Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/190408

Optimized Learning Methods for Classifying Breast Cancer Subtypes Based on Gene Expression Data

Ana Beatriz M. Valentin[1], *Glaucia M. Bressan*[1,*], *Elisângela Ap. S. Lizzi*[1] *and Leonardo Canuto Jr.*[2]

[1]Graduate Program on Bioinformatics, Universidade Tecnológica Federal do Paraná, 1640 Alberto Carazzai Avenue, Cornélio Procópio, 86300-000, Paraná, Brazil
[2]Academic Department of Computing, Universidade Tecnológica Federal do Paraná, 1640 Alberto Carazzai Avenue, Cornélio Procópio, 86300-000, Paraná, Brazil

Received: 7 Dec. 2024, Revised: 21 Jan. 2025, Accepted: 23 Mar. 2025
Published online: 1 May 2025

Abstract: Understanding the characteristics of tumors and breast cancer subtypes from gene expression data is crucial to aid in cancer type identification, obtain a more accurate diagnosis, and promptly direct appropriate treatment. In this context, the objective of this study is to apply machine learning and deep learning methods for the multi-class classification of genes associated with breast cancer, considering gene expression datasets, and to evaluate the predictive performance of these methods. The dataset used is obtained from The Cancer Genome Atlas repository and are preprocessed for data treatment and the application of dimensionality reduction techniques due to the high number of variables. Initially, principal component analysis was used to reduce the dimensionality of the data. Next, different traditional machine learning methods such as Logistic Regression, Support Vector Machine, and Random Forest are applied, along with deep learning models such as Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN). To enhance the performance of these models, the Optuna library is used for hyperparameter optimization, evaluating the performance of the algorithms both with and without optimization. Performance comparison among the algorithms showed that Support Vector Machine achieved high accuracy. However, the MLP and CNN models, especially when optimized with Optuna, also showed competitive results. Optimization adjusted crucial parameters such as learning rate and number of layers, which resulted in significant performance improvements. Although Random Forest was less affected by optimization, MLP and CNN showed substantial gains. The analysis highlighted that hyperparameter optimization can be essential to improve the accuracy of the classifier. An analysis of feature importance was conducted in order to study which genes have the greatest relevance in the classification task.

Keywords: Hyperparameters optimization, feature importance, machine learning, deep learning, breast cancer

1 Introduction

Breast cancer remains the most frequently diagnosed cancer in women globally, representing approximately 24.5% of all cancers in women [20]. According to the World Health Organization[1], more than 2.3 million new cases of breast cancer were recorded in 2020 alone, and it continues to be a leading cause of cancer-related mortality. Despite the availability of advanced screening tools, survival rates still vary significantly in different regions, particularly in low- and middle-income countries where access to early detection and treatment options is limited [4].

Early diagnosis is fundamental to improving survival outcomes and reducing the overall healthcare burden, highlighting the need for more precise diagnostic tools. The stage in which cancer is diagnosed plays a critical role in determining a patient's prognosis and survival outcomes [31]. Minimizing diagnosis delays is crucial to ensure timely treatment and improve the likelihood of a successful cure [12]. As such, understanding the nature of breast tumors and differentiating between benign and malignant types is the key to advancing breast cancer detection and diagnosis [7]. This distinction is fundamental in improving survival rates and the effectiveness of treatment.

One of the emerging areas in breast cancer research is the analysis of gene expression data, which provides a

[1] https://www.who.int/news-room/fact-sheets/detail/breast-cancer

* Corresponding author e-mail: glauciabressan@utfpr.edu.br

molecular-level understanding of the disease. Gene expression profiling has allowed researchers to identify different subtypes of breast cancer, each with unique molecular signatures that influence prognosis and response to treatment. However, analyzing these high-dimensional datasets, often containing thousands of genes, poses significant challenges due to the complex relationships between gene expressions and cancer phenotypes. This has led to the growing application of machine learning (ML) and deep learning (DL) techniques, which are capable of uncovering hidden patterns and making accurate predictions based on large volumes of data [30].

Machine learning methods, such as Support Vector Machines (SVM), Decision Trees, Random Forests, and Logistic Regression, have been extensively used to classify breast cancer subtypes based on gene expression profiles [38,16,32]. These models offer powerful classification capabilities, particularly in scenarios involving complex and multidimensional data. For instance, SVM is known for its robustness in separating classes with clear margins, while decision trees and Random Forests provide interpretable models with a focus on feature selection. Logistic Regression, on the other hand, offers a probabilistic approach to binary classification, making it an effective tool for distinguishing between benign and malignant tumor subtypes. The effectiveness of these models, however, is often dependent on the careful tuning of hyperparameters to optimize their performance.

Deep learning models, such as Multilayer Perceptrons and Convolutional Neural Networks, can be understood as a subset of machine learning techniques and have also gained traction in breast cancer research [34]. MLPs, which are fully connected neural networks, can model complex nonlinear relationships in the data, while CNNs, originally designed for image processing tasks, have been adapted for analyzing high-dimensional gene expression data [23]. These models excel at feature extraction, automatically learning hierarchical representations from the data, which can lead to more accurate classifications. Furthermore, deep learning techniques have shown great potential in reducing the need for manual feature engineering, a key advantage when working with gene expression datasets [30,27].

Given the complexity of breast cancer subtypes and the volume of gene expression data, optimizing model performance becomes critical. Hyperparameter optimization techniques, such as those provided by algorithms like Optuna, allow for systematic tuning of model parameters, improving the accuracy and generalization of ML and DL models. In addition to optimization, feature importance analysis is another valuable aspect of this study, as it helps to identify key genes or molecular features that are most predictive of cancer subtypes. This not only enhances model interpretability but also provides biological insights that can guide future research into cancer treatment and prevention strategies.

Faced to this context, the objective of this study is to apply both machine learning (Support Vector Machines, Random Forest and Logistic Regression) and deep learning (Multilayer Perceptron and Convolutional Neural Network) methods to classify breast cancer subtypes - labeled as Luminal A (LumA), Luminal B (LumB), Her2, Basal and Normal (no cancer) - using gene expression data, from The Cancer Genome Atlas (TCGA)[2] repository. The study aims to optimize the performance of these models through hyperparameter tuning and assess their effectiveness in accurately distinguishing between different subtypes of breast cancer.

By leveraging advanced machine learning and deep learning techniques, this study seeks to contribute to the development of more precise diagnostic tools and personalized treatment strategies in breast cancer care. The integration of gene expression data into classification models has the potential to revolutionize early detection and improve survival rates, particularly in regions where access to specialized medical resources is limited. Ultimately, the findings of this research may serve as a foundation for future work aimed at enhancing the understanding of breast cancer subtypes and guiding the development of targeted therapies.

2 Literature Review

Recent research on cancer classification, particularly breast cancer and other types, has intensively employed machine learning (ML) techniques and bioinformatics-inspired algorithms to improve accuracy in disease detection and prognosis. These approaches focus on optimizing feature selection and developing robust predictive models, which are challenging due to the high dimensionality of gene expression data [5, 9].This approach is essential for analyzing large-scale multi-omics data, as demonstrated by DeepProg, a deep learning tool designed to predict patient survival subtypes across various cancers. DeepProg shows promising results in risk stratification and association with genomic signatures [26].

For instance, triple-negative breast cancer (TNBC) classification underscores the importance of identifying subtypes based on immune signatures to stratify patients who may respond to immunotherapy. Using both supervised and unsupervised machine learning methods, researchers identified three TNBC immune subtypes—Immunity-H, Immunity-M, and Immunity-L—each with distinct signaling pathway activities and survival prognoses. These subtypes highlight the relevance of the immune environment in cancer classification and treatment [15].

[2] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

World Health Organization (WHO) data emphasize breast cancer's global prevalence as the most common cancer type, with a high mortality rate reported in 2020. Breast cancer accounted for approximately 12% of all diagnosed cases worldwide, affecting more than 2.3 million women and resulting in 685,000 deaths. Such statistics underscore the importance of enhancing diagnostic and stratification techniques [35], according to the use of supervised and optimized techniques [9,26,15,35,29,10] .

In this context, ML-based predictive techniques like Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Ensemble Classifiers (EC) have been tested to improve breast cancer diagnostic accuracy, utilizing datasets such as the Wisconsin Diagnostic Breast Cancer (WDBC) and Breast Cancer Coimbra Dataset (BCCD). These techniques, implemented with standard performance metrics like confusion matrices and cross-validation, achieved high levels of accuracy and efficiency, notably with SVM, which reached 99.3% accuracy [29].

Further studies highlight the application of algorithms like XGBoost, random forest, and KNN, with key evaluation metrics including recall, precision, and F1-score. Recall is particularly relevant for identifying malignant cancer cells. Hierarchical sampling techniques to address class imbalance were also notable, with XGBoost achieving superior results over other approaches, demonstrating high accuracy and sensitivity in diagnostic applications [10].

Collectively, these studies underscore the potential of artificial intelligence and machine learning techniques to revolutionize medical practice by providing more accurate and personalized diagnostics, which can guide clinical decision-making and improve outcomes for cancer patients.

3 Methodology

This section presents the methodology proposed in this work to perform the classification multiclass breast cancer-related gene expression data, as well as the dataset decription, preprocessing and the learning methods adopted for the classification task.

3.1 Methodological path

Figure 1 shows the methodological workflow employed in this study, outlining each step from data preparation and preprocessing to the application of classification models and the final evaluation of the results obtained.

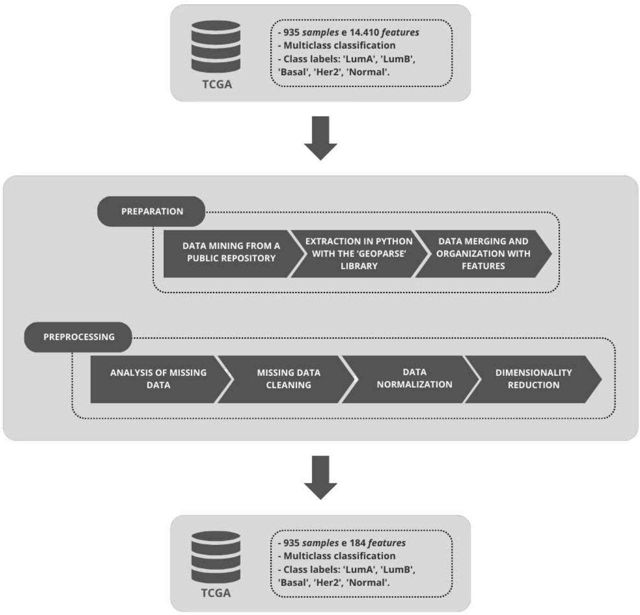
TCGA
- 935 *samples* e 14.410 *features*
- Multiclass classification
- Class labels: 'LumA', 'LumB', 'Basal', 'Her2', 'Normal'.
PREPARATION
DATA MINING FROM A PUBLIC REPOSITORY
EXTRACTION IN PYTHON WITH THE 'GEOPARSE' LIBRARY
DATA MERGING AND ORGANIZATION WITH FEATURES
PREPROCESSING
ANALYSIS OF MISSING DATA
MISSING DATA CLEANING
DATA NORMALIZATION
DIMENSIONALITY REDUCTION
TCGA
- 935 *samples* e 184 *features*
- Multiclass classification
- Class labels: 'LumA', 'LumB', 'Basal', 'Her2', 'Normal'.

Fig. 1: The methodological approach of the study

3.2 Data description and preprocessing

This section details the steps for acquiring, preprocessing, and reducing the dimensionality of the datasets, which is necessary due to the large number of features involved. The raw data from TCGA repository includes key information, such as the number of rows and columns, classification type, and class labels. The columns represent the descriptor variables, which serve as input features for the classification task.

The TCGA databases, which are well-established repositories of gene expression data, apply specific criteria to identify breast cancer samples using array expression profiling techniques. These databases are the result of collaborative projects aimed at comprehensively characterizing genomic alterations across various cancer types. The data is real and publicly available, enabling researchers worldwide to leverage it for advancing research across different fields. Data extraction was performed using the Python library "GEOparse", followed by data integration and organization of the columns.

The dataset used contains 935 samples (rows) and 14,410 features (columns). The final column includes five output classes representing the subtypes of breast cancer: Basal, Her2, LumA, LumB, and Normal (no cancer), making this a multiclass classification task.

Following data acquisition, the preprocessing phase was initiated. This involved handling missing data and performing a data cleaning process, which revealed outliers and values outside the expected range, thus requiring data normalization through z-score scaling [18].

This step prevents variables with significantly higher values from disproportionately influencing the analysis, ensuring the dataset is properly prepared for the modeling phase.

Additionally, due to the high dimensionality typical of gene expression data (where there are more columns than rows), it was necessary to apply a dimensionality reduction technique, such as Principal Component Analysis (PCA) [28,33]. PCA generates uncorrelated principal components that capture most of the variance in the original data. This method is valuable for identifying and interpreting relationships between variables, as well as exploring potential connections between samples [17]. In this study, PCA not only aids in understanding the data but also enhances the performance of statistical models and classification algorithms.

3.3 Classification Models: learning algorithms

This section describes the machine learning and deep learning techniques employed for the multiclass classification task outlined in this study. The core stages of applying supervised classification methods are highlighted. Given that the dataset contains five output classes—Basal, Her2, LumA, LumB, and Normal (no cancer)—representing genes associated with breast cancer, this task involves a multiclass classification challenge.

The supervised models used for breast cancer subtype classification in this study include Support Vector Machine, Logistic Regression, and Random Forest [1]. SVM is notable for identifying an optimal decision hyperplane to effectively separate different classes within a multidimensional space [19]. Logistic Regression, on the other hand, is a statistical method that models the relationship between a binary or multiclass dependent variable and its independent variables [13]. Random Forest [8] is based on constructing multiple decision trees, where each tree is built from a random sample of features, and the final classification is determined by a majority vote from all trees.

These models were selected based on their demonstrated effectiveness in the literature, offering robust performance across various scenarios and applications. Their proven track record, relevance to the domain, and resource availability made them ideal choices for this study [6].

Additionally, a Multilayer Perceptron was utilized, which is a type of feedforward neural network composed of an input layer, multiple hidden layers, and an output layer. Each neuron in one layer is fully connected to every neuron in the following layer, forming a dense network [1]. The MLP is trained using the backpropagation algorithm. The input layer consists of neurons representing features from the gene expression data, while hidden layers apply weighted sums and activation functions like ReLU, sigmoid, or tanh to capture complex patterns within the data. For breast cancer subtype classification, a softmax activation function is typically applied in the output layer to normalize the results into probabilities that sum to one, facilitating interpretation. MLPs are well-suited to identifying and classifying intricate patterns in gene expression data, making them effective for breast cancer subtype classification based on molecular profiles.

Convolutional Neural Networks were also used, consisting of multiple layers where each layer performs localized processing of neighboring features, resulting in more abstract representations as data progresses toward the output layer [1]. CNNs, widely regarded as the first successful deep learning architecture that leverages prior knowledge [21,22], have sparse weight spaces, with each neuron in a layer receiving input from only a subset of neurons in lower layers. This study leverages the local connectivity and weight-sharing characteristics of CNNs to identify complex patterns and interactions among gene expressions associated with different breast cancer subtypes.

At the core of CNNs are the convolutional layers, which apply filters (or kernels) over the input data to extract meaningful features. These filters slide across the input, generating feature maps that highlight specific patterns [21,11]. Following the convolution operations, an activation function, such as ReLU, is applied to introduce non-linearity, allowing the model to learn more complex patterns [22].

As the network progresses, dense layers are used to perform high-level reasoning. These layers connect every neuron in the flattened output from earlier layers to neurons in subsequent layers, helping to integrate the learned features for classification or prediction [11]. The final output layer, typically a dense layer, contains a neuron for each output class. A softmax activation function is used here to convert the model's raw output into probabilities, indicating the likelihood of each class.

3.4 Hyperparameter Optimization

The performance of classification models is significantly influenced by hyperparameters, making hyperparameter tuning an essential step in enhancing the efficiency of learning algorithms [36,37]. To optimize this process, a framework called "Optuna" was introduced [2]. Optuna is an open-source tool designed for hyperparameter optimization, offering flexibility through define-by-run programming, efficient sampling methods, and pruning mechanisms that enhance adaptability and ease of configuration.

In this study, Optuna was utilized to fine-tune the hyperparameters of the machine and deep learning models. By systematically exploring the hyperparameter space, Optuna identifies the best set of hyperparameters to improve model performance. It uses the Tree-structured Parzen Estimator (TPE) algorithm, which efficiently navigates the hyperparameter landscape by constructing probabilistic models of promising configurations. Additionally, Optuna supports pruning, which halts trials that show low potential early on, speeding up the optimization process.

For the MLP model, the following hyperparameters were optimized: the number of hidden layers (ranging from 1 to 5), the number of neurons per layer (between 32 and 256), learning rate (from 0.0001 to 0.1), batch size (from 16 to 128), and dropout rate (ranging from 0.0 to 0.5 to prevent overfitting). Similarly, for the CNN model, Optuna optimized the number of convolutional layers (from 1 to 4), the number of filters (ranging from 32 to 256), kernel size, learning rate (from 0.0001 to 0.1), batch size (from 16 to 128), and dropout rate (between 0.0 and 0.5).

For the SVM model, the hyperparameters optimized included the regularization parameter C explored on a log scale from 10^{-5} to 10^{2}, and the kernel type, selected from linear, polynomial, radial basis function (RBF), and sigmoid. Similarly, for the CNN model, the regularization parameter C he regularization parameter newton-cg, lbfgs, liblinear, and saga.

In the case of the Random Forest model, the process focused on tuning the number of estimators (ranging from 50 to 200), the maximum depth of the trees (from 10 to 50), and the minimum number of samples required to split an internal node (from 2 to 20).

The optimization process was conducted through multiple trials for each model, exploring various combinations of hyperparameters to identify those that maximize model performance. After 10 trials, Optuna returned the best set of parameters for each model, resulting in configurations that achieved the highest average accuracy across cross-validation folds.

3.5 Analysis and Validation

3.5.1 Evaluation Metrics

To train and evaluate the models' performance, the datasets were split into training and testing sets, with 80% allocated for training and 20% for testing, following the Pareto distribution. After the split, the k-fold cross-validation method with k=10 folds was applied [1]. Cross-validation provides a more reliable assessment of model performance on unseen data by using multiple data splits, which minimizes the influence of a single train-test division. During each fold, the selected evaluation metrics were calculated, and after 10 iterations, the average performance of the model was estimated based on these values.

The models' performance was evaluated using well-known metrics, including the confusion matrix [25], accuracy and cross-validation [1,6]. Python version 3.11 was chosen for this study due to its efficiency and high performance. All analyses were conducted with a significance level fixed at 5%. The Python scripts used for this analysis are available on GitHub, accessible through the following link[3].

3.5.2 Feature Importance Analysis

The feature importance analysis was conducted using the SHAP library[4], a widely used tool in Python to interpret the output of machine learning models. Based on Shapley values, derived from cooperative game theory, the library accurately quantifies the contribution of each input variable to the model's prediction, providing a reliable interpretation even for highly complex models [24].

The main features of SHAP (SHapley Additive exPlanations) include variable importance plots, dependence analysis, and prediction decomposition. These tools help to better understand the inner workings of models such as decision trees, neural networks, and other complex algorithms, providing a clear view of both the global and local effects of each variable on the final output.

SHAP unifies different explanation methods under a coherent approach, assigning an importance value to each feature based on its individual impact on the model's prediction. This method calculates the expected prediction by considering various combinations of features, allowing a precise identification of how each feature influences model performance, both locally and globally [24].

4 Results and Discussion

This section presents the findings of the study, highlighting the performance of classification models when applied to datasets from the TCGA repository. A comparative performance analysis is also provided, utilizing statistical metrics for evaluation, the classifiers were evaluated both without hyperparameter optimization and with the application of optimization techniques.

The original dataset contains 935 samples and 14,408 variables corresponding to gene expression. Due to the high dimensionality of the data, the PCA technique was applied, resulting in a reduction to 184 variables.

4.1 Results Without Hyperparameter Optimization

Table 1 presents the results of the models without any hyperparameter tuning. The results are shown for each of the five output classes, which represent different breast cancer subtypes: LumA, LumB, Basal, Her2 and Normal.

[3] https://github.com/anabev/breastcancer-classification

[4] Available at: https://shap.readthedocs.io/en/latest/

Table 1: Classification model performance

Logistic Regression			
	Precision	Recall	F1-score
LumA	0.84	0.82	0.83
LumB	0.55	0.66	0.60
Basal	0.80	0.77	0.79
Her2	0.78	0.33	0.47
Normal	0.18	0.50	0.27
Accuracy		0.7166	
Cross-validation		0.7347	

Multilayer Perceptron			
	Precision	Recall	F1-score
LumA	0.78	0.89	0.83
LumB	0.64	0.51	0.57
Basal	0.75	0.77	0.76
Her2	0.62	0.38	0.47
Normal	0.50	0.75	0.60
Accuracy		0.7273	
Cross-validation		0.7498	

Support Vector Machine			
	Precision	Recall	F1-score
LumA	0.75	0.98	0.85
LumB	074	0.49	0.59
Basal	0.85	0.94	0.89
Her2	0.78	0.33	0.47
Normal	1.00	0.00	0.00
Accuracy		0.7701	
Cross-validation		0.7925	

Convolutional Neural Network			
	Precision	Recall	F1-score
LumA	0.66	0.93	0.77
LumB	0.50	0.22	0.31
Basal	0.81	0.84	0.83
Her2	0.80	0.38	0.52
Normal	1.00	0.00	0.00
Accuracy		0.6791	
Cross-validation		0.7059	

Random Forest			
	Precision	Recall	F1-score
LumA	0.70	0.92	0.79
LumB	0.59	0.46	0.52
Basal	0.80	0.90	0.85
Her2	1.00	0.05	0.09
Normal	1.00	0.00	0.00
Accuracy		0.7005	
Cross-validation		0.7294	

The classifiers were evaluated with and without hyperparameter optimization. Logistic Regression was configured with `max_iter = 6500`. to ensure model convergence. For SVM, the RBF kernel was used without additional adjustments, and Random Forest was employed with its default settings. The MLP used an architecture with four hidden layers containing 100, 75, 50, and 25 neurons, respectively. It employed ReLU activation in the hidden layers, softmax in the output layer, and was trained for up to 500 iterations using the Adam optimizer. Cross-validation was performed using StratifiedKFold with 10 folds.

The CNN was configured with two Conv1D layers (with 32 and 64 filters), followed by pooling layers, a dense layer with 64 neurons, ReLU activation, and a dropout of 0.5. The output layer used softmax to classify the 5 classes. The model was trained using the Adam optimizer for 25 epochs, with 5-fold cross-validation (KFold).

To further analyze the classifiers' performance, a confusion matrix was also considered. Figure 2 provides a visual representation of the confusion matrix performance of all the classification models used, showing that the highest values are located along the diagonal of the matrix, indicating correct classifications on the test set. In the confusion matrix, the breast cancer subtypes are denoted as LumA (0), LumB (1), Basal (2), Her2 (3) and Normal (4).

To gain a deeper understanding of the feature importance in the classification process, Figure 3 illustrates the SHAP values for the following models: (a) SVM, (b) Logistic Regression, (c) Random Forest, and (d) MLP. These visualizations highlight the contribution of each principal component, derived from PCA, to the prediction of each sample. It is noted that, despite differences in the algorithms, certain principal components exhibit high relevance across various models.

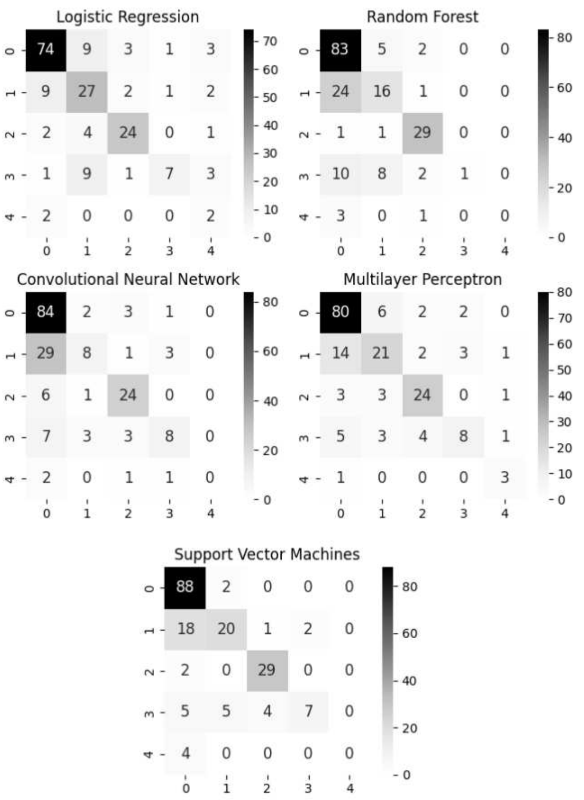

Fig. 2: Confusion matrix

Table 2: Optimized classification model performance

Logistic Regression			
	Precision	Recall	F1-score
LumA	0.74	0.98	0.84
LumB	0.80	0.49	0.61
Basal	0.85	0.94	0.89
Her2	1.00	0.43	0.60
Normal	0.00	0.00	0.00
Accuracy		0.7807	
Cross-validation		0.7887	

Multilayer Perceptron			
	Precision	Recall	F1-score
LumA	0.85	0.91	0.88
LumB	0.68	0.63	0.66
Basal	0.93	0.87	0.66
Her2	0.69	0.52	0.59
Normal	0.29	0.50	0.36
Accuracy		0.7914	
Cross-validation		0.7794	

Support Vector Machines			
	Precision	Recall	F1-score
LumA	0.82	0.93	0.88
LumB	0.64	0.61	0.62
Basal	0.85	0.90	0.88
Her2	0.70	0.33	0.45
Normal	0.33	0.25	0.29
Accuracy		0.7754	
Cross-validation		0.7767	

Convolutional Neural Network			
	Precision	Recall	F1-score
LumA	0.70	0.89	0.78
LumB	0.53	0.44	0.49
Basal	0.87	0.84	0.85
Her2	0.67	0.29	0.40
Normal	0.50	0.25	0.33
Accuracy		0.7005	
Cross-validation		0.9160	

Random Forest			
	Precision	Recall	F1-score
LumA	0.69	0.92	0.79
LumB	0.59	0.46	0.52
Basal	0.80	0.90	0.85
Her2	1.00	0.00	0.00
Normal	1.00	0.00	0.00
Accuracy		0.6952	
Cross-validation		0.7259	

4.2 Results With Hyperparameter Optimization

The results of the classification models, now with hyperparameter optimization, are presented in Table 2. For each classifier, the optimal hyperparameters were determined and applied.

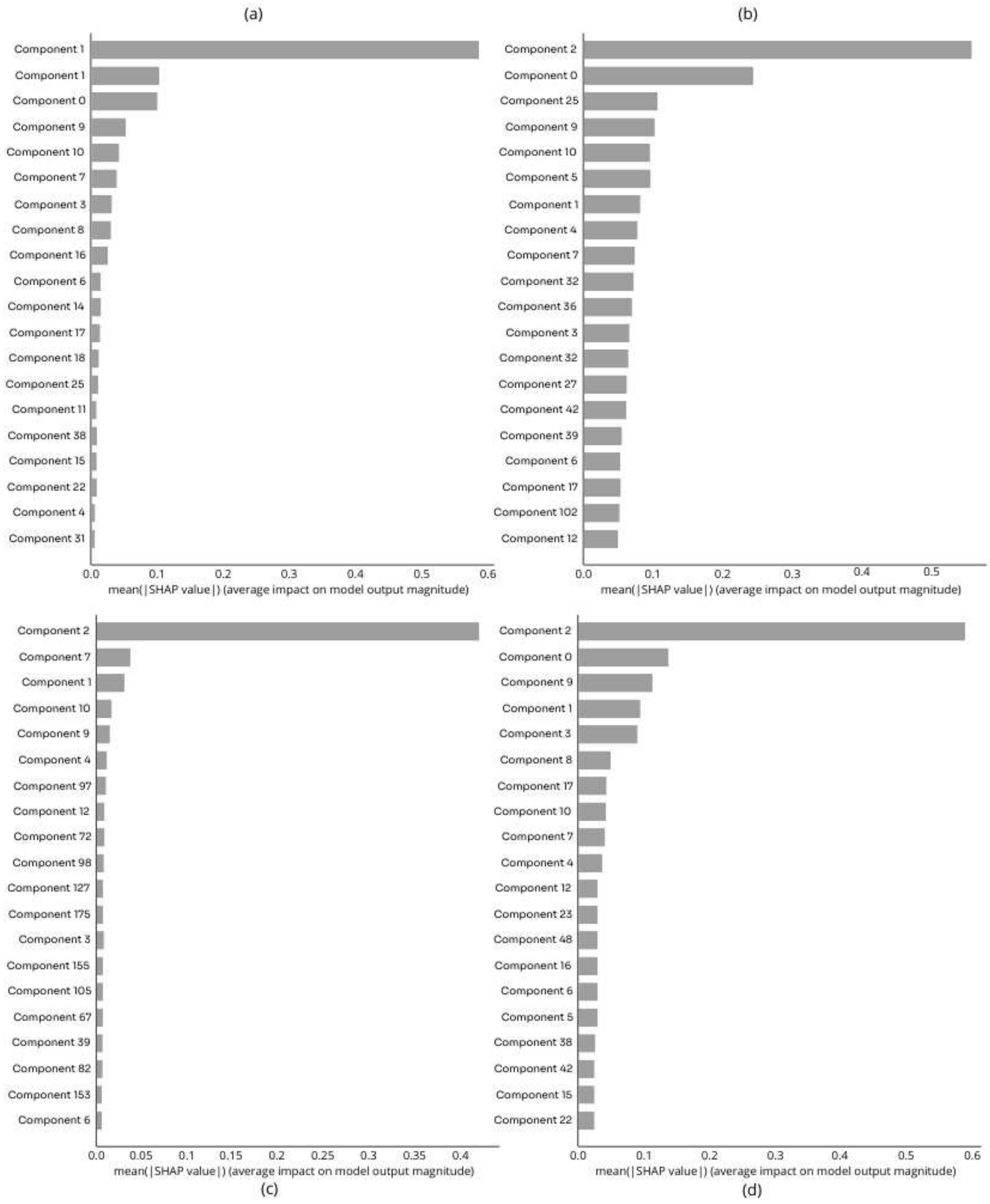

Fig. 3: Feature importance analysis

For the Logistic Regression model, the best value for C was set to $4.43 * 10^{-5}$, with the solver configured as `saga`. The SVM model was optimized with C set to 83.57 and the RBF kernel. In the Random Forest classifier, the best parameters includesde 196 trees, a maximum depth of 22, and a minimum of 15 samples required for a node to split.

The MLP was configured with an architecture of three hidden layers containing 138, 97, and 25 neurons, respectively. The activation function used was ReLU, the solver was `adam`, and the initial learning rate was set to 0.0074. Additionally, the regularization parameter alpha was adjusted to 0.0431. The CNN was optimized with three convolutional layers, one dense layer, 36 filters, a kernel size of 4, a pooling size of 3, and 127 units in the dense layer. A dropout rate of 0.231 was applied to prevent overfitting, and the model was trained for 12 epochs.

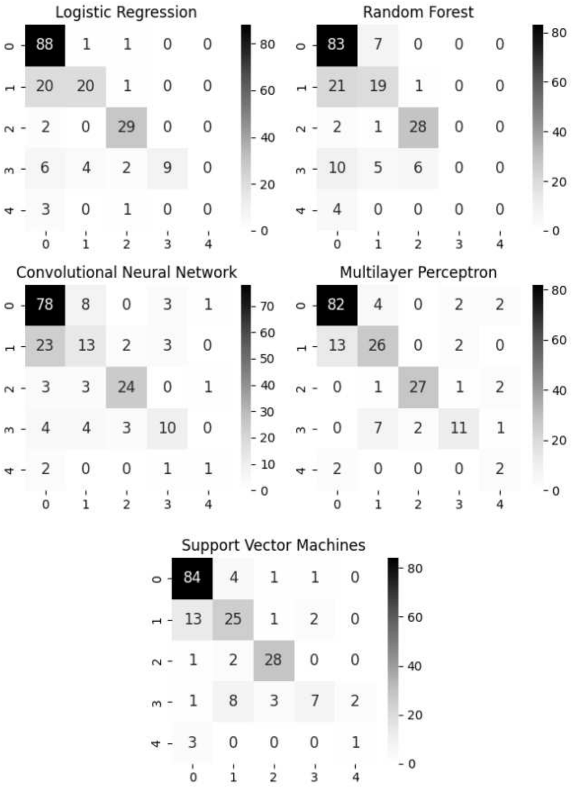

Fig. 4: Confusion matrix with hyperparameter optimization

Cross-validation for the Logistic Regression, SVM, Random Forest, and MLP models was performed using StratifiedKFold with 10 folds. For the CNN, cross-validation was conducted using KFold with 5 folds, due to the higher computational cost of training and evaluating convolutional neural networks.

Figure 4 illustrates the confusion matrix performance of the classification models with hyperparameter optimization. As before, in the confusion matrix, the breast cancer subtypes are denoted as LumA (0), LumB (1), Basal (2), Her2 (3) and Normal (4).

To further analyze variable importance after hyperparameter optimization with Optuna, Figure 5 shows the contributions of the principal components in the adjusted models: (a) SVM, (b) Logistic Regression, (c) Random Forest, and (d) MLP. It is important to note that "components" refers to the principal components obtained from the PCA method. The SHAP technique was once again employed to assign importance values to the variables reduced via PCA. The optimization process further refined the identification of relevant variables, with some variables showing higher consistency across different models, suggesting a positive impact on classifiers performance.

4.3 General Comparison of Results

The evaluation of the classification models' effectiveness is crucial for assessing their generalization capabilities and practical applicability. Notable differences in classifier performance were observed when comparing results with and without hyperparameter optimization using Optuna.

Without optimization, the CNN classifier showed acceptable but inferior results in terms of accuracy, precision, recall, and F1-score, suggesting that the initial parameters—such as learning rate, number of neurons, and network depth—were not optimally configured, which hindered the model's performance.

After applying Optuna, most classifiers showed significant improvements in performance metrics. For instance, Optuna optimized the learning rate and number of hidden layers in the MLP, leading to faster convergence and higher accuracy. In the case of the CNN, optimization was critical for adjusting parameters like the number of filters, kernel sizes, and dropout rates, improving its ability to capture meaningful patterns in the gene expression data.

Interestingly, the Random Forest classifier behaved unexpectedly. This method achieved an accuracy of 0.7005 without optimization and a slightly lower accuracy of 0.6952 after optimization. This can be explained by the inherent robustness of Random Forest, which is less sensitive to hyperparameter tuning compared to more complex models like MLP and CNN. Moreover, the search space defined by Optuna may not have explored the parameters that could improve Random Forest's performance efficiently. In some cases, optimization may not yield noticeable improvements, and this does not necessarily indicate a failure, as the model might already be well tuned without optimization.

The confusion matrix revealed that the most correct classifications were concentrated along the main diagonal, indicating good overall model performance. However, classification errors persisted in certain categories, particularly in distinguishing between the LumA and LumB subtypes. The difficulty in differentiating these two classes suggests they share similar characteristics, which makes it challenging for the models to distinguish between them, even after optimization.

The analysis of variable importance using SHAP provided additional information on which principal components had the most significant impact on the classifiers' predictions. Without hyperparameter optimization, principal component 2 emerged as the most relevant across all models, suggesting it captures critical information for data variation. Component 0 was also consistently important, especially in the Logistic Regression and MLP models, highlighting its relevance in the analysis.

After optimization, some classifiers showed changes in variable importance. In SVM, for example, component 0 became one of the most influential, alongside

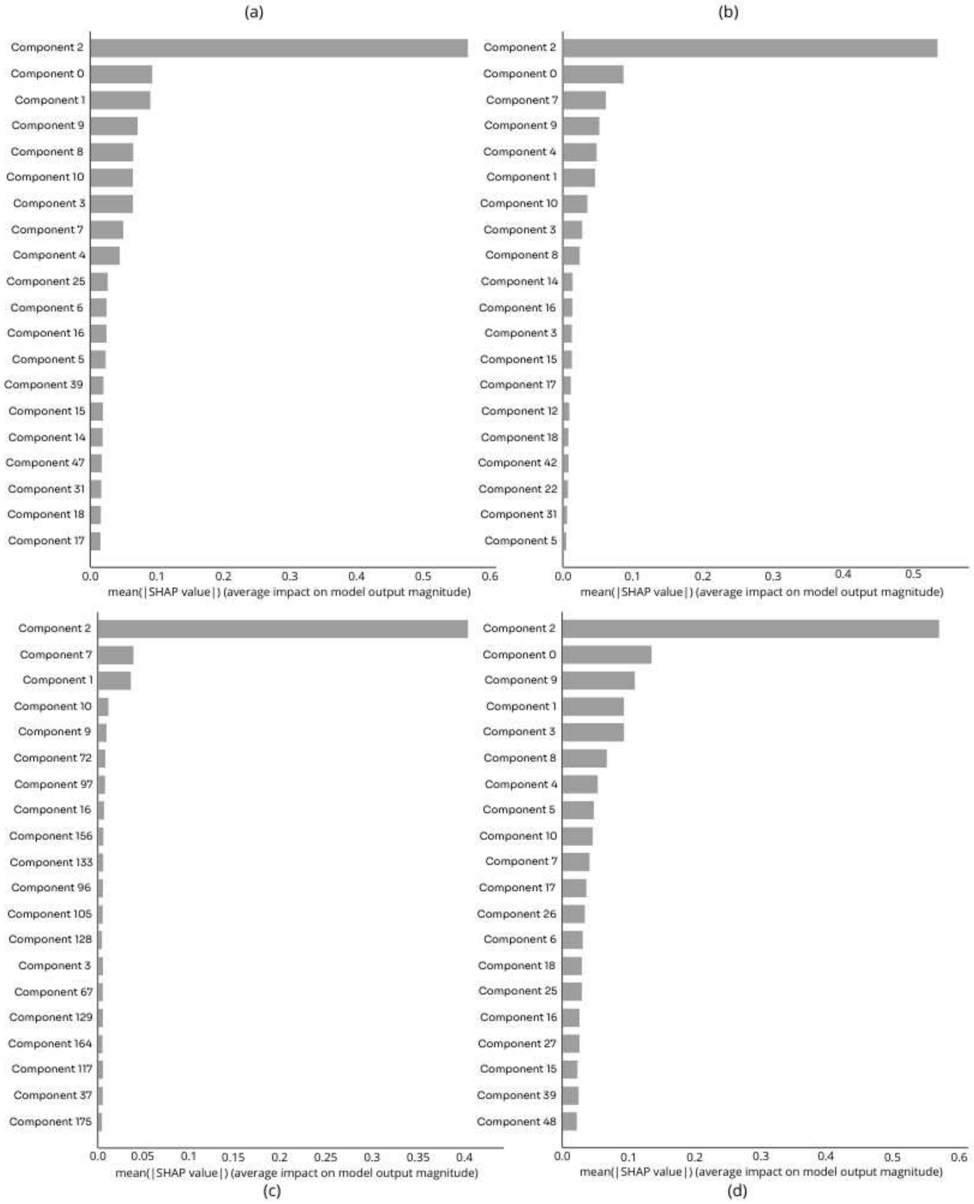

Fig. 5: Feature importance analysis with hyperparameter optimization

component 9, which had already been significant before optimization and gained more importance. In Logistic Regression, component 2 remained the most important, but component 7, previously not in the top five, became more prominent. For Random Forest, the importance pattern remained stable, with component 2 remaining the most crucial before and after optimization, indicating that hyperparameter tuning had little effect on this model.

The SHAP analysis results indicate that while models like MLP and SVM benefited more from optimization in terms of performance and variable importance refinement, Random Forest showed less sensitivity to hyperparameter adjustments, reflecting its robustness across various configurations but limiting the potential gains from optimization.

It is important to highlight that, for the deep learning CNN model, it was not possible to apply the SHAP library due to a dimension incompatibility error, as SHAP is generally designed to work with vectorized input formats [24]. This difference leads to incompatibilities since complex data must be properly transformed to be compatible with interpretation methods, which would likely make it incomparable with the other applied methods. Additionally, due to the architecture of CNNs, which includes multiple convolutional layers and nonlinear functions [14], not all SHAP methods are suitable for interpreting this type of model [3]. This error may indicate that the selected method is not appropriate for capturing the modeling of convolutional network layers, making SHAP optimization unfeasible.

5 Conclusion

This study demonstrated the effective use of machine learning and deep learning methods in classifying breast cancer subtypes based on gene expression data. The application of techniques such as SVM, Random Forest, MLP, and CNN, combined with hyperparameter optimization using Optuna, significantly improved model accuracy, especially in deep learning models.

The results obtained represent important advancements, highlighting specific methods that excel in certain datasets, underscoring the importance of selecting appropriate analytical approaches for each context. Notably, the analysis of feature importance allowed for the identification of variables with greater influence on classification, providing not only an improvement in model accuracy but also a deeper understanding.

This work reinforces the importance of combining model optimization and feature analysis to enhance diagnostic tools and potentially guide the development of more personalized treatment strategies.

Acknowledgement

The authors would like to thank CAPES and the Postgraduate Program in Bioinformatics (UTFPR) for the scholarship granted to student Ana Beatriz Miranda Valentin.

References

[1] Aggarwal, C. C. Data classification. Springer International Publishing, 2015.

[2] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. CoRR, abs/1907.10902, 2019.

[3] Ancona, M., Ceolini, E., Öztireli, C., Gross, M. Gradient-based attribution methods. Explainable AI: Interpreting, explaining and visualizing deep learning, Springer, p. 169–191, 2019.

[4] Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast, 66:15-23, 2022.

[5] Biecek, P., Burzykowski, T. Explanatory model analysis: explore, explain, and examine predictive models. Chapman and Hall/CRC, 2021.

[6] Bishop, C. M. Pattern recognition and machine learning. Springer, 2nd ed., 645-678, 2006.

[7] Braz Junior, G., da Rocha, S. V., de Almeida, J. D., de Paiva, A. C., Silva, A. C., Gattass, M. Breast cancer detection in mammography using spatial diversity, geostatistics, and concave geometry. Multimedia Tools and Applications, 78(10), 13005-13031, 2019.

[8] Breiman, L. Random Forests. Machine Learning, 45:5-32, 2001.

[9] Chatra, K., Kuppili, V., Edla, D. R., & Verma, A. K. Cancer data classification using binary bat optimization and extreme learning machine with a novel fitness function. Medical & Biological Engineering & Computing, 57(12), 2673-2682, 2019.

[10] Chen, H., et al. Classification prediction of breast cancer based on machine learning. Computational Intelligence and Neuroscience, 2023(1), 6530719, 2023.

[11] Chollet, F. Deep Learning with Python. New York: Manning Publication, 2nd ed., 2021.

[12] Dianatinasab, M., Mohammadianpanah, M., Daneshi, N., Zare-Bandamiri, M., Rezaeianzadeh, A., Fararouei, M. Socioeconomic factors, health behavior, and late-stage diagnosis of breast cancer: Impact of delay in diagnosis. Clinical Breast Cancer, 18(3), 239-245, 2018.

[13] Fávero, L. P. L., Belfiore, P. P., Silva, F. L., Chan, B. L. Análise de dados: modelagem multivariada para tomada de decisões, 2009.

[14] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. A survey of methods for explaining black box models. ACM computing surveys (CSUR), ACM New York, NY, USA, v. 51, n. 5, p. 1–42, 2018.

[15] He, Y., et al. Classification of triple-negative breast cancers based on immunogenomic profiling. Journal of Experimental & Clinical Cancer Research, 37, 1-13, 2018.

[16] Iparraguirre-Villanueva, O., Epifanía-Huerta, A., Torres-Ceclén, C., Ruiz-Alvarado, J., Cabanillas-Carbonel, M. Breast cancer prediction using machine learning models. International Journal of Advanced Computer Science and Applications, 14(2), 610–620, 2023.

[17] Johnson, R. A., Wichern, D. W. Applied Multivariate Statistical Analysis, 6th ed., Pearson Education, 2007.

[18] Kreyszig, E. Advanced Engineering Mathematics. John Wiley & Sons, New York, 10th ed., 2010.

[19] Lantz, B. Machine learning with R: Expert techniques for predictive modeling. Packt Publishing Ltd., 2019.

[20] Lei, S., Zheng, R., Zhang, S., Wang, S., Chen, R., Sun, K., Wei, W. Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. Cancer Communications, 41(11), 1183-1194, 2021.

[21] LeCun, Y., Bengio, Y., et al. Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks, 3361(10), 1995.

[22] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324, 1998.

[23] Lopez-Garcia, G., Jerez, J. M., Franco, L., Veredas, F. J. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. PloS One, 15(3), e0230536, 2020.

[24] Lundberg, S. M., Lee, S-I. A unified approach to interpreting model predictions. In 31st Conference on Neural Information Processing Systems (NIPS), 2017. arXiv preprint arXiv:1705.07874.

[25] Moore, D. S., McCabe, G. P., Craig, B. A. Introduction to the Practice of Statistics. 9th ed., W.H. Freeman and Company, 2017.

[26] Poirion, O. B., et al. DeepProg: An ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome Medicine, 13, 1-15, 2021.

[27] Rabiei, R., M., A. S., Sohrabei, S., Esmaeili, M., Atashi, A. Prediction of breast cancer using machine learning approaches. J Biomed Phys Eng, 12(3), 297–308, 2022.

[28] Rencher, A. C. Methods of multivariate analysis. Wiley Series in Probability and Mathematical Statistics, New York, 2002.

[29] Rasool, A., et al. Improved machine learning-based predictive models for breast cancer diagnosis. International Journal of Environmental Research and Public Health, 19(6), 3211, 2022.

[30] Sait, A. R. W., Nagaraj, R. An enhanced LightGBM-based breast cancer detection technique using mammography images. Diagnostics, 14(2), 227, 2024.

[31] Shieh, S. H., Hsieh, V. C. R., Liu, S. H., Chien, C. R., Lin, C. C., Wu, T. N. Delayed time from first medical visit to diagnosis for breast cancer patients in Taiwan. Journal of the Formosan Medical Association, 113(10), 696-703, 2014.

[32] Tewari, Y., Ujjwal, E., Kumar, L. Breast cancer classification using machine learning. In 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, Greater Noida, India, 01-04, 2022.

[33] Timm, N. H. Applied Multivariate Analysis. Springer Texts in Statistics, New York, 2002.

[34] Turgut, S., Dağtekin, M., Ensari, T. Microarray breast cancer data classification using machine learning methods. In Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 01-03, 2018.

[35] World Health Organization. Global breast cancer initiative implementation framework: assessing, strengthening, and scaling-up of services for the early detection and management of breast cancer. World Health Organization, 2023.

[36] Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., Deng, S.-H. Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17(1), 26–40, 2019.

[37] Yang, L., Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295–316, 2020.

[38] Yue, W., Wang, Z., Chen, H., Payne, A., Liu, X. Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13, 2018.

Ana Beatriz M. Valentin Ana Beatriz Miranda Valentin holds a degree in Mathematics from the Universidade Tecnológica Federal do Paraná (UTFPR), Cornélio Procópio campus. She is currently pursuing a Master's degree in Bioinformatics at the Graduate Program in Bioinformatics (PPGBIOINFO) of the Universidade Federal do Paraná (UFPR). Her research focuses on Computational and Systems Biology, with a strong interest in integrating computational techniques to address biological questions.

Glaucia M. Bressan Glaucia Maria Bressan received a teaching degree in Mathematics with emphasis in computation from Universidade Federal de São Carlos (UFSCar), Brazil, a Master's degree in Computational and Applied Mathematics from Universidade de São Paulo (USP) and her PhD in Electrical Engineering at Universidade de São Paulo (USP). She did a post-doctoral research in Electrical Engineering at Universidade de São Paulo (USP), studing Artificial Intelligence. She works as a professor and researcher at the Universidade Tecnológica Federal do Paraná (UTFPR) in the Mathematics Department. She is researcher and accredited professor in the Graduate Program in Bioinformatics and conducts research in Computational and Applied Mathematics, Biostatistics, Operational Research, and Artificial Intelligence.

Elisângela Ap. S. Lizzi Elisângela Lizzi holds a bachelor's degree in Statistics (UFSCar) and a master's and doctorate in Biostatistics/Public Health (USP), including a doctoral trainee period at Johns Hopkins University (JHU). She is currently an Adjunct Professor at UTFPR, Cornélio Procópio campus, and a researcher in the Graduate Program in Bioinformatics (PPGBIOINFO) at the same institution. Her research areas include applied statistics in public health and epidemiology, with a focus on machine learning, artificial intelligence, and Bayesian statistics. Additionally, she serves as a technical consultant for the Pan American Health Organization and the Ministry of Health, contributing to advanced statistical methods applied to health surveillance.

Leonardo Canuto Jr. Leonardo Canuto Junior holds a bachelor’s degree in Computer Engineering (UTFPR). His research areas include Bioinformatics, Data Engineering, Data Science and Software Development.