IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 June 2024, accepted 12 July 2024, date of publication 22 July 2024, date of current version 27 September 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3431998*

RESEARCH ARTICLE

# Breast Carcinoma Prediction Through Integration of Machine Learning Models

**ROSMERI MARTÍNEZ-LICORT[1], CARLOS DE LA CRUZ LEÓN[2,3], DEEVYANKAR AGARWAL[2], BENJAMÍN SAHELICES[1], ISABEL DE LA TORRE[2], JOSÉ PABLO MIRAMONTES-GONZÁLEZ[4,5], AND MOHAMMED AMOON[6]**

[1]GCME Research Group, Department of Computer Science, University of Valladolid, 47011 Valladolid, Spain
[2]Department of Signal Theory and Communications and Telematics Engineering, University of Valladolid, 47011 Valladolid, Spain
[3]CARTIF Technology Center, Boecillo, 47151 Valladolid, Spain
[4]Department of Medicine, Faculty of Medicine, University of Valladolid, 47005 Valladolid, Spain
[5]Internal Medicine Service, Río Hortega University Hospital, 47012 Valladolid, Spain
[6]Department of Computer Science, Community College, King Saud University, Riyadh 11437, Saudi Arabia

Corresponding author: Rosmeri Martínez-Licort (rosmeri.martinez@uva.es)

This work was supported by Researchers Supporting Project, King Saud University, Riyadh, Saudi Arabia, under Grant RSPD2024R968.

**ABSTRACT** Breast cancer poses a global health challenge, with high incidence and mortality rates. Early detection and precise diagnosis are crucial for patient prognosis. Machine learning (ML) models applied to mammary biopsy image data hold promise for achieving an efficient and accurate breast cancer diagnosis. In this study, we evaluated the performance of several ML algorithms, including Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB) and Support Vector Machine (SVM). We establish evaluation contexts by implementing data standardization and reducing the correlation between variables. Firstly, we select the best-performing parameters for each algorithm by building and evaluating the individual models. Then, we implement a combined model using weighted voting, where the weights of each model are determined based on its performance on the test dataset. The final model is constructed by combining the LR, RF and SVM models. We find that SVM is the best-performance individual model, so it has the highest weight in the final model. The final integrated model achieves an accuracy of 98%, a precision of 97%, a recall of 99%, an F1-score of 98% and an AUC of 0.98. Our weighted voting model compares favourably with the other models analysed. This approach demonstrates its efficiency and transparency in handling structured medical data. It is a prototype that will be refined and expanded to encompass larger real-world datasets.

**INDEX TERMS** Breast cancer, ensemble learning, machine learning, majority voting, principal component analysis.

## I. INTRODUCTION

Breast cancer begins in breast tissue and can spread to other regions of the body during its progression. Although breast cancer can affect both sexes, it is more common in women. It is the most prevalent form of cancer and the most commonly diagnosed type of cancer among women worldwide. According to the American Cancer Society, breast cancer is the second leading cause of cancer-related deaths after lung cancer [1], [2]. Early identification is crucial to improve survival rates and treatment effectiveness. Mammography is the predominant diagnostic method and uses low doses of X-rays to detect irregularities in breast tissue. Other techniques such as ultrasound, Magnetic Resonance Imaging (MRI) or biopsies can be used [3].

Health professionals face challenges related to training, such as the time required for the diagnosis process and human error in achieving an accurate diagnosis [4]. In response to these challenges, Artificial Intelligence (AI) serves as a tool to simulate, expand and optimize human cognitive

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

capabilities. AI is based on technologies such as ML and Deep Learning (DL) [5], with the latter being a technique within ML. The use of ML makes it possible to detect patterns that are difficult for humans to identify, reducing diagnostic errors and improving the prognosis of diseases such as cancer, coronary artery disease and Alzheimer's disease [6]. In the field of ML, it has been observed that results improve considerably when multiple variants of learning models are combined instead of selecting a single prominent variant [7]. This improvement is attributed to the greater effectiveness offered by the Ensemble Learning (EL) approach in terms of predictive performance. In [8], EL is defined as the construction of a predictive model by integrating the strengths of a series of simpler base models.

During the design and development of this study, publications published from 2013 to 2024 were reviewed. During this period, there has been an increase in research applying ML methods to predict breast cancer. Some studies focus on processing demographic risk factors, while others analyse mammographic patterns or biopsy data, particularly Fine Needle Aspiration (FNA) testing. Detailed research has been identified, such as that carried out by [9], [10], [11], [12], [13], [14], [15], and [16]. Specifically, research on the use of EL has been highlighted, such as [17], [18], [19], [20], [21], [22], [23], and [24].

Among the most recent works, is that of [25], where eight classification methods, were analysed, including DL models. After the evaluation, SVM stood out for achieving the highest precision (97.7%) with minimal classification errors, overcoming the DL method used. In [26], six methods were evaluated, including Extreme Gradient Boost (XGBoost), Multi-Layer Perceptron (MLP) and Ada Boost. XGBoost showed the best initial performance, achieving an accuracy of 96.48%, which was further improved through data transformations and the inclusion of optimizers. Similarly, [27] employed six ML models. DL with Artificial Neural Network (ANN) achieved the highest accuracy of 98.24%. In [28] they use DL methods such as stacked sparse autoencoders (SAE-Sparse autoencoder, FS-Feature Space, SM-Softmax classifier) and the Softmmax regression model (FE-SSAE-SM model). They achieve an accuracy of 98.6%, applying many techniques and complex variations of these. According to [6], the LR, RF and SVM models have demonstrated outstanding results in medical applications.

Prior research provided the foundation for developing our methodology, while concurrent studies enabled us to compare our results with the latest advances in the field. Table 1 shows the most notable studies addressed in the literature, which use tabular data derived from FNA. The studies [29], [30], [33] in Table 1 are highlighted in italics because they correspond to research carried out at the same time as ours. These investigations opted for different methods and methodologies from those employed and developed in our current work.

The review concludes that the processing of tabular data derived from FNA is very efficient in computational terms, has a solid evidence base and is compatible with ML algorithms. Furthermore, this data type is practical and applicable in real environments [35]. The preference of researchers to train ML models using tabular datasets instead of images is noted in research such as [9], [10], [11], [13], [21], [23], and [24]. Additionally, [36] points out that the use of DL on some types of images can affect cancer detection due to the variability in lesion locations and their different intensity distributions.

In this context, the present research aims to develop and validate a prototype ML classifier based on EL to predict the presence of breast cancer. This classifier analyses the characteristics of cell nuclei present in FNA images to distinguish between malignant and benign tumours. By focusing on improving accuracy in prognosis, we aim to contribute to the timely identification of breast cancer, using this prototype as a foundational step towards future applications in larger datasets.

## II. MATERIALS AND METHODS

In this section, the database used in the research is introduced. Subsequently, the characteristics of each process stage are detailed, from the initial data preprocessing to the final model evaluation. Fig. 1 presents a flowchart that visualizes this process clearly and systematically.

### A. DATASET

The study was developed using the database Breast Cancer Wisconsin (Diagnostic, known as WDBC), which is publicly accessible from the University of California Irvine (UCI Machine Learning Repository). This database was chosen for its extensive use in previous studies, facilitating comparative analysis. It contains information on the characteristics of cell nuclei present in FNA images of breast masses. According to [37], these characteristics, such as the size, shape and texture of cell nuclei, provide quantifiable and objective measurements that are easily interpretable. They can be related to the diagnostic criteria used by pathologists in the detection of cancers. The dataset consists of 32 variables and comprises a total of 569 patient records from the University of Wisconsin Hospital. Of these records, 357 are classified as benign tumours and 212 as malignant.

We recognize that the database used in this study, sourced from a public repository, contains a limited number of records. However, for our prototype development, these initial data are adequate to evaluate the feasibility and preliminary performance of our models. In future stages, we plan to expand and diversify our dataset to improve the generalization and robustness of the model in real clinical applications.

### B. DATA PREPROCESSING

Exploratory analysis revealed no null, missing, or outlier values. The variable "Diagnosis" was selected as a predictive variable, coding malignant cases as "1" and benign cases as "0". The variable "ID" was excluded due to lack of relevance.

**TABLE 1. Summary of the research studies addressed.**

| Study | ML Techniques | Methodologies | Achieved accuracy |
|---|---|---|---|
| 2016 [13] | SVM, NB, | Pearson correlation coefficient, PCA, dimension reduction, data discretization, EL | 98.88% with SVM without discretization and 19 features<br>97.39% with SVM with discretization and 5 features |
| 2017 [9] | SVM | Two-Step clustering | 95.23% with SVM without clustering<br>99.1% with SVM with clustering |
| 2018 [11] | SVM, AB, Bagging Classification Trees (BCT) | EL | 97.68% with SVM (Weighted Area Under the Receiver Operating Characteristic Curve Ensemble, WAUCE) |
| 2019 [21] | SVM, KNN, DT, GB, RF, LR, AB, GNB, LDA | Standardization, feature selection | 98.24% with AB and log-loss 0.39<br>95.57% with GB and log-loss 0.06<br>96.45% with RF and log-loss 0.09<br>97.34% with ET and log-loss 0.09 |
| 2019 [10] | SVM, ANN | Features selection | 96.99% with SVM (Sequential Minimal Optimization, SMO)<br>95.44% with ANN (Multi-Layer Perceptron, MLP) |
| 2021 [23] | SVM, RF, LR, DT, KNN | Features extraction | 97.2% with SVM |
| 2022 [24] | DT, AB, RF, Hierarchical Clustering Random Forest (HCRF, a model developed in the study) | Clustering, Variable Importance Measure (VIM) feature selection | 97.05% with HCRF and VIM |
| *2022* [29] | *Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting (LightGBM)* | *Explainable Artificial Intelligence (XAI): Shapley Additive exPlanations (SHAP)* | *99% with LightGBM* |
| *2023* [30] | *LR, SVM, KNN, Classification and Regression Tree (CART), NB* | *EL with Majority-Voting* | *99.3% with LR + SVM + CART* |
| *2023* [31] | *RF, Hybrid models: Genetic Algorithm (GA) + Fisher Score, Variance + GA, GA ∪ L1 (Lasso)* | *Hybridization methodologies, Filter methods, Wrapper and Embedded methods* | *99.12% with GA + Fisher Score, 97.37% with Variance + GA, 98.25% with Cases 1 + 2 (GA ∪ L1)* |
| *2023* [32] | *XGBoost, RF, SVM, KNN, LR* | *Synthetic Minority Oversampling Technique and Edited Nearest Neighbors (SMOTE-ENN), Recursive Feature Elimination (RFE), SHapley Additive exPlanations (SHAP)* | *99.52% with RF* |
| *2024* [33] | *RF, SVM, ETC, LR, GNB, KNN, GBM, DT, SGD* | *Features from the Convolutional Neural Network (CNN) model, EL (Majory-Voting)* | *99.99% with RF + SVM* |
| *2024* [34] | *ET, LightGBM, Ridge Classifier (RC), Linear Discriminant Analysis (LDA)* | *EL using an ELRL-E model, Voting classifier* | *97.66% with ELRL-E model* |

In this study, a strategy is followed to evaluate the performance of ML algorithms at different stages of data preprocessing (see Fig. 1) The models are applied to data that have undergone various preprocessing phases, such as:

- Data Standardization: All data is standardized to ensure all features are on the same scale.
- Principal Component Analysis (PCA): It is used to reduce the dimensionality of the dataset while preserving relevant information.
- Adjustment of Training Ratio: The proportion of data used for model training and validation is varied, exploring different configurations to optimize performance.

Data were standardized to maintain a consistent range of values and a shared scale. Eq. (1) shows the standardization of $x$, where $\bar{x}$ represents the average of $x$ whereas $\sigma$ denotes the standard deviation of $x$.

$$x_{\text{scaled}} = \frac{x - \bar{x}}{\sigma} \tag{1}$$

PCA identifies patterns and correlations between variables, transforming the original ones into new uncorrelated ones

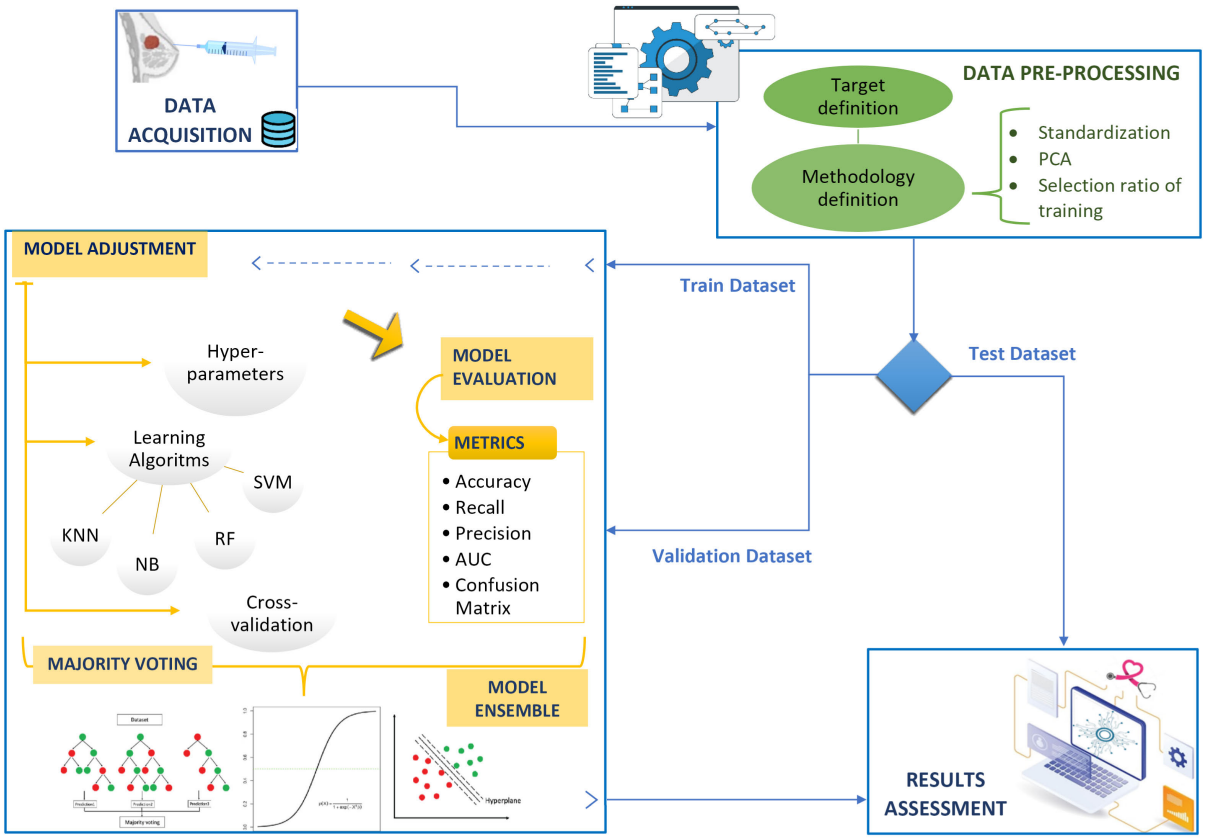


**FIGURE 1.** Methodology to identify the optimal configuration for each ML algorithm.

called Principal Component (PC) [38]. The appropriate number of PCs was determined by establishing a threshold that cumulatively explains between 80% and 90% of the variance in the data, as described in the study by [39].

The process of adjusting the proportion of training data is done using cross-validation (see Fig. 2). Which allows determining the best proportion of training data that optimizes the performance of the model without overfitting or underfitting. The dataset was divided into training and testing sets using different split ratios, ranging from 85% training - 15% testing to 50% training - 50% testing. These variations were used to assess their impact on the ML models Test set sizes lower than 15% and higher than 50% were discarded because they did not allow optimal model training (see [40]).

### C. BASE MODELS SELECTION

The selection of ML models for this study was based on results published by previous studies which also use the WDBC dataset. Each chosen method has its advantages and limitations, depending on the type of data and the classification objective. It is expected that the combination of these algorithms will provide more comprehensive and efficient processing. The following methods were selected:

- Logistic Regression (**LR**). Widely used statistical model for classification or predictive analysis. Estimate the probability of an event occurring using the independent variables in the dataset [42].
- Random Forest (**RF**) Combine multiple Decision Trees (DT) to obtain a more robust prediction with less risk of bias and variation. Its versatility and ease of use have made it a popular choice for classification and regression problems [36].
- Naive Bayes (**NB**) Probabilistic classification model based on Bayes' theorem. It is especially useful for data with continuous predictor variables [36], [43].
- Support Vector Machine (**SVM**) operates in a high-dimensional feature space to classify data points, even when they cannot be linearly separated. According to [36] and [44] this method leads to an increase in the accuracy of cancer prediction and a decrease in classification errors.

### D. CROSS-VALIDATION

We have implemented a thorough k-fold cross-validation approach to evaluate our ML models. After importing the dataset, it is initially split into input variables and the target variable. Subsequently, we establish a predefined list of values for $k$ ($k = [5, 7, 10, 13, 15]$) representing the number of folds for cross-validation.

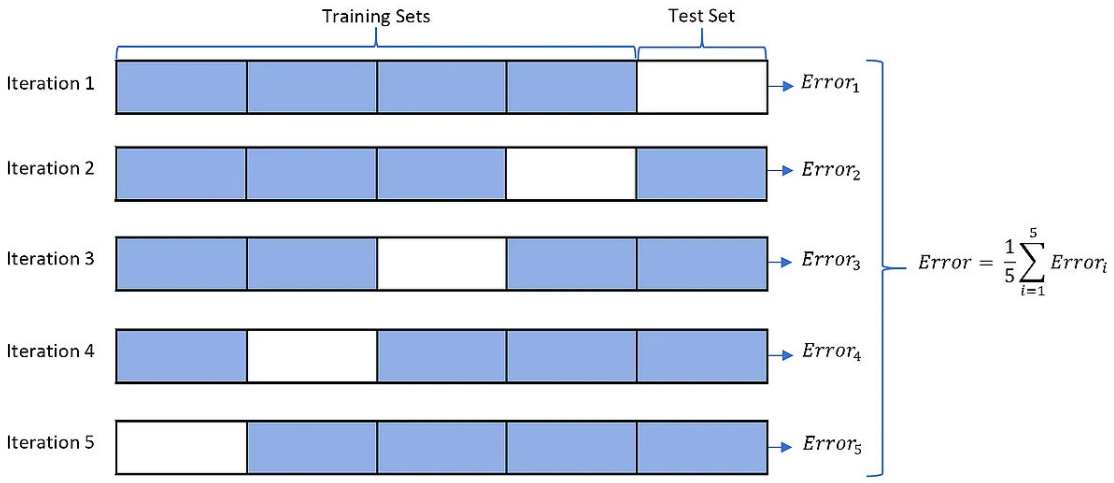


**FIGURE 2.** M5-Fold cross-validation diagram (Source: [41]).

During each iteration of the k-fold cross-validation process, the dataset is partitioned into k subsets of equal size (see Fig. 2). For each value of $k$, $k-1$ subsets are used for training the model, while the remaining subset is reserved for testing. Within our methodology, we construct a pipeline that includes standardization of the data and application of PCA with varying numbers of PCs, followed by training the respective ML model.

Throughout k iterations, each employing a distinct testing subset, we compute average metrics for accuracy, precision, recall and F1-score, alongside the standard deviation of predictions from the test set. The final evaluation of model performance is derived as the mean of these scores across all iterations.

### E. MODEL DEVELOPMENT

After completing the data preprocessing, we proceed to apply several ML models. Fig. 1 illustrates the steps followed in developing this study. We employ a methodology where we assess the performance of each ML model by adjusting values during preprocessing stages to achieve optimal outcomes. This approach allows us to define preprocessing methods that optimize model performance.

Model optimization is complemented by adjusting hyperparameter configurations. These hyperparameters are selected following an extensive review of previous studies in ML and medical classification. We validate these hyperparameters using cross-validation to maximize predictive performance in our study.

Subsequently, we fine-tune the hyperparameters and select models with superior performance, evaluating them using various metrics. Majority voting is used to determine the relative importance of each model in the final model construction (which is an ensemble model).

We detail below the parameters and hyperparameters used for configuring each ML model. These settings are crucial for optimizing the performance and ensuring the robustness of our models.

#### 1) SUMMARY OF LR MODEL CHARACTERISTICS BASED ON HYPERPARAMETERS

The LR model was set up using the following parameters and hyperparameters:

- Regularization:
  - **Penalty**: Type of regularization. In this case, *L2* helps prevent overfitting and improves model generalization.
  - **C**: The inverse of regularization strength. A value of 1.0 strikes a good balance between bias and variance.
- Convergence and Optimization:
  - **tol**: Tolerance for the stopping criterion ensures proper convergence of the algorithm (optimal value *1e-4*).
  - **solver**: The solver (in this case, the algorithm used to optimize the model) is efficient and suitable for data sets that are not extremely large (small to medium) and that use *L2* regularization. The optimal option for this case is the solver *lbfgs*.
  - **max_iter**: High number of iterations ensures algorithm convergence, especially with scaled and dimensionally reduced data (optimal value 2500).
- Intercept and Fitting:
  - **fit_intercept**: Adding an intercept term allows for better fitting to the data (value *True*).
- Class Balance and Weight:
  - **class_weight**: Specifies the weight that is assigned to each class in a classification problem. In this case, with a value of *None*, the model assumes that the classes are balanced or that the imbalance is not significant.

  - **multi_class**: Determines the strategy used to handle multiclass classification problems. In this case, the value *auto* is assigned, which means the model automatically selects the most appropriate strategy based on the type of data and the solver used. The chosen strategy is *ovr* (One-vs-Rest), where each class is treated as a positive class against all others treated as negatives. A separate classifier is trained for each class, and the prediction is made based on the class with the highest score
- Reproducibility:
  - **random_state**: Sets the seed used by the model's internal random number generator. By setting **random_state** to 2, we ensure reproducibility of the results. This means that every time the model is run with the same seed, the same results will be obtained, which is critical for consistency in model development and evaluation.

We selected the *L2* penalty to handle multicollinearity and the *lbfgs* solver due to its efficiency with small datasets. The regularization parameter **C** was set to 1.0 to provide a good balance between overfitting and underfitting. An intercept term was added to improve model accuracy.

#### 2) SUMMARY OF RF MODEL CHARACTERISTICS BASED ON HYPERPARAMETERS

The RF model was configured with the following parameters and hyperparameters:

- Number of Trees:
  - **n_estimators**: Determine the number of trees in the forest. A moderate number of trees (value 100) provides a good balance between model accuracy and calculation time.
- Splitting Criterion:
  - **criterion**: Determines the function used to measure the quality of a split at each node of the decision tree within the RF. The *Gini* criterion evaluates how pure the resulting child nodes are after a split, seeking to minimize the mixing of classes in each partition of the tree. It is a robust choice for most binary classification problems due to its effectiveness in class separation.
- Tree Depth and Splitting:
  - **max_depth**: This parameter controls the maximum depth of the trees in the RF. When set to *None*, trees grow until all leaves are pure or until they contain the minimum number of samples required to perform a split (**min_samples_split**).
  - **min_samples_split**: Sets the minimum number of samples required to split an internal node. With value 2, promote deeper trees that capture more details of the data set.
  - **max_features**: Determines the maximum number of features to consider when searching for the best split at each node. Using *sqrt* means considering the square root of the total number of features, which reduces variance and helps avoid model overfitting.
- Bootstrap and Sampling:
  - **bootstrap**: Indicates whether bootstrap samples should be used when building trees. Setting it to True helps improve the generalization of the model by introducing variability in the training sets.
- Performance and Reproducibility:
  - **random_state**: Set a random seed to initialize the random number generator in the RF. Setting this parameter to a specific number (in this case, 2) ensures that the model will produce the same results every time it is fitted to the same input data.

In configuring the RF model, we selected 100 estimators to balance performance and computational efficiency. The *Gini* criterion was chosen for its effectiveness in binary classification problems. We allowed the trees to grow to their maximum depth to capture the complexity of the dataset and used bootstrap samples to improve generalization.

#### 3) SUMMARY OF GAUSSIAN NB MODEL CHARACTERISTICS BASED ON HYPERPARAMETERS

The Gaussian NB model was configured with the following parameters:

- Priors:
  - **priors**: Specifies the a priori probabilities of the classes. When set to *None*, the model automatically adjusts these probabilities based on the actual distribution of classes in the data set. This allows the model to dynamically adapt to the class distribution observed during training.
- Smoothing Parameter:
  - **var_smoothing**: Improve numerical stability in the calculation of conditional probabilities in the NB model. Adds a small fraction of the largest variance among all features to the variances of each feature. The default value of *1e-9* ensures that the model is able to effectively handle features with very small variances, thus avoiding numerical instability problems during the training and prediction process.

We do not specify the prior probabilities of the classes, allowing the model to adjust these probabilities based on the data. We use a variance smoothing (**var_smoothing**) of *1e-9* to ensure numerical stability during the calculations.

#### 4) SUMMARY OF SVC MODEL CHARACTERISTICS BASED ON HYPERPARAMETERS

Support Vector Classifier (SVC) is a variant of SVM focused on classifying data into different categories. According to the study by [45] on ML models, this method achieved the best classification for breast cancer. The following hyperparameters were used for developing the SVC model:

- Regularization Parameter:

  - **C**: It is the regularization parameter that controls the penalty for classification errors in the SVM model. A higher value may lead to overfitting, while a lower value may result in underfitting. The value of 1.0 provides an appropriate balance as a starting point between fit and generalization.
- Kernel:
  - **kernel**: Defines what type of kernel function is used to transform the data. In this case, the kernel *RBF* (radial basis function) is suitable for handling nonlinear data and is commonly used due to its flexibility in capturing complex relationships between variables.
- Kernel Coefficient:
  - **gamma**: Adjusts the kernel coefficient. When set to *scale*, gamma is automatically adjusted to optimize model performance based on the scale of the input data. This allows SVM to adapt the complexity of the model automatically, ensuring better generalization ability without requiring specific manual adjustments.
- Shrinking Heuristic:
  - **shrinking**: Controls whether the reduction heuristic is used to speed up model training. When set to *True*, this heuristic is enabled, which can significantly improve the training time of the SVM model.
- Stopping Criterion:
  - **tol**: In SVM, this parameter refers to the tolerance for the stopping criterion during the optimization process. In simple terms, it determines when the SVM algorithm considers that it has converged sufficiently and can stop. A value of 0.001 means that the algorithm will continue to iterate until the difference between successive iterations of the objective function is less than or equal to 0.001.
- Cache Size:
  - **cache_size**: Specifies the size of the kernel cache in megabytes (MB). During training of the SVM model, a cache is used to store intermediate results of the kernel calculation. A value such as 200 indicates that 200 MB of memory is reserved for the kernel cache, thereby optimizing the performance and training efficiency of the SVM model.
- Multiclass Strategy:
  - **decision_function_shape**: Indicates the strategy used for multiclass classification. Using the *ovr* strategy, SVM trains a binary classifier for each class. In the context of binary classification, SVM creates one classifier to distinguish Class 1 from the rest of the classes (Class 2 in this case), and another classifier to distinguish Class 2 from the rest of the classes (Class 1 in this case). During the evaluation of an unknown instance, SVM uses both classifiers and determines the final prediction based on the class with the highest decision score

We selected the *RBF* kernel because of its ability to handle nonlinearity well. Parameter **C** was set to 1.0 to balance overfitting and underfitting. The **gamma** parameter was set to *scale* to automatically adjust this value based on the characteristics of the dataset. We also use the reduction heuristic to speed up the training process.

### F. FINAL MODEL

Once a final model is obtained for each ML algorithm, the area under the curve (AUC) is also evaluated. Finally, an ensemble model is built that combines the best of each model to improve the classification performance [41], [46]. For the final model, a weighted voting algorithm is used where each model contributes its rankings and a greater weight is assigned to those with better performance. In addition to the metrics used with the individual models, the performance of the final model is evaluated using the ROC curve for external validation and the confusion matrix metrics.

### G. EXPERIMENTAL ENVIRONMENT

The hardware configuration for this experiment included computers with an Intel Core i5 processor. A notable aspect of our study is that it did not require any specialized hardware, such as GPU. This is because the ML methods used and the dataset size were manageable with standard CPU resources. The absence of GPU requirements did not impact the efficiency or effectiveness of the models implemented. Table 2 shows the set of software tools that ensure an efficient and structured workflow. These libraries enable data manipulation and visualization, as well as the implementation and evaluation of ML algorithms.

## III. RESULTS AND DISCUSSION

This section explains the development and application of the ML methods described above, as well as the formation of the final model. To achieve this, as indicated in subsection II-B, a quantitative analysis was carried out to evaluate multicollinearity in the dataset. This analysis revealed significant correlations between most of the variables. To address this problem, PCA was implemented, resulting in the identification of six optimal PCs that explain 88.76% of the total variance. Table 3 presents each PC along with its contribution to the overall variance, individually and cumulatively.

### A. MODELS EVALUATION

In this section, we search for and obtain the best configuration for each ML model studied. The performance of these models when applying the transformations described in Section II is shown in Table 4, Table 5, Table 6, Table 7. The result of evaluating the best configuration of these individual models applying cross-validation is shown in Table 8.

**TABLE 2. Software and libraries used in the experimental environment.**

| Component | Description |
|---|---|
| Programming Language | Python 3.x |
| Python Libraries | |
| Pandas | Used for data manipulation and analysis, including importing and managing datasets. |
| Scikit-learn | Provides functionalities for model selection, classification, preprocessing, decomposition and metrics evaluation. |
| Matplotlib | Used for data visualization. |
| Seaborn | Used for statistical data visualization. |
| NumPy | Used for numerical operations and array manipulation. |

**TABLE 3. Principal components and percentage of explained variance.**

| Principal Component | Explained Variance Ratio (%) | Cumulative Explained Variance Ratio (%) |
|---|---|---|
| PC 1 | 44.27 | 44.27 |
| PC 2 | 18.97 | 63.24 |
| PC 3 | 9.39 | 72.64 |
| PC 4 | 6.60 | 79.24 |
| PC 5 | 5.50 | 84.73 |
| PC 6 | 4.02 | 88.76 |
| PC 7 | 2.25 | 91.01 |
| PC 8 | 1.59 | 92.60 |
| PC 9 | 1.39 | 93.99 |
| PC 10 | 1.17 | 95.16 |
| PC 11 | 0.98 | 96.14 |
| PC 12 | 0.87 | 97.01 |
| PC 13 | 0.80 | 97.81 |
| PC 14 | 0.52 | 98.34 |
| PC 15 | 0.31 | 98.65 |
| PC 16 | 0.27 | 98.92 |
| PC 17 | 0.20 | 99.11 |
| PC 18 | 0.18 | 99.29 |
| PC 19 | 0.16 | 99.45 |
| PC 20 | 0.10 | 99.56 |
| PC 21 | 0.10 | 99.66 |
| PC 22 | 0.09 | 99.75 |
| PC 23 | 0.08 | 99.83 |
| PC 24 | 0.06 | 99.89 |
| PC 25 | 0.05 | 99.94 |
| PC 26 | 0.03 | 99.97 |
| PC 27 | 0.02 | 99.99 |
| PC 28 | 0.01 | 100.00 |
| PC 29 | 0.00 | 100.00 |
| PC 30 | 0.00 | 100.00 |

### 1) LOGISTIC REGRESSION MODEL

Table 4 summarizes the evaluation of the developed LR models. Initially, we evaluate a model with different test set sizes. The same model was then evaluated with standardized data, also varying the size of the test set. Finally, we adjusted both the size of the test set and the number of PCs in the model with standardized data and evaluated it again.

The initial evaluation of the LR model (start of Table 4) shows accuracies between 91% and 94%, with variation in precision between 86% and 95%, a Recall and F1-score between 88% and 93%. We notice that test sets of 15% and 20% generate the lowest results, whereas values close to 30% show better results. In the second evaluation, data standardization significantly improves evaluation metrics, for a 20% test set. Achieving an accuracy and F1-score of 97%, precision of 96% and recall of 98%. Considering results, test sets of 20%, 25% and 30% will be used in subsequent analyses of the model.

The third evaluation (see Table 4) shows that by combining data standardization with PCA, the best performance is achieved with a test set of 25% and 10 PCs explaining 95.16% of the variance of the original data. This model obtains accuracy and precision of 98% and a recall and F1-score of 96% and 97%, respectively. The evaluation with k-fold cross-validation (see Table 8) shows similar results to the standard evaluation.

In conclusion, the best RL model studied is the one that uses data standardization together with dimensionality reduction using PCA with 10 PCs and a test set size of 25%.

### 2) RANDOM FOREST MODEL

Table 5 summarizes the evaluation of the RF models. For the first evaluation, we varied the size of the test set. Furthermore, we defined 100 estimators (trees that comprise it), as the standard value in proportion to the size of the dataset. In the second evaluation, the number of estimators is varied. In the third evaluation, the application of PCA to standardized data was explored.

We observed that the initial model scored highly on most performance metrics on a 30% test set. Specifically, it obtained 96% accuracy and recall, 94% precision and 95% F1-score. Due to the result of the first evaluation, 30% of the test set was also used in the second evaluation. There were no significant improvements above 60 estimators.

Next, we evaluated the RF with standardized data and found no significant differences in performance compared to the previous evaluation. However, we also explored the use of PCA on standardized data. Table 5 shows that with 14 PCs, the precision improved by 4%, whereas the recall decreased by 3%. Subsequently, we evaluated both the model with PCA and the model without PCA using k-fold cross-validation. The best performance with PCA was obtained for $k = 15$ and the best performance without PCA was obtained for $k = 7$. These results are shown in Table 8.

**TABLE 4.** Evaluation of the performance of the LR model in different stages of analysis.

| Test Size | Accuracy | Precision | Recall | F1-score | Nº of PC |
|---|---|---|---|---|---|
| Evaluation of the performance of the LR model, varying the size of the test set. | | | | | |
| 0.15 | 0.91 | 0.86 | 0.91 | 0.88 | |
| 0.20 | 0.93 | 0.91 | 0.91 | 0.91 | |
| 0.25 | 0.94 | 0.93 | 0.91 | 0.92 | |
| 0.30 | 0.94 | 0.93 | 0.93 | 0.93 | |
| 0.35 | 0.94 | 0.93 | 0.89 | 0.91 | |
| 0.40 | 0.94 | 0.95 | 0.89 | 0.92 | |
| 0.45 | 0.94 | 0.93 | 0.91 | 0.92 | |
| 0.50 | 0.94 | 0.93 | 0.93 | 0.93 | |
| Evaluation of the performance of the LR model with standardized data, varying the size of the test set. | | | | | |
| 0.15 | 0.97 | 0.94 | 0.97 | 0.96 | |
| 0.20 | 0.97 | 0.96 | 0.98 | 0.97 | |
| 0.25 | 0.97 | 0.95 | 0.98 | 0.96 | |
| 0.30 | 0.97 | 0.96 | 0.97 | 0.96 | |
| 0.35 | 0.97 | 0.97 | 0.95 | 0.96 | |
| 0.40 | 0.96 | 0.98 | 0.93 | 0.95 | |
| 0.45 | 0.97 | 0.97 | 0.96 | 0.96 | |
| 0.50 | 0.97 | 0.97 | 0.95 | 0.96 | |
| Evaluation of the performance of the LR model with standardized data and PCA analysis, varying the size of the test set and the number of PC. | | | | | |
| 0.20 | 0.96 | 0.95 | 0.93 | 0.94 | 6.00 |
| | 0.97 | 0.98 | 0.96 | 0.97 | 10.00 |
| | 0.96 | 0.93 | 0.96 | 0.95 | 14.00 |
| | 0.97 | 0.96 | 0.98 | 0.97 | 18.00 |
| | 0.97 | 0.96 | 0.98 | 0.97 | 22.00 |
| 0.25 | 0.97 | 0.96 | 0.95 | 0.95 | 6.00 |
| | 0.98 | 0.98 | 0.96 | 0.97 | 10.00 |
| | 0.97 | 0.95 | 0.98 | 0.96 | 14.00 |
| | 0.97 | 0.95 | 0.98 | 0.96 | 18.00 |
| | 0.97 | 0.95 | 0.98 | 0.96 | 22.00 |
| 0.30 | 0.97 | 0.97 | 0.96 | 0.96 | 6.00 |
| | 0.97 | 0.96 | 0.97 | 0.96 | 10.00 |
| | 0.97 | 0.96 | 0.97 | 0.96 | 14.00 |
| | 0.97 | 0.96 | 0.97 | 0.96 | 18.00 |
| | 0.97 | 0.96 | 0.97 | 0.96 | 22.00 |

The best performance with RF model was achieved with the standardized dataset, 30% of the data in the test set and 60 estimators.

### 3) NAIVE BAYES MODEL

Table 6 shows the results of the evaluation of the NB models. In the first evaluation, the size of the test set was varied and the best performance was achieved with a size of 30%. The highest scores observed across all performance metrics were an accuracy of 94% and 93% in the remaining metrics. This result indicates satisfactory performance in classification.

For the second evaluation, we applied the model to standardized data (see Table 6), without observing significant improvements compared to the previous evaluation. For the third evaluation, the standardized data is kept waiting for possible improvements in the performance of the model. In addition to standardization, PCA is applied with a 30% test set. With 6 PCs, improvements in precision and F1-score are observed, reaching a precision and accuracy of 95% and an F1-score of 94%. Next, we evaluate the best NB model so far (30% on the test set, standardized data and 6 PCs), using k-fold cross-validation. In Table 8, we observe that with cross-validation, the performance scores mentioned above decrease. The best NB model performs below the RF and LR models proposed above.

### 4) SUPPORT VECTOR MACHINE MODEL

Table 7 shows the result of the evaluation of the SVM models. In the first evaluation, we varied the size of the test set and observed that the best results were obtained with a size of 25%. The model achieves the following scores: 92% accuracy and precision, 86% recall and 89% F1-score.

In the second evaluation, SVM was applied to standardized data, observing significant improvements in the results. As Table 7 shows, for a 30% test set, an accuracy of 98%, a precision of 96%, a recall of 99% and an F1-score of 97% were achieved. In the third evaluation, SVM is applied to standardized data and with dimensionality reduction using PCA. Results similar to those of the previous model are presented in Table 7. Finally, we employed k-fold

**TABLE 5. Evaluation of the performance of the RF model in different stages of analysis.**

| Evaluation of the performance of the RF model, varying the size of the test set. | | | | |
|---|---|---|---|---|
| Test Size | Accuracy | Precision | Recall | F1-Score |
| 0.15 | 0.93 | 0.89 | 0.94 | 0.91 |
| 0.20 | 0.95 | 0.91 | 0.96 | 0.93 |
| 0.25 | 0.94 | 0.90 | 0.95 | 0.92 |
| 0.30 | 0.96 | 0.94 | 0.96 | 0.95 |
| 0.35 | 0.96 | 0.96 | 0.92 | 0.94 |
| 0.40 | 0.94 | 0.94 | 0.91 | 0.92 |
| 0.45 | 0.95 | 0.94 | 0.92 | 0.93 |
| 0.50 | 0.95 | 0.94 | 0.92 | 0.93 |
| Evaluation of the performance of the RF model, varying the estimators number. The second column shows the range of estimators tested. | | | | |
| Nº of Estimators | Accuracy | Precision | Recall | F1-Score |
| 50 | 0.95 | 0.93 | 0.96 | 0.94 |
| 60–150 | 0.96 | 0.94 | 0.96 | 0.95 |
| Evaluation of the performance of the RF model with PCA analysis, varying the number of PC. | | | | |
| Nº of PC | Accuracy | Precision | Recall | F1-Score |
| 6 | 0.95 | 0.94 | 0.94 | 0.94 |
| 10 | 0.95 | 0.95 | 0.93 | 0.94 |
| 14 | 0.96 | 0.98 | 0.93 | 0.95 |
| 18 | 0.96 | 0.95 | 0.94 | 0.95 |
| 22 | 0.96 | 0.95 | 0.94 | 0.95 |

**TABLE 6. Evaluation of the performance of the NB model in different stages of analysis.**

| Evaluation of the performance of the NB model, varying the size of the test set. | | | | |
|---|---|---|---|---|
| Test Size | Accuracy | Precision | Recall | F1-Score |
| 0.15 | 0.92 | 0.88 | 0.91 | 0.90 |
| 0.20 | 0.94 | 0.91 | 0.93 | 0.92 |
| 0.25 | 0.93 | 0.91 | 0.91 | 0.91 |
| 0.30 | 0.94 | 0.93 | 0.93 | 0.93 |
| 0.35 | 0.92 | 0.89 | 0.89 | 0.89 |
| 0.40 | 0.92 | 0.90 | 0.90 | 0.90 |
| 0.45 | 0.93 | 0.91 | 0.91 | 0.91 |
| 0.50 | 0.93 | 0.92 | 0.90 | 0.91 |
| Evaluation of the performance of the NB model with standardized data, varying the size of the test set. | | | | |
| Test Size | Accuracy | Precision | Recall | F1-Score |
| 0.15 | 0.92 | 0.88 | 0.91 | 0.90 |
| 0.20 | 0.94 | 0.91 | 0.93 | 0.92 |
| 0.25 | 0.93 | 0.91 | 0.91 | 0.91 |
| 0.30 | 0.94 | 0.93 | 0.93 | 0.93 |
| 0.35 | 0.92 | 0.88 | 0.89 | 0.89 |
| 0.40 | 0.92 | 0.89 | 0.90 | 0.89 |
| 0.45 | 0.93 | 0.90 | 0.91 | 0.90 |
| 0.50 | 0.93 | 0.91 | 0.90 | 0.90 |
| Evaluation of the performance of the NB model with PCA analysis, varying the number of PC. | | | | |
| Nº of PC | Accuracy | Precision | Recall | F1-Score |
| 6 | 0.95 | 0.95 | 0.93 | 0.94 |
| 10 | 0.93 | 0.94 | 0.88 | 0.91 |
| 14 | 0.93 | 0.92 | 0.90 | 0.91 |
| 18 | 0.88 | 0.84 | 0.85 | 0.84 |
| 22 | 0.85 | 0.78 | 0.85 | 0.81 |

cross-validation to evaluate the model identified as the best performer. The results did not show significant improvements in any of the metrics analysed.

**TABLE 7. Evaluation of the performance of the SVM model in different stages of analysis.**

| Evaluation of the performance of the SVM model, varying the size of the test set. | | | | |
|---|---|---|---|---|
| Test Size | Accuracy | Precision | Recall | F1-Score |
| 0.15 | 0.88 | 0.87 | 0.82 | 0.84 |
| 0.20 | 0.90 | 0.90 | 0.84 | 0.87 |
| 0.25 | 0.92 | 0.92 | 0.86 | 0.89 |
| 0.30 | 0.91 | 0.93 | 0.84 | 0.88 |
| 0.35 | 0.90 | 0.94 | 0.80 | 0.87 |
| 0.40 | 0.89 | 0.93 | 0.77 | 0.84 |
| 0.45 | 0.89 | 0.94 | 0.78 | 0.85 |
| 0.50 | 0.90 | 0.95 | 0.79 | 0.86 |
| Evaluation of the performance of the SVM model with standardized data, varying the size of the test set. | | | | |
| Test Size | Accuracy | Precision | Recall | F1-Score |
| 0.15 | 0.95 | 0.91 | 0.97 | 0.94 |
| 0.20 | 0.96 | 0.94 | 0.98 | 0.96 |
| 0.25 | 0.98 | 0.96 | 0.98 | 0.97 |
| 0.30 | 0.98 | 0.96 | 0.99 | 0.97 |
| 0.35 | 0.97 | 0.96 | 0.96 | 0.96 |
| 0.40 | 0.97 | 0.97 | 0.97 | 0.97 |
| 0.45 | 0.97 | 0.96 | 0.96 | 0.96 |
| 0.50 | 0.97 | 0.97 | 0.95 | 0.96 |
| Evaluation of the performance of the SVM model with PCA analysis, varying the number of PC. | | | | |
| Nº of PC | Accuracy | Precision | Recall | F1-Score |
| 6 | 0.96 | 0.94 | 0.97 | 0.96 |
| 10 | 0.97 | 0.94 | 0.99 | 0.96 |
| 14 | 0.98 | 0.96 | 0.99 | 0.97 |
| 18 | 0.98 | 0.96 | 0.99 | 0.97 |
| 22 | 0.98 | 0.96 | 0.99 | 0.97 |

We conclude that the SVM model with standardized data and a test set of 30% and more than 6 PCs presents the best performance among all those evaluated in the study of SVM.

### 5) PROPOSED MODEL

A final ensemble model is created that combines the optimal parameters discovered by implementing LR, RF, NB and SVM. The best-performing model is the SVM, followed by LR and RF, respectively. Due to its inferior performance compared to the other models, NB was excluded from the study. Below is the list of models, ordered from highest to lowest performance:

- SVM with standardized dataset and a test set that represents 30% of the total data.
- LR with standardized dataset, a test set covering 30% of the data and PCA analysis with 10 PCs.
- RF with standardized dataset, a test set covering 30% of the data and 60 estimators.

As it is a binary classification system with discrete results, it was decided to combine the three models, assigning weights to each one according to their individual performance. The best performance of the SVM model over the other two is considered and the following weights are assigned: 0.5 to SVM, 0.25 to LR and 0.25 to RF. The block diagram of the combined model is shown in Fig. 3. This figure illustrates that this model achieves an accuracy of 98%, a precision of 97%, a recall of 99% and an F1-score of 98%. Fig. 4

**TABLE 8.** Cross-validation results: mean and standard deviation of the evaluation of ML models.

| K | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| | Specific results for the LR model. | | | |
| 5 | $0.981 \pm 0.013$ | $0.986 \pm 0.037$ | $0.962 \pm 0.048$ | $0.974 \pm 0.018$ |
| 7 | $0.979 \pm 0.029$ | $0.981 \pm 0.047$ | $0.962 \pm 0.065$ | $0.971 \pm 0.039$ |
| 10 | $0.981 \pm 0.037$ | $0.982 \pm 0.060$ | $0.967 \pm 0.074$ | $0.974 \pm 0.050$ |
| 13 | $0.981 \pm 0.035$ | $0.986 \pm 0.050$ | $0.962 \pm 0.102$ | $0.973 \pm 0.050$ |
| 15 | $0.981 \pm 0.036$ | $0.982 \pm 0.060$ | $0.967 \pm 0.088$ | $0.973 \pm 0.050$ |
| | Specific results for the RF model. | | | |
| 5 | $0.937 \pm 0.021$ | $0.921 \pm 0.078$ | $0.910 \pm 0.046$ | $0.915 \pm 0.024$ |
| 7 | $0.951 \pm 0.029$ | $0.950 \pm 0.084$ | $0.920 \pm 0.077$ | $0.933 \pm 0.039$ |
| 10 | $0.953 \pm 0.044$ | $0.947 \pm 0.102$ | $0.929 \pm 0.086$ | $0.936 \pm 0.058$ |
| 13 | $0.949 \pm 0.074$ | $0.945 \pm 0.119$ | $0.920 \pm 0.130$ | $0.931 \pm 0.101$ |
| 15 | $0.953 \pm 0.048$ | $0.946 \pm 0.090$ | $0.929 \pm 0.090$ | $0.936 \pm 0.064$ |
| | Specific results for the NB model. | | | |
| 5 | $0.923 \pm 0.021$ | $0.914 \pm 0.066$ | $0.878 \pm 0.074$ | $0.894 \pm 0.027$ |
| 7 | $0.924 \pm 0.048$ | $0.922 \pm 0.075$ | $0.873 \pm 0.088$ | $0.896 \pm 0.066$ |
| 10 | $0.928 \pm 0.060$ | $0.924 \pm 0.098$ | $0.882 \pm 0.143$ | $0.900 \pm 0.088$ |
| 13 | $0.928 \pm 0.089$ | $0.923 \pm 0.116$ | $0.883 \pm 0.173$ | $0.900 \pm 0.126$ |
| 15 | $0.928 \pm 0.065$ | $0.930 \pm 0.107$ | $0.877 \pm 0.169$ | $0.899 \pm 0.097$ |
| | Specific results for the SVM model. | | | |
| 5 | $0.97 \pm 0.029$ | $0.97 \pm 0.034$ | $0.96 \pm 0.069$ | $0.96 \pm 0.040$ |
| 7 | $0.98 \pm 0.036$ | $0.97 \pm 0.053$ | $0.97 \pm 0.071$ | $0.97 \pm 0.049$ |
| 10 | $0.98 \pm 0.050$ | $0.97 \pm 0.062$ | $0.96 \pm 0.093$ | $0.97 \pm 0.070$ |
| 13 | $0.97 \pm 0.053$ | $0.97 \pm 0.074$ | $0.96 \pm 0.109$ | $0.96 \pm 0.073$ |
| 15 | $0.97 \pm 0.043$ | $0.97 \pm 0.068$ | $0.96 \pm 0.088$ | $0.96 \pm 0.059$ |

**TABLE 9.** Performance of the best individual models and the proposed model.

| ML Models | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Proposed Model | 0.98 | 0.97 | 0.99 | 0.98 |
| SVM | 0.98 | 0.96 | 0.99 | 0.97 |
| LR | 0.98 | 0.98 | 0.96 | 0.97 |
| RF | 0.96 | 0.98 | 0.93 | 0.95 |
| NB | 0.95 | 0.95 | 0.93 | 0.94 |

illustrates the ROC curve corresponding to these results, with an AUC value equal to 0.98, indicating a high level of accuracy in distinguishing between benign and malignant cases.The ensemble model has the best performance among the models created in this study in terms of accuracy, recall and F1-score metrics (see Table 9).

Fig. 5 illustrates the confusion matrix with relates the actual classification of the cases in the dataset and the classification performed by the model. The model correctly identifies 102 benign cases as benign (out of 104 benign cases in the test set), whereas it misclassifies one malignant case as benign. On the other hand, the model correctly classifies 66 malignant cases as malignant (out of 67 malignant cases in the test set), but misclassifies 2 benign cases as malignant. This suggests that the model is suitable for application as a diagnostic aid.

We acknowledge that our study has been conducted with a limited dataset sourced from a public repository. Despite this limitation, the development of our prototype has yielded promising results, demonstrating the potential effectiveness of our models. These initial findings provide a strong foundation for further evaluation using more diverse and comprehensive real-world data. The positive outcomes achieved thus far encourage us to continue refining our approach, ultimately aiming to enhance the model's performance and reliability in actual clinical settings.

### B. LIMITATIONS

Our study presents several limitations that should be considered. First, the dataset size is relatively small. This can lead to overfitting issues, affecting the models' ability to generalize well to new data. Additionally, our data come from a public repository. These data may not represent the diversity of global clinical cases. This limits the generalization of the findings to different clinical contexts or populations. The lack of diversity in the data could bias the results, making the models less effective in underrepresented populations.

Furthermore, although the ML methods used have proven effective in various contexts, they have their limitations. RF can be prone to overfitting if not managed properly, especially with small datasets. SVM performance can strongly depend on the choice of hyperparameters and kernel functions. Incorrect selection can significantly reduce model accuracy. LR can have issues if the relationships between variables are not linear. It can also be sensitive to outliers and multicollinearity among predictor variables. Recognizing these limitations, we aim to provide a transparent and comprehensive evaluation of our study findings and their implications.

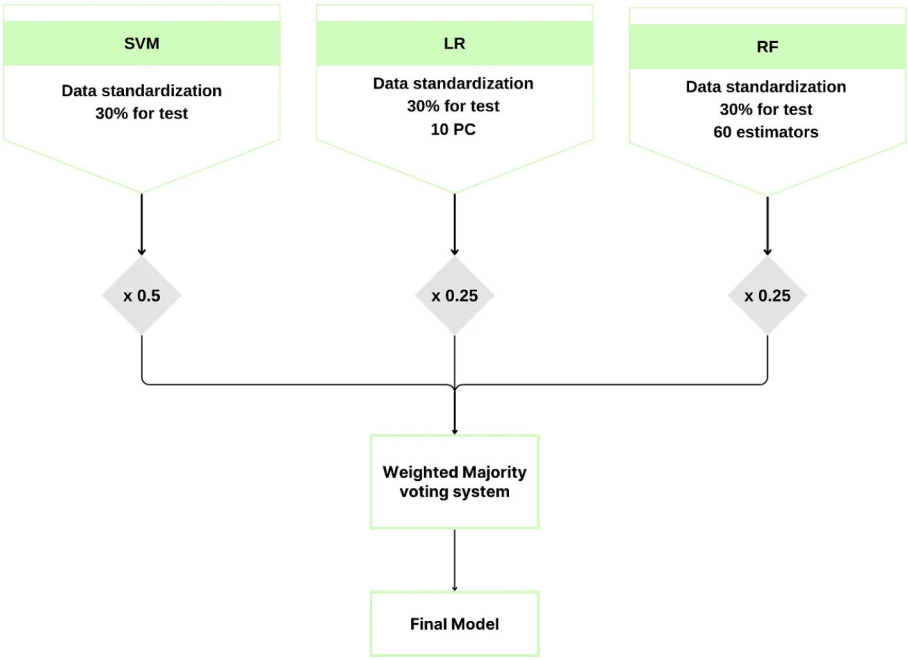


**FIGURE 3.** Diagram illustrating the composition of proposed model.

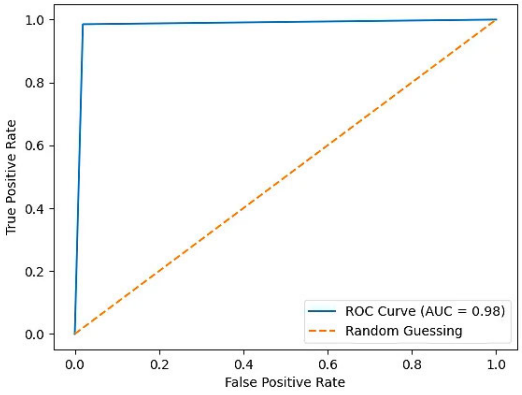


**FIGURE 4.** ROC curve showing the performance of the proposed final model applied to the test data.

### C. CLINICAL INTEGRATION ANALYSIS

To transition our prototype to practical use, it is essential to first implement and validate our findings with a larger, more diverse data set that reflects real-world clinical settings. This step is crucial to ensure the robustness and generalizability of the model in different clinical contexts and populations. Once validated, we must adapt the model to accommodate the variability and complexities of clinical settings. It is essential to ensure the accuracy, reliability and interpretability of the model by health professionals. The next step is to evaluate its integration into existing diagnostic workflows.

Integration can occur in several stages, from FNA data entry into the ML system to automated preprocessing (normalization and handling of missing values) and presentation of interpretable predictions to healthcare professionals. The development of an API will facilitate communication with hospital information systems (HIS) and electronic medical records (EMR). An intuitive user interface will simplify data entry and results display. Implementing an automated process from data acquisition to report generation will minimize errors and improve clinical workflow efficiency.

However, there are significant barriers to its adoption in real clinical settings. Variability in data quality between hospitals can affect model performance. Integrating the model with existing hospital systems could pose technical challenges. Maintaining and updating the model will require resources and a robust machine learning operations (MLOps) strategy. Adequate training of medical personnel is crucial from an organizational and cultural point of view. Ultimately, the implementation of ML solutions in the healthcare sector must comply with strict regulations and respect the privacy of patient data.

### D. COMPARISON OF THE PROPOSED MODEL WITH SIMILAR PROPOSALS

Our model implements a weighted voting system that combines LR, RF and SVM classifiers. We used data standardization and applied PCA to reduce the correlation between the original variables. We achieved an accuracy of 98%, indicating that our approach is effective. Simultaneously with this research, other studies have been developed using the same dataset and applying different ML techniques for the same objective. These works are presented in Table 1 in italics ( [29], [30], [31], [32], [33], [34]).

In [29], the authors compare the predictions of three ML methods: GBM, XGBoost and LightGBM (see Table 1). They apply feature adjustment and perform parameter

optimization, achieving an accuracy of 99% using LightGBM which stands out for its efficiency in training time. To implement the final model, they develop a mobile application. The integration of such a mobile application experimentally, or in a clinical setting, is not mentioned or discussed. The computational cost of our final model is similar due to the small size of the data set. In our study, we obtained similar results by using lightweight models, which are fast and flexible.

In [30], LR, SVM and CART, are combined (see Table 1) by applying weighting to the classes, achieving an accuracy of 99.3%. They apply ensemble models just like us and on the same dataset. However, there is no mention of the possible implementation of the final model in clinical settings or its integration with existing systems. Therefore, both proposals likely have similar processing times. In our study, we optimized our proposal by applying hyperparameters such as standardization, PCA and varying the size of the training set. Finally, we obtained similar results to [30] under similar conditions.

In [33], the authors merged tabular data prediction and image prediction. For tabular data, they use CNN model for feature extraction and combine RF and SVM models for classification. With these approaches, they achieve an accuracy of 99.99%. Their processing time is higher than ours due to the use of the CNN model for feature extraction. While DL and XAI are used to improve model interpretability, the usability of the model in clinical environments is not evaluated. Without using CNN feature extraction and with less computational time, we obtain a fairly similar result.

In [32], models like XGBoost, RF, SVM, KNN and LR are utilized. This study shows strong performance metrics, achieving 99.52% accuracy using the RF model. Additionally, SHAP is employed to interpret the impact of different features, thereby enhancing model interpretability. The use of multiple models and the combination of feature selection techniques (SMOTE-ENN and Mutual Information RFE based on XGBoost) suggest a high computational cost. Furthermore, parameter optimization through grid search can also be computationally intensive. Although the study provides a good solution in terms of interpretability and demonstrates metrics superior to those of the current study, the high computational cost remains a disadvantage due to the complexity of methods and intensive optimization required.

In [34], models like ET, LightGBM, RC and LDA are developed, proposing an ensemble model named ELRL-E. This study achieves good results with 97.66% accuracy. However, no specific techniques for improving model interpretability are mentioned, nor is the feasibility of clinical implementation addressed. The use of EL with multiple base algorithms (ET, LightGBM, RC, LDA) and a voting system entails significant computational costs. Additionally, no specific methods to mitigate this cost are discussed. Although the results are slightly lower than those obtained in our study, the computational cost remains considerable.

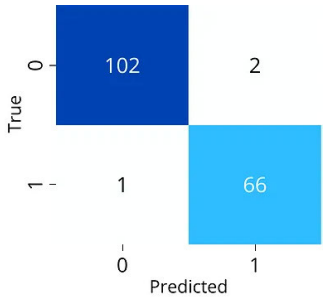


**FIGURE 5.** Confusion matrix of the proposed final model applied to the test data.

In [31], Hybrid Models like GA + Fisher Score, Variance + GA and GA ∪ L1 (Lasso) are employed. The study presents competitive metrics, highlighting that the best combination (GA + Fisher Score) achieves 99.12% accuracy. Hybrid feature selection (GA + Fisher Score) is used to prevent overfitting and enhance generalization. However, the use of hybrid feature selection methods can be computationally expensive, particularly due to the need to evaluate multiple feature combinations. Although the results are slightly superior to those obtained in our approach, [31] acknowledges a high computational cost associated with hybrid methods and the evaluation of multiple combinations. Additionally, they identify limitations in interpretability due to the complexity of these methods and recognize scalability issues for large datasets.

## IV. CONCLUSION AND FUTURE LINES

We have developed a combined model using LR, RF and SVM algorithms. Incorporating PCA has been shown to mitigate the excessive impact of individual variables in the model. Improvements in results have been observed when using a test set that represents approximately 30% of the total data. This model achieves an accuracy of 98% in classifying the WDBC dataset. As a prototype utilizing traditional ML models, it offers an efficient and transparent solution for processing structured medical data. This is important in medical applications, where interpretation of results is essential.

To address the identified limitations and enhance our ML models for breast cancer detection, we suggest several avenues for future research. Firstly, exploring the use of deep neural networks and advanced ensembles such as boosting and stacking. Additionally, considers graph-based models and Bayesian networks. It's crucial to expand the dataset. This can be achieved by collaborating with various institutions and accessing different repositories to include data from diverse geographic regions. Conducting prospective studies in real clinical settings is essential to validate model effectiveness. Moreover, longitudinal studies that track patients over time can observe model performance with sequential data. Integrating other data types can also enhance accuracy. This includes genetic information, electronic medical records and medical imaging data. Improving model interpretability is

paramount. Applying techniques like SHAP and LIME can help explain predictions from complex models. Developing hybrid models that combine rule-based approaches with ML may also be beneficial. Implementing these strategies can mitigate current limitations and advance the accuracy and clinical acceptance of ML models for breast cancer detection.

## ETHICAL STATEMENT

In this study, we are committed to ensuring respect for ethical and legal principles in medical research, particularly regarding the use of patient data. The data is sourced from WDBC, a recognized and widely used source in medical research involving ML. It is important to note that all data in this repository has been anonymized by established policies to protect patient privacy. Therefore, no individual is identified in our analysis.

## REFERENCES

[1] Formatted B. S. Chhikara and K. Parang, "Global Cancer Statistics 2022: The trends projection analysis," *Chem. Biol. Lett.*, vol. 10, no. 1, p. 451, 2023.

[2] A. C. Society. (2024). *Cancer Facts and Figures 2024*. [Online]. Available: https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html

[3] T. B. Bevers, "Breast cancer screening and diagnosis, version 3.2018, NCCN clinical practice guidelines in oncology," *J. Nat. Comprehensive Cancer Netw.*, vol. 16, no. 11, pp. 1362–1389, 2018.

[4] L. Guo, Y. Xie, J. He, X. Li, W. Zhou, and Q. Chen, "Breast cancer prediction model based on clinical and biochemical characteristics: Clinical data from patients with benign and malignant breast tumors from a single center in South China," *J. Cancer Res. Clin. Oncol.*, vol. 149, no. 14, pp. 13257–13269, Jul. 2023.

[5] S. Yan, J. Li, and W. Wu, "Artificial intelligence in breast cancer: Application and future perspectives," *J. Cancer Res. Clin. Oncol.*, vol. 149, no. 17, pp. 16179–16190, Sep. 2023.

[6] G. D. Magoulas and A. Prentza, *Machine Learning and Its Applications: Advanced Lectures*. Berlin, Germany: Springer, 2001, ch. Machine learning in medical applications, pp. 300–307.

[7] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *Ensemble Learning*. New York, NY, USA: Springer, 2009, pp. 605–624, doi: 10.1007/978-0-387-84858-7_16.

[9] A. Hamza, "An enhanced breast cancer diagnosis scheme based on two-step-SVM technique," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, pp. 158–165, 2017.

[10] E. A. Bayrak, P. Kirci, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," in *Proc. Sci. Meeting Elect.-Electron. Biomed. Eng. Comput. Sci. (EBBT)*, Apr. 2019, pp. 1–3.

[11] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, Jun. 2018.

[12] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using genetically optimized neural network model," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4611–4620, Jun. 2015.

[13] A. Hazra, S. Kumar, and A. Gupta, "Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms," *Int. J. Comput. Appl.*, vol. 145, no. 2, pp. 39–45, Jul. 2016.

[14] B. E. Bejnordi, J. Lin, B. Glass, M. Mullooly, G. L. Gierach, M. E. Sherman, N. Karssemeijer, J. van der Laak, and A. H. Beck, "Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 929–932.

[15] W. Sun, T.-L. Tseng, J. Zhang, and W. Qian, "Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data," *Computerized Med. Imag. Graph.*, vol. 57, pp. 4–9, Apr. 2017.

[16] C. Militello, L. Rundo, M. Dimarco, A. Orlando, R. Woitek, I. D'Angelo, G. Russo, and T. V. Bartolotta, "3D DCE-MRI radiomic analysis for malignant lesion prediction in breast cancer patients," *Academic Radiol.*, vol. 29, no. 6, pp. 830–840, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1076633221003858

[17] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015.

[18] M. Learning, "Heart disease diagnosis and prediction using machine learning and data mining techniques: A review," *Adv. Comput. Sci. Technol*, vol. 10, no. 7, pp. 2137–2159, 2017.

[19] M. Tanveer, B. Richhariya, R. U. Khan, A. H. Rashid, P. Khanna, M. Prasad, and T. C. Lin, "Machine learning techniques for the diagnosis of Alzheimer's disease: A review," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 1, pp. 1–35, Apr. 2020.

[20] X. Shu, L. Zhang, Z. Wang, Q. Lv, and Z. Yi, "Deep neural networks with region-based pooling structures for mammographic image classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2246–2255, Jun. 2020.

[21] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani, and M. F. Nagi, "Automated breast cancer diagnosis based on machine learning algorithms," *J. Healthcare Eng.*, vol. 2019, pp. 1–11, Apr. 2019.

[22] P. E. Jebarani, N. Umadevi, H. Dang, and M. Pomplun, "A novel hybrid K-means and GMM machine learning model for breast cancer detection," *IEEE Access*, vol. 9, pp. 146153–146162, 2021.

[23] M. A. Naji, S. E. Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis," *Proc. Comput. Sci.*, vol. 191, pp. 487–492, Jan. 2021.

[24] Z. Huang and D. Chen, "A breast cancer diagnosis method based on VIM feature selection and hierarchical clustering random forest algorithm," *IEEE Access*, vol. 10, pp. 3284–3293, 2022.

[25] M. A. Elsadig, A. Altigani, and H. T. Elshoush, "Breast cancer detection using machine learning approaches: A comparative study," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 13, no. 1, p. 736, Feb. 2023. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85143852215&doi=10.11591%2fijece.v13i1.pp736-745&partnerID=40&md5=db1fe7b690e386b82cfb622b4fe7dcb3

[26] T. Chen, X. Zhou, and G. Wang, "Using an innovative method for breast cancer diagnosis based on extreme gradient boost optimized by simplified memory bounded A*," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105450. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809423008832

[27] P. Gupta and S. Garg, "Breast cancer prediction using varying parameters of machine learning models," in *Proc. 3rd Int. Conf. Comput. Netw. Commun. (CoCoNet)*, vol. 171, 2020, pp. 593–601. https://www.sciencedirect.com/science/article/pii/S1877050920310310

[28] V. J. Kadam, S. M. Jadhav, and K. Vijayakumar, "Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression," *J. Med. Syst.*, vol. 43, no. 8, pp. 1–11, Aug. 2019, doi: 10.1007/s10916-019-1397-z.

[29] K. M. M. Uddin, N. Biswas, S. T. Rikta, S. K. Dey, and A. Qazi, "XML-LightGBMDroid: A self-driven interactive mobile application utilizing explainable machine learning for breast cancer diagnosis," *Eng. Rep.*, vol. 5, no. 11, Nov. 2023, Art. no. e12666, doi: 10.1002/eng2.12666.

[30] T. R. Mahesh, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Anal.*, vol. 4, Dec. 2023, Art. no. 100247. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772442523001144

[31] J. A. Ayoola and T. Ogunfunmi, "A comparative analysis of hybridized genetic algorithm in predictive models of breast cancer tumors," *IEEE Access*, vol. 11, pp. 87111–87119, 2023.

[32] H. Chen, K. Mei, Y. Zhou, N. Wang, and G. Cai, "Auxiliary diagnosis of breast cancer based on machine learning and hybrid strategy," *IEEE Access*, vol. 11, pp. 96374–96386, 2023.

[33] R. M. Munshi, L. Cascone, N. Alturki, O. Saidani, A. Alshardan, and M. Umer, "A novel approach for breast cancer detection using optimized ensemble learning framework and XAI," *Image Vis. Comput.*, vol. 142, Feb. 2024, Art. no. 104910. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885624000131

[34] A. Batool and Y.-C. Byun, "Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm," *IEEE Access*, vol. 12, pp. 12869–12882, 2024.
[35] L. M. Abokaff, "Classification of breast cancer diagnosis systems using artificial intelligence techniques: Survey," *Social Netw. Comput. Sci.*, vol. 3, no. 5, Jul. 2022, Art. no. 368, doi: 10.1007/s42979-022-01275-x.
[36] M. Radak, H. Y. Lafta, and H. Fallahi, "Machine learning and deep learning techniques for breast cancer diagnosis and classification: A comprehensive review of medical imaging studies," *J. Cancer Res. Clin. Oncol.*, vol. 149, no. 12, pp. 10473–10491, Jun. 2023.
[37] Y. Feng, L. Zhang, and Z. Yi, "Breast cancer cell nuclei classification in histopathology images using deep neural networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 2, pp. 179–191, Feb. 2018, doi: 10.1007/s11548-017-1663-9.
[38] D. Peña, *Análisis de Datos Multivariantes*. New York, NY, USA: McGraw-Hill, 2002, ch. Componentes Principales. https://books.google.es/books?id=TrVlAAAACAAJ
[39] J. Aldas Manzano and E. Uriel Jimenez, *Análisis Multivariante Aplicado Con R*. 2nd ed., Madrid, Spain: Ediciones Paraninfo, S.A, 2017. [Online]. Available: https://books.google.es/books?id=FyE3DwAAQBAJ
[40] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Mining: ASA Data Sci. J.*, vol. 15, no. 4, pp. 531–538, Apr. 2022.
[41] R. Patro. (2021). *Cross-Validation: K Fold Vs Monte Carlo*. Medium. [Online]. Available: https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b
[42] IBM. (2023). *What is Logistic Regression*. IBM. [Online]. Available: https://www.ibm.com/topics/logistic-regression
[43] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *J. Algorithms Comput. Technol.*, vol. 12, no. 2, pp. 119–126, Jun. 2018, doi: 10.1177/1748301818756225.
[44] N. I. R. Yassin, S. Omran, E. M. F. El Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Comput. Methods Programs Biomed.*, vol. 156, pp. 25–45, Mar. 2018. https://www.sciencedirect.com/science/article/pii/S0169260717306405
[45] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *Proc. Int. Conf. Comput. Techn., Electron. Mech. Syst. (CTEMS)*, Dec. 2018, pp. 114–118.
[46] J. A. Maat and A. Iqbal, "Introduction to support vector machines and kernel methods," Tech. Rep., Apr. 2019. [Online]. Available: https://www.researchgate.net/publication/332370436

**ROSMERI MARTÍNEZ-LICORT** received the B.Sc. degree in telecommunications and electronics engineering from the University of Pinar del Rio, Cuba, in 2015, and the M.Sc. degree in computer engineering from the University of Valladolid, Spain, in 2022, where she is currently pursuing the Ph.D. degree in computer science.

Since 2023, she has been a Research Teaching Staff with the Computer Science Department, University of Valladolid. From 2015 to 2022, she was an Instructor Professor with the Department of Telecommunications and Electronics, University of Pinar del Rio. She received the international collaboration scholarship between Banco Santander and the University of Valladolid, within the "Iberoamérica + Asia" Program, for the master's degree. From 2022 to 2023, she held the position of a Computer Researcher as part of a research grant funded by the European project "E + DIETing_LAB.–Digital Laboratory for Education in Dietetics Combining Experimental Learning and Community," representing the University of Valladolid.

**CARLOS DE LA CRUZ LEÓN** received the B.Sc. degree in telecommunications engineering and the B.S. degree in business administration from the University of Valladolid, Spain, in 2023. From 2020 to 2023, he was a Telematics Systems Engineer with the education sector. In 2023, he was a Researcher with the eHealth and Telemedicine Group (GTe), University of Valladolid. He is currently an AI Researcher with the CARTIF Technology Center, Valladolid, where he actively contributes to innovative projects developing AI models to enhance efficiency in energy generation, distribution, and consumption.

**DEEVYANKAR AGARWAL** received the Ph.D. degree in applied artificial intelligence from the Department of Information Technologies and Telecommunications, University of Valladolid, Spain.

With an extensive experience in ML, the individual has conducted research focusing on applied DL, and has authored papers on automatic detection of COVID-19 and early detection of Alzheimer's using state-of-the-art CNN architectures and medical images, published in Q1 journals, such as *Applied Soft Computing* (Elsevier) and *Journal of Medical Systems* (Springer Nature). As an educator with 22 years of experience, he has been a Senior Lecturer with the Computer Engineering Department, University of Technology and Applied Sciences, Muscat, since September 2013. In addition to teaching, he is a research coordinator, responsible for managing funding processes within the engineering department; facilitating communication between funding agencies, enterprises, and the department; and striving to secure optimal funding for developing impactful products and projects benefiting society.

**BENJAMÍN SAHELICES** received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Valladolid, Spain, in 1989, 1991, and 1998, respectively.

He is a Professor with the Department of Informatics, University of Valladolid. He has carried out stays and collaborations with research teams from the University of Tennessee in Knoxville, the University of Illinois at Urbana–Champaign, the University of Edinburgh, and in different Spanish universities. He is a member of the GCME Research Group, University of Valladolid. He has involved in collaborations with research groups and companies in the areas of industrial manufacturing, astronomy, economics, finance, and biomedical engineering. He has published in prestigious journals and conferences in the field of computer architecture and artificial intelligence. His main areas of interests include high-performance computing, new memory hierarchies' designs, new paradigms of processing architectures, and heterogeneous computing for DL and its applications.

**ISABEL DE LA TORRE** received the B.Sc. and Ph.D. degrees in telecommunication engineering from the University of Valladolid, Spain, in 2002 and 2010, respectively.

Currently, she is a Full Professor with the Department of Signal Theory and Communications, University of Valladolid. She is a Leader of the GTe Research Group (http://sigte.tel.uva.es). She is the author of more than 300 papers in SCI journals, peer-reviewed conference proceedings, books, and international book chapters. She has co-authored 22 registered innovative software. She has been involved in more than 150 program committees of international conferences until 2024. She has participated/coordinated in 48 funded European, national, and regional research projects.

**JOSÉ PABLO MIRAMONTES-GONZÁLEZ** received the B.Sc. degree in medicine and surgery from the University of Alcala de Henares, Spain, and the University Expert Degree in vascular risk units and vascular risk management from the University of Cordoba, Spain. He received the title of Specialist in Internal Medicine and the Doctor of Medicine from the University of Salamanca, Spain, in 2011.

Currently, he is an Associate Professor with the Department of Medicine, University of Valladolid. He is an Assistant in the internal medicine service with multiple hospitals, including the University Hospital of Salamanca, the University Clinical Hospital of Valladolid, and the Río Hortega University Hospital. He has been a Professor with several universities in Spain. He has received prestigious fellowships, including the Martn Escudero Foundation Postdoctoral Fellowship at the University of California-San Diego, in 2011; the Novo Nordisk Independent Research Fellowship; and the IP GRS Fellowships, in 2020 and 2021. He has contributed to numerous publications and has presented at 107 national organizations and international conferences. He participates in research as a member of the Spanish Society of Internal Medicine (SEMI); and the SEMI Diabetes, Obesity, and Nutrition Working Group.

**MOHAMMED AMOON** received the B.Sc. degree in electronic engineering and the M.Sc. and Ph.D. degrees in computer science and engineering from Menoufia University, in 1996, 2001, and 2006, respectively. He is currently a Professor of computer science and engineering with the Department of Computer Science and Engineering, Menoufia University. He is also a Professor of computer science with the Department of Computer Science, King Saud University. His research interests include agent-based systems, fault tolerance techniques, scheduling algorithms, green computing, cloud computing, fog computing, and the Internet of Things (IoT).