

Comparative Study of Machine Learning Algorithms for Breast Cancer Diagnosis: A Clinician–Engineer Collaborative Approach

Nargiza Pulatova^{1,*}, *Ibrokhim Pulatov*²

¹Department of Clinical Pharmacology, Tashkent State Medical University

²Department of Computer Science, Specialised School named after Al-Khwarizmi

Abstract. Breast cancer is the most common cancer and the second leading cause of cancer-related deaths among women globally. The early and precise diagnosis of malignant breast tumours is beneficial for increasing the survival rate of cancer patients [1]. In the current investigation, we propose a multidisciplinary clinician–engineer collaboration to demonstrate the potential of ML in the diagnosis of breast cancer. A publicly available dataset consisting of 569 fine-needle aspirate samples (212 malignant, 357 benign) [2] and 30 quantitative cytological measures was employed to train and evaluate four classification models: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting. The data were randomly divided into 70% training and 30% testing with standard normalisation. Performance of models was evaluated based on accuracy, sensitivity, specificity and F1-score. SVM had the best of all the highest accuracy (96.5%) at a sensitivity of 93.7% and a specificity of 98.1%, which performed slightly better than the other models. For clinical applications, the high sensitivity means the model won't miss many cancers, and the high specificity reduces false alarms. The analysis of feature importance identified that cell size and shape-based features (e.g., “worst” radius, perimeter, and area features) contributed the most to the prediction of malignancy, consistent with known pathology guidelines. We debate the clinical impact of such ML tools and potential pitfalls (dataset bias, absence of external validation), as well as future directions such as prospective validation and image integration. In conclusion, according to our results, ML models are reliable classifiers to differentiate benign from malignant breast cytological lesions and point to a promising prospect to complement the clinical decision-making in oncology.

Keywords: Breast cancer, Machine learning, Support vector machine, Clinical application

*Corresponding Author: nargiza.pulatova1984@gmail.com

1 Introduction

1.1 Clinical Background of Breast Cancer

Breast cancer is the most frequent cancer among women and a leading public health problem.

It's also now higher on the list of most diagnosed cancers than lung cancer is — 2.3 million women were diagnosed and 685,000 died of it around the world in 2020. The condition develops due to rapid multiplication of cells in the breast tissue, which results in lumps that are either benign or malignant in nature. Early detection is key: the earlier that breast cancer is detected at a localised stage, the better the survival rates and treatment options[3]. Early breast cancer detection is sought in clinical practice by screening the breast at regular intervals or through an aggressive diagnostic investigation of lumps or abnormalities of the breast. Breast cancer is often screened for using X-ray mammography to detect suspicious lesions that may arise long before symptoms appear. If a lesion is found, further investigations, such as ultrasound, fine-needle aspiration (FNA) cytology, or core biopsy, are carried out to ascertain whether it is malignant. The aim is to accurately differentiate malignant tumours from benign tumours as early as possible, with potential for timely intervention and avoidance of unnecessary invasive procedures for benign cases.

1.2 Barriers to Detection and Diagnosis

However, limitations remain, and early detection of breast cancer is not without challenges. Mammography, the gold standard in screening, has a notorious moderate accuracy – studies show only ~65–78% accuracy rate in actual practice. This, in turn, means that a large proportion of the cancers cannot be detected (false negatives) or that the benign lesions are falsely named malignant (false positives). In addition, mammographic interpretation is also reader-dependent (i.e., multiple radiologists would interpret the same mammogram differently). This variety can confuse diagnosis and anxiety for the patient. FNA/biopsy should be done when imaging is indeterminate. Fine-Needle Aspiration (FNA) Cytology is a procedure that removes cells from a breast lump using a very thin needle, after which the cells are viewed under a microscope to check for cell appearance. FNA, although less invasive compared to a surgical biopsy, needs to be read by an expert cytopathologist and may be nondiagnostic in certain circumstances. The gold standard is a surgical biopsy, which has a near 100% accuracy, but it is expensive, labour-intensive, and invasive, and not every patient can undergo this procedure[4]. These obstacles accentuate the requirement for better diagnostic tools. A perfect solution would be more sensitive (to catch more true cancers early) without being any less specific (to avoid false alarms and unnecessary treatment). This is where AI and ML tools can come in – relying on patterns learned from data to help clinicians make more consistent and accurate diagnoses.”

1.3 AI and Computational Methodologies in Cancer Research

Artificial Intelligence (AI) is expanding at an unprecedented rate, reshaping the field of medical diagnosis, as well as changes to the seasoned face of oncology. Machine learning algorithms can interpret intricate patterns in data (imaging, genetic and clinical) beyond human visual perception, which may increase the accuracy and reproducibility of diagnostics. In breast cancer, ML has been used in a wide variety of tasks, such as reading mammograms and pathology images, to predict the risk of cancer and response to treatment[5]. For the diagnosis of breast lesions, conventional ML models as well as newer methods, such as deep learning, have been implemented. We note that the Wisconsin Diagnostic Breast Cancer (WDBC) (applied in this study) is a standard dataset frequently used to assess ML algorithms. One of the largest differences is the observation that in previous studies in this and similar data, accuracies reached >95–99% using well-chosen algorithms or ensemble approaches [6]. These findings exemplify how ML can elevate clinical decision support. Yet translating these AI tools from bench-to-bedside in the real world involves a rigorous process of validation, interpretability, and collaboration between technical and clinical experts. Here, a clinician–

engineer co-design team collaboratively prepared an ML pipeline for breast cancer diagnosis and discussed the findings to make certain that the algorithm's performance and clinical relevance were simultaneously scrutinised.

2 Materials and Methods

2.1 Dataset and Feature Selection

We employed the pretrained Breast Cancer Wisconsin (Diagnostic) dataset, which is a widely used dataset in the field of machine learning and medical science. The dataset has been originally prepared by Wolberg et al. and is obtained from digitalized images of FNA breast mass samples. Each instance represents a tumour sample from a patient's breast (here 569) represented by a 30-dimensional vector which describes some features of the cell nuclei present in the FNA cytology image. These functions describe ten aspects of the character of cell nuclei, and 3 statistics (about the mean, standard error and so-called worst value) are calculated on each of these functions. As such, the features essentially encapsulate the size, shape and composition of cells in the tumour. The dataset comprises a total of 569 instances; of these, 212 are malignant (positive class) and the remaining 357 are benign (negative class). In the data, the diagnosis field is text corresponding to "malignant" or "benign" for each sample, and we encoded the diagnosis as 1 and 0, respectively, for modelling. We selected this dataset due to its established use in benchmarking diagnostic algorithms and the clinical interpretability of its features. Feature selection was not made beyond the 30 attributes provided as input; these have already been carefully selected, representing important tumour features[2]. Key statistics for benign vs. malignant cases are summarised in Appendix 1 - Table 4.

2.2 Data Preprocessing

Before modelling, we performed pre-processing of the data for robust and unbiased assessment. The dataset was further checked for missing values or outliers; however, no missing data were found in the parsed dataset. We further divided the data into a training (398 samples) and testing (171 samples) set using a 70/30 ratio. The split was performed at random for reproducibility using a fixed seed (random state = 15). This ratio kept a reasonably balanced class distribution in train vs. test sets (the training set consisted of 148 malignant and 250 benign cases, while the test set had 64 malignant and 107 benign ones) and allowed both classes to be nicely represented during the model fitting and testing process. Then, z-score scaling was applied to all the feature variables. In particular, we trained a StandardScaler on the feature_train_data (calculating the mean and standard deviations for each feature on the training set), and applied the same transformation to the training and test features. This scaling brings the features onto the same level of magnitude, which is a requirement for machine learning algorithms that are not scale invariant, such as logistic regression and SVM. In particular, scaling factors were estimated on the training data only and then applied to the test data to prevent any information flow from the test data during the training process. For this baseline pipeline, there was no further feature engineering or dimensionality reduction applied, since we wanted to compare the algorithms in the same feature space.

2.3 Development and Training of the Model

We applied and tested four supervised ML techniques relevant to binary classification (benign vs. malignant): Logistic regression (LR), Random forest (RF), Support vector machine (SVM), and Gradient boosting (GB). The algorithms were chosen as a mix between model architectures (Linear vs. Nonlinear, SingleEstimator vs. Ensemble):

Logistic Regression: a linear model used to learn a weighted sum of features to predict the log-odds of malignancy. We applied an L2-regularised logistic regression (`sklearn.linear_model.LogisticRegression`) was employed with up to 1000 iterations, while repeating the solution 10 times to guarantee the convergence. This baseline model is simple and easy to interpret in terms of feature coefficients.

Random Forest: an ensemble of decision trees trained with bagging. We used `sklearn.ensemble.RandomForestClassifier` (`n_estimators=100`, `random_state=15`). Since random forests can represent feature interactions in a non-linear fashion and usually assume resilience with feature noise, we use random forests to obtain a natural importance measure of features.

Support Vector Machine: a nonlinear classifier that computes the hyperplane that separates classes in the feature space. We used `sklearn.svm.SVC` with a RBF kernel (the default), with probability estimates enabled (`probability=True`), in other words, Outputs the probability of the sample for each class. We fixed the C parameter and the kernel width of the SVM at their default values; these parameters could be tuned, but we were interested in comparing performance baselines. SVMs have been demonstrated to be efficient, especially for high-dimensional biomedical datasets having obvious class separation.

Gradient Boosting: an ensemble method in which trees are added sequentially to correct the errors of the trees added previously. We used `sklearn.ensemble.GDBosting` (`n_estimators=100`) with `random_state=15`. For instance, gradient boosting tends to achieve high accuracy via learning complex patterns and might overfit without proper tuning.

All models were implemented with Python using the library scikit-learn (version 1. x). Hyperparameters of models follow the common setup and related literature for this dataset[1], and no extensive hyperparameter search was performed due to time limitations. Both models were trained using the scaled training set. The training of ensemble learning models (RF and GB) uses its own internal bootstrap sampling (RF) or stage-wise fitting (GB), and the cost function is optimised on the training set for LR and SVM. We also saved the trained models and trained scaler for possible use in the future or for reproducibility. The training pipeline was scripted such that data-loading, preprocessing, model training, and evaluation could be run end-to-end, facilitating easy re-runs and modifications to the pipeline by the engineering team, under the supervision of the clinical collaborator, to produce clinically meaningful output.

2.4 Performance Metrics

The model was assessed on the test set (30% hold-out data on which no training was done). We concentrated on clinically diagnostic relevant metrics, accuracy, sensitivity, specificity, and F1-score. These are defined as follows:

Accuracy: Ratio of total samples that were correctly predicted by the model (overall correctness). It is given by $(TP+TN)/(TP+TN+FP+FN)$ or $(TP + TN) / (TP + TN + FP + FN)$, where TP is true positives (malignant correctly predicted),

TN is true negatives (benign correctly predicted).

Sensitivity (Recall) –: The true positive rate for malignant cases. $\text{Sensitivity} = \frac{TP}{TP+FN}$. Sensitivity is the proportion of true cancer instances that the model nominates. High sensitivity is essential to prevent missed cancers (false negatives).

Specificity: The proportion of true negatives among benign cases. $\text{Specificity} = \frac{TN}{TN+FP}$. This reflects the fraction of benign cases classified as benign (e.g., instead of being classified in error as cancer). High specificity translates to fewer false positives and less needless follow-up among well patients.

F1-Score: The weighted average of precision for the positive class (malignant) and recall for the positive class (malignant). Precision is $\frac{TP}{TP+FP}$, the fraction of the predicted “malignant” which are actually malignant. F1-score is the harmonic mean of precision and recall and provides a single score that represents a good trade-off between precision and recall of the test for the positive class, and when the class distribution is skewed. We mainly use the F1-score for the malignant class to summarise its detection performance, with the macro-averaged F1 also considered.

We also looked at the complete classification report for each model (including precision and class-based precision) and the confusion matrix to gain an understanding of how the models made mistakes. While accuracy provides an overall measure, sensitivity and specificity offer more clinically applicable information on test performance. During our review, we focused on sensitivity for detection of malignancies (to prevent false negatives, which could delay treatment) and specificity (to prevent overdiagnosis of benign disease). All the numbers are presented with units, percentages for better understanding. Statistical significance testing of model performance was not performed due to the small set of tests; we focus here on differences observed and their significance in practice. We also conducted 10-fold stratified cross-validation across the dataset, determining the accuracy, sensitivity, specificity, and F1-score for each fold to increase statistical reliability. The mean \pm standard deviation is used to report the results. Additionally, we calculated 95% bootstrap CIs using out-of-fold predictions, which lessens reliance on a single train-test split and provides uncertainty ranges for the performance metrics.

3 Results

3.1 Comparative Model Performance

The four ML models were trained on a training set and subsequently tested on a 171-sample test set. Table 1 shows the classification performance in terms of accuracy, sensitivity, specificity and F1-score for the malignant class of each method. All models obtained high accuracy (>94%), indicating that the characteristics of our dataset were very discriminating between non-malignant and malignant cases. The SVM performed slightly better than other models (with 96.5% accuracy) and correctly classified 165 /171 test samples. Logistic Regression and Gradient Boosting closely followed, both with about 95.9% accuracy, and Random Forest with 94.2%. The accuracies are only about 1–2 percentage points off, demonstrating similar performance overall, but subtle differences can still be significant for some critical diagnoses.

Table 1. Comparison of Machine Learning Models on the Test Database (n = 171) for Breast Cancer Detection

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (malignant) (%)
Logistic Regression	95.9	93.7	97.2	94.0
Random Forest	94.2	88.9	97.2	92.0
SVM (RBF kernel)	96.5	93.7	98.1	95.0
Gradient Boosting	95.9	90.5	99.1	94.0

The ability to correctly identify benign cases was excellent for each model, with a specificity of greater than 97% for all models (Table 1). The sensitivity and specificity of benign cases as malignant cases. This was with some loss of sensitivity (90.5% for GB, i.e. 9.5% of malignancies were missed by the model). The Random Forest was, in contrast, the least sensitive (88.9%), and as such was less effective at catching all cancers. The Logistic Regression and SVM achieved a more balanced result - a sensitivity of over 93.7% and a specificity of over 97%. The best performance metrics in terms of balancing sensitivity and specificity were observed for SVM (sensitivity 93.7%, specificity 98.1%), and therefore, we considered SVM as the best model in this comparison. The F1-scores for the malignant class parallel these results- SVM had the greatest F1 (95.0%), meaning a very solid precision-recall tradeoff overall, while RF had a slightly lower F1 (~92.0% because the recall is lower). These results indicate that SVM may be very effective in sensitivity for this type of tabular diagnostic data, just because it can be tuned very well, which seems to be consistent with some literature studies, that SVM performs better in the WDBC dataset [7]. It is remarkable to notice that a simple model (logistic regression) achieved an impressive performance (accuracy =96%) with the features, and even features in our dataset separately describe benign vs. malignant cases.

In context, the accuracy we were able to achieve by our models (~94–96%) was much higher than the accuracy of a mammographic screening (approximately 65–78%[1]). It should be noted that our study is conducted using FNA retrospectively extracted cytology with outcome, which cannot be directly compared to prospectively reading mammograms; nonetheless, the high performance of ML on the dataset emphasises the potential of data-driven approaches in aiding diagnostic precision. The test set of 63 malignant and 108 benign cases resulted in 96.5% SVM model accuracy, where 59 and 106 cases were correctly predicted as malignant and benign, respectively. In other words, it had 4 false-negative errors (malignant tumour predicted to be benign) and 2 false-positive errors (benign predicted to be malignant) when tested with 171 cases. Logistic Regression and Gradient Boosting had a comparable error profile to SVM (both 4 false negatives, and LR had 4 false positives, GB 1. Random Forest, being the least sensitive, missed 7 malignancies and had 3 false positives. Because the importance of avoiding missing cancers is paramount, the ability of the SVM to still keep false positives as low as possible while also minimising false negatives is particularly attractive. We also performed a 10-fold cross-validation analysis accompanied by bootstrap 95% confidence intervals (Table 2). Using this analysis to confirm these results, SVM was also the best with a mean accuracy of $97.7\% \pm 1.4\%$ (95% CI: 96.5–99.0), a sensitivity of

96.3% ± 2.8% (95% CI: 93.4–98.6), a specificity of 98.6% ± 1.4% (95% CI: 97.3–99.7), and an F1-score of 96.9% ± 1.8% (95% CI: 95.1–98.5). Generally, Logistic Regression produced similar results while Random Forest and Gradient Boosting had somewhat lower sensitivities. These results indicate that SVM is the most accurate, reliable, and balanced classifier.

Table 2. Ten-fold cross-validation results (mean ± standard deviation, with 95% bootstrap confidence intervals).

Model	Accuracy	Sensitivity	Specificity	F1-score
Logistic Regression	0.979 ± 0.015 (96.7–99.0%)	0.963 ± 0.028 (93.2–98.6%)	0.989 ± 0.014 (97.7–99.7%)	0.972 ± 0.020 (95.3–98.6%)
Random Forest	0.960 ± 0.026 (94.2–97.4%)	0.948 ± 0.040 (91.7–97.5%)	0.966 ± 0.030 (94.6–98.4%)	0.946 ± 0.035 (92.2–96.6%)
SVM (RBF)	0.977 ± 0.014 (96.5–99.0%)	0.963 ± 0.028 (93.4–98.6%)	0.986 ± 0.014 (97.3–99.7%)	0.969 ± 0.018 (95.1–98.5%)
Gradient Boosting	0.960 ± 0.028 (94.2–97.4%)	0.944 ± 0.046 (90.9–97.2%)	0.969 ± 0.032 (95.0–98.6%)	0.946 ± 0.038 (92.1–96.6%)

3.2 Statistical implications for the clinical interpretation of diagnostic accuracy

Clinically, the above findings suggest that ML models are useful adjuncts in the diagnosis of breast cancer by FNA. The best model (SVM) could reach about 93.7% sensitivity, which is equivalent to when we would have 100 cases of malignant and in these 100 cases, the model would be able to detect, correctly, 94 and ~6 cases can be missed. That false-negative rate (of around 6%) is contrasted with screening mammography in population screening, which may miss over 20–30% of cancers in some age groups[8]. A false negative, that is, overlooking a malignant tumour, is the most serious error in cancer diagnostics since the patient may lose the possibility of having a life-saving treatment in time. Thus, finding a sensitivity in the mid 90% would seem promising, but for a diagnostic test to be used with confidence at the bedside, clinicians would likely want a sensitivity closer to 99%, if not higher, if possible. The high specificity of 98.1% for SVM (and textasciitilde97-99% for other models) is also

significant. A high specificity means that the model hardly ever predicts that a benign lesion is malignant. In the real world, few people with benign conditions would needlessly be alarmed or subjected to unnecessary aggressive treatment as a result of an AI misdiagnosis. For instance, SVM’s 98.1% specificity in our test means only 2 benign cases (from a total of 100 cases) could have been predicted as wrongly predicted cancer and hence could be subjected to two biopsies or anxiety for those patients. By comparison, no more than 16 out of every 1,000 asymptomatic women screened annually are falsely identified as having cancer by human screening radiologists, that is, wheeled into biopsy or further imaging because of their mammographic false alarms.[9] So, for AI with equally high specificity, it may aid in lifting the weight of unnecessary procedures.

These statistics need to be contextualised. Our ML models were used on cytological data, starting with a clinical suspicion (lump present, FNA done). In this type of diagnostic situation, specificity is important to avoid overtreatment of benign masses, whilst keeping high sensitivity to not miss cancers. As a second reader for pathologists, or a kind of triage system, this AI tool could, for example, flag the relatively small percentage of cases it finds malignant with high confidence, and perhaps enable fast-tracking those patients for further testing. Conversely, in a screening setting, heavy emphasis may be placed on sensitivity (even at the cost of higher false-positive rates) to capture as many early CAs as possible. Our results suggests that by changing the balance of the model or the decision threshold, different trade-offs could be obtained: for example, the Gradient Boosting classifier preferred specificity (99.0% specificity and 90.0% sensitivity), a good fit if clinical applications would want to be very strict at calling a malignancy (no false alarm) whereas, one could tune the SVM or use an ensemble to push sensitivity further if clinical use-case demands it. Clinically, with our co-author doctor hat, SVM – if the overall percentage of smears missing is lowered to 6%, as it is now, or even lower – it would be useful as an aid (especially as most of the ones falsely missed were likely borderline cases), but from an auto-diagnosis perspective, sensitivity would have to be ideally improved. However, reaching >95% accuracy and F1 95% in discriminating malignant vs. benign is a strong proof-of-concept that ML can at least equal, if not surpass, diagnostic accuracy in contrast to human experts on common standardised cytological data. We also observe that the interpretability of the model decisions and the pathological plausibility of the features used are important for clinician acceptance, leading to the explainability analysis in the next section.

3.3 Explainability and Analysis of Errors

In addition to its raw performance characteristics, we investigated the models for interpretation and analysed the failures for insights. For decision-tree-based models, such as Random Forest and Gradient Boosting, there exists an intuitive way to evaluate feature importance. Table 3 shows the top 10 features according to the importance obtained with the Random Forest model (Gini importance: How much each feature decreased the impurity/tot impurity (over all at final node (over all trees))).

Table 3. Top 10 feature importances (normalised Gini) from the Random Forest classifier on the Breast Cancer Wisconsin (Diagnostic) dataset.

Feature	Importance Score
Worst perimeter	0.165
Worst radius	0.142

Worst concave points	0.124
Worst area	0.096
Mean concave points	0.086
Mean area	0.057
Mean radius	0.051
Mean concavity	0.045
Area error	0.044
Mean perimeter	0.037

It is interesting to point out that the most important features in the best 3 models are all "worst" (largest) value for the feature: 1) worst perimeter; 2) worst radius; 3) worst concave points.

4) worst area; it's consistent with what we know clinically that the larger the tumour, the more likely it is malignant. Worst perimeter and worst radius (the size of the cell nuclei of the tumour), for example, are both highest: this seems to imply that tumour size is very significant. Similarly, the worst concave points (the number of concave points on the nuclear contour) are highly scored as malignant cells commonly exhibit irregular and spiculated shapes. These features and features such as mean concave points and worst texture, also indicate that the models are using measurements indicative of larger and more irregularly shaped nuclei to differentiate between malignancies and benign lesions. These types of correlations provide face validity to the model: it makes sense to a pathologist by analogy of how cancer is currently recognised (malignant cytology often displays larger nuclei with more variation and less regular shapes).

From a clinical perspective, it is reassuring that the top features of the ML model are biologically plausible. So "worst" (largest) for these features is important to the prediction, since the largest possible cell nuclei, say, would represent an even more abnormal, more cancer-like cell type. This reflects the pathology practice of scavenging all potential malignant cells. Secondly, we observed that the benign samples have, on average, a less variable profile in these feature values, while the malignant cases tend to have, for any given feature type, extremely small and relatively large values of the features. That's why the model can have such a high specificity: the benign samples hardly contain any extremely large or irregular nuclei, so very few of them are misclassified as malignant.

In error analysis, we inspected a list of misclassified patterns by the best model (SVM) to observe any patterns. Of the 4 binomial prediction false negative examples (malignant cases predicted as benign), we observed these cases to have feature values closer to the benign distribution: these were tumours, by cytology, that appear relatively small and regular (and are potentially either low grade or early stage of high grade disease). These are potentially clinically well-differentiated carcinomas, which may be more difficult to distinguish from benign cells. The 2 false positives (benign classified as malignant) were in fact benign lesions, but consisted of a few very large or irregular cells (for example, a case of benign tumour like fibroadenoma with focal cellular atypia), in which most were evaluated as malignant. "Such

errors underscore the fact that, while the model overall is very strong, it's not perfect -- there is overlap in feature space between some of these benign and malignant cases (as there also is between human interpretations). No clear single feature was absent to separate those problematic cases perfectly; instead, these wrong calls might simply need additional information or context (age or imaging findings) to clear up. This hints at a possible effect of multi-modal data, and they could contribute to carrying semantics in the future.

We also compared the errors between the models: random forest and gradient boosting, both of which had a bit lower sensitivity – they likely discarded a few of all the malignant cases that were closer to the borderline value weeding out feature wise than logistic regression, they probably wanted to make sure that whatever they call as malignant is malignant, as a consequence keeping up specificity a bit (still, both of them had lower sensitivity). The logistic model's errors were in line with those of the SVM, this being coherent since they also had a similar performance, both being results of fairly balanced classifiers. Fortunately, none of the models made any spectacularly incorrect predictions on the very obvious cases; the duds were restricted to these more ambiguous samples, which may vex even specialist pathologists. Our clinical co-author independently reviewed a subset of cases he believed to be SVM false negatives or false positives, and we found the SVM cases he believed were incorrect to be similarly subtle. Explainability and error analysis are so important because trust in an AI system must be established. Having a sense of what the model renders its prediction on (eg nuclear size and irregularity of shape) as well as where it fails (eg cases with borderline features) can make clinicians to understand, trust and interact with the tool appropriately: For example, they could use how certain a model is as a “second opinion” and pay particular attention to cases where the model is less certain. Overall, for the ML pipeline, the engineer–clinician team observed that the behaviour was mostly in line with knowledge of the domain, a necessary step to enable safe deployment. Additionally, while we considered overall feature importance, we can integrate recent frameworks like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) into future research. They can offer case-wise explanations of predictions (e.g., which of the cytology features were responsible for a specific diagnosis), adding to transparency and clinician trust in the tools assisted by AIs.

4 Discussion

4.1 Clinical Relevance and Workflow Integration

The excellent diagnostic ability of these ML models has important implications for clinical applications. An AI system performing a distinction between malignant and benign breast lesions with ~95% accuracy might be integrated with the diagnostic process in different ways. One possible application is as a decision aid for cytopathologists reviewing FNA slides. Currently, pathologists visually inspect cell morphology from these slides analytically; this process can be labour-intensive and subjective. A trained ML model (say, our SVM classifier) would quickly process the quantitative features derived from a FNA slide image and produce an immediate preliminary assessment—such as categorising cases into “High likelihood of malignancy” vs “Likely benign.” This would not supersede, but still complement the physician's judgement with additional views: Clinicians could look at the AI-predicted result and the top features and then determine to confirm the diagnosis. Such an instrument could potentially enhance precision and repeatability in diagnosis, particularly in borderline cases or in areas with more junior pathologists. It could also facilitate workflow by helping to prioritise the cases for review; for example, AI calling cases predicted to be malignant could be double-read or fast-tracked for confirmatory testing. Another use case is application in developing regions or telemedicine: if the features can be extracted from a digital pathology

system, then the ML model could be used by clinicians who are diagnosing remotely or who don't have access to a specialist. High specificity would mean the tool wouldn't "cry wolf" that often, which is important if we want the tool to keep trust – you'd want to get a tool that won't be annoying and shout 'Fire!' when there is none. Its high sensitivity suggests it might also serve as a safety net to capture cancers that might otherwise be missed.

To adopt such an AI solution into clinical practice, both user acceptance and regulatory implications must be considered. To oval), so black box algorithms are likely to be viewed with scepticism by clinicians; so the interpretability we argue for (e.g highlighting the cytological features that are important for the model's decision) is crucial. A doctor can confirm (or not) those conclusions under the microscope, if the system can spare reasoning or conclusions (for example, "malignancy predicted, due to very high nuclear perimeter and many concave points detected"). That type of transparency would help assure that the AI model was using real pathological cues to make its decisions. More importantly, any ML diagnostic tool would likely be prospectively trialled before utilisation. A controlled introduction of that could happen – say, by running them in parallel with a human diagnosis in a pilot study to figure out how much better (or if at all) you could do – could be made to happen. If proven, such a tool might limit diagnostic errors and shorten the diagnostic delay. For patients, this translates into potentially earlier diagnosis and treatment of cancer, along with fewer unnecessary biopsies for benign tumours — all in keeping with the aims of precision medicine.

4.2 Technical Constraints and Applicability

Despite these encouraging findings, our study has some technical limitations. Firstly, the models proposed were built and evaluated on a single dataset (WDBC), albeit a small one, and as widely used as it may be, it is very small (569 samples) and also comes from a single source/laboratory. Missing an external validation set, their generalisation is in question: the model might learn specific feature distributions and class patterns of this dataset. In practice, sample preparation, imaging systems, or patient population differences could introduce changes in model performance. Therefore, a WDBC-trained model may not generalise well to data from FNA from a different Hospital without retraining or fine-tuning. Secondly, we did not carry out a thorough hyperparameter search or alternative modelling methods. More advanced methods might even be able to further improve the capability of accuracy, for example, some methods as ensemble methods, or neural networks (which are done on this dataset), have achieved 98-100% [10]. By choosing default or common hyperparameters are models will likely not be as tuned as possible. For instance, the improvement might not yet be saturated in an SVM with properly tuned kernel parameters or a random forest that has more trees and deeper. However, we need to be aware: high accuracies found in the literature might overfit or come from non-independent data (e.g., cross-validation instead of a true hold-out). In ours, having an additional test set gives us a more realistic estimate of performance, at the cost of not training on all of our available data. We rescored the analysis with inclusion of 10-fold cross-validation and 95% bootstrap confidence intervals, while earlier results were primarily generated using a single training and test set split. It thus fortifies the statistical basis and, to an extent, minimises overfitting risk; however, to achieve clinical translation, external verification on independent datasets will still be rotational.

A further limitation is that this analysis was based only on the 30 pre-calculated features. These features are indicative of the cytology slides, but are a summary of them. Some subtle expressions that exist in the original images probably can not be caught by these features. In particular, state-of-the-art deep learning methods might be able to analyse the original FNA slide images (not available here) to potentially find additional predictive patterns (e.g., texture nuances, cell clustering patterns) that are currently not captured by the feature set. Although

we have not investigated deep learning in this work, since we worked on the structured data and classical ML, this is an aspect for improvement. Furthermore, our model neither considers other clinical variables a clinician might use for diagnosis (age of patient, size of tumour in imaging, etc.). The addition of such data might enhance performance or at least increase the clinical utility of the tool (by providing a more detailed evaluation).

Explainability is a separate technical problem – we gave some insights with the feature importances, but since we were not using more advanced (e.g. SHAP or LIME explanations for individual predictions. Such approaches would be required to meticulously debug the model's decision-making process and guarantee it does not depend on any artefact or random correlation. Last, we admit that we have not considered class imbalance beyond the use of say F1-like metrics; there are fewer malignant cases, but the imbalance in say WDBC (37% malignant) is not very strong. When the proportion of presents is a type of deployment where its prevalence is lower (e.g., screening population), we may need to calibrate the output probabilities and the decision threshold in the model in this way, adjusting the trade-off sensitivity/specificity to the clinical context more adequately. In conclusion, our findings are promising but are based on a controlled retrospective study. For these findings to be translated into a reliable clinical tool, they would need to overcome the limitations mentioned above by collecting additional data, refining the algorithm and validating.

4.3 Future Directions: External Validation and Prospective Research Several aspects of the present study suggest promising approaches for future studies.

Building on this work, several future steps are warranted to move closer to real-world application. External validation in independent data sets is the need of the hour. For instance, we might apply the trained models to similar breast cytology datasets from different institutions or to the original images processed using the same feature extraction pipeline. Good performance on external data would indicate the model's generality. If you find the model under-performing, you should investigate transfer learning or domain adaptation to fine-tune the model with new data. Further, a prospective study could be performed where clinicians are using the model in real time. Such a study would allow assessment of how the model works on completely new data (prospective FNA samples) and how physicians use the model's predictions. It would also enable measurement of what the end-to-end impact is: e.g., if we use the ML tool, do we have fewer misdiagnoses or fewer unnecessary biopsies? Do clinicians move quickly with the help of AI? These results are crucial to quantify the value of the technology.

Technically, future work is to try the ensemble and sophisticated algorithms. Because each one had partly different good/weak points (e.g., GB was very specific, SVM was very balanced), an ensemble over all the outputs could likely yield a better classifier. For example, a dumb ensemble like a voting or stacking ensemble may achieve increased sensitivity without a compromise in specificity by capitalising on the diversity of the model decision boundaries. An additional aspect is feature engineering: while we took the features as given, one could generate new features (e.g., polynomial mixtures or domain-motivated indices) or perform dimensionality reduction by feature selection. Another advantage of feature selection is that it can also increase interpretability (and perhaps accuracy) by eliminating irrelevant noise. We could further include cost-sensitive learning such that we penalise false negatives (missing a cancer) more than false positives, allowing us to directly optimise our models for clinician preference for high sensitivity. This could include modifying classification thresholds or relying on algorithms that account for imbalanced costs.

The next obvious step in the future will be to add the imaging information through deep

learning. Since deep neural networks have achieved great success in image-based diagnostics, one can also input the FNA-slide images into a CNN for classification, either independently or jointly with these hand-crafted features. Such a model would be able to learn morphological features not represented by summary statistics. But it would need a larger dataset and proper validation, yet could offer an end-to-end solution (from image to diagnosis) and even bring into consideration the regions of interest contributing to the malignant prediction of the image. Another direction is that we can extend our discussion beyond binary classification. In practice, pathologists may categorise FNA results discretely (benign, atypical, suspicious, malignant). An ML model may be developed for a multi-class outcome or to measure the probability of malignancy, which could be more informative for borderline cases where there is uncertainty.

Lastly, any eventual deployment should comply with regulatory recommendations for AI in medicine, which consider transparency, validation, as well as ongoing monitoring. For deploying the model on new data, the model's input may undergo some distribution shift in contrast (e.g. a new staining technique of slides). It would be essential to have some method of continuing re-training or calibration to maintain performance. User feedback loops could also be built in – say, if a physician disputes the AI on a case, and is ultimately proven right, that case could be fed back into the training set for future betterment. In summary, the direction forward is one not only towards the enhancement of accuracy but also reliability, interpretability, and integration within the clinical workflow, leading to an enhancement of patient care.

5 Conclusion

In this research, we showed the effective combination of clinical experience and AI approaches for addressing the breast cancer diagnosis problem. We developed and tested several ML models using a publicly available breast cancer cytology dataset, all of which were effective in discriminating malignant from benign lesions with high accuracy. The optimal model (SVM) achieved an accuracy of 96.5% with a sensitivity of 93.7% and specificity of 98.1%, which means it can identify most of the cancer and misclassifies very few benign cases. Such performance indicates that ML-based diagnostic aid tools could be used as a complement to classical diagnosis, and may even be able to achieve higher levels of consistency than human experts in some tasks. That the models tend to hinge on features (e.g., deviation from normal nuclear size and shape) that are concordant with known medical information is comforting for trust and adoptability. The joint clinician–engineer co-authored the design of the work, guaranteed that both the algorithmic performance and clinical translations were appropriately taken into account: the engineer applied all his focus on optimising the pipeline and performance metrics, whereas the clinician interpreted results of concrete significance for the patient care purpose.

The results are encouraging; however, we stress that this is a proof of concept under controlled conditions. Real usage will need more validation, getting over limitations, and most of the time, it will be a progressive integration of a tool into practice, still supervised. If those steps can be done successfully, the impact could be profound — earlier and more accurate breast cancer detection, fewer unnecessary procedures for women with benign disease and a data-driven decision support system for busy clinicians. More generally, this work is illustrative of how AI can be used in medicine: by combining technical insight with domain expertise, we can produce solutions that are not only computationally powerful but also compatible with clinical relevance and values. Further development will focus on model improvements, testing in broader clinical settings, and ultimately regulatory approval and integration into routine practice. If developed thoughtfully, AI-based diagnostic tools for breast cancer could

potentially be adopted in the broad fight against breast cancer with great benefit by enabling earlier and more accurate diagnosis.

References

Journal articles

1. E. Strelcena and S. Prakoonwit, "Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study," *BioMedInformatics*, vol. 3, no. 3, Art. no. 3, Sep. 2023, doi: 10.3390/biomedinformatics3030042.
2. "Breast Cancer Wisconsin (Diagnostic) Dataset," *GeeksforGeeks*. Accessed: Jun. 24, 2025. [Online]. Available: <https://www.geeksforgeeks.org/machine-learning/breast-cancer-wisconsin-diagnostic-dataset/>
3. N. A. Elhawary et al., "Descriptive epidemiology of female breast cancer around the world: incidence, mortality, and sociodemographic risks and disparities," *Int. J. Environ. Health Res.*, Apr. 2025, Accessed: Jun. 24, 2025. [Online]. Available: <https://www.tandfonline.com/doi/abs/10.1080/09603123.2025.2492826>
4. Xin Wen, Xing Guo, Shuihua Wang, Zhihai Lu, Yudong Zhang. "Breast cancer diagnosis: A systematic review," *Biocybern. Biomed. Eng.*, vol. 44, no. 1, pp. 119–148, Jan. 2024, doi: 10.1016/j.bbe.2024.01.002.
5. Oliver Díaz, Alejandro Rodríguez-Ruíz, Ioannis Sechopoulos. "Artificial Intelligence for breast cancer detection: Technology, challenges, and prospects." Accessed: Jun. 24, 2025. [Online]. Available: [https://www.ejradiology.com/article/S0720-048X\(24\)00173-6/pdf](https://www.ejradiology.com/article/S0720-048X(24)00173-6/pdf)
6. Mohsen Ghorbian, Saeid Ghorbian. "Usefulness of machine learning and deep learning approaches in screening and early detection of breast cancer." Accessed: Jun. 24, 2025. [Online]. Available: [https://www.cell.com/heliyon/pdf/S2405-8440\(23\)09635-4.pdf](https://www.cell.com/heliyon/pdf/S2405-8440(23)09635-4.pdf)
7. S. Das, S. Koley, and T. Saha, "Machine Learning Approaches for Investigating Breast Cancer," *Biosci. Biotechnol. Res. Asia*, vol. 20, no. 4, pp. 1109–1131, Dec. 2023, doi: 10.13005/bbra/3163.
8. M. M. Hossin, F. M. J. M. Shamrat, M. R. Bhuiyan, R. A. Hira, T. Khan, and S. Molla, "Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset," *Bull. Electr. Eng. Inform.*, vol. 12, no. 4, Art. no. 4, Aug. 2023.
9. H. D. Nelson, K. Tyne, A. Naik, C. Bougatsos, B. K. Chan, and L. Humphrey, "Screening for breast cancer: An update for the U.S. Preventive Services Task Force," *Ann. Intern. Med.*, vol. 151, no. 10, pp. 727–737, Nov. 2009, doi: 10.7326/0003-4819-151-10-200911170-00009.
10. Mahmoud Darwich, Magdy Bayoumi. "An evaluation of the effectiveness of machine learning prediction models in assessing breast cancer risk," *Inform. Med. Unlocked*, vol. 49, p. 101550, Jan. 2024, doi: 10.1016/j.imu.2024.101550.

Appendix 1

Table 4. Feature-wise mean \pm sd for benign and malignant cases
 (Breast Cancer Wisconsin Diagnostic Dataset)

Feature	Mean \pm SD (Benign)	Mean \pm SD (Malignant)
mean radius	12.15 \pm 1.78	17.46 \pm 3.20
mean texture	17.91 \pm 4.00	21.60 \pm 3.78
mean perimeter	78.08 \pm 11.81	115.37 \pm 21.85
mean area	462.79 \pm 134.29	978.38 \pm 367.94
mean smoothness	0.09 \pm 0.01	0.10 \pm 0.01
mean compactness	0.06 \pm 0.05	0.16 \pm 0.10
mean concavity	0.03 \pm 0.03	0.10 \pm 0.09

mean concave points	0.03 \pm 0.02	0.06 \pm 0.04
mean symmetry	0.18 \pm 0.02	0.19 \pm 0.03
mean fractal dimension	0.06 \pm 0.01	0.06 \pm 0.01
radius error	0.40 \pm 0.28	0.73 \pm 0.68
texture error	1.21 \pm 0.69	1.33 \pm 0.91
perimeter error	1.63 \pm 1.15	2.55 \pm 2.19
area error	21.10 \pm 17.77	46.57 \pm 94.52
smoothness error	0.01 \pm 0.00	0.01 \pm 0.00
compactness error	0.03 \pm 0.02	0.06 \pm 0.04
concavity error	0.02 \pm 0.02	0.03 \pm 0.03
concave points error	0.01 \pm 0.01	0.02 \pm 0.02
symmetry error	0.02 \pm 0.01	0.02 \pm 0.01
fractal dimension error	0.00 \pm 0.00	0.01 \pm 0.00

worst radius	14.11 ± 2.52	25.82 ± 4.73
worst texture	25.68 ± 6.14	29.16 ± 5.01

worst perimeter 92.87 ± 13.80 159.26 ± 34.10 worst area 669.89 ±
217.42 2019.79 ± 885.24 worst smoothness 0.16 ± 0.03 0.20 ± 0.06
worst compactness 0.25 ± 0.21 0.45 ± 0.32 worst concavity 0.12 ± 0.10
0.27 ± 0.20 worst concave points 0.05 ± 0.04 0.16 ± 0.09 worst
symmetry 0.29 ± 0.06 0.29 ± 0.07 worst fractal dimension 0.08 ± 0.02
0.09 ± 0.03