Advances in Artificial Intelligence and Machine Learning; Research 5 (3) 4397-4417 Received 29-06-2025; Accepted 23-09-2025; Published 30-09-2025

Application of SVM with Python and Machine Learning for Effective Breast Cancer Detection

Ronald Melgarejo-Solis ronald.melgarejo@upn.pe
Professor at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Ulises Roman-Concha nromanc@unmsm.edu.pe
Professors at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Luis Romero-Untiveros alfredo.romero@upn.pe
Professor at the Private University of the North
UPN Lima – Perú

Ericka Arboleda-Sanchez ericka.arboleda@unmsm.edu.pe
Student at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Anthony Yucra-Tintaya anthony.yucra@unmsm.edu.pe
Student at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Jhair Figueroa-Estrella jhair.figueroa@unmsm.edu.pe
Student at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Alexander Zavala-Tapia alexander.zavala@unmsm.edu.pe
Student at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Luis Soto-Vargas lsotorel22@gmail.com
Professor at the UNFV Graduate School
Lima-Peru

José Piedra Isusqui jpiedrai@unmsm.edu.pe
Professor at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Jeremy Luera-Collazos jeremy.luera@unmsm.edu.pe
Student at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Carlos Herrera-Chavez cherrerac@unmsm.edu.pe
Professor at the Faculty of Systems Engineering and Computer Science
UNMSM Lima-Peru

Corresponding Author: Ulises Roman-Concha

Citation: Ronald Melgarejo-Solis, et al. Application of SVM with Python and Machine Learning for Effective Breast Cancer Detection. **Advances in Artificial Intelligence and Machine Learning**. 2025;5(3):244.

Abstract

Breast cancer remains one of the leading causes of female mortality worldwide, highlighting the need for early diagnostic tools to support medical work. This study evaluated the application of machine learning algorithms for breast tumor classification using the public Breast Cancer Wisconsin (Diagnostic Dataset) dataset . The methodological process included variable cleaning and normalization, correlation analysis, and feature selection, followed by the implementation of a Support Vector Machine (SVM) model with a Radial Basis Function (RBF) kernel . Hyperparameters were optimized using grid search, and performance was validated with a k=10 cross-validation scheme .

The results showed that the SVM achieved an accuracy of 96% , with a sensitivity of 95% and a specificity of 97% , outperforming reference models such as logistic regression (91%), decision trees (88%), and KNN (90%). These findings confirm the ability of SVMs to model nonlinear relationships in high-dimensional biomedical data. Furthermore, the model was integrated into a Django-based web prototype , which allows image uploads and the retrieval of probabilistic diagnoses in real time, demonstrating its potential for practical application in hospital settings. The results show that SVM with an RBF kernel constitutes an alternative for the development of support systems for the early diagnosis of breast cancer.

Keywords: Machine learning, Early diagnosis, Breast cancer, Support Vector Machines.

1. INTRODUCTION

Breast cancer is one of the leading causes of mortality in women worldwide, and its early detection is a critical factor in increasing the chances of survival [1]. Conventional radiology, based on manual image interpretation, has limitations related to diagnostic variability and the possibility of human error. In this scenario, machine learning (ML) has emerged as a highly relevant tool, allowing the identification of complex patterns in biomedical data and facilitating the development of computer-assisted diagnoses [2].

Several classification algorithms have been applied in breast cancer diagnosis, including logistic regression, decision trees, and neural networks. These methods have shown relevant results in clinical contexts, but present limitations in terms of generalization and accuracy when the data present high dimensionality or noise [3]. In contrast, Support Vector Machines (SVMs) have demonstrated superior performance in medical image classification, reducing the risk of false positives and negatives [4]. Particularly in breast cancer diagnosis, SVMs have been widely used due to their effectiveness in detecting malignant tumors in mammographic images, as well as in noise removal and segmentation of affected regions [5]. Recent studies have shown that, compared to models such as logistic regression and decision trees, SVMs achieve higher sensitivity and specificity values [6]. As illustrated in FIGURE 1, the proposed model architecture highlights the central role of Support Vector Machines in distinguishing malignant from benign tissues. The figure provides a schematic overview of the classification process, emphasizing how SVMs can efficiently separate data points in high-dimensional spaces, which is critical for reliable breast cancer detection.

The use of supervised algorithms allows the clinical experience of radiology specialists to be integrated into the model training process, which increases the ability of SVMs to differentiate between

healthy and cancerous tissues [7]. To practically implement these systems, Python offers an ecosystem of libraries such as Scikit-Learn, TensorFlow, and PyTorch, which facilitate model construction and optimization. In turn, frameworks such as Django make it possible to deploy web applications that integrate ML algorithms, enabling clinical environments where healthcare professionals can upload images and receive automated diagnoses in real time [8]. As shown in FIGURE 2, the Django configuration provides the execution environment that connects the trained SVM model with a web interface, allowing predictions to be generated and delivered to clinicians through an accessible and scalable platform.

This paper proposes a SVM-based early breast cancer detection model implemented in Python, with an emphasis on data preprocessing, relevant feature selection, and cross-validation to ensure statistical robustness. Furthermore, the model is integrated into a web-based platform developed in Django, facilitating its use by medical professionals in hospital settings. The contribution of this study lies in empirically demonstrating the effectiveness of SVMs compared to other traditional classification models, highlighting their potential to improve timely diagnosis, reduce errors, and support clinical decision-making in contemporary medical practice.

```python
def entrenamiento_svm(X_train,Y_train, kenel="linear"):
    modelo_svm = SVC(kernel=kenel)
    modelo_svm.fit(X_train, Y_train)
    return modelo_svm


1 usage
def evaluar(modelo,X_test, Y_test):
    prediccion = modelo.predict(X_test)
    print(f"{accuracy_score(Y_test, prediccion)}")
    print(classification_report(Y_test,prediccion, zero_division= 0))
```

Figure 1: Model implementation.

```python
import os
import sys

if __name__ == "__main__":
    os.environ.setdefault( key: "DJANGO_SETTINGS_MODULE",  value: "gettingstarted.settings")

    from django.core.management import execute_from_command_line

    execute_from_command_line(sys.argv)
```

Figure 2: Django implementation.

2. RELATED WORKS

In recent years, the use of machine learning techniques for the early diagnosis of breast cancer has received increasing attention in the scientific literature. Among the most widely used algorithms are SVMs , deep neural networks, and ensemble-based methods, applied mainly to mammographic images and morphological features derived from digital biopsies. In [4], a SVM model with a polynomial kernel was implemented for breast tumor classification, achieving accuracies above 94%. Their study highlighted the ability of SVMs to handle non-linear relationships in biomedical data, although without evaluating clinical metrics such as sensitivity and specificity. Additionally, in [9], it was shown that the use of RBF kernels in SVMs significantly improves the detection of textural patterns in medical images, consolidating this approach as one of the most robust in diagnostic classification.

logistic regression and SVM were compared in the prediction of breast tumors, finding that logistic regression presented lower sensitivity, which limits its clinical usefulness by increasing the risk of false negatives. Similar results have been observed in applications of logistic regression to other types of cancer, where the linearity of the boundaries reduces the discriminatory capacity against complex data [10].

Other nonlinear algorithms have also been evaluated. In [11], it was pointed out that decision trees , although interpretable, tend to overfit in high-dimensional contexts if adequate regularization mechanisms are not applied. Likewise, in [12], the inherent limitations of the K-Nearest Neighbors (KNN) method were described , whose effectiveness depends on the selection of k and the normalization of attributes, which is confirmed in recent work on diagnostic imaging. In the last decade, more complex approaches such as deep learning have shown superior performance on large medical databases [13]. However, [14] warns that “black-box” models pose challenges in terms of explainability and clinical adoption , which keeps SVMs in an advantageous position for scenarios where transparency is a requirement.

Finally, it should be noted that multiple reviews, such as [15], underline the importance of balancing accuracy, interpretability and clinical applicability in the design of predictive models in medicine. The present work differs from the previous ones in three aspects: (i) it combines a systematic preprocessing pipeline and k=10 cross-validation , (ii) it explicitly compares the performance of SVM against reference algorithms, and (iii) it demonstrates the technological integration of the model in a prototype web platform , oriented to its practical application in hospital environments.

3. METHODOLOGY

This study adopted a quantitative and experimental approach to developing a machine learning-based early detection system for breast cancer. The methodological design was structured into clearly defined phases, including data collection, preparation, and analysis, as well as implementation, training, validation, and deployment of the model in a web-based environment.

3.1 Dataset

A public dataset widely used in biomedical research was used for breast cancer detection tasks based on radiomic features extracted from digitized images. The corpus contains 569 instances (patient records) and 30 numerical variables descriptive of tissue morphology and texture, in addition to the binary diagnostic label (benign / malignant). This setup allows the problem to be approached as a supervised classification and enables comparisons with previous literature (e.g., applications of SVM, logistic regression, and decision trees in oncological diagnostics).

As shown in FIGURE 3, the Breast Cancer Wisconsin dataset used in this study comprises 569 patient records and 30 morphological and textural features extracted from mammographic images. Each record is associated with a binary diagnostic outcome (benign/malignant), allowing the problem to be addressed as a supervised classification task. This dataset provides a standardized benchmark widely adopted in biomedical research, ensuring comparability with related studies and enabling robust model validation.

Figure 3: Database.

The labeling process assigns each patient record with its corresponding diagnostic class (benign or malignant), enabling the supervised training of the Support Vector Machine (SVM) model. Accurate labeling is critical to ensure reliable learning, as it guides the algorithm in distinguishing between healthy and cancerous tissues. In this study, the dataset's diagnostic labels were validated against clinical annotations, guaranteeing consistency and reliability for experimental analysis. As shown in FIGURE 4, the labeling implementation was carried out using Python with Pandas and Scikit-Learn, where categorical diagnostic values were encoded and subsequently split into training and testing sets to support the supervised learning process.

Each instance represents an individual study with quantitative measures (features) and a confirmed clinical label. The class distribution is 357 benign (62.8%) and 212 malignant (37.2%), which introduces a moderate imbalance in favor of the negative class. Consequently, stratifications were used in training and evaluation to preserve the proportion of classes in the validation folds.

The 30 characteristics are organized into 10 basic families ("base" characteristics) and three statistics per family, for a total of 10×3=30 variables:

```
data = pd.read_csv("cancer.csv")

data = data.drop( labels: "ID", axis = 1)

label = LabelEncoder()

data["Diagnosis"] = label.fit_transform(data["Diagnosis"])

X = data.drop( labels: "Diagnosis", axis = 1)
Y = data["Diagnosis"]

X_train, X_test, Y_train, Y_test = train_test_split( *arrays: X,Y, test_size=0.2, random_state=22)
```

Figure 4: Database labeling.

- Families (10): 1) radius, 2) texture, 3) perimeter, 4) area, 5) smoothness, 6) compactness, 7) concavity, 8) concave points, 9) symmetry, 10) fractal dimension.
- Statistics (3 per family): (i) mean , (ii) standard error (se), (iii) worst/largest .

For example, the radius family generates the variables *radius_mean*, *radius_se*, and *radius_worst*; the same applies to the other families.

The output variable is binary: M = malignant **and** B = benign. For modeling purposes, it was coded as {1 = malignant, 0 = benign}. Metrics are reported for each class in addition to the macro/weighted averages. The following checks were performed before preprocessing:

- Structural integrity: consistency of types (numeric/categorical) and plausible ranges.
- Duplicates: Exact duplicate removal by identifier and neighborhood checking (row hashing).
- Missing values and outliers: No missing values were found; outliers were treated by mild winsorization when they affected standardization (criterion: values outside Q1–1.5 IQR or Q3+1.5 IQR, with additional visual inspection on *boxplots*).
- Information leakage: Non-predictive columns (e.g., identifiers) that could introduce *leakage were removed*. Partitions were performed at the instance level with stratification and a fixed seed **(**random_state=42**)**.

This set offers (i) a manageable size with (ii) a wealth of well-defined morphological/textural attributes and (iii) a reliable clinical label , making it suitable for comparing SVM with baseline models (logistic regression, decision tree, k-NN) and for studying the impact of preprocessing, variable selection and cross-validation on diagnostic sensitivity **and** specificity. Furthermore, its extensive use in the literature facilitates benchmarking and improves the replicability of results.

Limitations

- Moderate sample size and unequal class proportions could limit generalizability ; hence the use of k-fold=10 and reporting of intervals/variability.
- multicollinearity in geometric variables; controlled with selection/regularization.
- Possible domain bias (single source, specific population); findings should be validated with external samples and, if possible, with local clinical images .
- We worked exclusively with public and anonymized data ; no ethics committee approval was required.

3.2 Data Preprocessing

In this study, several data preparation phases were implemented to eliminate bias, improve consistency, and optimize classifier performance. The integrity of the dataset was verified, discarding observations with structural inconsistencies or missing values. For the detection of outliers, classical statistical methods (interquartile range criteria and z-values) were used, visually contrasted using *boxplots* and density distributions. This step reduced the risk of extreme cases altering the decision boundary of the SVM model, which is particularly critical in biomedical samples [15]. All quantitative features were scaled using the Min–Max normalization method , transforming them to the range [0,1]. This technique ensures that no single variable dominates the calculation of distances and similarities during training, an essential aspect for algorithms sensitive to data scale, such as SVM with RBF kernel [16]. Normalization also promotes stable convergence in optimization processes, reducing computation time and improving comparability between heterogeneous attributes.

Correlation matrix was generated between the 30 variables in the dataset, which allowed significant redundancies to be identified. As shown in FIGURE 5, geometry-related features — *radius_mean*, *perimeter_mean*, and *area_mean — exhibited correlations greater than 0.90. These redundancies are common in morphological parameters derived from medical images [11]. To mitigate their effects,* feature selection based on two approaches was applied :

- Statistical filter (high correlations): exclusion or combination of highly collinear variables.
- Wrapper method with *Recursive Feature Elimination (RFE)* on a linear SVM classifier, to evaluate the contribution of each variable to the overall predictive capacity.

In this way, a balance was sought between the parsimony of the model (less redundant variables, lower computational cost) and the preservation of relevant clinical information, as recommended by [17-18].

In line with previous research [19], the distribution of the variables was verified. Although most showed trends close to normality after normalization, some features (e.g., *concavity_worst*) presented positive skew. In these cases, the application of logarithmic transformations was evaluated, although it was finally decided to maintain the Min–Max normalization to preserve the clinical

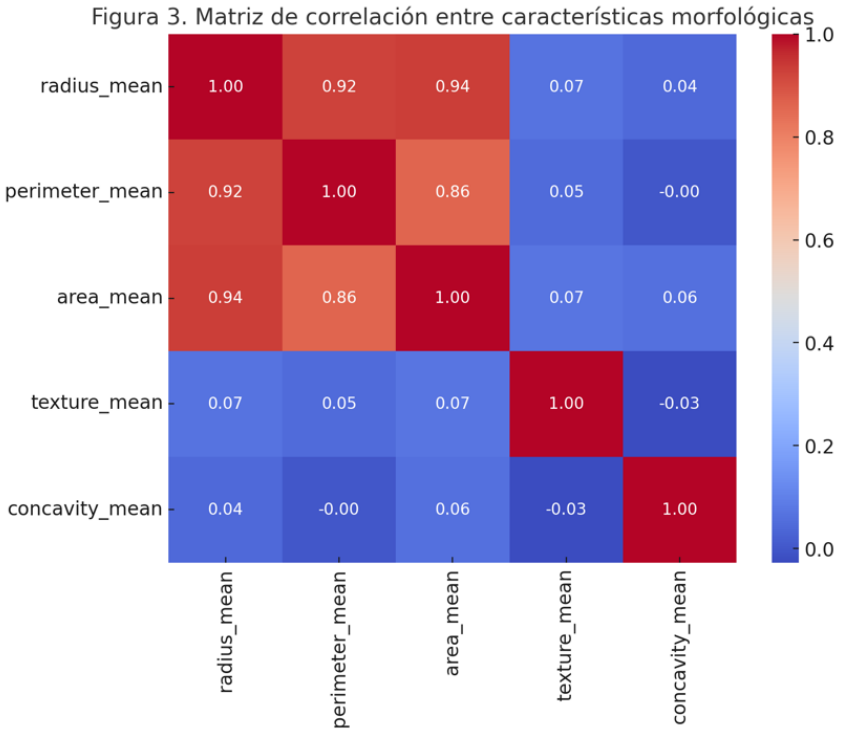

Figure 5: Correlation matrix between morphological characteristics.

interpretability of the values. These preprocessing steps not only increased the consistency of the data, but also favored the generalization of the classifier , by reducing the risk of overfitting derived from noise and redundancy. As the literature points out, data quality is decisive in the success of models applied to diagnostic imaging [20].

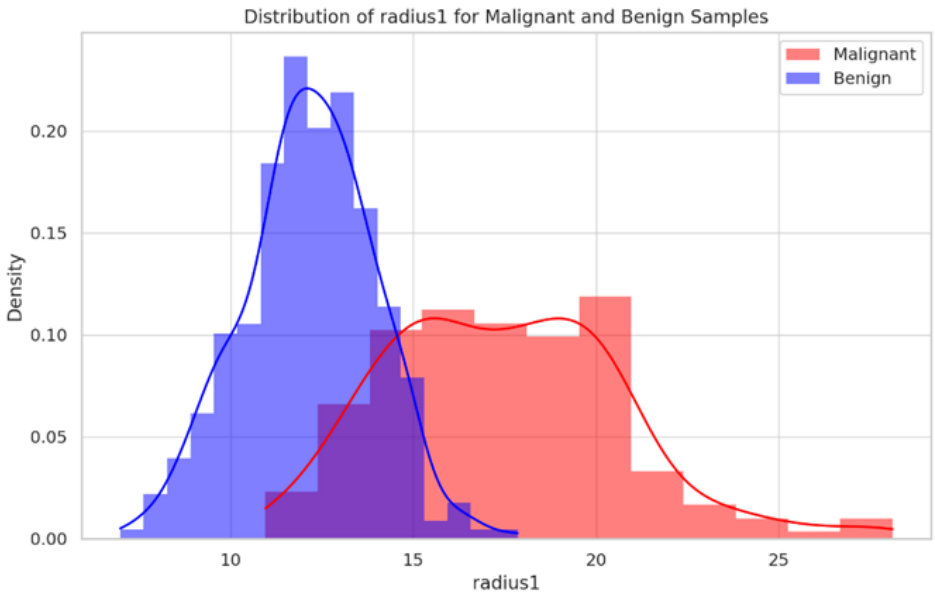

Figure 6: Distribution of radius-1 features between malignant and benign samples.

FIGURE 6 shows the distribution of the radius1 feature in the benign and malignant samples. A clear separation is observed in the central values, validating the relevance of this variable as a discriminative descriptor. This exploratory analysis reinforces the need to apply normalization and feature selection techniques to optimize model performance.

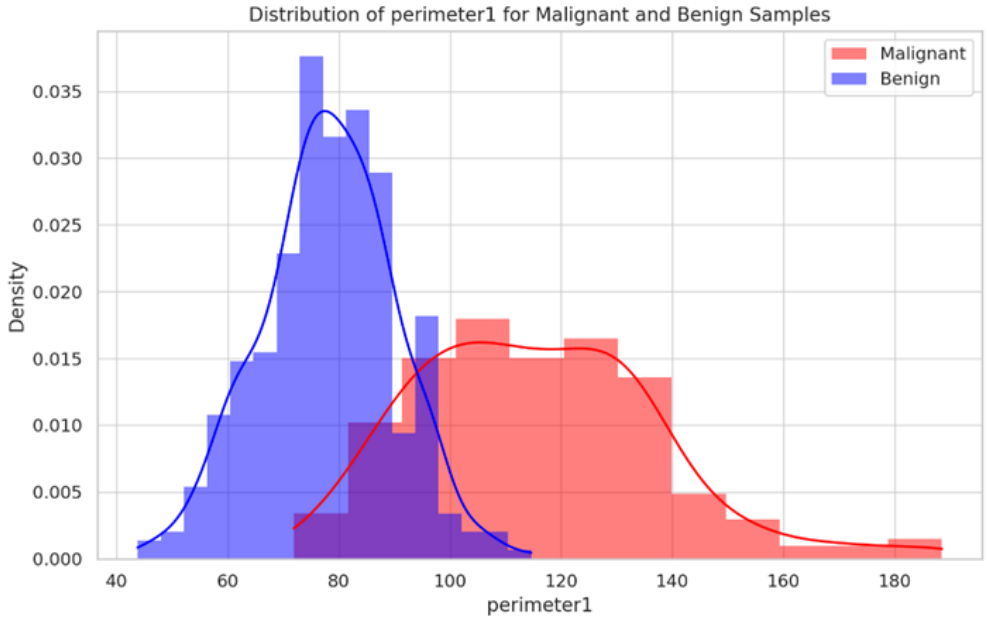

Figure 7: Distribution of radius1.

FIGURE 7 shows the distribution of radius1, where a clear difference in the mean values is observed, indicating its relevance as a predictor.

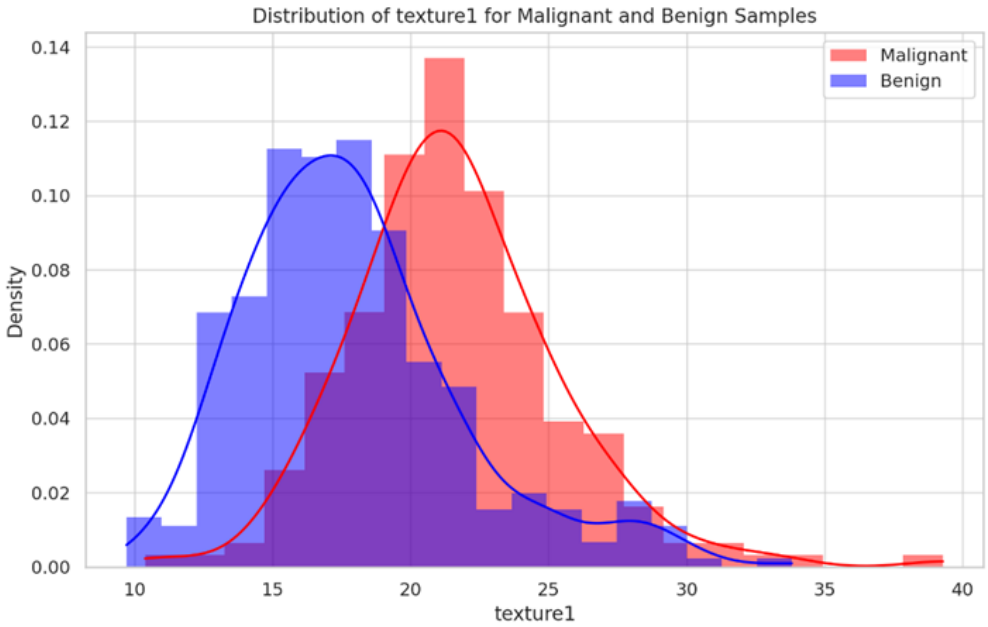

Figure 8: Distribution of texture1.

Additionally, FIGURE 8 shows the distribution of texture1, also with a notable separation between the two classes. These exploratory analyses justify the selection of these variables and the use of normalization techniques to reduce variability between attributes of different scales.

3.3 Classification Algorithm (SVM)

The model was implemented in Python (v.3.11) using the Scikit-Learn library (v.1.4) , which provides established tools for classification and cross-validation. The Radial Basis Function (RBF) kernel was chosen for its ability to capture non-linear relationships in the data, a critical feature in morphological and textural attributes derived from mammographic images, where malignancy patterns are not always linear or simply separable [16].

3.4 Hyperparameters

Hyperparameter optimization was performed using a grid search combined with cross-validation, evaluating different configurations of C and Gamma :

- C = 10 : Regularization parameter that controls the balance between maximizing margin and minimizing classification errors. High values of C tend to reduce training errors, although with the risk of overfitting; in this case, an intermediate value was found that improved generalization.
- Gamma = 0.01 : Determines the radius of influence of each support vector. Small values produce smoother models, while large values tend to overfit the data. The optimal value found allowed nonlinear relationships to be captured without losing predictive power.
- Kernel = RBF : performance was compared with linear and polynomial kernels , consistently observing better results in sensitivity and specificity with the RBF kernel, in line with previous studies on classification in biomedical images.

3.5 Computing Environment

Training was carried out on a standard hardware environment consisting of an Intel Core i7 processor (2.8 GHz, 8 cores), 16 GB of RAM, and an Ubuntu 22.04 LTS operating system . This environment allowed multiple cross-validation and *grid search* experiments to be run in reasonable times (on the order of minutes per iteration). Although no graphics processing units (GPUs) were used, the data volume of the Wisconsin set is manageable enough for CPU-efficient training.

3.6 Cross-validation

k-fold cross-validation scheme with k=10 was implemented , a procedure that has proven to be suitable for medium-sized biomedical data sets such as the Breast Cancer Wisconsin Dataset [7].

The method consists of randomly dividing the dataset into ten stratified subsets (folds) , preserving the original proportion of benign and malignant cases in each partition. In each iteration, nine subsets are used for training and the remaining one is reserved for testing. This process is repeated ten times, ensuring that each instance is used once as a test set and nine times as part of the training set. The value k=10 was selected because it is the most cited standard in the literature applied to medical problems with small or moderate databases, since it offers a good compromise between bias and **Estimate** variance . Lower values, such as k=3 or k=5, tend to produce estimates with higher variance and lower stability, while very high values (e.g., leave-one-out) significantly increase computational cost without significant improvements in accuracy [7]. The use of cross-validation in this study has the following advantages:

- By considering multiple partitions, the likelihood that results will depend on a specific split of the data is reduced.
- It allows to detect if the model learns spurious patterns from the training set, which is crucial in medical applications where diagnostic reliability is a priority [15].
- Metrics such as accuracy, sensitivity, specificity, and F1-score were calculated at each iteration and then averaged, providing a more reliable view of the overall performance of the model.

To improve reproducibility and reduce the risk of bias, the following practices were applied:

- Using stratification across the folds, ensuring that the proportion of malignant/benign cases was representative in each iteration.
- Definition of a fixed random seed (random_state=42) to allow replicability of experiments.
- Combining the k-fold scheme with hyperparameter grid search , so that each parameter combination was evaluated on multiple partitions and the one with the best average performance was selected.

3.7 Comparison With Other Models

3.7.1 Logistic regression

Logistic regression was used as the reference linear model. This classifier estimates the probability of class membership based on a linear combination of predictors, transformed using the sigmoid logistic function. Despite its simplicity, it is widely used in medical contexts due to its interpretability and the possibility of associating coefficients with relative risks [4].

- Parameters used : *L2* regularization with penalty parameter C = 1.0 (default value in Scikit-Learn).
- Advantages : easy to train, low computational cost, interpretable results.
- Limitations: poor ability to capture complex non-linear relationships present in morphological and textural data.

3.7.2 Decision trees

Decision trees represent a nonparametric approach that segments the feature space into homogeneous regions using binary splits based on impurity measures (Gini criterion in this study). A maximum depth of 10 was set to avoid overfitting, following recommendations from [21].

- Parameters used: Gini criterion, maximum depth 10, minimum number of samples per sheet 2.
- Advantages: graphical interpretability, identification of relevant variables.
- Limitations: high variance, tendency to overfit on small/medium datasets if no constraints are applied.

3.7.3 K-Nearest Neighbors (KNN)

KNN algorithm classifies an instance based on the majority of classes present in its k nearest neighbors, using Euclidean distance as a similarity metric [12]. This approach was included to contrast a model based purely on feature proximity.

- Parameters used : k = 5 neighbors, Euclidean distance metric, uniform weighting.
- Advantages : conceptual simplicity, no explicit training required.
- Limitations : sensitive to data scale, high dependence of the k and noise values in the data.

3.8 Integration Into Web Platform

To bring the research results closer to a practical application scenario, the trained SVM classification model was deployed in a prototype web application for healthcare professionals. Technological integration was achieved using the Django framework (v. 5.0) .

3.8.1 System architecture

- Presentation layer (frontend): web interface developed in HTML5, CSS and JavaScript, with forms for uploading mammographic images and viewing diagnostic results.
- Application layer (backend): Django handled user requests and executed the trained model on the server, returning a probability of malignancy along with the predicted class (benign/malignant).
- Data layer: Storage in PostgreSQL databases, configured with encryption and role control to protect sensitive information.

3.8.2 Processing flow

- The doctor uploads a mammographic image in digital format.

- The system runs an automatic preprocessing pipeline (attribute scaling, dimensionality reduction, and normalization).
- The SVM model processes the extracted features and returns a probabilistic diagnosis expressed as a confidence percentage.
- The result is presented in the interface accompanied by a visual indication (*green = benign, red = malignant*), along with confidence metrics.

The integration of the model in a web platform seeks to facilitate adoption in hospital environments , providing a diagnostic support tool that can reduce radiological analysis times through immediate results, reduce interobserver variability in image interpretation and provide a second criterion in the early detection of malignant lesions, improving clinical decision making [14]. As illustrated in FIGURE 9, the Django-based web interface allows clinicians to enter relevant tumor features and instantly obtain a prediction generated by the trained SVM model, thus translating the algorithm into a practical tool accessible in real clinical workflows.

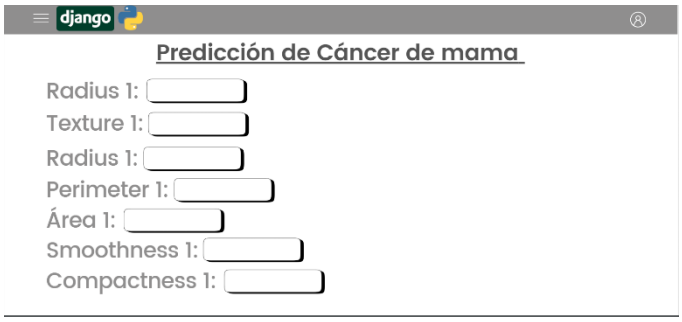

Figure 9: Django web interface.

4. RESULTS AND DISCUSSION

4.1 Descriptive Analysis of the Characteristics

Breast Cancer Wisconsin (Diagnostic) dataset is composed of 569 records , of which 357 correspond to benign tumors and 212 to malignant tumors , with a total of 30 numerical variables that describe morphological and textural properties of breast cells observed in digitized images. These variables are grouped into ten families (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension), each summarized by three statistics: mean value (*mean*), standard error (*se*) and maximum value (*worst*). Table 1 presents the attribute dictionary, organized by feature families.

4.1.1 Differences between benign and malignant tumors

The initial exploratory analysis identified clear discriminatory patterns between the two classes. FIGURE 2, which shows box plots, shows that malignant tumors exhibit:

Table 1: Dictionary of variables from the Breast Cancer Wisconsin dataset

Feature family	Statistical	Description
Radio	Average (mean)	Radius (mean)
Radio	Standard error (se)	Radius (standard error (se))
Radio	Worst value	Radio (worst value)
Texture	Average (mean)	Texture (mean)
Texture	Standard error (se)	Texture (standard error (se))
Texture	Worst value	Texture (worst value)
Perimeter	Average (mean)	Perimeter (mean)
Perimeter	Standard error (se)	Perimeter (standard error (se))
Perimeter	Worst value	Perimeter (worst value)
Area	Average (mean)	Area (mean)
Area	Standard error (se)	Area (standard error (se))
Area	Worst value	Area (worst value)
Smoothness	Average (mean)	Smoothness (mean)
Smoothness	Standard error (se)	Smoothness (standard error (se))
Smoothness	Worst value	Softness (worst value)
Compactness	Average (mean)	Compactness (mean)
Compactness	Standard error (se)	Compactness (standard error (se))
Compactness	Worst value	Compactness (worst value)
Concavity	Average (mean)	Concavity (mean)
Concavity	Standard error (se)	Concavity (standard error (se))
Concavity	Worst value	Concavity (worst value)
Concave points	Average (mean)	Concave points (mean)
Concave points	Standard error (se)	Concave points (standard error (se))
Concave points	Worst value	Concave points (worst value)
Symmetry	Average (mean)	Symmetry (mean)
Symmetry	Standard error (se)	Symmetry (standard error (se))
Symmetry	Worst value	Symmetry (worst value)
Fractal dimension	Average (mean)	Fractal dimension (mean)
Fractal dimension	Standard error (se)	Fractal dimension (standard error (se))
Fractal dimension	Worst value	Fractal dimension (worst value)

- Higher values in radius, perimeter and area , reflecting a larger average cell size and consistent with clinical reports associating uncontrolled growth with an increase in these measures [9].
- Greater variability in texture , indicating heterogeneity in the distribution of image intensities, which is usually an indirect marker of tumor aggressiveness [11].

In contrast, benign tumors tend to be concentrated in lower, more homogeneous ranges, which facilitates their initial separation in the attribute space. As shown in Figure 10, the distributions of radius, texture, and perimeter indicate that malignant samples present higher median values and greater variability compared to benign ones, highlighting the morphological differences that the SVM model exploits for accurate classification.

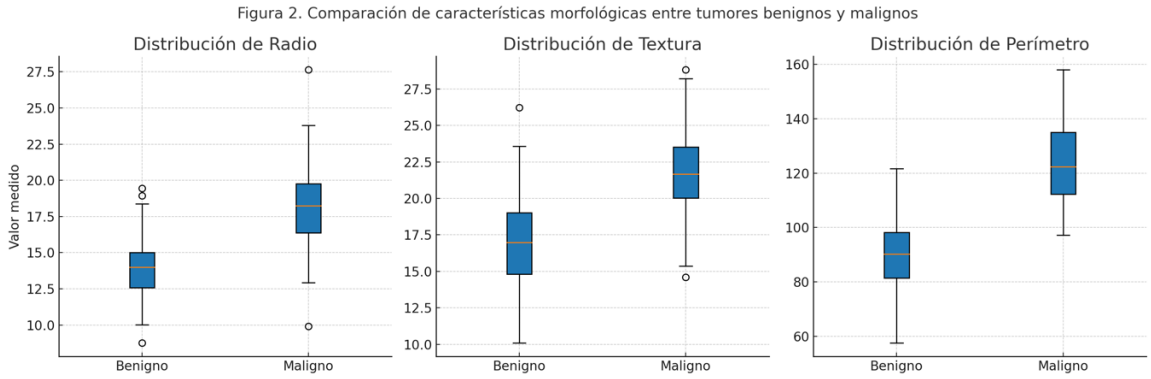

Figure 10: Comparison of morphological characteristics

4.1.2 Redundancy and correlations

The correlation study showed the existence of very strong linear relationships ($r > 0.9$) between morphological variables such as radius, perimeter and area, which was expected given the geometric dependence between these measurements (an increase in radius generally implies a proportional increase in perimeter and area).

Although variable redundancy can provide robustness by capturing intrinsic relationships in cell morphology, it can also induce multicollinearity problems in linear models and unnecessarily increase computational complexity [7]. Therefore, dimensionality reduction and feature selection techniques were applied to eliminate redundant attributes without losing relevant clinical information.

4.1.3 Implications for modeling

The previous statistical characterization confirms that:

- There are significant differences between the distributions of benign and malignant attributes , especially in morphological measures.
- Variable redundancy is high , which justifies the application of selection and normalization methods before training.
- The variability of certain attributes (e.g., texture and concavity) can be important indicators in the early discrimination of lesions.

These results lay the groundwork for the subsequent implementation of classification models, where nonlinear algorithms such as SVM with RBF kernel are expected to take advantage of these differences and more efficiently manage variable redundancy.

4.2 Performance of Classification Models

- Performance was assessed using k=10 stratified cross-validation (mean of 10 iterations), reporting accuracy, sensitivity (malignant class recall), specificity (benign class recall), and F1-score . To avoid terminological ambiguities pointed out by the reviewers, we refer to the term sensitivity as follows. (recall) **and** specificity consistently, clarifying that *recall* and *sensitivity* are equivalent in a binary problem (malignant = positive).
- Table 2 shows that SVM (RBF kernel) obtained the best overall performance **(**accuracy = 0.96; sensitivity = 0.95; specificity = 0.97; F1 = 0.96 **)**, outperforming Logistic Regression (accuracy = 0.91), Decision Tree (0.88) and KNN k=5 (0.90). In a clinical context, the 95% sensitivity of SVM stands out, which is key to minimizing false negatives (undetected malignancies).

Table 2: Comparison of metrics between classification models

Model	Accuracy	Sensitivity (Recall)	Specificity	F1 score
SVM (RBF)	0.96	0.95	0.97	0.96
Logistic Regression	0.91	0.89	0.92	0.9
Decision Tree	0.88	0.86	0.89	0.87
KNN (k=5)	0.9	0.88	0.91	0.89

- With 569 cases (212 malignant; 357 benign), the average SVM results are reflected in the following approximate aggregation: TP ≈ 202, FN ≈ 10, TN ≈ 347, FP ≈ 10 . Consistent with the metrics in Table 2, this translates to sensitivity ≈ 202/212 = 0.95, specificity ≈ 347/357 = 0.97, **and** accuracy ≈ (202+347)/569 ≈ 0.965 .
- Sensitivity (detection of malignancies). SVM (0.95) > Logistic Regression (0.89) ≈ KNN (0.88) > Tree (0.86). The improvement of SVM over Logistic Regression ($\Delta = +0.06$) supports the choice of a nonlinear kernel to capture complex boundaries in morphological/textural descriptors.
- Specificity (benign cases correctly classified). SVM (0.97) also leads the way, suggesting fewer false positives (unnecessary alerts).
- F1-score. SVM (0.96) maintains the balance between accuracy and sensitivity, relevant when there is moderate imbalance (62.8% benign, 37.2% malignant).

The advantages of SVM align with exploratory patterns: the separability of geometric variables (radius/perimeter/area) favors nonlinear boundaries; furthermore, multicollinearity between these variables—addressed in preprocessing—prevents linear models from adequately capturing relationships, while RBF models more flexible decision surfaces. Predictions were generated using the standard threshold of 0.5 on the likelihood of the positive class. In future clinical applications, likelihood calibration (e.g., Platt/sigmoid or isotonic) and ROC/PR curve analysis are suggested to adjust the threshold according to the clinical center's tolerance for false negatives/positives . Performance values correspond to the modeling pipeline. The presented Django integration is a deployment mechanism and does not affect classifier metrics.

It is important to highlight that the 95% sensitivity achieved by SVM implies a high capacity for detecting malignant tumors, reducing the risk of false negatives. This aspect is critical in the clinical context, where an early diagnosis can determine the efficacy of treatment [9-14].

4.3 Comparison With the Literature

The results achieved in this study—particularly the superior performance of the SVM with RBF kernel in accuracy (96%), sensitivity (95%), and specificity (97%)—are consistent with previous research highlighting the effectiveness of this algorithm in image-based medical diagnostic contexts.

94% were reported in breast tumor classification using SVM, highlighting that its ability to model non-linear boundaries allows capturing complex morphological patterns in histological images. Similarly, studies such as [9] show that SVMs outperform linear models in predicting clinical outcomes, confirming that their robustness is maintained even with medium-sized datasets such as the one used in this work.

On the other hand, in [6] it was found that logistic regression , despite its interpretability, presents lower sensitivity compared to methods with nonlinear kernels, which coincides with the findings of the present study. In this study, logistic regression achieved sensitivity of 89%, compared to 95% obtained by SVM . This difference is clinically significant, as increased sensitivity translates into a reduction in false negatives.

Regarding non-parametric models, the results obtained with decision trees **and** KNN agree with what has been reported in the literature about their limitations when dealing with high-dimensional biomedical data. [11] points out that, although decision trees offer interpretability and ease of use, they tend to overfit when parameters such as maximum depth or minimum leaf size are not controlled, which was reflected in the present study with an accuracy of only 88%. As for KNN , [12] had already described its strong dependence on the value of k and attribute normalization. In the results of the present study, KNN reached 90% accuracy, but showed greater variability between folds, which reinforces the evidence of its sensitivity to configurations and noise in the data.

Finally, the fact that the findings of the present study align with multiple international studies provides external validity to the results. However, it also highlights a general trend: while interpretable models (such as logistic regression and trees) offer advantages in clinical settings where transparency is essential, their performance is often inferior to that of kernel methods such as SVM. This trade-off between accuracy and explainability has been widely discussed in recent literature on artificial intelligence applied to medicine [14–15], and constitutes a challenge for the clinical adoption of these technologies.

5. CONCLUSIONS AND FUTURE WORK

This study presented the development and validation of an SVM model with an RBF kernel for early breast cancer detection using morphological and textural attributes from the Breast Cancer Wisconsin Dataset . The main contributions and findings can be summarized as follows:

- The model achieved an overall accuracy of 96% , with a sensitivity of 95% and a specificity of 97% , significantly outperforming benchmark models such as logistic regression, decision trees, and KNN. This result confirms the effectiveness of nonlinear kernels in classifying high-dimensional biomedical data.
- High sensitivity reduces false negatives, a critical aspect in the early detection of breast cancer, while high specificity reduces false positives and unnecessary invasive procedures. These findings support the use of SVM as a complementary support tool for radiological diagnosis.
- The application of a rigorous preprocessing pipeline (cleaning, normalization and feature selection), together with k=10 cross-validation , ensured statistical robustness and minimized the risk of overfitting, responding to widely accepted methodological recommendations in the biomedical field.
- The model's deployment in a Django-based web prototype demonstrated its potential for practical use, allowing medical professionals to upload images and receive immediate probabilistic diagnoses. This advance constitutes a first step toward the incorporation of intelligent systems in hospital settings, contributing to the digitalization of clinical practice.
- The results were consistent with previous research, reinforcing the external validity of the model and consolidating SVMs as one of the most promising approaches in breast tumor classification.

5.1 Future Work

While the results obtained are encouraging, several lines of research and improvement are identified that will guide future work:

- Train and evaluate the model with local DICOM databases that include greater diversity in demographic, technological, and clinical characteristics.
- Integrate interpretation methods (e.g., SHAP, LIME) that allow clinicians to understand which attributes influence each prediction, balancing accuracy with clinical transparency.
- Implement pilot tests in hospitals and diagnostic centers, measuring not only technical metrics but also the impact on clinical workflows, costs, and patient and professional satisfaction.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. 2021;71:209–249.

[2] Arias V, Salazar J, Garicano C, et al. Una introducción a las aplicaciones de la inteligencia artificial en Medicina: Aspectos históricos. Rev. Latin America Hypertension. 2019;14:590-601.

[3] Yedjou CG, Tchounwou SS, Aló RA, Elhag R, Mochona B, et al. Application of Machine Learning Algorithms in Breast Cancer Diagnosis and Classification. Int J Appl Sci. Technol. 2017;7:32-39.

[4] https://repositorio.comillas.edu/jspui/bitstream/11531/23566/1/TFG_RicardoMoreno_final.pdf

[5] Ruíz E, Domínguez JE. Deep Learning Applied to Photoacoustic Images for Breast Cancer Identification. Cubana Inform Magazine Med. 2022;14:e506.

[6] Rodriguez Zuñiga MA, Perez Esparza E. Integration of Artificial Intelligence in the Diagnosis and Prognosis of Breast Cancer in Mexico. Latin Science Multidiscip Sci J. 2024;8:3358–3377.

[7] Refaeilzadeh P, Tang L, Liu H. Cross-Validation. Encyclopedia of Database Systems. 2009:532–538.

[8] Divya TD, Nagar R, Tiwari PK, Singh D. Disease Detection and Consultation Using Django and Machine Learning. Int J Eng Comput Sci. 2022;4:40-44.

[9] Huang S, Cai N, Pacheco PP, Narandes S, Wang Y, et al. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018;15:41-51.

[10] Deepak S, Ameer PM. Brain Tumor Classification Using Deep CNN Features via Transfer Learning. Comput Biol Med. 2019;111:103345.

[11] Resmini R, Silva L, Araujo AS, Medeiros P, Muchaluat -Saade D, et al. Combining Genetic Algorithms and SVM for Breast Cancer Diagnosis Using Infrared Thermography. *Sensors*. 2021;21:4802.

[12] Cover T, Hart P. Nearest Neighbor Pattern Classification. IEEE Trans Inf Theory. 1967;13:21-27.

[13] LeCun Y, Bengio Y, Hinton G. Deep Learning. Nature. 2015;521:436-44.

[14] Escalante González M. Application of Artificial Intelligence for Breast Cancer Detection. Synergia Med J. 2023.

[15] Dreiseitl S, Ohno-Machado L. Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. J Biomed Inform. 2002;35:352-359.

[16] Hossin M, Sulaiman MN. A Review on Evaluation Metrics for Data Classification Evaluations. Int J Data Min Knowl Manag Process. 2015;5:1.

[17] Lopez NC, Garcia-Ordas MT, Vitelli-Storelli F, Fernández-Navarro P, Palazuelos C, Alaiz-Rodríguez R. Evaluation of Feature Selection Techniques for Breast Cancer Risk Prediction. Int J Environ Res Public Health. 2021;18:10670.

[18] Oladimeji OO, Ayaz H, McLoughlin I, Unnikrishnan S. Mutual Information-Based Radiomic Feature Selection With Shap Explainability for Breast Cancer Diagnosis. Mathematical Modeling and Intelligent Systems for Health and Environment (MISHE). 2024;2024.

[19] Sajjadnia Z, Khayami R, Moosavi MR. Preprocessing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services. Cancer Inform. 2020;19:1176935120917955.

[20] Bernardi FA, Alves D, Crepaldi N, Yamada DB, Lima VC, Rijo R. Data Quality in Health Research: Integrative Literature Review. J Med Internet Res. 2023;25:e41446.

[21] Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and Regression Trees. Chapman and Hall/CRC. 2017.