Prediction of Pancreatic Cancer Through Biomarkers Using Machine Learning Techniques: An Approach for Early Diagnosis

Bianca L.S.M. Cardoso[1*], João Vitor S. Mendes[2]
[1]Ana Nery Hospital, Clinical Research Department; [2]SENAI CIMATEC, Robotics Department; Salvador, Bahia, Brazil

This article explores the prediction of pancreatic cancer using CA 19-9 and CA 125 biomarkers with three machine learning models: Gradient Boosting, Random Forest, and Logistic Regression. The study evaluates their effectiveness through 10-fold cross-validation. Results showed competitive performance, with the Logistic Regression model displaying the highest accuracy, precision, and F1-score, indicating its potential for early diagnosis. Integrating biomarkers and machine learning promises for improving pancreatic cancer prediction and patient outcomes.

Keywords: Pancreatic Cancer. Machine Learning. Cancer Prediction.

Introduction

Pancreatic cancer is a highly challenging and devastating disease that poses a severe public health problem. Its aggressive nature and the absence of distinctive symptoms in the early stages make early diagnosis difficult, resulting in significantly elevated mortality rates [1]. Moreover, pancreatic cancer has the lowest survival rate compared to other cancers, with approximately 80% of cases being inoperable, and 74% of patients succumbing within the first year [2]. In this context, it is crucial to investigate effective approaches for predicting and detecting this type of cancer to improve prognoses and increase patients' chances of survival.

One promising line of research in this field involves the use of biomarkers. Biomarkers are biological substances that can be measured and evaluated as indicators of normal or pathological biological processes [3], including cancer development [4]. Two tumor markers commonly associated with pancreatic cancer are CA 19-9 and CA 125. The presence of these proteins at elevated levels in the blood may indicate the existence of malignant tumors in the pancreas, making them potential candidates for early disease detection [5,6].

Significant advancements have occurred in machine learning, greatly benefiting many medical applications. Machine learning allows algorithms to learn complex patterns in data and make accurate predictions. Integrating biomarker information with machine learning models can be a promising approach to enhance the prediction and diagnosis of pancreatic cancer, increasing the sensitivity and specificity of the detection process [7].

Machine Learning Models

Machine learning is a subfield of artificial intelligence that focuses on developing algorithms and models capable of learning patterns and making decisions from data without being explicitly programmed to perform specific tasks. These models are trained on previously collected datasets, allowing them to recognize relevant features and make predictions or classify new data based on this learning [9].

The models proposed in this study are three popular machine-learning approaches applied to pancreatic cancer prediction.

Gradient Boosting Model

Gradient Boosting is a machine learning technique based on decision trees, in which several weak trees are combined to form a robust and more

Received on 21 October 2023; revised 20 November 2023.
Address for correspondence: Bianca L.S.M. Cardoso. SENAI CIMATEC University Center. Av. Orlando Gomes, 1845 - Piatã, Salvador – Bahia, Brazil. E-mail: bianca.cardoso.farma@gmail.com.

J Bioeng. Tech. Health 2023;6(4):260-264

accurate model. This method works in sequential steps, that each tree is adjusted to correct the errors of the previous model. The result is a weighted combination of predictions from all the trees, which tends to be more robust and accurate in predicting cases of pancreatic cancer based on biomarkers [10].

Random Forest Model

Random Forest is a technique based on decision trees but with a different approach. In this model, multiple decision trees are constructed from random subsets of the original dataset, and their predictions are combined through voting to reach a final decision. This approach helps to reduce the probability of overfitting (excessive fitting to the training data). It increases the model's generalization to unseen data, making it a viable option for predicting pancreatic cancer based on biomarkers [11].

Logistic Regression Model

Unlike tree-based models, Logistic Regression is a machine learning method that aims to predict a binary categorical variable (in this case, pancreatic cancer or not). It uses a logistic function to calculate the probability of belonging to a specific class based on the values of biomarkers. The model is trained to adjust coefficients that weigh the influence of each biomarker on the probability of pancreatic cancer occurrence. From this, predictions can be made, and patients with a higher risk of developing the disease can be identified. When combined with biomarkers CA 19-9 and CA 125, these models can offer a promising approach to improve pancreatic cancer prediction and early diagnosis, enabling more effective treatment and increasing the chances of survival for affected patients [12].

Materials and Methods

The data used in this research were obtained from a dataset by Wieand and colleagues [13], containing information about patients with a history of pancreatic cancer. The dataset includes records of the biomarkers CA 19-9 and CA 125 and the classification of patients into positive and negative cases for pancreatic cancer.

Data Preprocessing

Before proceeding with the analysis, a data preprocessing step was performed. In this phase, possible missing or inconsistent values were handled, and normalization techniques were applied to standardize the scale of the biomarkers. The objective was to ensure data integrity and reliability for subsequent experiments and avoid the disproportionate influence of higher numerical values on the models' outcomes. Data normalization was executed to adjust the values of the biomarkers CA 19-9 and CA 125 to the exact numerical scale. This way, significant differences between the magnitudes of the biomarkers were avoided, preventing them from unduly influencing the performance of the machine learning models. Normalization allowed the algorithms to focus on analyzing relationships and patterns within the data, contributing to more accurate and consistent results.

Cross-Validation

The technique of 10-fold cross-validation was used to evaluate the models' performance robustly and avoid training bias. The dataset was divided into 10 equal parts, where each model was trained on 9 folds and tested on the remaining fold. This process was repeated 10 times, alternating the test folds. Performance metrics were recorded at each iteration, and at the end, the averages were calculated to obtain more accurate estimates of the model's performance.

Metrics of Evaluation

The metrics used to assess the models' performance were accuracy, precision, recall, F1-

score, and the area under the curve (AUC). Accuracy measured the proportion of correct predictions out of the total predictions made. Precision evaluated the models' ability to avoid false positives, meaning the proportion of correctly identified positive cases among those predicted as positive. Recall measured the models' ability to correctly find all positive cases of pancreatic cancer. The F1 score provided a measure of the balance between precision and recall. The AUC metric estimated the models' discriminative capacity.

Results and Discussion

This section presents the evaluation results of the three machine learning models (Gradient Boosting - GB, Random Forest - RF, and Logistic Regression - LR) applied to pancreatic cancer prediction based on the biomarkers CA 19-9 and CA 125.

The experiments used a 10-fold cross-validation to ensure robust results and minimize training bias. Before discussing the model performance metrics, we analyzed the importance of each feature (Figure 1). For the Gradient Boosting Model, both biomarkers, CA 19-9 and CA 125, played a significant role in the prediction, with the relative importance of 71.1% and 28.9%, respectively. The Random Forest Model also showed considerable relevance for both biomarkers, with values of 62.2% for CA 19-9 and 37.8% for CA 125. On the other hand, the Logistic Regression Model attributed higher importance to CA 19-9 (71.9% of the total) compared to CA 125 (28.1%).

The results of the 10-fold cross-validation revealed that all three models showed a solid overall performance in predicting pancreatic cancer. The average accuracy (Figure 2A) obtained was 75.1% for the GB and RF models and slightly higher at 79.4% for the LR model. Around 75% to 79% of the predictions were correct. Furthermore, the precision metric (Figure 2B) demonstrated the models' ability to avoid false positives, meaning their capacity to correctly identify actual positive cases of pancreatic cancer. The LR model achieved the highest precision, reaching 89.9%, followed by the RF model with 83.3% and the GB model with 82.5%. Regarding the recall metric (Figure 2C), which indicates the models' ability to find all positive cases of pancreatic cancer, the LR model achieved 77.8%. In comparison, the GB and RF models obtained a slightly lower rate of 78.9%. It indicates that all three models performed similarly concerning this metric. The F1-score (Figure 3A), which presents a

Figure 1. Feature importance graph.

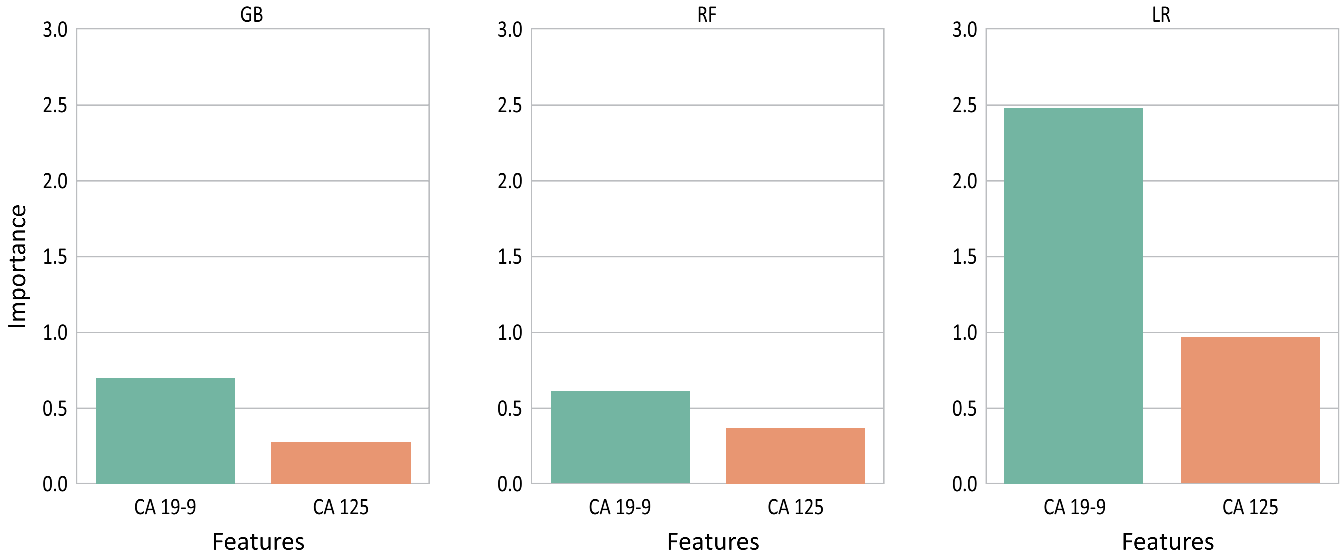


Figure 2. Performance metrics: A) accuracy; B) precision; C) recall.

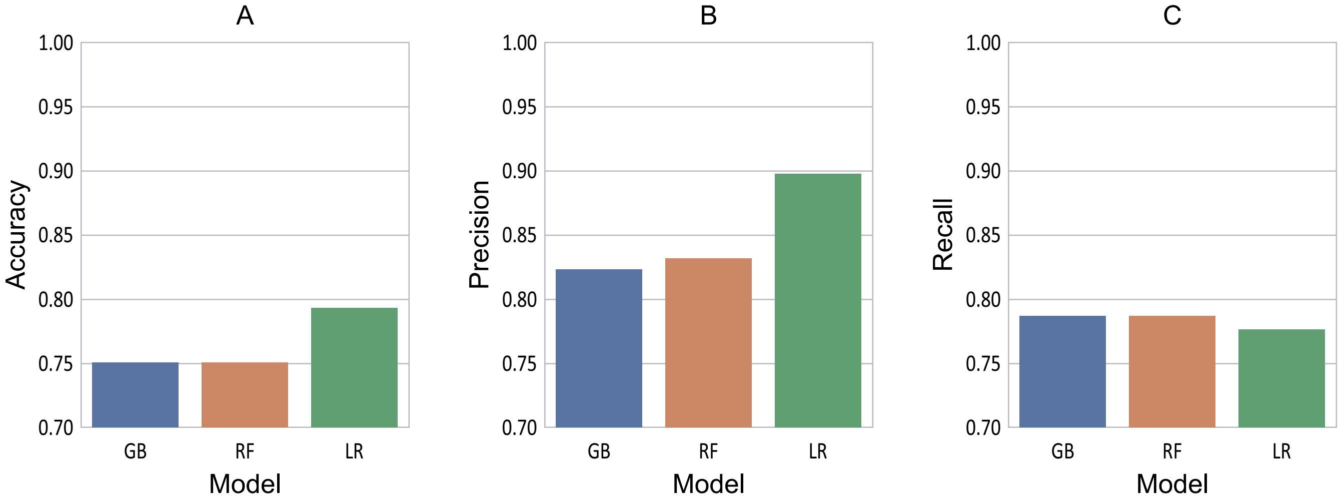
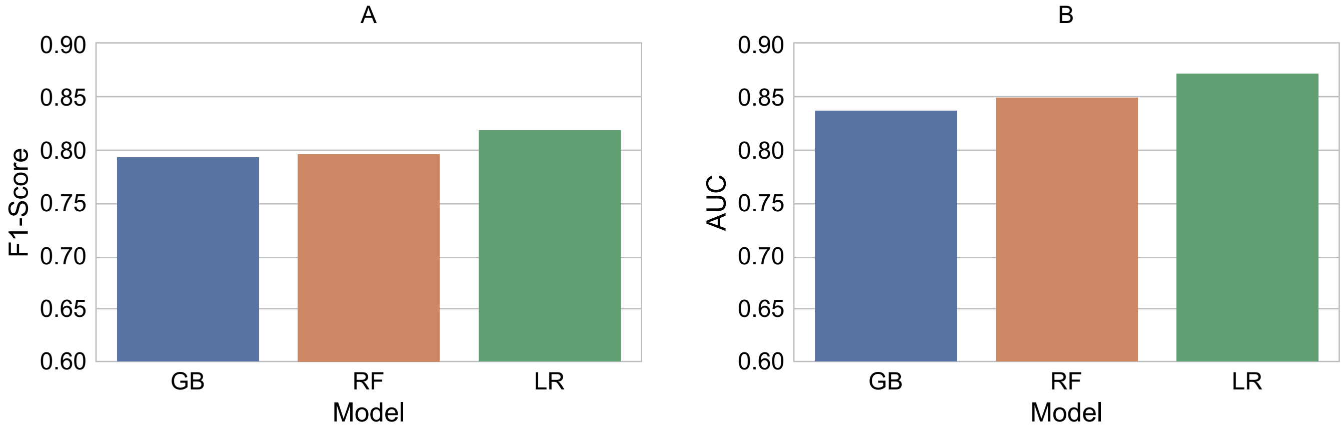


Figure 3. Performance metrics: A) F1-score B) AUC.



harmonic mean between precision and recall, revealed that the LR model achieved the best balance between the two measures, reaching 81.9%. The GB and RF models also performed satisfactorily, with F1 scores of 79.4% and 79.7%, respectively. Finally, we evaluated the AUC (Area Under the Curve) metric (Figure 3B), which represents the area under the Receiver Operating Characteristic (ROC) curve and measures the models' discriminative capacity. Once again, the LR model performed the best, at 87.3%. The GB and RF models also showed promising results, with AUCs of 83.8% and 84.9%, respectively.

The results obtained in this study demonstrate that the three machine learning models, when integrated with the biomarkers CA 19-9 and CA 125, show promising performance in predicting pancreatic cancer. The Logistic Regression model was the most accurate approach, offering a solid balance between precision and recall. However, the Gradient Boosting and Random Forest models also demonstrated efficacy in early disease identification. These results suggest that the machine learning approach based on biomarkers can be a valuable ally in the diagnosis and timely treatment of pancreatic cancer, significantly improving human health.

Conclusion

This study investigated three machine learning models (Gradient Boosting, Random Forest, and

Logistic Regression) for predicting pancreatic cancer based on CA 19-9 and CA 125 biomarkers. The results of the 10-fold cross-validation demonstrated that the models showed solid performance in predicting the disease. The Logistic Regression Model achieved the highest precision, avoiding false positives. The Gradient Boosting and Random Forest models showed promising results with good F1-scores. CA 19-9 and CA 125 biomarkers were identified as essential factors in the prediction. These findings indicate that the machine learning approach integrated with biomarkers can be a valuable tool for early diagnosis of pancreatic cancer, contributing to timely medical interventions and potentially improving the survival of patients affected by the disease. Future studies and clinical validation are recommended to consolidate these findings and enhance diagnostic tools in medical practice.

Acknowledgments

The authors thank Wieand S. and colleagues for providing the data on pancreatic cancer diagnostic markers CA 19-9 and CA 125. We also appreciate the reference to the Biometrika paper (Wieand S, Gail MH, James BR, and James KL. Biometrika 1989;76(3):585-92), where the dataset was cited. This dataset was instrumental in our research on applying machine learning models for early pancreatic cancer detection. We sincerely appreciate the valuable contribution of this dataset to our study.

References

1. de Braud F, Cascinu S, Gatta G. Cancer of pancreas.Critical reviews in oncology/hematology 2004;50(2):147-155.
2. da Fonseca AA, Rêgo MAV. Tendência da mortalidade por câncer de pâncreas em Salvador-Brasil, 1980 a 2012. Revista Brasileira de Cancerologia 2016;62(1):9-16.
3. Strimbu K, Tavel JA. What are biomarkers? Current Opinion in HIV and AIDS 2010;5(6):463.
4. Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. Nature Reviews Cancer 2005;5(11):845-856.
5. Liu L et al. A preoperative serum signature of CEA+/CA 125+/CA 19-9≥ 1000U/mL indicates poor outcome to pancreatectomy for pancreatic cancer. International Journal of Cancer 2015;136(9):2216-2227.
6. Duraker N et al. CEA, CA 19-9, and CA 125 in the differential diagnosis of benign and malignant pancreatic diseases with or without jaundice. Journal of Surgical Oncology 2007;95(2):142-147.
7. Chang JC, Kundranda M. Novel diagnostic and predictive biomarkers in pancreatic adenocarcinoma. International Journal of Molecular Sciences 2017;18(3):667.
8. Liu L et al. Serum CA125 is a novel predictive marker for pancreatic cancer metastasis and correlates with the metastasis-associated burden. Oncotarget 2016;7(5):5943.
9. Zhou Z-H. Machine learning. Springer Nature 2021.
10. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 2013;7:21.
11. Biau G. Analysis of a random forests model. The Journal of Machine Learning Research 2012;13:1063-1095.
12. Su X, Yan X, Tsai C-L. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics 2012;4(3):275-294.
13. Wieand S et al. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 1989;76(3):585-592.