Comparative Evaluation of ANN, Decision Tree, and SVM Models for Breast Cancer Classification Using Cytological Features

Rabia Akhtar[1], Nauman Khalid*[2]
[1]School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, United Kingdom.
[2]Govt. Graduate College Samanabad Faisalabad, Pakistan
*Correspondence: drnaumankhalid@gcuf.edu.pk

Citation | Akhtar. R, Khalid. N, "Comparative Evaluation of ANN, Decision Tree, and SVM Models for Breast Cancer Classification Using Cytological Features", IJIST, Vol. 8 Issue. 5 pp 1917-1937, August 2026
DOI | https://doi.org/10.33411/IJIST/1982
Received | July 25, 2026 **Revised |** Aug 06, 2026 **Accepted |** Aug 06, 2026 **Published |** Aug 15 2026.

Breast Cancer (BC) is still one of the deadliest causes of death for women, and timely and precise diagnosis is essential for achieving better clinical results with the use of reliable machine learning-based decision support. All three models, namely the optimized Artificial Neural Network (ANN), Decision Tree (DT) and Support Vector Machine (SVM), are evaluated in relation to their predictive performance in a uniform experimental set-up and the best model was identified to classify BC. The data set used to build and test the three supervised machine learning models is called the BC Wisconsin (Original) data set. Data preprocessing techniques such as missing value imputation, target encoding, feature scaling and stratified data partitioning are used to improve the generalization and predictive ability of the model, respectively, whilst hyperparameter optimisation and cross-validation are adopted to obtain the best model. Results and Discussion showed that the SVM model performed best overall in terms of classification accuracy (96.19%) with a balanced precision of 94.44% in terms of precision, 94.44% in terms of recall, 94.44% in terms of F1-score and 98.83% in terms of AUC, which means its performance was good in terms of generalisation. The ANN model also showed promising results with a high accuracy of 95.24% and the highest AUC of 99.03%, indicating that the model is suitable for modelling nonlinear relationships. The highest recall (97.22%) was achieved by DT model, indicating that this model had excellent sensitivity in detecting malignant cases, but the low precision (81.40%) led to a relatively high false positive rate. The results suggest that model optimisation and careful choice of the algorithm significantly impact the reliability of diagnosis. From the evaluated models, the optimized SVM proved to be the most robust model with fair performance, and it can be used as an effective decision support tool for the classification of BC.

Keywords: Breast cancer classification; Machine learning; Support Vector Machine; Artificial Neural Network; Decision Tree

Introduction:

BC remains the most prevalent cancer in women and is one of the major causes of cancer-related death in the world. In 2022 the estimates for new cases and deaths worldwide were 2.3 million and 666,103 respectively and highest in Asia, Europe and Northern America [1][2]. However, the mortality-to-incidence ratios do tell another story: they are still the poorest in Africa and parts of Asia, not so much because the disease behaves differently biologically, but because access to screening and diagnosis is also different there. Stage at diagnosis is a key determinant of survival: A cancer with a localized stage at diagnosis has a much better prognosis than a cancer that has spread [3]. The avoidance of time and uncertainty in diagnosis has the same value in reducing mortality as any treatment advance.

The usual diagnostic pathway is still clinical examination, imaging (mammography, ultrasound, MRI) and histopathology, in which pathologists examine the cytological characteristics of the cells, such as clump thickness, uniformity of cell size and adhesion to the margins to distinguish between a benign and a malignant tumour. It does work, but not completely: inter-observer agreement among pathologists is about 75% for some diagnostic categories [4], and in areas where there is not a lot of screening, or where there are fewer specialists, the delays are greater, and the cancers are more likely to be diagnosed late [5]. Machine learning (ML) has been touted to solve all sorts of problems that are subjective, people are scarce, turnaround time is long, and there's a lot of research on it.

The bulk of that research is divided into two categories. One is imaging, where systems trained on mammograms have matched or even surpassed expert radiologists in some tests [6], a trend broadly supported by a later systematic review of test-accuracy studies for AI in breast screening [7]. It is the other camp that is closer to what this study does, and here a standard testbed is the "Wisconsin Breast Cancer Dataset" (WBCD) [8]. [9] achieved accuracy of 97.13% with SVM in comparison to Naïve Bayes, k-NN and C4.5 on this dataset; Chaurasia, Pal and Tiwari [10] recorded 96.84% with SVM using an RBF kernel; [11] further showed the superiority of SVM over other classifiers tested; and [12] with a hybrid ANN–deep belief network achieved accuracy of 98.14%. So, there is no question that they can support very high accuracy.

But beyond the reported accuracy figure, it's not clear which is the better choice for clinical use. Many of those comparisons include only one accuracy metric without any details about different cross-validation, hyperparameter tuning or the specific features being used to make the prediction which makes it difficult to discern how much of this accuracy is likely to persist on another data split, let alone another data set. What's also very surprising is that there is little direct engagement with the accuracy–interpretability trade-off. ANN and SVM have shown good performance but are black box models; DT has traded off a bit of accuracy for a model a clinician can reason on and audit [13]. This trade-off isn't merely theoretical because there is a difference in the cost of false positive and false negative rates in a diagnostic context, and the stability of a model as it's trained with less information is arguably as important for real-world deployment as its best performance on a held-out test set.

Recently, comparative studies have also continued to show high accuracy for SVM- and ensemble-based classifiers on the WBCD and related breast cancer datasets [14][15][16][17]; however, few of these studies have correlated the accuracy of the classifiers with an explanation of the impact of the feature on the prediction. To address this, Explainable AI (XAI) methods like SHAP [18] and LIME [19] have started to be used specifically to improve breast cancer classification, providing insights into which features or characteristics are most important for a particular prediction, beyond just the global order of feature importance [20][21][22][23].

In this study, an effort is made to bridge this gap by applying ANN, DT, and SVM to WBCD in the same pipeline, which includes the same preprocessing, the same cross validation

scheme, and the same tuning process, so that any differences in performance are due to differences in the models themselves and not due to differences in how the models are evaluated. It also goes beyond accuracy and monitors precision, recall, F1-score, AUC and learning curve behaviour for a more comprehensive understanding of the potential clinical performance of each model. In particular, this paper makes the following contributions: it performs a controlled, apples-to-apples comparison between three structurally distinct ML models, instead of comparing models constructed in different ways and across different studies; it evaluates interpretability and stability not as a side effect or an indirect consequence of the model's performance, but as a first-class outcome of the model; and it provides a direct answer to which model provides an optimal tradeoff between performance and clinical usability in this type of diagnostic task.

Literature Review:

With the recent advances in the field of AI, ML techniques have been developed for BC classification to a great extent. Various supervised learning algorithms have been explored with the use of benchmark datasets like WBCD to achieve the accuracy of the diagnosis and to minimize the reliance on manual interpretation. While many studies have reported good prediction accuracy, variations in preprocessing, feature engineering, hyperparameter optimisation, validation and evaluation methods make it hard to measure the superiority of one algorithm over another one. A thorough review of recent studies is thus necessary to highlight the current accomplishments and outstanding research questions.

The authors in [11] used the dataset WBCD to perform a comparative study of five supervised learning techniques: SVM, DT, Random Forest (RF), Logistic Regression (LR) and K-Nearest Neighbour (KNN) on the data. The experimental results showed that SVM outperforms the other classifiers with the highest classification accuracy (96.12%) and AUC (95.6%). The study identified that margin-based classifiers are promising in the context of structured clinical data [24]. However, the experimental design was based mainly on a single train–test partition, and less on hyperparameter optimisation and feature contribution. Therefore, the strength and repeatability of the performance reported is not guaranteed in various validation situations.

[25] furthered the comparative analysis by adding conventional ML algorithms and deep learning models. They have used the Extended Wisconsin Breast Cancer dataset to evaluate network models such as Naïve Bayes, DT, LR, RF, AdaBoost, and VGG16 convolutional neural network. The highest classification accuracy (98.9%) was obtained using the RF model, which was followed by the VGG16 model with 98.2%, thus demonstrating the effectiveness of ensemble and deep learning models on breast cancer prediction. The study, however, was mainly concentrated on the accuracy of the predictions with little discussion of feature importance, parameter optimisation, and model interpretability. Also, the low accuracy of DT was not explained in detail, although it is widely used in medical decision making.

Gupta et al. presented a hybrid deep learning algorithm for breast cancer classification, which is an integration of the ANN and Deep Belief Network [12]. In their methodology, they used several advanced preprocessing techniques that preceded the training of the model; they used KNN-based missing value imputation, Synthetic Minority Oversampling Technique (SMOTE), and feature correlation analysis. The proposed hybrid architecture was successfully tested and has an accuracy of 98.14% which is better than some of the reported models. While the framework showed good predictive ability, the complexity of the architecture led to a higher computational expense and reduced transparency of the model, thus hindering its use in resource-limited healthcare settings [26][27].

To examine how the architecture of a neural network affects the prediction of BC, Nasien et al. systematically increased the number of hidden neurons in a back propagation neural network [28]. The results demonstrated that the optimized ANN attained a high

accuracy of 96.93% which outperformed the other conventional ML algorithms tested in this study. These encouraging results were obtained, but evaluation of the models was mainly done with respect to the overall classification accuracy, while only a limited use was made of precision, recall, F1 score and receiver operating characteristics (ROC) analysis. Moreover, no cross validation was used, limiting the model generalization assessment across the different data partitions.

In recent years, several studies have been conducted on hybrid and ensemble learning methods to further enhance diagnostic performance. These methods usually enable better classification accuracy than classical ML algorithms but tend to have increased complexity and decreased interpretability. For clinical decision-support systems, interpretability also plays a role, as health care workers need to understand and validate prediction results and need transparent models. As a result, there is no guarantee that the most sophisticated deep learning architectures are the most feasible, even if there are slight gains in prediction accuracy.

These trends are further supported by more recent comparative studies on various imaging modalities and feature-selection strategies. [14] tested ANN and SVM on breast ultrasound images from a hospital dataset along with a public dataset where ANN marginally outperformed SVM in terms of accuracy (87.78%) and sensitivity while SVM outperformed ANN in terms of specificity, emphasizing the importance of the accuracy–sensitivity trade-off that applies to other data modalities too. On image data of breast cancer, [15] compared SVM with KNN, logistic regression, random forest and DT and once again SVM was the best performer. Similarly, [16] compared feature-selection methods with DT, RF, SVM, XGBoost and ANN on the Wisconsin Diagnostic dataset and additionally reported the training time taken by each of the classifiers and found that SVM had the highest AUC with the least training time and ANN had the longest training time. [17] went a step further by performing SHAP-based explainability on a primary clinical dataset consisting of Bangladeshi patients—showing that XAI can be incorporated into a supervised breast cancer classification pipeline without significantly adding to model complexity. All these studies contribute to the methodological perspective of present research, i.e., reporting accuracy along with interpretability, and, if possible, computational costs.

In general, previous studies show that SVM, ANN, RF and hybrid learning models all consistently yield high classification accuracy on the WBCD. There are, however, some features that are common to many of the existing studies. The accuracy is often the only metric used for evaluating the results of many investigations, neglecting the precision, recall, F1 score, AUC, and the false-positive rate, which are essential in medical diagnosis. Additionally, variations in the preprocessing techniques, data splitting strategies, hyperparameter tuning, and evaluation processes make it difficult to make direct comparisons between algorithms.

Table 1. Comparison of Recent Machine Learning Studies for Breast Cancer Classification

Ref	Methods, Dataset and Key Points	Critical Analysis
[11]	**BC Detection** • WBCD with 699 cases • 75:25 train-test split **Algorithms**: RF, LR, T(C4.5), SVM, K-NN **Metrics:** Precision, recall accuracy, AUC **Key Results:** ○ **DT(C4.5)** Accuracy: 95.1%, AUC: 94.5%, F_1 score: 0.94	**Strengths:** Thorough comparison of various algorithms; utilization of various evaluation metrics; well-defined methodology. **Limitations:** The data sets were not cross-validated, the details of the hyperparameter optimization are limited, there is no feature importance analysis, and the train-test split was basic. **Research Gap:** The study did not explore in-depth the features and/or approaches that resulted in the high accuracy and AUC achieved with SVM. No other hybrid model or feature

	○ **RF** Accuracy: 96.5%, AUC: 96%, F_1 score: 0.95 ○ **SVM** Accuracy: 96.12%, AUC: 95.6%, F_1 score: 0.95 ○ **K-NN** Accuracy: 93.7%, AUC: 95.2%, F_1 score: 0.91 ○ **LR** Accuracy: 95.8%, AUC: 94.7%, F_1 score: 0.94	optimization method was considered for further improving predictive accuracy. The assessment of real-world reliability was not carried out with the same rigor as the accuracy metrics.
[25]	**BC Detection** • Extended WBCD with 32 features **Algorithms**: RF, LR, DT, VGG16, Naïve Bayes, AdaBoost, **Metrics:** Accuracy, recall **Key Results:** ○ **RF** Accuracy: 98.9%, F_1 score: 0.99 ○ **DT** Accuracy: 83%, F_1 score: 0.75 ○ **VGG16** Accuracy: 98.2%, F_1 score: 0.982 **AdaBoost** Accuracy: 83%, F_1 score: 0.75 ○ **LR** Accuracy: 97.7%, F_1 score: 0.97 ○ **Naïve Bayes:** Accuracy: 74%, F_1 score: 0.76	**Strengths:** Traditional and deep learning methods; robust cross-validation; enlarged dataset with more features. **Limitations:** Poor DT performance contradicts previous research without explanation; insufficient model optimization information; no feature significance analysis despite enlarged feature set. **Research Gap:** The study results were presented using six classifiers including VGG16, RF, but it does not provide an impact of features or optimization of hyperparameters. The comparison doesn't provide an explanation of why the RF model did better than others. Real world diagnostic situations and clinical interpretation were also not considered.
[12]	**BC Classification** • Hybrid ANN+DBN model with 5 layers and ReLU activation • SMOTE for balancing class • KNN imputation for missing values • Feature correlation analysis **Key Results:** Accuracy: 98.14%	**Strengths:** Innovative hybrid architecture, pre-processing, different assessment metrics, correlation analysis of features. **Limitations:** Very complex design, significant computational cost; small performance gain (1-2%) as design complexity increases; some debate about the meaning of this design. **Research Gap:** The hybrid ANN+DBN model has achieved excellent accuracy, but it was affected by the layer design, the activation functions (e.g., ReLU) and dropout. Moreover, there was no comparison between the deep learning models with common ML techniques. Practical reliability, sensitivity and specificity were not clinically validated.
[28]	**BC Prediction** • Extended WBCD • ANN with backpropagation	**Strengths:** Advanced preprocessing and hybrid learning architecture.

	• Single hidden layer architecture contains 2-15 neurons • 80:20 train-test split **Key Results:** Accuracy: 96.93%, Error rate: 3.07%	**Limitations:** High computational complexity and reduced interpretability. **Research Gap:** Limited comparison with simpler ML models under identical experimental settings.

Research Aim and Objectives:

To conduct a comprehensive evaluation and comparison of the performance of different ML models for BC detection.

To encourage the use of the best ML models in the clinical setting for improved diagnosis.

To assess the predictive capability of the specified machine learning models such as ANN, DT and SVM.

To find out the important factors and data preprocessing techniques which help further enhance the prediction accuracy of these models.

To assess the reliability of the ML model and its possibility for clinical use based on its accuracy, interpretability, and operability.

Novelty of the Study:

The present study addresses several limitations identified in previous comparative studies by establishing a standardized and reproducible evaluation framework for breast cancer classification using machine learning techniques.

Standardized comparison between ANN, DT and SVM with the same preprocessing pipeline, training/testing strategy and evaluation protocol.

Systematic hyperparameter optimization to compare models fairly and reproducibly.

Beyond relying on just one metric, comprehensive evaluation (accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices and misclassification analysis), with statistical significance testing to confirm the reliability of performance differences between models.

Discussions about the interpretability of the model, its clinical applicability and practical considerations in deployment for future AI-aided breast cancer diagnosis.

Material and Methodology:

Figure 1 presents the overall workflow adopted in this study for BC classification. The proposed framework is divided into five stages: Dataset acquisition, Exploratory data analysis, Data preprocessing, Model development, and Performance evaluation. Every stage had been carefully thought out to guarantee objective comparisons of models and complete assessment of their performance within a standardized experimental framework and reliable data preparation.

Data Acquisition:

The experiments were done with the BCWD provided by the UCI ML Repository. Due to its high-quality data, representation of cytological characteristics and suitability for the evaluation of supervised ML algorithms, this benchmark data set has been widely used in BC diagnosis research. There are 699 patient records, each representing a fine-needle aspiration biopsy (FNA) of a breast mass. All observations are given as either benign (2) or malignant (4) by pathological diagnosis.

The dataset comprises 699 observations, with 11 attributes, including a non-predictive ID number, nine integer attributes (scored 1-10) for cytological features, and a binary class (2 = benign, 4 = malignant); these integer attributes are: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses. The WBCD was chosen for several reasons: it serves as a baseline in this literature and can be directly compared to previous studies [29]; it has been

well curated, with few missing data, limited to one feature; and it is structured in tables, allowing for a controlled comparison of ANN, DT, and SVM without the added confounder of image-based feature extraction.

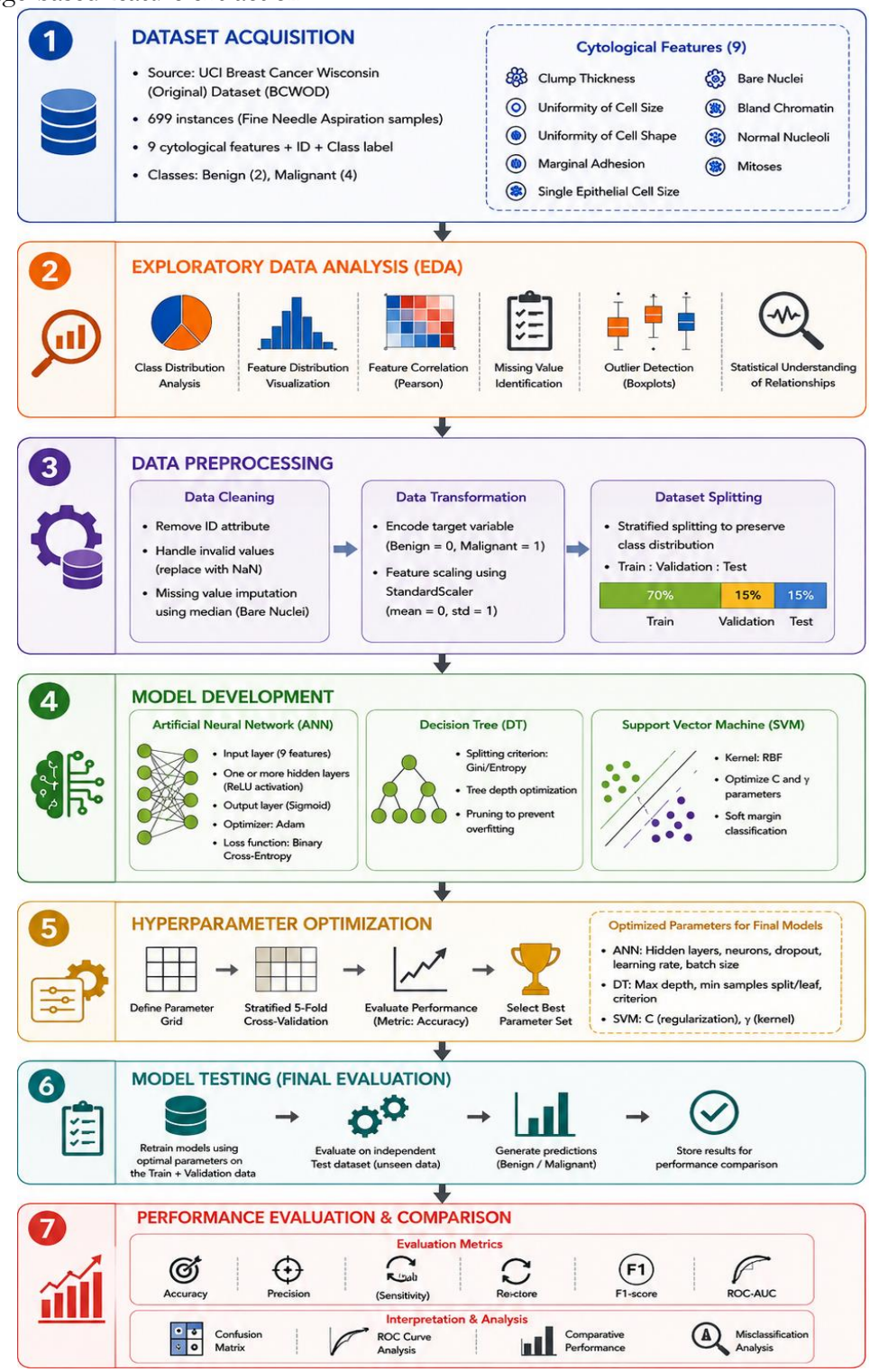

Figure 1. Proposed Project Model

Exploratory Data Analysis (EDA):

Before developing the models, EDA was conducted to investigate the dataset's attributes and to check for any data-related problems that might impact model accuracy. A class distribution, feature relationships and data completeness have been assessed.

The dataset is moderately imbalanced with 458 benign cases (65.52%) and 241 malignant cases (34.48%) as seen in Figure 2. The imbalance was not large, but it was considered when evaluating the model by using a variety of performance metrics other than the accuracy of the model, such as precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC).

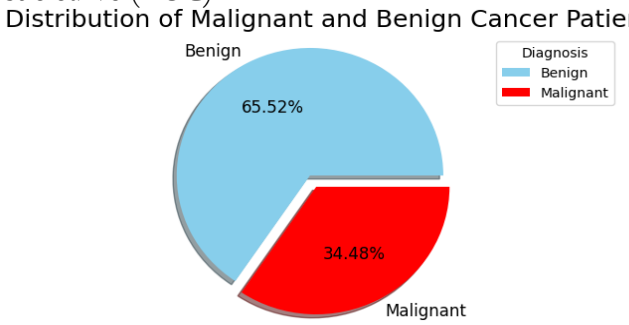

Figure 2. Distribution of Benign and Malignant Patients in Pie Chart

Most variables (marginal adhesion, uniformity of cell size, uniformity of cell shape, bare nuclei) were found to be right-skewed and have most values between 1 and 10 on the scale (see Figure 3). This pattern makes good clinical sense: benign samples tend to fall into a relatively compact area in the lower range of scores on almost all of the nine features, whereas the malignant samples tend to be more scattered toward the higher-end of the scores on almost all the nine features which correlates with the higher extent of cellular irregularity seen in malignancy.

Marginal adhesion (normal nucleoli; mitoses highest at 120), single epithelial cell size, normal nucleoli, and marginal adhesion were found to have outliers (inspected using a boxplot; see Figure 4) in four features: marginal adhesion (120), single epithelial cell size (77), normal nucleoli (60), and mitoses (120). These were not discarded but retained; in a diagnostic context, such extreme values may be a true indicator of aggressive malignant presentation rather than a measurement artifact, and their removal can lead to loss of the signal that a classifier might require the most.

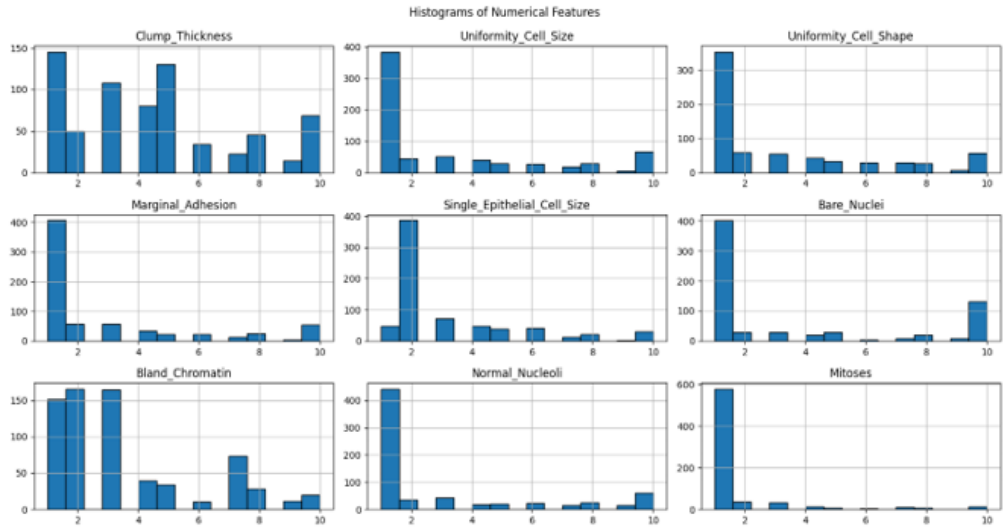

Figure 3. Plot for All Nine Feature Distributions

Data Preprocessing:

Data was preprocessed using a structured pipeline to ensure their quality and exclude potential sources of bias in the pre-processing phase before training the model. The first one, the identifier attribute, was dropped as it contained no predictive information useful for breast cancer classification. The values corresponding to the Bare Nuclei feature that were invalid

were replaced with nulls and then the nulls were imputed with the median of the attribute. Median imputation was chosen because it maintains the distribution of the data, but it is still resistant to outliers.

The diagnosis labels were converted to a binary representation for binary classification purposes; benign tumours (0); malignant tumours (1). Using stratified sampling the data set has been split into three subsets of 70% training, 15% validation, and 15% testing, with the same class distribution maintained in each of these subsets.

Finally, feature standardization was performed using StandardScaler, where the scaling parameters were only calculated on the training set and the same values were used on the validation and testing sets. This procedure helped prevent information leakage and ensured that all the predictor variables had an equal contribution during model optimization, especially for algorithms that are sensitive to the magnitude of the features.

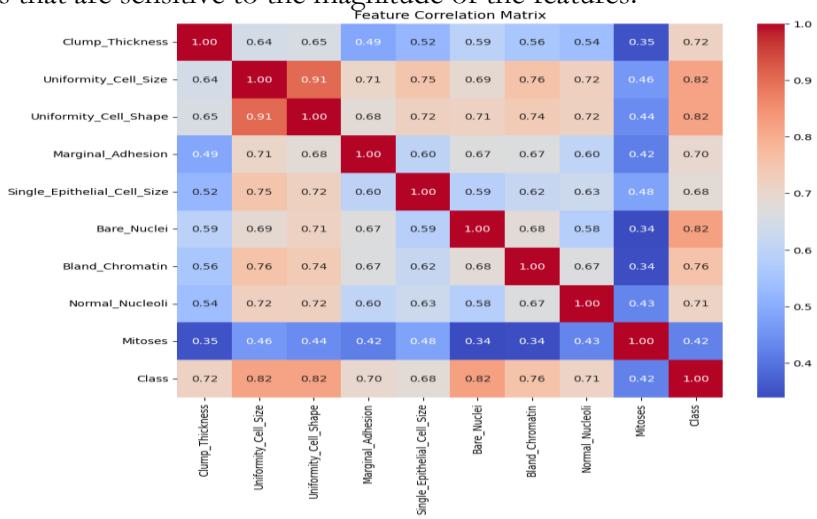

Figure 4. Plot for Pearson correlation matrix of the predictor variables.

Ethical Considerations:

The WBCD was downloaded from the UCI ML Repository, and this study used it for its analysis. The information contained in the data set is fully anonymized and personally identifiable information is not included. The study was entirely secondary and de-identified data openly available for academic use and no further ethical approval and informed consent was required.

Model Development:

Three supervised machine learning techniques, each with a different classification paradigm were chosen to be compared in the evaluation: ANN, DT and SVM. The selected models have a wide range of learning strategies, computational complexity, and interpretability, allowing for a comprehensive evaluation of the models' suitability for breast cancer diagnosis.

Artificial Neural Network (ANN):

ANNs have been shown to learn intricate non-linear relationships from biomedical data with great ability. A feed-forward neural network model was designed in this study to classify breast tumours as benign or malignant based on cytological features extracted from the WBCD.

The ANN architecture consisted of an input layer corresponding to the nine cytological features, one or more hidden layers optimized through hyperparameter tuning, and a single-neuron output layer with a sigmoid activation function for binary classification. Rectified Linear Unit (ReLU) activation was employed in the hidden layers. Model optimization was performed using the Adam optimizer with candidate learning rates of 0.001 and 0.01, while the binary cross-entropy loss function was adopted for training. A manual grid

search was conducted over 60 combinations of hidden-layer configuration ((16,), (32,), (64,), (64,32), (64,32,16)), dropout rate (0.0, 0.2, 0.4), learning rate (0.001, 0.01), and batch size (8, 32), each evaluated on a held-out validation set (15% of the data) rather than via k-fold cross-validation, with Early-Stopping applied identically across all trials.

The learning behaviour of the optimized ANN is shown in Figure 5, with the training and validation accuracy climbing quickly until the epochs reach 70, then leveling off, and the training and validation loss dropping steadily throughout the optimization process. The close agreement between the training and validation curves is a sign that the architecture being used is learning discriminative feature representations without overfitting.

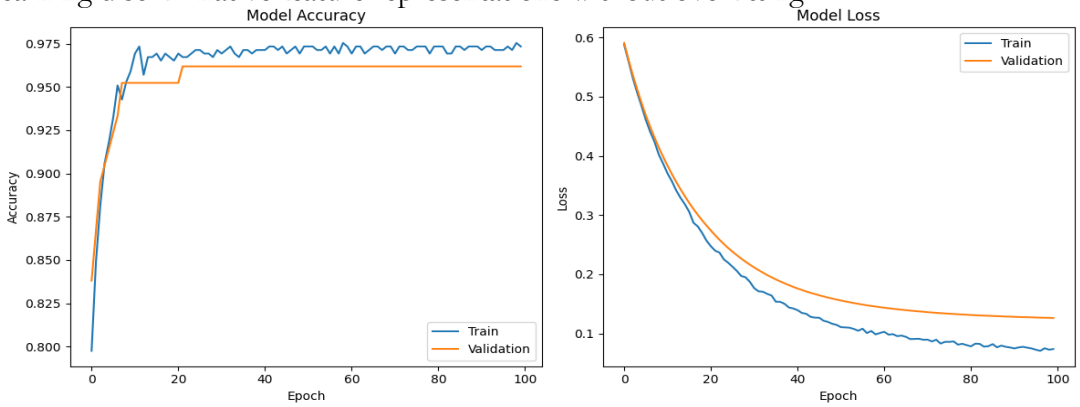

Figure 5. Training and validation accuracy and loss curves of the optimized ANN.

Decision Tree (DT):

DTs were chosen as they have a clear decision-making process and are used in many medical decision support systems. Unlike black-box models, DTs provide explicit classification rules that are easily understood by a clinician and thus appealing for explainable AI [30].

DT is a recursive partitioning algorithm that successively splits the feature space to obtain homogeneous subsets at each node by choosing the feature which has the largest class separation. To control the complexity of the tree, some hyperparameters were optimized: splitting criterion, max depth, minimum number of samples needed for splitting a node, minimum number of samples at a leaf, and the strategy for selecting features. The hyperparameter optimization ensured that the trees did not grow too large but still performed well in terms of classification.

DTs typically have slightly less predictive accuracy than more complex nonlinear classifiers, but they have a clear advantage when interpretability is crucial for a clinical application in which transparency in decision making is important.

Support Vector Machine (SVM):

In high dimensional biomedical classification problems, SVMs always have been found to be very effective. The algorithm aims to find an optimal separating hyperplane between different classes that minimizes classification errors and maximizes the margin between the classes. Since the cytological features of the breast tumours have nonlinear relationships between the features, Radial Basis Function (RBF) kernel was chosen to map the original feature space to higher dimensional space in which the two diagnostic classes can be easily separated.

To find the optimal values for the regularization parameter (C), kernel coefficient (γ), kernel function and class weighting strategy, a grid-search optimization was performed. The optimization process significantly enhanced the model's generalization ability without leading to over-complexity. The resulting classifier showed excellent predictive performance in all evaluation metrics, thus validating the suitability of SVMs for breast cancer diagnosis in structured cytological information.

Hyperparameter Optimization:

The predictive performance of each ML model is highly dependent on the selection of appropriate hyperparameters. So, systematic hyperparameter optimization was performed for each individual classifier. The classifier was trained in 4 folds and tested on the remaining fold for all the parameter combinations. This process was repeated until each fold has been used as a validation set once. The combination of the parameters that obtained the highest mean validation accuracy was chosen for final model construction.

Optimization for the ANN involved the architecture of the hidden layer, dropout rate, learning rate and batch size. DT optimization was done based on splitting criterion (Gini and Entropy), maximum depth, minimum samples required for splitting the nodes, minimum samples per leaf and maximum selection strategy. In the same way, the SVM optimization using GridSearchCV with StratifiedKFold (five folds) considered various kernel functions, different values for the regularization parameter (C), kernel coefficient (**γ**), and class weighting configuration. After optimization of the parameters, each classifier was re-trained with its optimum parameter setting and then tested with the independent test data set. The same optimization method was used for each of the models to maintain the objectivity and experimental consistency of the results of the comparisons among classifiers.

Table 2. Final Hyperparameter configuration of the evaluated ML models

Model	Optimized Parameters	Final Selected Values
ANN	Hidden Layers, Dropout, Learning Rate, Batch Size	Hidden Layers: 64–32–16; Dropout: 0.20; Learning Rate: 0.001; Batch Size: 32
DT	Criterion, Maximum Depth, Minimum Samples Split, Minimum Samples Leaf	Entropy; Depth = 5; Split = 2; Leaf = 1;
SVM	Kernel, C, Gamma, Class Weight	RBF Kernel; C = 10; Gamma = 0.01; Class weight: Balanced

Experimental Environment and Reproducibility:

All experiments were performed in Python, with the Jupyter Notebook development environment. Matplotlib, Seaborn, and Plotly libraries were used to visualize the data, and Pandas and NumPy were employed for data preprocessing and statistical analysis. The DT and SVM classifiers were developed using Scikit-learn library and the ANN was created using the TensorFlow–Keras deep learning framework. The experiments were carried out on Intel Core i7 processor, 16 GB RAM, Windows 11, Google GPU.

For better reproducibility, a random state of 42 was provided to split the dataset and to carry out the conventional machine learning models. Seeds were also set for the TensorFlow random number generator prior to the optimization of the ANN, to guarantee that the model is initialized and trained in the same way every time. The ANN was trained using Adam optimizer and binary cross-entropy loss function with a maximum of 200 epochs during hyperparameter optimization. Early Stopping was designed to monitor the validation loss and automatically recover the best model weights to prevent overfitting. Each classifier was re-trained with the optimum set of parameters and tested on the separate test set after hyperparameter optimization.

Performance Evaluation:

Multiple complementary performance metrics were used to assess the diagnostic capability of the optimized machine learning models and to give an overall assessment of the models. The accuracy of the classification and the reliability of the prediction of the malignant samples were measured by classification accuracy and precision, respectively. For clinical diagnosis, where false prediction of malignancy may result in a delay of treatment, Recall (Sensitivity) was a major evaluation of the model's ability to correctly predict malignant tumours.

The F1-score was used to calculate the harmonic mean of precision and recall gaining a balanced measure of predictive performance when the classes are imbalanced. Additionally, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was computed to measure the capacity of each classifier to discriminate between benign and malignant tumours at a range of classification thresholds.

Confusion matrices were also created for all optimized models on the independent testing data to further understand the behavior of classifiers. These matrices show how many true positives, true negatives, false positives, and false negatives each classifier has, giving insight into both the strengths and weaknesses of each classifier. Another visualization technique was used to demonstrate the discrimination power of the classifiers and to compare the predictive performance: ROC curve.

Finally, the performance of the ANN, DT and SVM was summarized in a comparative performance chart where accuracy, precision, recall, F1 score and ROC-AUC values are presented in a single graph. This visualization helps to make direct comparisons among the classifiers evaluated and to provide objective interpretation of the experimental results.

Results and Discussions:

The performance of the three classifiers, ANN, DT, and SVM, was compared using the independent testing dataset. The accuracy, precision, recall, F1-score, and ROC-AUC were used to evaluate the performance of the model. Furthermore, the confusion matrices and ROC curve were analysed to have a comprehensive understanding of each classifier's diagnostic capability. All models are trained and evaluated within the same experimental setting to ensure that the differences in performance in the experiment are due to the learning algorithm rather than some differences in data preprocessing or in the evaluation methodology.

The classification behaviour of the three optimized models is presented in the confusion matrices displayed in Figure 6. The ANN classified 67 benign cases correctly and 33 malignant cases correctly with only 5 misclassifications. Two non-cancer samples were misdiagnosed as cancer and three cancer samples were misclassified as non-cancer. The results show that ANN was able to classify with high accuracy with a balance between the number of false positive and false negative predictions.

The classification pattern of the DT was different. The model was able to correctly classify 61 benign and 35 malignant samples as seen in Figure 6. It had the best recall of the classifiers evaluated, but it made many more false positive predictions, misdiagnosing eight of the benign cases as malignant. In contrast, only one malignant sample was misclassified as benign. This behaviour is the reason for the excellent sensitivity while the overall accuracy was relatively low for the DT. A high sensitivity is desirable from a clinical point of view as it will reduce the likelihood that a case of malignancy will not be detected, but it can also lead to a higher rate of false-positive results and increased worry among patients.

The SVM had the most balanced classification results of the three models. As can be seen in Figure 6, the classifier has correctly classified 67 benign samples and 34 malignant samples, and made 4 wrong predictions, two false positives and two false negatives. Almost equal distribution of the classification errors suggests the SVM performed well in achieving a balance between sensitivity and specificity which led to the overall highest diagnostic reliability.

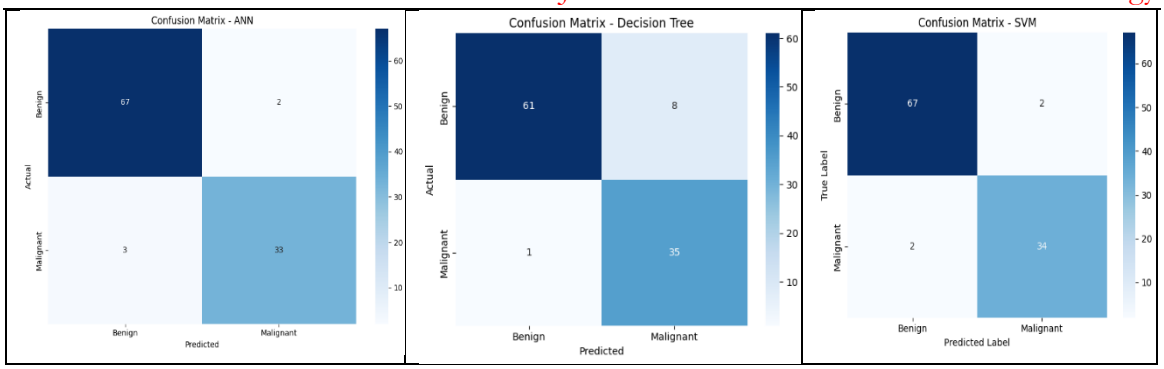

Figure 6. Confusion matrices of optimized ANN, DT and SVM classifiers.

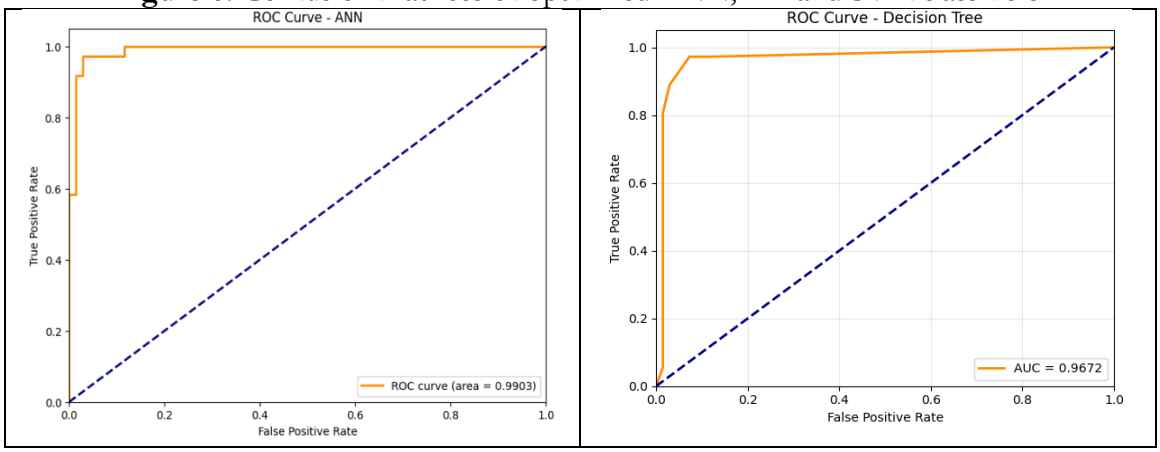

Figure 7. Optimized ROC curves comparison of ANN, Decision Tree classifiers.

The ROC curves shown in Figure 7 also indicate the discriminative power of the classifiers analyzed. The ROC curves of all three models were well above the reference line, indicating good separation between benign and malignant tumours. The highest ROC-AUC value of 99.03% was obtained by the ANN model among the evaluated models, with a remarkable ranking performance in different classification thresholds. The SVM performed almost as well with a ROC-AUC of 98.83% and the Decision Tree generated an ROC-AUC of 96.72%. The ANN had the highest ROC-AUC, however, the SVM had a better balance between ROC-AUC, accuracy, precision and recall making it the more reliable overall classifier for this dataset.

Figure 8 shows a quantitative comparison of the evaluation metrics. SVM outperforms all other classifiers in terms of overall accuracy (96.29%) and in terms of balanced precision (94.44%), recall (94.44%), and F1 score (94.44%). The results show similar predictive capabilities for benign and malignant breast tumour classification.

The ANN obtained an accuracy of 95.24%, precision of 94.29%, recall of 91.67%, F1-score of 92.96%, and the highest ROC-AUC (99.03%). The recall rate for the ANN was less than the SVM, but its performance in discriminating between the two classes was excellent, and its convergence in training was stable, demonstrating that the ANN could learn the nonlinear relationships between the cytological characteristics of breast cells effectively.

The DT achieved an overall accuracy of 91.43%, precision of 81.40%, recall of 97.22%, F1-score of 88.61%, and ROC-AUC of 96.72%. It has an exceptionally high recall which means that it is better at detecting malignant tumours, minimizing the risk of cancer not being detected. The relatively poor accuracy, however, indicates that more benign tumours were misclassified as malignant and the number of false positives would be higher. This decision-making approach does not improve classification accuracy overall, but decision trees have an important advantage, that is their inherent interpretability which can be crucial for clinical decision support systems where interpretability is a key requirement for the system.

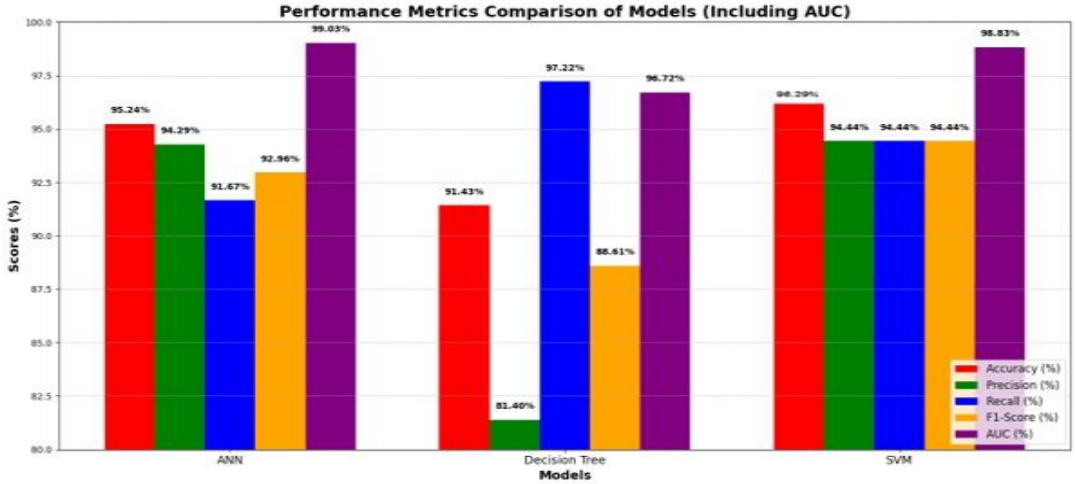

Figure 8. Comparative performance of the ML models evaluated with accuracy, precision, Recall, f1, and ROC-AUC.

Statistical Significance Testing:

To test the statistical significance of performance differences between ANN, DT, and SVM beyond point estimate comparisons, McNemar's exact test was administered for each model as a pairwise test on the independent test set (N = 105) [31]. Differences among the labels predicted by the three classifiers were paired, and therefore, the McNemar's test was used, as the classifiers operated on the same instances [32]. Test results are reported in Table 3.

Table 3. McNemar's Test Results (Pairwise Model Comparison)

	Model A correct only	Model B correct only	Test statistic	p-value	Result
ANN vs. DT	6	2	2.00	0.289	No significant difference
ANN vs. SVM	0	1	0.00	1.000	No significant difference
DT vs. SVM	1	6	1.00	0.125	No significant difference

The lack of a significant pairwise difference across the three cases ($\alpha = 0.05$) is informative in and of itself. It shows that, structurally different from what they are, all three classifiers reach a similar level of predictive reliability after optimization. This finding reaffirms the primary objective of the study, which is that, following careful adjustment of hyperparameters and standardized methods of evaluation, the gap in performance of structurally different algorithms is practically negligible. It is also worth noting that out of 105 test cases, ANN and SVM only disagreed in one, which demonstrates the strength of the standardized pipeline in processing and optimization to produce consistent and reproducible results in classification. This analysis also strengthens the model selection argument in favor of clinical defensibility over accuracy in the periphery; it indicates that in spite of SVM's slight advantage over ANN (Table 2), statistically both are equally reliable, and the remaining considerations in the selection of the models would be based on constraints of deploying the models, interpretability, and inference speed, in that order.

Misclassification Analysis:

By analyzing errors in classification from all three models, as opposed to just reviewing the confusion matrices, it is possible to better understand the behavior of the classifiers. All three models misclassified three test cases, and a fourth case was misclassified by two of the three models (Table 4). This indicates that there are a few cases of cytological ambiguity that

are misclassified and are not the result of a decision boundary of one of the models. Given the convergence of a rule-based tree, a margin-based classifier, and a nonlinear neural network, shows that these cases are truly misclassified by the models and are not an artifact of the models.

Table 4. Misclassification Overlap Across Models

Category	Count
Misclassified by ANN	5
Misclassified by DT	9
Misclassified by SVM	4
Misclassified by all three models	3
Misclassified by at least two models	4

The remaining errors tended to be specific to each model. For example, 6 of the 9 errors that DT made, which the ANN and SVM did not make, were expected due to the less complex nature of DT's decision boundary. Errors made by DT were strongly asymmetric and, consistent with Table 2, 8 of the 9 DT errors were false positives (in which benign cases were misclassified as malignant). ANN and SVM had more error distributions. This asymmetry is of clinical importance. DT's false positive errors explain why it has a high recall (97.22%); thus, its errors are benign (over-caution) rather than false negatives (missed malignancies). This would likely be a more acceptable error-type concern for a screening test, despite it reducing the model's overall results.

Computational Complexity and Training Time:

Table 5 shows the training and search time for each classifier, measured during the same session and on the same hardware, which assesses the practical use and deployment of each model in comparison to each other.

Table 5. Computational Time Comparison

Model	Hyperparameter search time	Final model training time	Inference time (105 samples)
DT	8.52 s (648 combinations, 5-fold CV)	0.0015 s	< 0.01 s
SVM	23.86 s (602 valid kernel-specific combinations, 5-fold CV)	0.009 s	0.0004 s
ANN	60-combination search (Table 2)	3.86 s (16 epochs, EarlyStopping)	< 0.01 s

The three models are all lightweight enough for real-time clinical inference because the time it takes to make a single prediction is less than a hundredth of a second. Therefore, the major cost difference happens at the offline hyperparameter tuning step and not at deployment. DT is the fastest to both search and train due to its shallow and simple structure. Since SVM has a larger search space for kernels and parameters, it takes longer to search but does not translate into a slower model deployment. This shows that at deployment time; computational cost is not significant enough to differentiate the three classifiers. Thus, factors like predictive performance, interpretability, and the robustness of the models should dictate the choice of a classifier to be used.

The clinical costs of false positives and false negatives for breast cancer screening differ and are handled differently in the three models. DT's profile (8 false positives to 1 false negative) prioritizes minimizing the risk of missed malignant cases and ends up generating a lot of unnecessary follow-up biopsies for benign cases. In the context of a first-line screening tool, it is a justifiable risk to delay a diagnosis in the case of a benign case over the cost of a false positive, but in the case of a confirmatory diagnostic tool, it is poor to have a lot of false positives, because it means unnecessary procedures, costs to the healthcare system, and false positive screening anxiety to the patients. ANN and SVM are better because they have a more

balanced profile of false positives and false negatives. Because of this, DT, ANN, and SVM should be considered for the different screening roles. DT would be the first line of screening, encumbering the least amount of cost, and followed by its confirmatory role by SVM or ANN, which would be more cost-effective for the healthcare system.

Explainability analysis (SHAP and LIME):

To tackle the accuracy-interpretability trade-off addressed in this study, SHAP and LIME were implemented to calculate the contribution to predictions of each feature in all three classifiers, and a case-level explanation was generated for a single prediction, respectively. Compared to the built-in feature importance in DT, which indicates the importance of features only based on which splits the tree made, SHAP values can be computed for all three model types to provide a direct and fair comparison for feature importance between the fully transparent DT model and the two black-box models (ANN and SVM).

For DT, exact SHAP values were calculated with Tree-Explainer. For SVM and ANN, Kernel-Explainer, with a 20-cluster background sample of k-means, was used to compute SHAP values on an illustrative subset of the test set, since calculating SHAP values for non-tree-based models. Mean absolute SHAP values for each feature are shown in Figure 9 and Table 6.

Table 6. Mean Absolute SHAP Feature Importance by Model

Feature	DT	SVM	ANN
Uniformity_Cell_Size	0.2483	0.0427	0.0555
Bland_Chromatin	0.1135	0.0352	0.0713
Bare_Nuclei	0.0631	0.1073	0.0892
Normal_Nucleoli	0.0629	0.0361	0.0331
Clump_Thickness	0.0290	0.0926	0.0812
Uniformity_Cell_Shape	0.0142	0.0515	0.0489
Single_Epithelial_Cell_Size	0.0077	0.0393	0.0287
Marginal_Adhesion	0.0000	0.0424	0.0322
Mitoses	0.0000	0.0158	0.0262

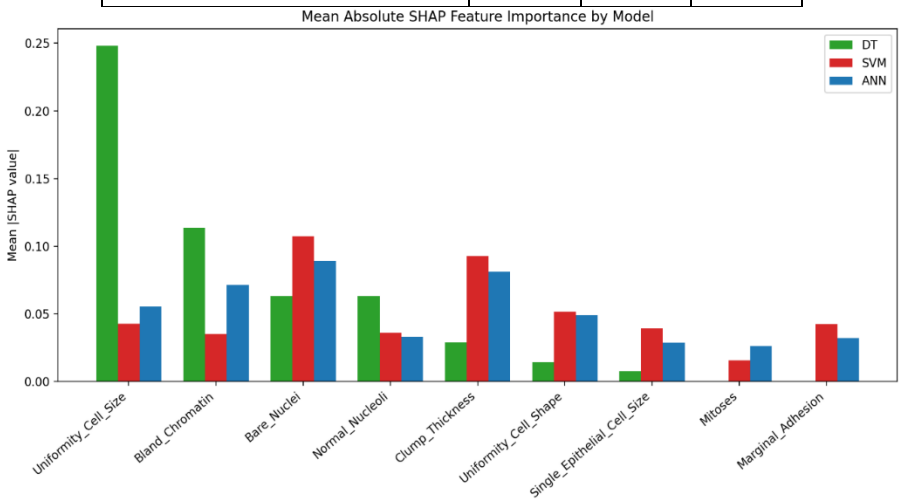

Figure 9. Mean Absolute SHAP Feature Importance by Model

A distinct pattern can be observed when comparing these models. Most of DT's predictions are based on Uniformity_Cell_Size (the most dominant variable in DT's SHAP value), which is in line with its ranking. This is also likely due to the tree's shallow and shallow tree structure. Unlike DT, Bare_Nuclei and Clump_Thickness rank highest of all features in both SVM and ANN models and are representative of the black-box models. Here, DT, SVM and ANN learn more sophisticated structures and thus represent a more complex, multidimensional view of malignancy, which is supported by the higher accuracy of these

models in comparison to DT. This difference is of clinical value, as even though DT is clear and transparent, its predictions are based on very limited evidence compared to SVM and ANN, which are less clear but possess a high degree of accuracy. Thus, choosing between the models means choosing between decision clarity and the evidence breadth.

To show how global patterns apply to individual predictions, LIME was used on one of the three test cases where all classifiers misclassified. For this particular case (true label: malignant; SVM predicted: benign), LIME showed local explanation that all nine cytological features were in the below-average range compared to the training data, and that the three features (Single_Epithelial_Cell_Size, Bland_Chromatin, and Clump_Thickness) were the primary contributors to the (incorrect) benign prediction. This local explanation is consistent with the misclassification analysis provided above. This case did not contain any cytological marker that was strongly elevated, thus, from the model's point of view, it was a truly borderline case, as compared to other cases in which there may have been a mismeasurement of a particular feature, an outlier, or a case of misclassification. This type of explanation at the level of a single case, that goes beyond just showing an error, provides the type of clinically useful auditability that is structurally offered by DT and is also provided by SHAP/LIME for the other high performing, yet less transparent, ANN and SVM models.

Discussion:

These results are consistent with the findings of similar machine learning applied studies focused on breast cancer diagnosis. Naji et al. demonstrated SVM with a classification accuracy of 97.2%, and Gupta et al. demonstrated the hybrid model of ANN–DBN with a classification accuracy of 98.14%. The numbers these authors provide are marginally better than the best classification accuracy achieved in this study. However, the described methodology of the experimental procedure in this study provides a more robust and more objective comparison. Every model analyzed in this study was built following the same preprocessing pipeline, the same training/validation/testing data, and systematic hyperparameter optimization and was assessed by a variety of complementary metrics, many of which were not considered in previous studies that focused on the evaluation of diverse classifiers. Consequently, the performances that were reported truly reflect the behaviour of the algorithms, and the rest of the reported findings are consistent with this conclusion and are supported by the statistical significance testing described earlier in that no pairwise difference among the three optimized models was significant.

An important dimension of this work is bringing predictive performance, statistical robustness, and clinical applicability together, rather than focusing on accuracy alone. Beyond classification accuracy, clinical deployment of models calls for dependable sensitivity, equal specificity, stable behaviour of the model, and consistent performance. Among the three models, SVM achieved the best balance of the aforementioned properties, but the significance testing shown earlier indicates that ANN is, if anything, slightly better or at least on par with SVM, meaning that all else being equal, the selection between the two models may be based on other considerations such as inference speed and the evidence of features built by each model (Table 6). Although DT has the lowest accuracy of the models, it is the most interpretable by design, and is supported by the above SHAP analysis, which indicates that it is, in fact, not really interpretable: DT relies, more than anything else, on one feature of the cytology model, while SVM and ANN rely on a wider, more statistically supported set of features for their predictions.

Several caveats must be addressed. Firstly, the Wisconsin Breast Cancer Original Dataset is a small benchmark dataset and cannot replicate the real diversities of the clinical populations. With 699 samples and nine features, the dataset cannot capture the imaging-based or genomic biomarker datasets integrated into the modern diagnostic systems. Secondly, the proposed models were only tested using one dataset, and the external validation using

independent multicenter data was not performed. Therefore, the comparable performance of ANN and SVM, as shown in Table 3, may not hold true for other clinical populations. Thirdly, while the SHAP and LIME analyses in this study contribute to feature-level interpretability of ANN and SVM, with respect to computational considerations, explainability techniques were applied in a representative small subset of the testing data, and while it is a common practice to use kernel-based SHAP as approximations, they are not exact. The Tree-Explainer values for DTR are exact, and the broad validation of these explainability techniques is a key aspect to focus on in the future. Lastly, the high recall of DTR was obtained at a substantial cost to precision (81.40%). Since this trade-off was designed intentionally, it is not a shortcoming, and it is a flexible design that can be applied at different points of the clinical pathway as discussed earlier.

In general, the experiments presented indicate that breast cancer classification through the application of machine learning significantly benefits from careful data pre-processing, optimal hyperparameter selection, and application of a uniform assessment methodology. Of all the machine learning techniques evaluated, SVMs support their selection as the most appropriate for the automation of the diagnostic process based on cytological data. Overall, SVMs exhibited the highest estimate of accuracy, comparable statistics to ANNs, and a clearer and more distributed rationale of evidence when compared to Decision Trees. ANNs are a defensible alternative if the classification systems are designed for a higher level of recall, or if more data are collected in the future.

Conclusion:

This study compared three supervised ML techniques ANN, DT and SVM, in BC classification on the WBCD. The classifiers were trained and tested in the same experimental setting, and all classifiers were trained using the same pipeline of pre-processing steps, which was optimized step-by-step by systematic hyperparameter tuning.

The performance of the experimental models was evaluated and the results showed that the SVM model had the highest classification accuracy of 96.29%, Precision 94.44%, Recall 94.44%, F1-score 94.44%, and ROC-AUC 98.83%. The ANN also showed very high predictive power with a rate of accuracy of 95.24 % and the highest ROC-AUC (99.03 %), which means that it has very good discriminative power and generalization. The DT has demonstrated very good results in terms of recall (97.22%) and good overall accuracy (91.43%), and it is also easy to interpret, making it a desirable feature for an explainable clinical decision support system.

Statistical validation with McNemar's test determined that none of the pairwise performances of ANN, DT, and SVM were significantly different ($p > 0.05$) from one another, indicating that all three optimized classifiers, regardless of their architectural differences, are likely to exhibit comparable reliability. Misclassification analysis showed that a small number of test cases that were misclassified by all three models are likely the result of ambiguous cytological presentations rather than weakness of the models. SHAP and LIME provided a basis for explainability with SHAP showing that DT relies on a single dominant feature, Uniformity of Cell Size, while ANN and SVM rely on several features, among which are Bare Nuclei and Clump Thickness, therefore making the explainability for these black-box models SHAP and LIME provided easier to interpret. Interestingly, all three classifiers were lightweight during inference, suggesting that ease of deployment does not meaningfully separate these classifiers.

The study suggests that the choice of ML model for BC diagnosis should not just be based on predictive accuracy but also on other factors such as robustness, interpretability, and clinical applicability. The use of a standardized experimental framework in this study provides a solid foundation for making objective comparisons, and the results show that careful

hyperparameter tuning and pre-processing techniques can greatly enhance the diagnostic capabilities.

SVM proved to be the best classifier for BC diagnosis using cytological features, as it showed the highest trade-off between sensitivity and specificity in addition to being the most appropriate in terms of its overall accuracy and performance. The ANN is a highly competitive option, even though it has been supplanted in some instances by deep learning models capable of harnessing more complex data representations, especially for more extensive data sets; on the other hand, DT provides useful interpretation for situations where transparent decision-making is required.

Future Work:

The proposed models should be further validated with larger multi-institutional clinical datasets for better external generalizability in future research. Also, the use of ensemble learning methods based on ANN, DT and SVM can improve diagnostic performance. Lastly, the incorporation of explainable AI techniques into the high performing predictive models will enhance transparency and clinician confidence and enable the safe deployment of machine learning systems in clinical care for breast cancer.

References:

[1] M. Arnold *et al.*, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, Dec. 2022, doi: 10.1016/J.BREAST.2022.08.010.

[2] A. J. D. Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay ME, Rebecca L. Siegel , Isabelle Soerjomataram, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, 2024, doi: https://doi.org/10.3322/caac.21834.

[3] "Breast cancer - Symptoms and causes - Mayo Clinic." Accessed: Apr. 23, 2026. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[4] P. A. Carney *et al.*, "Identifying and processing the gap between perceived and actual agreement in breast pathology interpretation," *Mod. Pathol.*, vol. 29, no. 7, pp. 717–726, Jul. 2016, doi: 10.1038/MODPATHOL.2016.62.

[5] S. Naseer *et al.*, "Enhanced network anomaly detection based on deep neural networks," *IEEE Access*, vol. 6, pp. 48231–48246, Aug. 2018, doi: 10.1109/ACCESS.2018.2863036.

[6] S. M. McKinney *et al.*, "International evaluation of an AI system for breast cancer screening," *Nat. 2020 5777788*, vol. 577, no. 7788, pp. 89–94, Jan. 2020, doi: 10.1038/s41586-019-1799-6.

[7] K. Freeman *et al.*, "Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy," *BMJ*, vol. 374, Sep. 2021, doi: 10.1136/BMJ.N1872.

[8] "Breast Cancer Wisconsin (Original) Data Set." Accessed: Aug. 06, 2026. [Online]. Available: https://www.kaggle.com/datasets/mariolisboa/breast-cancer-wisconsin-original-data-set

[9] Hiba Asri, Hajar Mousannif, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, 2016, doi: https://doi.org/10.1016/j.procs.2016.04.224.

[10] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *J. Algorithms Comput. Technol.*, vol. 12, no. 2, pp. 119–126, Jun. 2018, doi: 10.1177/1748301818756225.

[11] 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 Comput. Sci.*, vol. 191, pp. 487–492, Jan. 2021, doi: 10.1016/J.PROCS.2021.07.062.

[12] V. Gupta, S. Wadhawan, V. Kumar, H. Sethi, and G. Gupta, "Breast Cancer Classification with ANN and DBN," *Proc. - 2nd Int. Conf. Adv. Comput. Comput. Technol. InCACCT 2024*, pp. 246–251, 2024, doi: 10.1109/INCACCT61598.2024.10550987.

[13] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, and L. Li, "Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning," *Sci. Reports 2016 61*, vol. 6, no. 1, pp. 27327-, Jun. 2016, doi: 10.1038/srep27327.

[14] M. O. Abdullah, Y. Altun, and R. M. Ahmed, "Leveraging Artificial Neural Networks and Support Vector Machines for Accurate Classification of Breast Tumors in Ultrasound Images," *Cureus*, vol. 16, no. 11, p. e73067, Nov. 2024, doi: 10.7759/CUREUS.73067.

[15] P. Dinesh, A. S. Vickram, and P. Kalyanasundaram, "Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest and decision tree to measure accuracy," *AIP Conf. Proc.*, vol. 2853, no. 1, May 2024, doi: 10.1063/5.0203746/3290220.

[16] A. Kumar, R. Saini, and R. Kumar, "A Comparative Analysis of Machine Learning Algorithms for Breast Cancer Detection and Identification of Key Predictive Features," *Trait. du Signal*, vol. 41, no. 1, Feb. 2024, doi: 10.18280/TS.410110.

[17] C. S. Rajpoot, G. Sharma, P. Gupta, P. Dadheech, U. Yahya, and N. Aneja, "Feature Selection-based Machine Learning Comparative Analysis for Predicting Breast Cancer," *Appl. Artif. Intell.*, vol. 38, no. 1, Dec. 2024, doi: 10.1080/08839514.2024.2340386.

[18] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 4766–4775, May 2017, Accessed: Aug. 14, 2024. [Online]. Available: https://arxiv.org/abs/1705.07874v2

[19] T. Khater, A. Hussain, S. Mahmoud, and S. Yasen, "Explainable AI for Breast Cancer Detection: A LIME-Driven Approach," *Proc. - Int. Conf. Dev. eSystems Eng. DeSE*, pp. 540–545, 2023, doi: 10.1109/DESE60595.2023.10469341.

[20] Amirehsan Ghasemi, Soheil Hashtarkhani, "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *Cancer Innov.*, vol. 3, no. 5, 2024, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11488119/

[21] T. Alelyani, M. M. Alshammari, A. Almuhanna, and O. Asan, "Explainable Artificial Intelligence in Quantifying Breast Cancer Factors: Saudi Arabia Context," *Healthc. 2024, Vol. 12, Page 1025*, vol. 12, no. 10, p. 1025, May 2024, doi: 10.3390/HEALTHCARE12101025.

[22] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996/METRICS.

[23] T. K. Murugan, P. Karthikeyan, and P. Sekar, "Efficient breast cancer detection using neural networks and explainable artificial intelligence," *Neural Comput. Appl. 2024 375*, vol. 37, no. 5, pp. 3759–3776, Dec. 2024, doi: 10.1007/S00521-024-10790-2.

[24] C. J. Kelly *et al.*, "Diagnostic accuracy, fairness and clinical implementation of AI for breast cancer screening: results of multicenter retrospective and prospective technical feasibility studies," *Nat. Cancer 2026 73*, vol. 7, no. 3, pp. 494–506, Mar. 2026, doi: 10.1038/s43018-026-01127-0.

[25] A. Khalid *et al.*, "Breast Cancer Detection and Prevention Using Machine Learning," *Diagnostics*, vol. 13, no. 19, p. 3113, Oct. 2023, doi:

10.3390/DIAGNOSTICS13193113.

[26] A. Uwimana, G. Gnecco, and M. Riccaboni, "Artificial intelligence for breast cancer detection and its health technology assessment: A scoping review," *Comput. Biol. Med.*, vol. 184, Jan. 2025, doi: 10.1016/J.COMPBIOMED.2024.109391.

[27] S. MacHeka, P. Y. Ng, O. Ginsburg, A. Hope, R. Sullivan, and A. Aggarwal, "Prospective evaluation of artificial intelligence (AI) applications for use in cancer pathways following diagnosis: a systematic review," *BMJ Oncol.*, vol. 3, no. 1, May 2024, doi: 10.1136/BMJONC-2023-000255.

[28] D. Nasien, V. Enjeslina, M. Hasmil Adiya, and Z. Baharum, "Breast Cancer Prediction Using Artificial Neural Networks Back Propagation Method," *J. Phys. Conf. Ser.*, vol. 2319, no. 1, p. 012025, Aug. 2022, doi: 10.1088/1742-6596/2319/1/012025.

[29] M. H. Alshayeji, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomed. Signal Process. Control*, vol. 71, p. 103141, Jan. 2022, doi: 10.1016/J.BSPC.2021.103141.

[30] Z. A. Ansari, M. M. Tripathi, and R. Ahmed, "The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models," *Discov. Artif. Intell. 2025 51*, vol. 5, no. 1, pp. 75-, May 2025, doi: 10.1007/S44163-025-00307-8.

[31] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[32] M. Mitu, S. M. M. Hasan, A. H. Efat, M. F. Taraque, N. Jannat, and M. Oishe, "An Explainable Machine Learning Framework for Multiple Medical Datasets Classification," *2023 Int. Conf. Next-Generation Comput. IoT Mach. Learn. NCIM 2023*, 2023, doi: 10.1109/NCIM59001.2023.10212821.