Breast Cancer Detection using Machine Learning Approach: A Comparative Study of Different Methods (December 2025)

Faraz Heydar, Hesameddin Bitarafan Rajabi, and Jinan Fiaidhi, *Senior Member, IEEE*

***Abstract*— Breast cancer remains a critical global health challenge and a major cause of death among women worldwide, with mortality rates continuing to rise annually. Consequently, achieving high accuracy in early prediction and diagnosis is of paramount importance for optimizing treatment plans and improving patient survivability standards. While Machine Learning (ML) has emerged as a research hotspot and a powerful technique to assist in this process, selecting the optimal model requires rigorous validation. This paper presents a comprehensive comparative analysis of five supervised machine learning algorithms—Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (C4.5), and K-Nearest Neighbors (KNN)—applied to the Breast Cancer Wisconsin Diagnostic (WBCD) dataset. Unlike previous studies that often rely on single-split evaluations, this research employs a rigorous methodology integrating Exploratory Data Analysis (EDA), feature scaling via standardization pipelines, and a multi-metric evaluation strategy comprising a 75/25 train-test split and 5-fold cross-validation. Our experimental results demonstrate that the SVM model, when integrated into a scaling pipeline, outperforms other classifiers with a testing accuracy of 97.90%, an F1-Score of 0.9714 for the malignant class, and a robust 5-fold cross-validation average of 96.66%. Furthermore, we introduce a practical Graphical User Interface (GUI) based predictive system to demonstrate the clinical applicability of the trained models. These findings validate the robustness of SVM for this domain.**

***Index Terms*— Breast cancer prediction, machine learning, support vector machine, cross-validation, F1-Score, medical informatics.**

I. Introduction

BREAST cancer is the most frequently diagnosed malignancy among women worldwide and a leading cause of cancer-related mortality [1]. According to statistics released by the International Agency for Research on Cancer (IARC) in December 2020, breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer globally. In the past two decades, the overall number of people diagnosed with cancer nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020 [2].

Anatomically, the breast is composed of a diffusion of tissues ranging from very fatty to very dense tissue. Within these tissues lies a community of lobes, each made of small tubular structures called lobules that include mammary glands. Small ducts join these glands to carry milk to the nipples. Cancer occurs when healthy cells within the chest change, grow uncontrollably, and form a mass or sheet of cells referred to as tumors. These tumors may be benign (non-spreading) or malignant (cancerous), meaning they can grow and spread to other body parts through vessels or lymphatics—a process called metastasis. A breast structure is shown in Fig. 1.

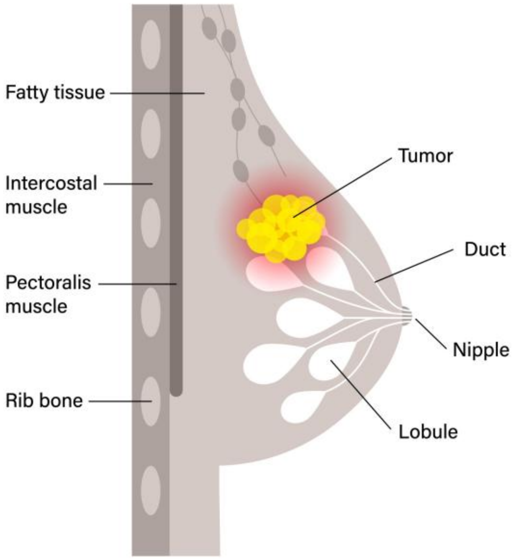


Fig. 1. Breast Structure.

The statistics are sobering. Today, one in 5 people worldwide will develop cancer during their lifetime, and projections

Faraz Heydar is with the Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: fheydar3@lakeheadu.ca).

Hesameddin Bitarafan Rajabi is with the Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: hbitaraf@lakeheadu.ca).

Jinan Fiaidhi is with the Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: jfiaidhi@lakeheadu.ca).

suggest that diagnoses will increase by nearly 50% in 2040 compared to 2020 [2]. Mortality trends are equally concerning; the number of cancer deaths increased from 6.2 million in 2000 to 10 million in 2020, with more than one in six deaths now attributed to cancer [2]. In the United States alone, 2005 estimates predicted 211,240 new cases of invasive breast cancer for women and approximately 1,690 for men [3]. While mortality declined significantly between 1992 and 1998 [3], breast cancer remains the second leading cause of cancer-related death among women [3].

Early diagnosis is paramount; survival rates improve significantly when the cancer is detected at a localized stage [4]. Conventional diagnostic techniques, such as mammography and Fine Needle Aspirate (FNA) biopsy, generate vast amounts of complex data that can be challenging for radiologists and pathologists to interpret consistently. The detection of ductal carcinoma in situ, which accounts for about 88% of cases [3], is often a direct output of mammography screening. However, the sheer volume of unstructured, heterogeneous, and non-standard healthcare data necessitates advanced analytical tools.

Computer-Aided Diagnosis (CAD) systems leveraging Artificial Intelligence (AI) and Machine Learning (ML) have emerged as vital support tools [5]. The successful introduction of information and communication technologies (ICT) in medical practice—specifically Big Data and ML—has revolutionized healthcare by analyzing this data to improve patient care and reduce costs. By identifying non-linear patterns in cytological features, ML algorithms can provide an objective "second opinion," reducing false negatives and inter-observer variability [6]. Data mining algorithms applied in the healthcare industry play a significant role due to their high performance in predicting and diagnosing diseases and making real-time decisions to save lives.

While numerous algorithms have been applied to this task, discrepancies in preprocessing methods (particularly feature scaling) and evaluation metrics often lead to conflicting conclusions regarding the "best" model. Several algorithms, such as Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN), are considered among the most influential data mining algorithms by the research community [2].

This study aims to resolve existing ambiguities and identify the most effective classifier by:

1) Conducting a rigorous Exploratory Data Analysis (EDA) to identify feature correlations and outliers.
2) Implementing and comparing five distinct ML algorithms (LR, SVM, RF, DT, KNN) using a standardized pipeline to prevent data leakage during cross-validation.
3) Evaluating models not just on Accuracy, but also on F1-Score, Sensitivity, and Confusion Matrices to address the cost of false negatives in medical diagnosis.
4) Developing a deployable Predictive System (GUI) to bridge the gap between theoretical modeling and clinical application.

The rest of this paper is organized as follows: Section II introduces related work and methods from previous research. Section III describes the proposed methodology, including dataset description and preprocessing. Section IV and V present the experimental setup and discuss the results in detail, and Section VI concludes the paper.

II. Literature Review

The integration of Artificial Intelligence (AI) into breast cancer diagnosis has transformed the landscape of medical informatics. Over the past two decades, a multitude of studies have investigated the efficacy of Machine Learning (ML) algorithms using the Wisconsin Breast Cancer Dataset (WBCD) [7], aiming to automate the classification of cytological features. Table II in the reference literature often summarizes these extensive efforts, providing a benchmark for new studies.

A. Foundational Studies and Baseline Algorithms

Early research focused on establishing baselines for standard algorithms. Quinlan [8], using the C4.5 Decision Tree algorithm with 10-fold cross-validation, achieved an accuracy of 94.74%. This established decision trees as a viable, interpretable method, although later studies sought to improve upon this accuracy.

A significant milestone was the work by Bennett & Blue [9], who applied Support Vector Machines (SVM) with 5-fold cross-validation, achieving an accuracy of 97.20%. This result is particularly relevant to our study, as it highlights the robustness of SVMs even in earlier implementations and serves as a direct point of comparison for our own SVM results.

B. Evolution of Neural Networks and KNN

Parallel to SVMs, Neural Networks and K-Nearest Neighbors (KNN) were extensively explored. Setiono [10] utilized a Neuro-rule ANN with 10-fold cross-validation, reaching 97.97% accuracy. Similarly, Sarkar & Leong [11] demonstrated that K-NN and Fuzzy K-NN could achieve high accuracies of 98.25% and 98.83% respectively using a 50-50 train-test split. While these methods showed high performance, they often required careful tuning of parameters like the number of neighbors (k) or network architecture.

C. Optimization and Hybrid Approaches

As the field advanced, researchers moved towards optimized and hybrid models to push accuracies beyond the 98% threshold. Polat & Güneş [12] introduced the Least Squares SVM (LS-SVM), which achieved 98.53% accuracy under 10-fold cross-validation. This demonstrated that modifying the standard SVM formulation could yield marginal but clinically significant improvements.

Ensemble methods also gained traction. Seera & Lim [13] proposed the FMM-CART-RF model, a hybrid approach involving Random Forests, which achieved 97.29% accuracy. This aligns with modern trends favoring ensemble methods for their ability to reduce variance and overfitting, a challenge often observed in single decision trees.

D. State-of-the-Art and Complex Models

More recent studies have focused on complex, highly optimized architectures. Akay [14] utilized F-score feature selection combined with SVM (F-score-SVM) to reach a remarkable 99.51% accuracy. Similarly, Abdel-Zaher & Eldeib [15] employed Deep Belief Networks (DBN-ANN) to achieve 99.68%. While these models offer superior accuracy, they often come with increased computational complexity and reduced interpretability compared to standard SVM or RF implementations.

E. Contribution of This Study

Building upon these foundations, this research aims to validate whether standard, efficient implementations of core algorithms—specifically SVM and Random Forest—can still offer competitive performance comparable to these historical benchmarks. By employing a rigorous 5-Fold Cross-Validation strategy and a standardized preprocessing pipeline, we seek to confirm if our SVM implementation can match the 97.20% benchmark set by Bennett & Blue [9] and how it compares to the ensemble approaches like those of Seera & Lim [13]. Additionally, we address the practical gap identified by Jones et al. [6] by developing a deployable Graphical User Interface (GUI) to make these high-performing models accessible for clinical use.

III. Methodology

The main objective of our experiment is to identify the effective and predictive algorithm for the detection of breast cancer, therefore we applied machine learning classifiers Support Vector Machine (SVM), Random Forests, Logistic Regression, Decision tree (C4.5), K-Nearest Neighbors (KNN) on Breast Cancer Wisconsin Diagnostic dataset and evaluate the results obtained to define which model provides a higher accuracy. The proposed architecture is detailed in Fig. 2.

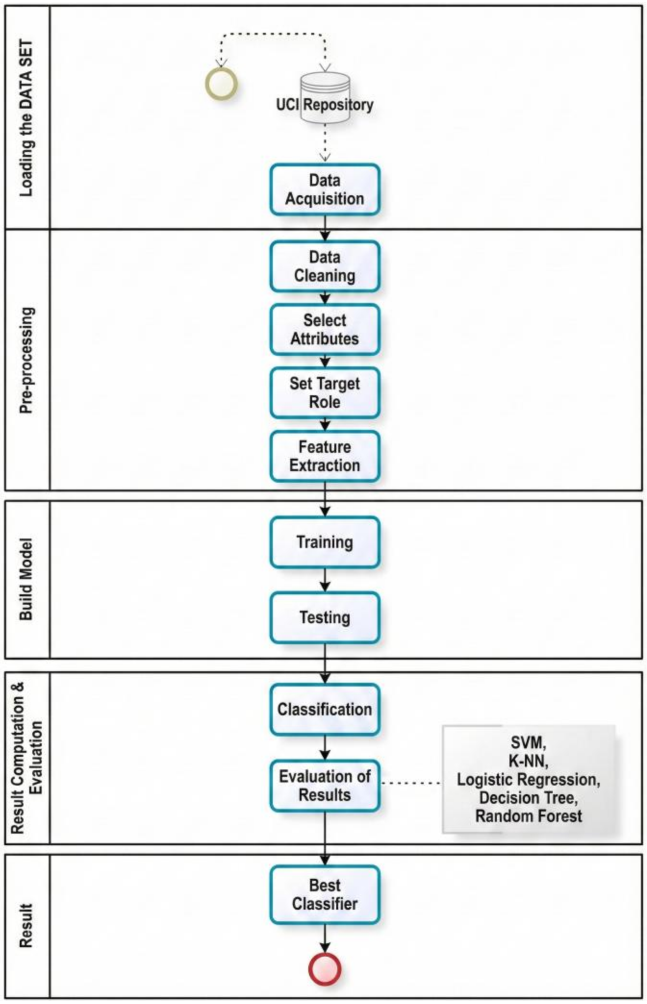


Fig. 2. Process Flow Diagram.

A. Dataset Description

The study utilizes the WBCD dataset [7], containing 569 instances. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Each instance represents a patient and includes 30 real-valued features computed from ten characteristics of cell nuclei: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points, Symmetry, and Fractal Dimension. For each characteristic, the Mean, Standard Error (SE), and Worst (largest) values are recorded. A sample dataset is shown in Table I.

TABLE I
Sample Dataset

mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	...	worst radius	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	...
17.99	10.38	122.80	1001.0	0.11840	0.27760	...	25.38	17.33	184.60	2019.0	0.1622	0.6656	...
20.57	17.77	132.90	1326.0	0.08474	0.07864	...	24.99	23.41	158.80	1956.0	0.1238	0.1866	...
19.69	21.25	130.00	1203.0	0.10960	0.15990	...	23.57	25.53	152.50	1709.0	0.1444	0.4245	...
11.42	20.38	77.58	386.1	0.14250	0.28390	...	14.91	26.50	98.87	567.7	0.2098	0.8663	...
20.29	14.34	135.10	1297.0	0.10030	0.13280	...	22.54	16.67	152.20	1575.0	0.1374	0.2050	...

Class Distribution: The target variable is binary: Malignant (M) or Benign (B). As shown in Fig. 3, the dataset is moderately imbalanced, with 357 benign cases (62.7%) and 212 malignant cases (37.3%).

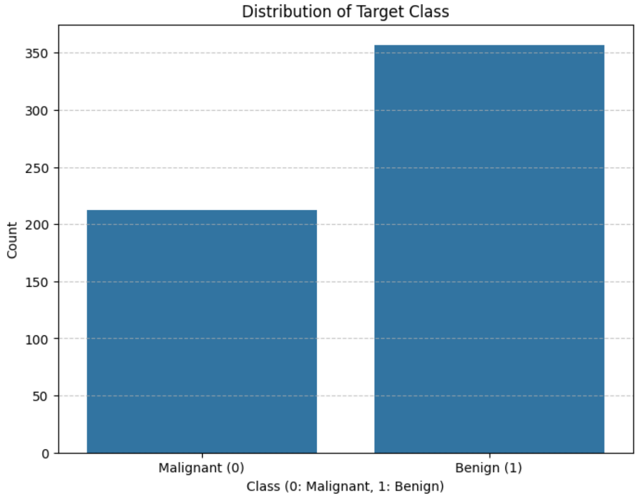


Fig. 3. Class Distribution of the Wisconsin Diagnostic Dataset (Benign vs. Malignant).

B. *Exploratory Data Analysis (EDA)*

To understand feature interactions, we computed the Pearson correlation coefficient matrix. The heatmap presented in Fig. 4 reveals significant multicollinearity. Specifically, geometric features such as *Radius Mean*, *Perimeter Mean*, and *Area Mean* exhibit correlation coefficients exceeding 0.95. This redundancy suggests that dimensionality reduction could be applied in future iterations, though all features were retained for this baseline comparison.

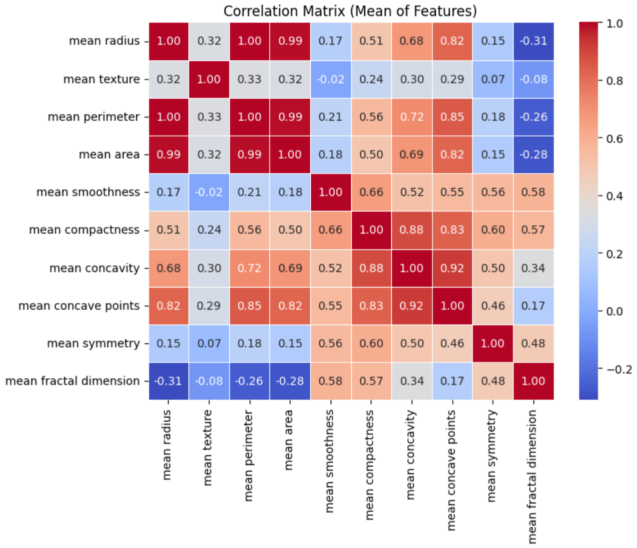


Fig. 4. Correlation Heatmap of the 30 input features. Dark red indicates strong positive correlation, highlighting multicollinearity among geometric features.

C. *Data Preprocessing*

1) Cleaning: The dataset was inspected for missing values. No missing values were found.
2) Encoding: The categorical target 'diagnosis' was mapped to numerical values: $M \rightarrow 0$ and $B \rightarrow 1$.
3) Splitting: The data was split into Training (75%) and Testing (25%) sets using train_test_split with a fixed random seed (random_state=2).
4) Feature Scaling (Crucial Step): Algorithms like SVM and KNN compute distances between data points. Features with large magnitudes (e.g., Area ≈ 1000) can dominate those with small magnitudes (e.g., Smoothness ≈ 0.1). To prevent this bias, we applied StandardScaler to normalize features to a mean of 0 and variance of 1. Crucially, for Cross-Validation, this scaling was embedded within a Pipeline to ensure statistics are computed only on the training folds, preventing data leakage.

D. *Experimental Setup*

The experiments were conducted in a Python environment using Scikit-learn. Five algorithms were trained:

1) Logistic Regression (LR): A linear model utilizing the sigmoid function (max_iter=5000).
2) Support Vector Machine (SVM): Implemented with the Radial Basis Function (RBF) kernel.
3) Random Forest (RF): An ensemble of decision trees.
4) Decision Tree (DT): A standard CART implementation.
5) K-Nearest Neighbors (KNN): Implemented with k=5 neighbors.

Evaluation Metrics included Accuracy, F1-Score (Malignant Class), and Confusion Matrix. Validation was performed using both a single split and 5-Fold Cross-Validation.

IV. Results and Discussion

A. *Quantitative Analysis*

Table II summarizes the comprehensive performance metrics for all five classifiers. The table compares the accuracy on the training and testing sets (75/25 split), the F1-Score specifically for the Malignant class (which is clinically more significant), and the mean accuracy derived from 5-Fold Cross-Validation.

TABLE II
Final Performance Summary Of Machine Learning Models

Algorithm	Training Accuracy	Testing Accuracy	F1 Score (Malignant)	5-Fold CV Mean
SVM	98.36%	97.90%	0.9714	96.66%
Random Forest	100.00%	93.71%	0.9143	95.08%
Logistic Regression	98.59%	97.20%	0.9615	92.79%
Decision Tree	100.00%	92.31%	0.8958	91.91%
K-NN	97.42%	97.20%	0.9623	92.79%

B. *Visual Analysis*

The performance disparity is further visualized in Fig. 5, which contrasts the Testing Accuracy against the Cross-Validation Mean. SVM shows the most consistent high performance across both metrics.

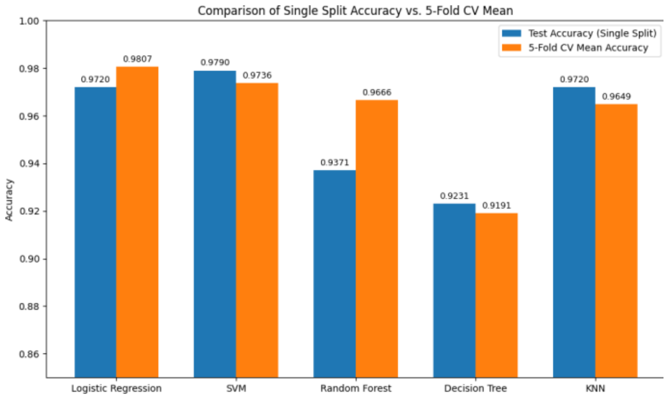


Fig. 5. Comparison of Single Split Accuracy vs. 5-Fold Cross-Validation Mean Accuracy. SVM demonstrates superior stability.

Furthermore, the Confusion Matrices in Fig. 6 provide a granular view of prediction errors. For the Malignant class (labeled '0'), SVM minimized False Negatives (FN) and False Positives (FP) more effectively than tree-based models.

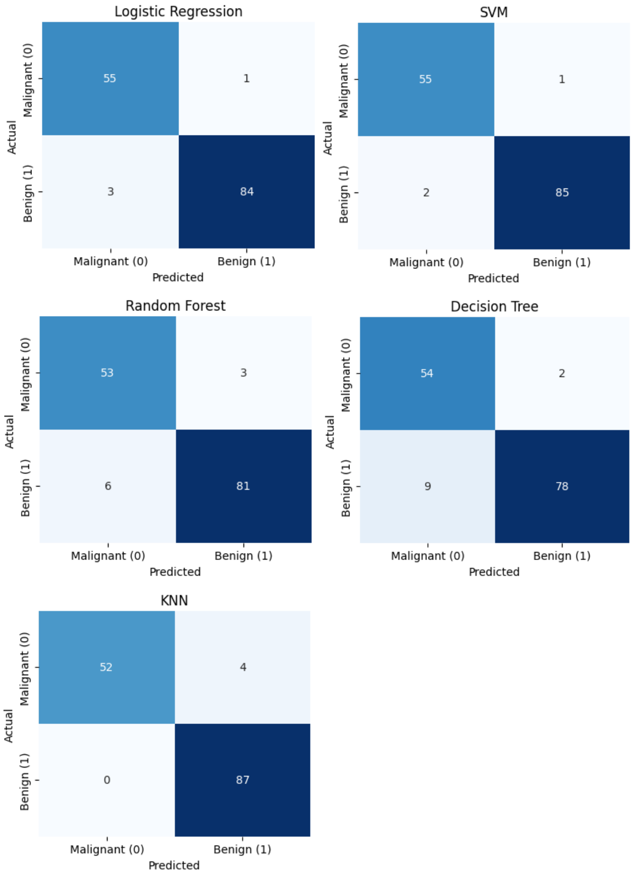


Fig. 6. Confusion Matrices for all five algorithms. The top-left quadrant of each matrix represents True Negatives (Correctly identified Malignant cases).

Confusion matrix is that the summary of predictions on a classification problem. It gives us actuality understanding about the errors that are made by our classifiers and also shows their type.

C. Discussion

The Support Vector Machine (SVM) emerged as the superior model. In the single split, it achieved an impressive 97.90% accuracy and a 0.9714 F1-Score, indicating it successfully balanced Precision and Recall for detecting cancer. Its 5-fold CV score of 96.66% confirms that this performance is robust and not due to a lucky data split.

Random Forest achieved perfect training accuracy (100%), but its testing accuracy dropped to 93.71%, a clear sign of overfitting. While it performed better in Cross-Validation (95.08%), it still lagged behind SVM. Logistic Regression and KNN proved to be strong contenders, with KNN notably achieving a high F1-Score (0.9623) on the test split, but their average performance in cross-validation was lower than SVM.

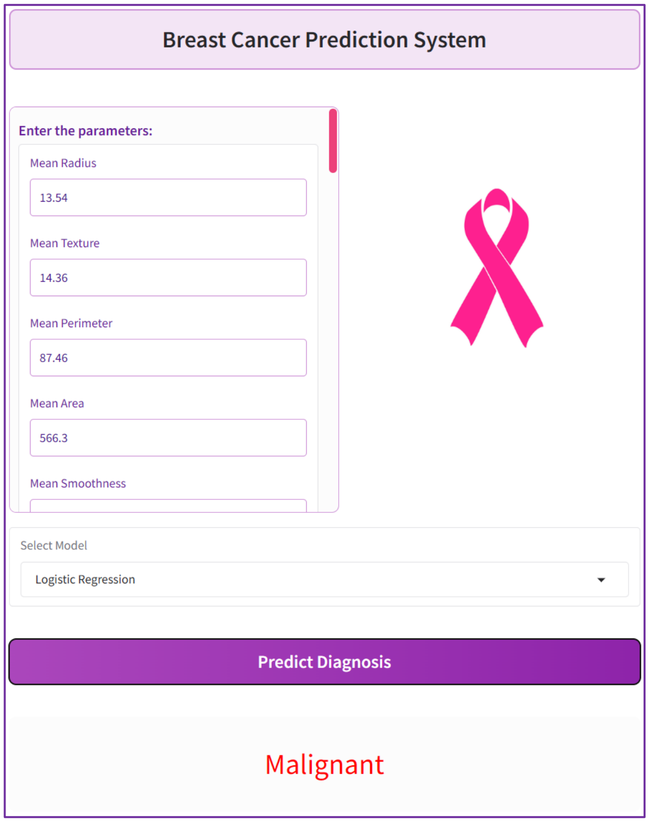
V. Predictive System Implementation

To translate these theoretical results into a practical and accessible tool, a web-based Graphical User Interface (GUI) was developed using the Gradio library within a Google Colab environment. This approach allows for rapid prototyping and easy sharing of the model without requiring local software installation.

As shown in Fig. 7, the system provides a clean, user-friendly interface where clinicians can input the 30 cellular features manually. The interface includes:

1) Input Fields: Thirty distinct numerical input fields corresponding to the dataset features (e.g., Radius Mean, Texture Mean), pre-filled with default values to facilitate testing.
2) Model Selection: A dropdown menu allowing the user to choose between the five trained algorithms (Logistic Regression, SVM, Random Forest, Decision Tree, K-NN) for the prediction.
3) Real-time Prediction: Upon clicking the "Predict Diagnosis" button, the system preprocesses the input data (reshaping it to the correct 2D array format) and feeds it into the selected model.
4) Visual Feedback: The diagnosis ("Benign" or "Malignant") is displayed prominently, color-coded (Blue for Benign, Red for Malignant) to ensure immediate and clear interpretation of the results.

This implementation demonstrates the feasibility of deploying machine learning models as cloud-based diagnostic aids, making advanced AI tools more accessible to medical professionals.



Breast Cancer Prediction System

Enter the parameters:

Mean Radius: 13.54

Mean Texture: 14.36

Mean Perimeter: 87.46

Mean Area: 566.3

Mean Smoothness

Select Model: Logistic Regression

Malignant

Fig. 7. The Web-based Breast Cancer Prediction System built with Gradio. The interface allows parameter input, model selection, and displays the color-coded diagnostic result.

VI. Conclusion

This study successfully implemented and validated a machine learning pipeline for breast cancer diagnosis using the WBCD dataset [7]. Through rigorous evaluation involving feature scaling pipelines and cross-validation, we demonstrated that the Support Vector Machine (SVM) is the optimal classifier, achieving 97.90% test accuracy and 96.66% mean CV accuracy.

The study highlights two critical insights:

1) Tree-based models like Random Forest are prone to overfitting on this dataset without extensive tuning, and
2) Distance-based models like SVM and KNN require proper scaling to reach their full potential.

Future work will focus on hyperparameter tuning (Grid Search) to potentially push the SVM accuracy beyond 99% as suggested by advanced studies in [14][15], and employing feature selection techniques to reduce the redundancy identified in the correlation analysis [2].

Data and Code Availability

The source code used for all experiments and constructed for this study as well as link to the dataset, are publicly available at: https://github.com/FarazHeydar/Breast-Cancer-Detection.

References

[1] T. Arravalli, K. Chadaga, H. Muralikrishna, N. Sampathila, D. Cenitta, R. Chadaga, and K. S. Swathi, "Detection of breast cancer using machine learning and explainable artificial intelligence," *Scientific Reports*, vol. 14, 2024.

[2] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis," *Procedia Computer Science*, vol. 191, pp. 487-492, 2021.

[3] R. B. Mim, A. B. Islam, A. Sattar, and S. Roy, "Breast Cancer Detection using Machine Learning Approach," in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Prague, Czech Republic, 2022, pp. 1-6.

[4] S. Hussain, M. Ali, U. Naseem, F. Nezhadmoghadam, M. A. Jatoi, T. A. Gulliver, and J. G. Tamez-Peña, "Breast cancer risk prediction using machine learning: a systematic review," *Frontiers in Oncology*, vol. 14, 1343627, 2022.

[5] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast cancer detection using artificial intelligence techniques: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 127, 102276, 2022.

[6] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *5th International Symposium on Health Informatics and Bioinformatics*, pp. 114-120, 2010.

[7] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear Feature Extraction For Breast Tumor Diagnosis," *Computer Sciences Technical Report #1131*, University of Wisconsin, 1992.

[8] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996.

[9] K. P. Bennett and J. A. Blue, "A support vector machine approach to decision trees," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 3, pp. 2396-2401, 1998.

[10] R. Setiono, "Extracting M-of-N rules from trained neural networks," *IEEE Transactions on Neural Networks*, vol. 11, no. 2, pp. 437-449, 2000.

[11] M. Sarkar and T. Y. Leong, "Application of K-nearest neighbors algorithm on breast cancer diagnosis problem," in *Proceedings of the AMIA Symposium*, pp. 759-763, 2000.

[12] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digital Signal Processing*, vol. 17, no. 4, pp. 694-701, 2007.

[13] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2239-2249, 2014.

[14] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3240-3247, 2009.

[15] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Systems with Applications*, vol. 46, pp. 139-144, 2016.

Faraz Heydar was born in Tehran, Iran, in 1998. He received the B.S. degree in Computer Engineering - Software from the Islamic Azad University (Central Tehran Branch), Tehran, Tehran, Iran, in 2022. He is currently pursuing the M.Sc. degree in computer science at Lakehead University, Thunder Bay, ON, Canada.
His research interests include artificial intelligence, software engineering, machine and deep learning, and computer vision.

Hesameddin Bitarafan Rajabi was born in Tehran, Iran, in 1998. He received the B.S. degree in Computer Engineering (Software) from the University of Science and Culture, Tehran, in 2022. He is currently pursuing the M.Sc. degree in Computer Science at Lakehead University, Thunder Bay, ON, Canada.
His academic interests include artificial intelligence, computer vision, machine learning, and related areas of intelligent systems.

Jinan Fiaidhi received the Pg.D. degree in computer science from Essex University, Colchester, U.K., in 1983, and the Ph.D. degree in computer science from Brunel University, London, U.K., in 1986.

From 1986 to 2001, she served in academic positions at the University of Technology (Associate Professor and Chairperson), Philadelphia University, Applied Science University, and Sultan Qaboos University. She is currently a Full Professor of Computer Science and the Graduate Coordinator of the Ph.D. program in Biotechnology at Lakehead University, Thunder Bay, ON, Canada, where she has served since 2001. She was also the Graduate Coordinator for the Computer Science M.Sc. program from 2009 to 2018 and is an Adjunct Research Professor with the University of Western Ontario. Her research focuses on mobile and collaborative learning utilizing emerging technologies such as deep learning, cloud computing, calm computing, learning analytics, social networking, crowdsourcing, enterprise mashups, and the semantic web. Her research is supported by major granting associations in Canada, including NSERC and CFI.

Dr. Fiaidhi is a Senior Member of IEEE, a Professional Software Engineer of Ontario (PEng), a member of the British Computer Society (MBCS), and a member of the Canadian Information Society (CIPS) holding the ISP designation. She serves as the Chair of the Big Data for eHealth special interest research group with the IEEE ComSoc eHealth TC and is the Editor-in-Chief of the IGI Global International Journal of Extreme Automation and Connectivity in Healthcare (IJEACH).