An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers USMAN NASEEM, JUNAID RASHID, LIAQAT ALI , JUNGEUN KIM , QAZI EMAD UL HAQ , MAZHAR JAVED AWAN , AND MUHAMMAD IMRAN .

[11] Intelligent Hybrid Deep Learning Model for Breast Cancer Detection https://www.mdpi.com/2079-9292/11/17/2767

[12] Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method Ganjar Alfian, Muhammad Syafrudin, Imam Fahrurrozi , Norma Latif Fitriyani , Fransiskus Tatas Dwi Atmaji , Tri Widodo , Nurul Bahiyah , Filip Benes and Jongtae Rhee

[13] BM-Net: CNN-Based MobileNet-V3 and Bilinear Structure for Breast Cancer Detection in Whole Slide Images Jin Huang †, Liye Mei Mengping Long †, Yiqiang Liu, Wei Sun, Xiaoxiao Li, Hui Shen, Fuling Zhou, Xiaolan Ruan, Du Wang, Shu Wang, Taobo Hu and Cheng Lei .

[14] Automated Breast Cancer Detection Models Based on Transfer Learning Madallah Alruwaili * and Walaa Gouda

[15] BreaCNet: A high-accuracy breast thermogram classifier based on mobile convolutional neural network Roslidar Roslidar1, Mohd Syaryadhi, Khairun Saddami, Biswajeet Pradhan , Fitri Arnia , Maimun Syukri7 and Khairul Munadi.