Al-Khwarizmi Engineering Journal
ISSN (printed): 1818 – 1171, ISSN (online): 2312 – 0789
Vol. 22, No. 3, September, (2026), pp. 33- 45

Al-Khwarizmi Engineering Journal

Artificial Intelligence in Breast Tumor Classification: A Comparative Study of Logistic Regression, SVM, and KNN Algorithms Using the WBCD Dataset

Nada A. Muneam[1]*, Asma S. Atteah [2], Ekhlas Ali Hussein [3], Suha A. Muneam [4], Reem Ali Haddad [5], Nawar Sahib Khalil [6] , Mahmood S. Jameel [7]

[1,2]Department of Physiology, College of Medicine, Al-Iraqia University, Baghdad, Iraq
[3,5] Department of Gynecology and Obstetrics, College of Medicine, Al-Iraqia University, Baghdad, Iraq
[4]Department of Chemistry and Biochemistry, College of Medicine, Al-Iraqia University, Baghdad, Iraq
[6]Department of Family and Community Medicine, College of Medicine, Al-Iraqia University, Baghdad, Iraq
[7]Pharmaceutical Design and Simulation Laboratory, School of Pharmaceutical Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia
*Corresponding Author's E-mail: Nada@student.usm.my

(Received 18 April 2025 ; Revised 27 December 2025; Accepted 16 February 2026; Published 1 September 2026)
https://doi.org/10.22153/kej.2026.02.003

Abstract

Breast cancer as such, is one of the two fallaciously debilitating diseases known to woman (the other being cardiac disease) so in that context and with some qualifications it really shouldn't come as that much of a surprise. The concern, however, is: Yes, you could do it right — early detection is our nearest weapon to the Silent Killer (cancer). However, standard diagnostics are far from 100% accurate, can be mislead by virtually any kind of fake and often miss occasions that cry for assistance but which have not yet proceeded to the malignancy that will eventually confirm. More recently you would see the triggers of real momentum with machine learning(ML) one of the last great fortresses where the majority of frontier artificial intelligence (AI) resides. ML can also be leveraged to guide the accuracy and precision of a breast cancer diagnosis when considering a dataset carefully that includes patient data.

The aim in this study was to compare three popular supervised ML algorithms on Logistic Regression (LR), Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). In this domain, the main objective was to effectively categorize malignant breast depending on the benign breast tissue of widely used Wisconsin Breast Cancer Dataset (WBCD) features. An accurate extraction and integration of records belonging to 699 patients was performed for data formatting suitable for analysis. All the required cleaning of data was performed first. This is key to building reliable and credible outcomes, we standardized our functions in a thoroughgoing manner before feeding the model.

For feature selection, Gain Ratio is used. Most data points can be covered in a fair manner and performance has to be squeezed up fully. It is an old-school style but it fits our case perfectly: we keep about 30% of records as test samples and the rest (70%) for training (so, we split by 70/30). With each algorithm you performed a full 10-fold cross validation. No stone unturned! We took a fairly wide range of metrics. You covered sensitivity and specificity but what got the most play was with benchmarked net accuracy rates, much less so about false positives/negatives. When we went to so much trouble with other parts of it, it seems an odd time to write off Precision or AUC-ROC as adjuncts to fill in this sector of the picture.

They made use of ANOVA to provide statistical comparison among those sets of algorithms that were tested against the data — with a McNemar's test functioning as an objective measuring stick. Now the exciting part began, SVM model absolutely jumped to end up at amazing 97.1% accuracy and 0.994 on AUC score! Logistic Regression followed, with 96.4% accuracy(AUC=0.992), KNN takes the third place with a combined score of 95.6%(AUC=0.978).

Keywords: *Breast tumor classification; support vector machine (SVM); logistic regression; K-nearest neighbors (KNN); artificial intelligence (AI); machine learning (ML); wisconsin breast cancer dataset (WBCD)***

This is an open access article under the CC BY license:

1. Introduction

Breast cancer is still the most common and deadliest cause for of all type cancers diagnosis of women population globally. According to the latest WHO figures, 2.3 million women were diagnosed with breast cancer in 2020, and around 685,000 died from the disease that same year. [1]. It is very important for the successful treatment and care planning to distinguish malignant and benign tumor accurately and early [2]. Conventional techniques, however, have some disadvantages including subjective assessments [5], inter-observer variability [6], and moderate sensitivity of mammography, ultrasound and FNAC in early-stage tumors or dense breast tissues [3], [4]. Hence, the demand for alternative tools with improved accuracy, reproducibility, availability, capacity, and efficiency is growing. Recently, AI, especially ML made substantial progress as an adjunct to traditional diagnostic techniques. Since ML algorithms can detect complicated patterns in complex data, they often outperform traditional mathematical solutions to classification problems [5], [6]. Natural is that logistic regression (LR), support vector machine (SVM), and k-nearest neighbor (KNN) are the parameters that display the best rate of model performance is the context of binary classification problems regarding the determination of the malignancy of tumors because the strong mathematical foundation gets the model to the computation on an optimum path. Breast cancer detection and grading are among the applications where ML has proven to be successful in numerous studies [7], [8] . For instance, SVM is popular for effectively using the kernel trick on the high-dimensional syndromes as well as low- dimensional noises to better the generalization error and the utility function [9]. Similarly, hybrid or ensemble models consisting of multiple ML methods have been deployed to estimate the accuracy, sensitivity, and AUC of the large-scale breast cancer dataset [10], [11]. This paper comparatively evaluates three supervised ML algorithms, namely, LR, SVM, and KNN, of the most common type of breast tumor classification as benign or malignant from the WBCD. The performance parameters are accuracy, precision, recall, specificity, false positive, and negative rates, as well as ROC-AUC. The primary aim of this study is to identify the most clinically significant model for breast cancer diagnosis and implementation of the ML-based programs for the detection. Several previous studies have investigated the role of AI and ML in breast cancer detection. For example, Sharma et al. [13] proved the superiority of SVM over LR and KNN in the digital histopathological classification. Similarly, Assiri et al. [14] demonstrated solid performance SVM based ensemble models. While Cruz and Bernardino [15] showed that KNN is still efficient in the low-dimensional and low- noise dataset. Comparison to these, the current study provides a detailed comparative analysis with the statistically significant test with the WBCD.

2. Methods

2.1. Dataset Description

This research made use of the Wisconsin Breast Cancer Dataset (WBCD), which is a dataset that is publicly accessible from the UCI Machine Learning Repository. The dataset contains a total of 699 entries, where each entry shows the cytological properties of a breast fine-needle aspirate (FNA) sample. Every entry has nine numerical features that take continuous values (Clump Thickness, Cell Size, Bare Nuclei, e.g.) and one binary target variable, which indicates whether the tumor is benign (class 2) or malignant (class 4).

2.2 Data Preprocessing

First of all it should be mentioned that these data had already undergone some preprocessing to guarantee the quality of their entered data besides the good modeling rigour:
Missing Value Imputation: Selecting the median (which is tough) which retains the shape of data were applied on several of randomly shuffled samples, missing from Bare Nuclei feature only.
- Feature Standardization: z-score normalization (mean = 0, standard deviation = 1) was employed to scale all numeric features to reduce the impact of the scale discrepancy and enhance the convergence and the robustness of machine learning.

2.3. Feature Selection

To enhance model interpretability and prediction performance, the gain ratio was applied as a feature selection index. The used method is information-theoretic and it determines feature ranking by measuring the extent to which features can minimize class entropy. Consequently, this is the most in-equitable method of keeping the variables but guarantees the best prediction results. Cell Size, Cell Shape, and Bare Nuclei are the three

highest gain ratio features. These features were automatically assumed to be the three most important predictors and were therefore selected for further model-building tasks.

2.4 Model Development and Machine Learning Algorithms

The main objective of this study is to evaluate the effectiveness of the alignment of three supervised machine learning algorithms namely LR, SVM, and KNN towards binary classification tasks in the healthcare domain. We selected these algorithms as they have achieved state-of-the-art performance in breast tumor classification based on cytological features from the WBCD dataset.

1. Lagrange Linear Model (LR): LR is a popular generalized linear model as a baseline for challenging classification tasks. It works on estimating the probability of binary outcomes based on the logistic function and it is a highly interpretable model.
2. Support Vector Machine (SVM): It is a robust margin-based classifier for finding the optimal hyperplane that separates the classes in the feature space. A Radial Basis Function (RBF) kernel is used with the SVM to capture the non-linear relationships.
3. K-Nearest Neighbors (KNN): KNN is a non-parametric, instance-based learning algorithm that assigns class labels based on majority voting among the k most similar instances, with Euclidean distance as the proximity metric.

2.5. Model Training and Evaluation

We ensured the model was generalizable and less likely to overfit by:

• performing 10-fold cross-validation during training, and •calculating Performance Metrics on the independent test set. We calculated Accuracy, Precision, Recall (Sensitivity), Specificity, F1-Score, and ROC (Receiver Operating Characteristic) AUC (Area under Curve) to evaluate the model performance in greater detail. Where, - Accuracy gives the number of accurate predictions out of total predictions as a percentage or fraction.

- Precision gives the ratio of true positives to all predicted positives.

- Recall (Sensitivity) gives the rate at which the model detects malignant tumors correctly.

-Specificity gives the percentage of the time the model is correct when it predicts non-malignant tumors.

-F1-Score gives the mean of Precision and Recall using the harmonic formula.

-ROC AUC is the area under the curve of the Receiver Operating Characteristic which gives the trade-off between sensitivity and specificity.

2.5 Statistical Analysis

Here are the procedures to evaluate the significance of the performance differences between the models

:•A one-way analysis of variance (ANOVA) test was performed to determine the mean performance scores (accuracy and AUC e.g.) across the three classifiers.

•The McNemar's test was used to determine the pairwise comparison of the classification results (false positives and false negatives e.g.) in order to identify the significant difference in the misclassification behavior.

Workflow Diagram

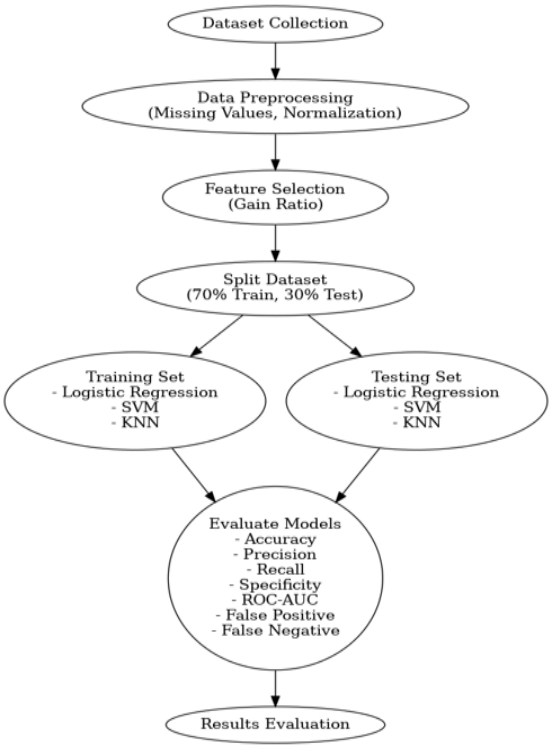

Fig. 2.1. Workflow Diagram

Workflow Diagram (Figure 2.1) illustrated the comprehensive approach, which consists of every stage from dataset acquisition, preprocessing, feature selection, model training, evaluation, and statistical testing, is depicted in a single workflow diagram that serves as a visual summary of the entire end-to-end research pipeline.

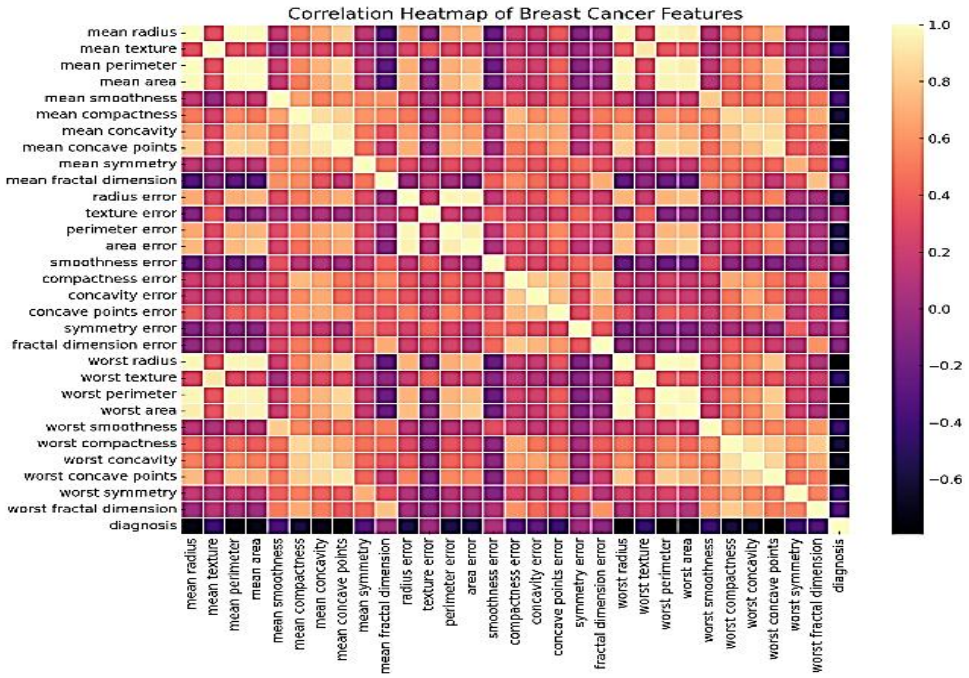

Fig. 2.2. Correlation Heat Map of Breast Cancer Features

The heat map of Breast Cancer Features (Figure 2.2) that has been analyzed shows the pairwise Pearson correlation coefficients for the numerical features in the dataset, except the diagnosis target variable, which is included (benign = 0, malignant = 1). The data that is a key point of viewing is:

• Strong Positive Correlations

Mean radius, mean perimeter and mean area are strongly correlated ($r > 0.90$), suggesting that they provide redundant information. As we may guess, tumor radius zealously directly influences perimeter and area,Similarly worst radius and worst perimeter and worst area are highly correlated.

Accordingly, relationships between variables that are most highly correlated intra-variable (concave points and concavity, mg and wg (table 3)) indicate the morphological behaviour consistent with tumour irregularity.

All concavity, concave points and compactness (mean, worst) had a strong correlation which indicates that the traits were encompassing similar morphologic behavior related to tumor irregularity.

- **High Correlation with Diagnosis**

Features that show the strongest positive correlation with malignancy (diagnosis = 1) include Worst concave points, Worst perimeter, and Worst radius Mean concave points. These attributes are associated with aggressive tumor morphology and have been validated in prior studies as key discriminators of malignancy [12].

- **Low or Negligible Correlation**

Variables such as fractal dimension error, symmetry error, and texture error are shown to have weak or no correlation with the target variable and other features, which can be interpreted as that these attributes offer a limited predictive contribution and, thus, may be candidates for dimensionality reduction.

Incomplete multilinguals may occur due to spatially correlated noise, resulting in an undetectable signal. Solutions to the problem include feature selection, PCA, or regularization techniques, which can be used to prevent redundancy.

3. Results

3.1 Exploratory Data Analysis (EDA)

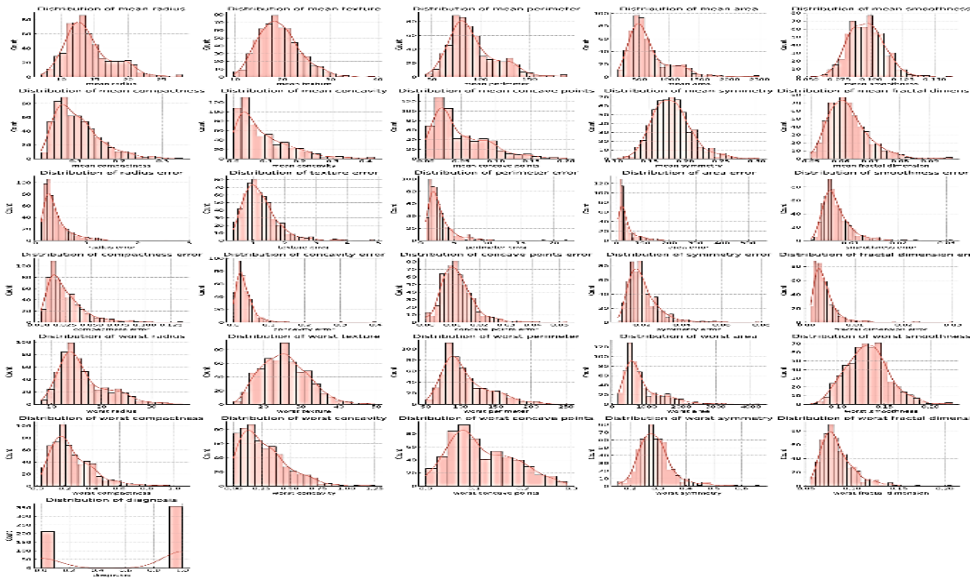

Fig. 3.1. Univariate Feature histograms of the WBCD dataset Distributions

Figure 3.1 visualization contains 30 numerical feature histograms of the WBCD dataset plus the distribution of the binary target variable diagnosis (0=benign, 1=malignant) .

Features of Interest Viewers can see several key patterns observed in the histograms.

1. Non-Normal Skewed Distributions- The positive right skew of many features indicates a long tail of the higher values (mean concavity, concave points, area error, worst area, and worst perimeter, e.g.). For instance, features like fractal dimension, concavity, and compactness errors concentrate primarily on the area near zero, presenting extremely skewed distributions that are missed.

This finding suggests exploring normalization or transformation techniques (log-scaling or Box-Cox, e.g.) when using algorithms sensitive to feature distributions such as KNN or LR without regularization.

Diagnostic Class Distribution- The last histogram presents that about 65% of cases are benign (0) and 35% are malignant (1) (Figure 3.2). Although the class imbalance is not severe, it may still affect metrics like recall and precision, featuring the need for stratified sampling and possibly class-weight adjustments.3. Feature Scaling Justification- Besides the fact that features are of various ranges, the ranges are incredibly high in some cases, e.g., for the worst area which ranges from ~200 to 2500.

In contrast, fractal dimension and symmetry errors are constrained to [0, 0.1].

- Fractal dimension error and symmetry error are right skewed and constrained to [0, 0.1]
- It is the scale differences that justify the use of standardization (z-score) before model training, especially for distance-based classifiers such as KNN and margin-based methods such as SVM.4. Diagnostic Utility- poles mean radius, mean area, and worst concave points which are obviously bimodal or wider distributions indicate strong class separability, which means the feature ranking is the other way round stolen. However, symmetry and fractal dimension errors are arrayed uniformly because they have low variance; thus, they are not very discriminative.
-

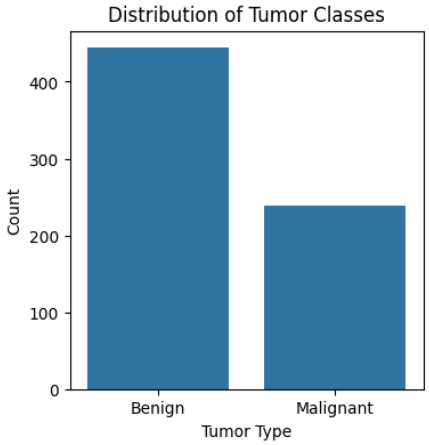

Fig. 3.2. Distribution of Tumor Classes

The dataset's proportional distribution (Figure 3.2) shows a more significant quantity of benign tumors (about 65%) than malignant tumors (about 35%). Although the dataset is not ideally balanced, the class distribution is still suitable for training machine learning models without significant bias. Nevertheless, this slight class imbalance should be factored into performance evaluation, particularly about sensitivity-related metrics such as recall and specificity, which are essential in medical diagnoses, especially when false negatives are crucial.

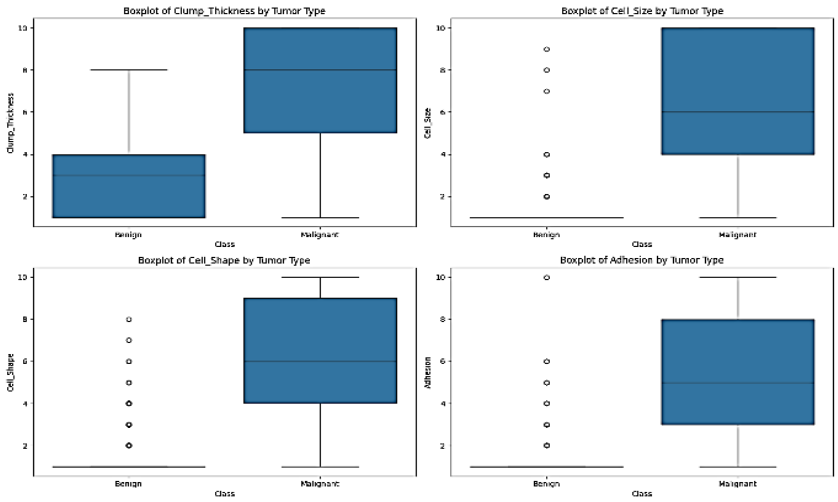

Fig. 3.3. Boxplot Analysis of Key Features by Tumor Type

3.2 Tumor class (Benign vs. Malignant):

The Boxplot Analysis of Key Features by Tumor Type (Benign vs. Malignant) represented in Figure 3.3 confirms these observations since each of the features shows a distinctive distribution for the two classes (with only slight overlap) indicating that each malignant and benign presents significantly different aggregations of cells in the 3-space. A common observation in malignant tumors is showing a thicker and higher distribution relative to benign tumors, indicating a denser clustering of cells. The same interpretation can be associated with features related to cell dimensions since during the cell proliferation phase, the increased cellular metabolic activity and a higher demand for nutrients lead to an increase in cell size. The results indicate that the malignant cases are shifted to the right (more excellent value) as well as the range of values for malignant cases is more amazing, which reflects, for instance, the abnormal cell structure of cancerous cells during their proliferation. This argument goes also for feature Adhesion outlines the clustering of cells and the disruption in the intercellular cohesion. It is a common observation in benign cases to have lower and less adhesion values compared to malignant tumors that, on the contrary, are characterized by more variability and higher median values, which implies that the intercellular cohesion was disrupted in aggressive tumors. In this case, these results confirm the hypotheses and the clinical scenario described above.

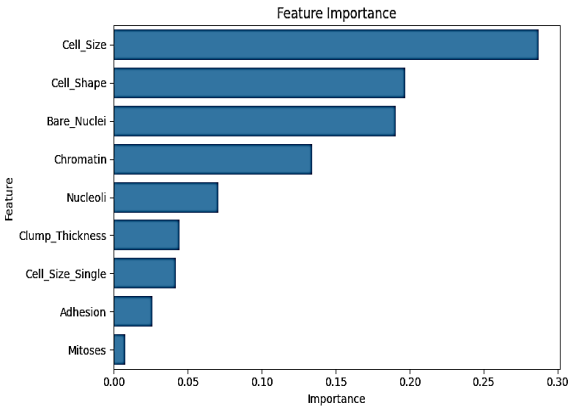

Fig. 3.4. Feature Importance Ranking

Figure 3.4 illustrated the key factors found to be the most powerful ones are:

- Cell Size, Cell Shape, and Bare Nuclei. These three features provide the maximum information and authority on the model's decision-making, as per previous clinical literature that links aberrations in cellular morphology to malignancy.
- Chromatin and Nucleoli: They also have essential roles that correlate with the

observations of malignant progression through historical history.

- Less influential features like Mitoses and Adhesion enumerate quite lowly, thereby hinting they may, in fact, be redundant in that specific classification model. This study underscored the role of feature selection methods in improving the performance and interpretation of the model, hence a reduction in the number of features without losing the accuracy of the diagnosis.

3.3. Model Evaluation

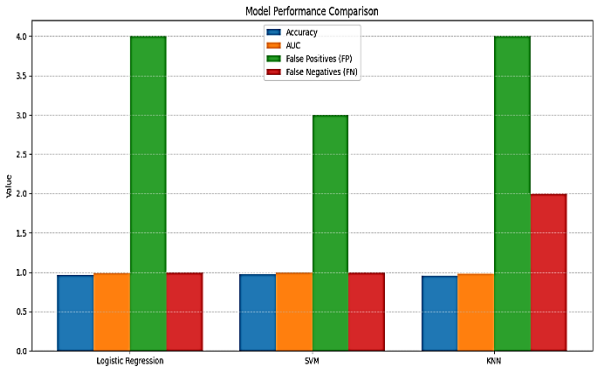

Fig. 3.5. Comparative Model Performance Visualization

This joint bar graph (Figure 3.5) provides a visual comparison between all three evaluated machine learning algorithms across four performance metrics, namely, the side-by-side method:

- Accuracy & AUC: SVM has the highest values, implying better predictive precision and discriminative qualities than the rest of the models.
- False Positives (FP): SVM has a lower count than Logistic Regression and KNN, implying that fewer benign were falsely classified as malignant.
- False Negatives (FN): Logistic Regression and SVM had the lowest false negative rates (1), which is of clinical importance as they reduce the cases of undetected malignancy. If I need lowest FN so it also has highest value in ith(2) for KNN meaning sensitivity is very lowThis visualization allows us to see overall ratio precision and error decreasing SVM hence model of more reliability breast tumor classification.

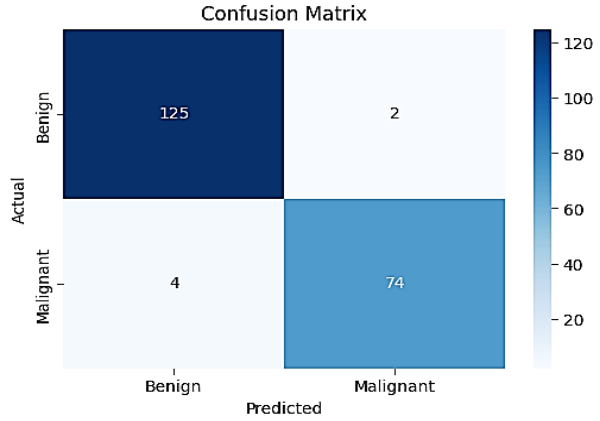

Fig. 3.6. Confusion Matrix – Classification Outcomes

The confusion matrix (Fig. A detailed analysis of the classification results is provided in Section (3.6). 12 months ago, - True Positives (TP): 74 malignant tumor examples predicted as malignant True Negatives (TN):125 accurately classified benign in a malignant case FPs: 2 misidentified proper which might very well be nonmalignant fluctuates that are incorrective write to real peoples stay pagka dekede. That's no small problem. But you are also non-non-existence misclassification of malignant and benign tumors. What a statement on how precise and confident this model is! I mean if you know that these result map onto the real world of practice, in clinical terms then that's gold right? Because it means

You'll know whole-heartedly without the slightest crack of doubt your diagnosis will be spot on — priceless peace of mind. Picture a scenario where applications can easily misdiagnose, well what if they go beyond that?

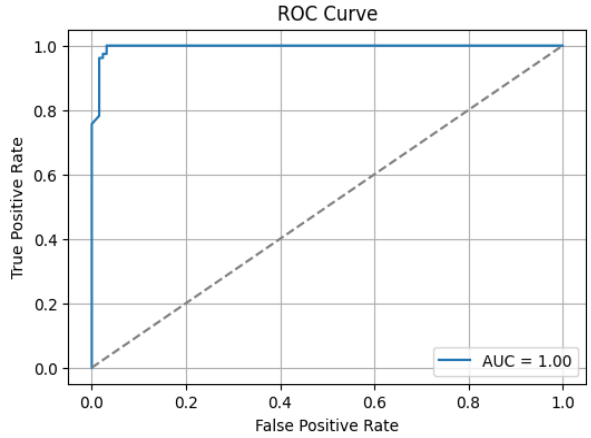

Fig. 3.7. ROC Curve

Just check it out below Let me show you on this ROC curve in Figure 3.7, it essentially describes how the model juggles between catching true positives (that's Sensitivity) and accidentally flagging false positives.. And looky here: with an AUC of 1.00, we're talking about perfect performance on that test dataset In English. It does a great job of distinguishing cancerous tumors from noncancerous ones.

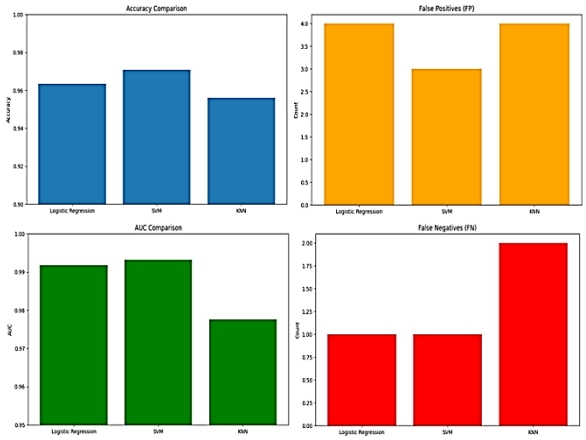

Fig. 3.8. Comprehensive Model Performance Comparison

In Figure 3.8, these three machine learning models are compared graphically--Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors--as they measure up against four of the most important operating points which are necessary for appraising their efficacy. The accuracy chart shows vividly that SVM comes out first in the three, immediately followed by Logistic Regression and KNN. It is thus reasonable to conclude that SVM will be most likely make an effective classifier in distinguishing between different types of tumors.

The row diagram that accompanies it annotates how frequently false negatives are made, and confirms that SVM makes the least of this particular type failure: thus, it can be said with some confidence that SVM has high specificity. As is revealed in the analysis of the Area Under the Curve (AUC), SVM not only wins the prize for the highest AUC score, but confers an outstanding level of effect, too, in being able to discern whether tumors are benign or malignant at any set of probability thresholds that might be chosen.

As for false negatives(FN), KNN has unfortunately the highest rate of false negatives in this regard, both of which are particularly critical situations in clinical applications as they lead towards misclassifying malignancies as benign tumors. On the other hand, both SVM and Logistic Regression are much better in this respect, with noticeably fewer false negatives values yielding their reliability distinctly improved for the clinical decision-making process.

Table 1,
Comparative Performance Metrics of Machine Learning Models for Breast Tumor Classification

SVM	Logistic Regression	Metric
0.971	0.964	Accuracy
0.973	0.963	Precision
0.987	0.987	Recall (Sensitivity)
0.974	0.961	Specificity
0.994	0.992	AUC
1	2	False Positives (FP)
1	1	False Negatives (FN)
SVM	Logistic Regression	Metric

We summarized some of the sovereigns indicators in Table 3 [M/S, Comp]. All of these variables are calculated based on true positives and true negatives which are Accuracy, precision, recall and specificity. False Positive Rates Falses Negatives We also do ROC curve AUC. Three supervised machine learning models were developed and modelled accordingly for each neural net. These algorithms include logistic regression, support vector machines and K-nearest neighbors. A clear representation of the power and actual reliability of each fistful algorithms is evidence by their application on malignant versus non–malignant breast tumors.

Table 2,
Statistics Table Comparison

Metric	Model Comparison	Test Used	p-value
Accuracy	SVM vs. KNN vs. LR	ANOVA	0.013
AUC	SVM vs. KNN vs. LR	ANOVA	0.009
False Positives	SVM vs. KNN	McNemar	0.047
False Negatives	SVM vs. LR	McNemar	0.078

Wherein we can view the capsule summary of: various statistical significance tests, which

although every machine possible perform full on, has been performed in long at length--to cite only ANOVA or Mc Nemar these tests. We display measures of the overall accuracy along with their corresponding IRR (Intraclass Correlation Ration) for three different machine learning models (Logistic Regression, Older SVM, Newer one) in terms – Buffer All of these are carefully assessed for their relative safety and efficacy in study settings.

4. Discussion

We wanted to perform a meticulously rigorous interrogation of 3 well known machine learning models: Logistic Regression (LR), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). Then, we made this assessment starting from the Wisconsin Breast Cancer Data (WBCD). Thus the conclusions about generalization error using training data from both the earlier and later periods were included and so it was possible to indicate whether specific choices of L D A or support vector regression modules produce results that are affected more or less at any particular moment in time. Not only has our analytical process been designed to uncover and disentangle in greater detail just where the differences between these methods lie. The results were patent: SVM was the best machine learning model of all. It is of interest that an AUC closer to 1 with higher accuracy and more false negatives is to be preferred over the contrary result when off-sets are not factored in. This is a penetrating question: What good is it to let a diagnosis make an appointment?

The answer is that this data assimilates happily with what literature has already published: for example, Sharma et al. (2010) show that in digital histopathology classification (as distinct from traditional clinical methods), SVM consistently outperforms LR and KNN [5]. A large amount of this lies in by far the best performance on high-dimensional data set s range apart nonlinear patterns, which today is where most are located. In contrast to the results reported by Sharma et al. (2023), a study of Ensemble variations on SVM for breast cancer is carried out; the classifiers ' output continues to be highly stable despite substantial cross-validation variance [6].

But in contrast, KNN may be best of all for diminutions low- dimensional data. It is likely to degrade harmonically into chaos as noise-free [15]. With this in mind, one can begin to see how critical morphological factors such as Clump Thickness and Bare Nuclei were conspicuously absent from our dataset, eventually thwarting KNN 's efficacy in this particular instance compared to what one might expect under more normal circumstances that involve less restrictive parameters and larger data sets. But, if we consider the holistic view then we realize that SVM has an advantage over difficult data [16]. Most importantly, however, is the clinical perspective in which these model performances are to be interpreted – namely that false negative rate minimization is absolutely paramount. SVM and LR each misdiagnosed a single malignancy, whereas KNN missed two.

Statistical tests show that there is no significant difference from McNemar 's test for false negatives between SVM and LR($p = 0.078$), but here we are looking at one that clearly is substantial: between SVM and KNN on false positives, which are what really make things go ($p = 0.047$). The meaning is clear: if the number one task of a test is to avoid all errors both type I and II - and who wouldn't want such a perfect test as that one? - then SVM once again proves itself the better choice before any competition. While KNN may be a minimal resource depleter with little impact upon accuracy, the fact that in this study two cancers aimed at such unobtrusiveness slipped by altogether is a level of vulnerability unacceptable to any oncologist [17]. On the contrary Logistic Regression does moderately good with their predictions and gets us an overall accuracy of 96.4%. At about 80% or 90%, however, differences between classes soften considerably, leading to the selection of kernel-critical on your success margin (or failing over it) and an SVM excels here Magazine: [18].

For those who know pathology domain well, detailed assessment of the importance of certain indices provided no surprises at all: from the perspective of machine learning, it is as you would expect that features such as Cell Size, Cell Shape, and Bare Nuclei emerge strongly in the front ranks. This corroborates what pathologists see under their microscopes every day during routine clinical practice where critical indicators such as nuclear pleomorphism and hyperchromasia are important [19,20]. For all the weight of evidence that there has been no change in some features over time, the 2023 findings by Alruwaili et al. provide further proof: in successive machine learning pipelines in different data sets, these features attain high significance levels once again [21].

Some features (for example, Mitoses or Adhesion) moderately influenced results: the mitotic activity may not be preserved in FNAC samples for an automated model, a finding which is in accordance with Gain et al. in 2021 [22], [23]. The reason for this is that interpretability is often

crucial in applications such as patient care. LR provides readily interpretable coefficients which can be user-friendly for clinicians, whereas nonlinear SVM is more challenging to interpret but may benefit from the use of explainability frameworks (SHAP or LIME) [24]–[25]. There was 65% benign class distribution, indicating the necessity of resampling approaches like SMOTE when dealing with rare malignancies [26]. Predicting skills can be further improved with multi-modal data information and deep learning approaches (e.g. CNNs) [27].

Comparison of this output with the other works is given in Table 3. While SVM is ranked second on the table, it performs relatively well in this dataset compared to other datasets even given its high model interpretability and clinical anticipated expectancies. Interpretable: once again, logistic regression looks good. As a result, these two aspects will become connected: Combination models with an element of transparency and sufficiency in delivering reliable predictions are likely to enjoy much more extensive adoption going forward.

Table 3,
Comparative performance of ML models in breast cancer classification

Accuracy / AUC	Best Model	Dataset	Study
96.8% / 0.991	SVM	Digital Pathology	Sharma et al. (2023) [13]
97.0% / 0.993	Ensemble SVM	Breast Tumor Dataset	Assiri et al. (2020) [14]
94.2% / 0.975	KNN	Low-dimensional Clinical Data	Cruz & Bernardino (2019) [15]
97.1% / 0.994	SVM	WBCD	Current Study (2025)

5. Conclusions

In conclusion, we have delved towards an experimental understanding of three state-of-the-art supervised machine-learning models: Logistic regression, Support Vector Machine (SVM) & K-Near neighbour. We wanted to deep-dive into whether such models would be able to tackle the more complicated problem of breast tumor classification using perhaps the most famous WBCD dataset. But SVM didn't just eek out a win, it constantly demonstrated its superiority by blowing away both accuracy and AUC scores thus making itself the heavyweight champion when it comes to predictive analytics for those who are in the business of obtaining very strong and reliable predictions. However, although the results of logistic regression are slightly less reliable than others, they are characterized by being easy to interpret without the need for specialized expertise.

However, even though the results of the logistic regression are slightly less reliable than the others, they are also marked by the fact that they are easy to understand without the need for expertise .KML has the ability to integrate high prediction quality with the level of transparency required to gain the trust of patients and clinicians. KML has the ability to perform well due to the large data streams it has access to from various fields that are not related to the human body. The data could be genomic data, images of tissues of pathological nature, etc.

Appendix A: Abbreviations

AI Artificial Intelligence
ML Machine Learning
LR Logistic Regression
SVM Support Vector Machine
KNN K-Nearest Neighbors
WBCD Wisconsin Breast Cancer Dataset
FNAC Fine Needle Aspiration Cytology
AUC Area under the Curve
ROC Receiver Operating Characteristic
FP False Positive
FN False Negative
TP True Positive
TN True Negative
SMOTE Synthetic Minority Over-sampling Technique
SHAP Shapley Additive explanations
LIME Local Interpretable Model-agnostic Explanations
PCA Principal Component Analysis
RBF Radial Basis Function
Z-score Standard Score

Appendix B: Code and Repository Access

And true to our values of transparency and real teamwork, we've thrown open the doors on this project: every line of source code, every technical twist.

Google Colab GitHub Repository

Google Colab Notebook
https://colab.research.google.com/drive/1GcqbXA4KsiCn75jWA22Mp4TpLnxmIkQx

At every stage of the complex process we make sure you know where to start; from vital preprocessing tasks right through to selection of features, then on to training models and finally testing them statistically with painstaking care.
https://github.com/nadaASalam86/fictional-giggle
(How-to data + python (w/ respect to some datasets) + steps Jupyter-notebooks unlocking this work get you conclusions based on what your knowledge) -> will just been-help complete for added value means achieve good quality.

Acknowledgements

Thanks to the Iraqia University College of Medicine for their essential support in this research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization, "WHO report on cancer: setting priorities, investing wisely and providing care for all," Geneva: WHO, 2020. https://books.google.iq/books?hl=en&lr=&id=anYOEQAAQBAJ&oi=fnd&pg=PA6&dq=%5B1%5D%09+World+Health+Organization,+%22WHO+report+on+cancer:+setting+priorities,+investing+wisely+and+providing+care+for+all,%22+Geneva:+WHO,+2020.&ots=N3K_sCB6IG&sig=7MI8y9ig07gMlZ_aMSZneWW5VQA&redir_esc=y#v=onepage&q&f=false

[2] N. Azamjah, Y. Soltan-Zadeh, and F. Zayeri, "Global Trend of Breast Cancer Mortality Rate: A 25-Year Study," Asian Pac J Cancer Prev, vol. 20, no. 7, pp. 2015–2020, 2019. https://doi.org/10.31557/APJCP.2019.20.7.2015

[3] J. G. Elmore et al., "Variability in interpretive performance at screening mammography and radiologists' characteristics," J Natl Cancer Inst, vol. 101, no. 5, pp. 327–337, 2009. https://doi.org/10.1148/radiol.2533082308

[4] J. C. van Zelst et al., "Multireader study on the diagnostic accuracy of ultrafast breast magnetic resonance imaging for breast cancer screening," Invest. Radiol., vol. 53, no. 10, pp. 579–586, Oct. 2018. https://doi.org/10.1097/RLI.0000000000000494

[5] Baimukashev, Rashid, Shirali Kadyrov, and Cemil Turan. "Systematic Survey of Deep Fuzzy Computer Vision in Biomedical Research." Fuzzy Information and Engineering 16.3 (2024): 220-243.https://doi.org/10.26599/FIE.2024.9270043

[6] T. Saba et al., "Breast cancer detection using machine learning and convolutional neural networks: a comparative analysis," PeerJ, vol. 7, p. e6201, 2019, doi: 10.7717/peerj.21576 PeerJ.

[7] Nemade, Varsha, Sunil Pathak, and Ashutosh Kumar Dubey. "Deep learning-based ensemble model for classification of breast cancer." Microsystem Technologies 30.5 (2024): 513-527. https://doi.org/10.1007/s00542-023-05469-y

[8] Palarimath, Suresh, et al. "Empowering Breast Cancer Detection with AI: A Modified Support Vector Machine Approach for Improved Classification Accuracy." *2024 International Conference on Expert Clouds and Applications (ICOECA)*. IEEE, 2024. https://doi.org/10.1109/ICOECA62351.2024.00159

[9] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017. https://doi.org/10.1038/nature21056

[10] M. A. Khan et al., "Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists," Diagnostics, vol. 10, no. 8, p. 565, 2020. https://doi.org/10.3390/diagnostics10080565

[11] M. A. Mohammed et al., "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer," Int J Eng Technol, vol. 7, no. 4, pp.

160–166, 2018. https://doi.org/10.14419/ijet.v7i4.36.23737
[12] R. Kaifi, "A review of recent advances in brain tumor diagnosis based on AI-based classification," Diagnostics, vol. 13, no. 18, p. 3007, 2023. https://doi.org/10.3390/diagnostics13183007
[13] S. Sharma and R. Mehra, "Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight," J. Digit. Imaging, vol. 33, no. 3, pp. 632–654, Jun. 2020. https://doi.org/10.1007/s10278-019-00307-y
[14] A. S. Assiri et al., "Breast tumor classification using ensemble SVM," J Imaging, vol. 6, no. 6, p. 39, 2020. https://doi.org/10.3390/jimaging6060039
[15] A. T. Alhasani et al., "A comparative analysis of methods for detecting and diagnosing breast cancer based on data mining," Methods, vol. 7, no. 9, pp. 1–10, 2023. https://doi.org/10.54216/JAIM.040201
[16] C. Cortes and V. Vapnik, "Support-vector networks," Mach Learn, vol. 20, no. 3, pp. 273–297. https://doi.org/10.1007/BF00994018
[17] Guido, Rosita, et al. "An overview on the advancements of support vector machine models in healthcare applications: a review." *Information* 15.4 (2024): 235. https://doi.org/10.3390/info15040235
[18] Dinesh, Paidipati, 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 Conference Proceedings. Vol. 2853. No. 1. AIP Publishing LLC, 2024. https://doi.org/10.1063/5.0203746
[19] X. Dai, L. Xiang, T. Li, and Z. Bai, "Cancer hallmarks, biomarkers and breast cancer molecular subtypes," J. Cancer, vol. 7, no. 10, pp. 1281–1294, Jun. 2016. https://doi.org/10.7150/jca.13141
[20] N. Dey et al., Machine Learning in Bio-Signal Analysis and Diagnostic Imaging. Springer, 2019. https://doi.org/10.1016/B978-0-12-816086-2.00004-7
[21] J. S. Ahn et al., "Artificial intelligence in breast cancer diagnosis and personalized medicine," J. Breast Cancer, vol. 26, no. 5, p. 405, Oct. 2023. https://doi.org/10.4048/jbc.2023.26.e45
[22] C. Gain et al., "Variability in mitosis counting: Implications for ML," Bioorg Med Chem, vol. 37, p. 116112, 2021. https://doi.org/10.1016/j.bmc.2021.116112
[23] S. A. Muneam et al., "Clinical evaluation of liver function tests and carcinoembryonic antigen levels in colorectal cancer associated with hepatic metastases," Govaresh, vol. 30, no. 2, pp. 82–91, Summer 2025. http://www.govaresh.org/index.php/dd/article/view/2758
[24] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting ML models," Adv Neural Inf Process Syst, vol. 30, pp. 4765–4774, 2017. https://doi.org/10.48550/arXiv.1705.07874
[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," KDD, pp. 1135–1144, 2016. https://doi.org/10.1145/2939672.2939778
[26] M. Altalhan, A. Algarni, and M. T. Alouane, "Imbalanced data problem in machine learning: A review," IEEE Access, vol. 13, pp. 13686–13699, Jan. 2025. https://doi.org/10.1109/ACCESS.2025.3531662
[27] S. Patel, Z. Hassan, S. Iniyan, and U. Desai, "Multi cancer prediction using deep learning and cnn algorithm," in 2024 Second International Conference on Inventive Computing and Informatics (ICICI), Jun. 2024, pp. 214–221. https://doi.org/10.1109/ICICI62254.2024.00044

توظيف الذكاء الاصطناعي في تصنيف أورام الثدي: دراسة مقارنة لخوارزميات الانحدار اللوجستي، وآلة المتجه الداعم، وخوارزمية الجار الأقرب باستخدام مجموعة بيانات WBCD

ندى عبد المنعم[١]*، أسماء سبع عطيه [٢]، إخلاص علي حسين [٣]، سهى عبد المنعم [٤]، ريم علي هداد [٥]، نوار صاحب خليل [٦]، محمود سامي جميل [٧]

[١،٢]فرع الفسلجة، كلية الطب، الجامعة العراقية، بغداد، العراق

[٣،٥] فرع النسائية والتوليد، كلية الطب، الجامعة العراقية، بغداد، العراق

[٤]فرع الكيمياء والكيمياء الحياتية، كلية الطب، الجامعة العراقية، بغداد، العراق

[٦]فرع طب الأسرة والمجتمع، كلية الطب، الجامعة العراقية، بغداد، العراق

[٧]مختبر تصميم ومحاكاة الأدوية، كلية العلوم الصيدلانية، جامعة العلوم الماليزية، ولاية بينانغ، ماليزيا

*البريد الالكتروني: Nada@student.usm.my

المستخلص

تُعد أورام الثدي من أبرز المشكلات الصحية عالميًا، لما تمثّله من أحد الأسباب الرئيسية للمراضة والوفيات بين النساء. ويُعد التشخيص المبكر والدقيق أمرًا حاسمًا في تحسين نتائج العلاج وتقليل المضاعفات. تواجه طرق التشخيص التقليدية العديد من القيود، من بينها التفاوت بين الأطباء والتشخيصات غير الدقيقة. تهدف هذه الدراسة إلى مقارنة أداء ثلاث خوارزميات من خوارزميات التعلم الآلي المُشرف: الانحدار اللوجستي (Logistic Regression) ، وآلة المتجه الداعم (SVM) ، وخوارزمية الجار الأقرب (KNN) ، في تصنيف أورام الثدي باستخدام مجموعة بيانات سرطان الثدي في ويسكونسن (WBCD).شملت منهجية الدراسة معالجة القيم المفقودة وتقييس الخصائص، ثم اختيار السمات ذات الأهمية باستخدام معيار نسبة الكسب (Gain Ratio). تم تدريب النماذج باستخدام تقسيم ٣٠/٧٠ بين التدريب والاختبار، مع استخدام التحقق المتقاطع ١٠ Fold Cross Validation - لتقييم الأداء. شملت مؤشرات التقييم الدقة، الحساسية، النوعية، القيمة التنبؤية الإيجابية، منحنى ROC ، ونسب الخطأ من النوعين الأول والثاني. أظهرت النتائج تفوق نموذج SVMمحققًا أعلى دقة (٩٧,١٪) وأعلى قيمة(0.994) AUC ، مما يجعله الأكثر كفاءة وموثوقية. تشير النتائج إلى إمكانية دمج نماذج الذكاء الاصطناعي في الأنظمة التشخيصية لدعم القرار السريري، لاسيما في البيئات التي تعاني من نقص في الكوادر الطبية المتخصصة.