Breast Cancer Risk Estimation with Intelligent Algorithms for Cuban Women

Eugene Munyaneza
Department of Computer Science
Makerere University Kampala
Student Number: 2400721936
Registration Number: 2024/HD05/21936U
munyaneza.eugene@students.mak.ac.ug

Khadija Athman
Department of Computer Science
Makerere University Kampala
Student Number: 2400721918
Registration Number: 2024/HD05/21918U
athman.hadi@gmail.com

***Abstract*—Breast cancer is a major health concern globally, it is one of the leading causes of cancer morbidity and mortality [14]. This report investigates breast cancer risk prediction using machine learning models with a special focus on specific demographic and local cultural factors. Traditional models like Gail [13] and Barlow [4] have shown limitations in predicting cancer risks outside their original demographics, usually underestimating or overestimating risk for non-Caucasian populations. Using our dataset of 1,697 cases from Cuban women, this report expands the traditional models by adding population-specific models using machine learning algorithms such as Decision Trees, Support Vector Machine and Random Forest.**

The methodology includes extensive data wrangling, feature selection, model training using binary classification approach. Methods such as Local Interpretable Model-agnostic Explanations (LIME) [16] and SHapley Additive exPlanations (SHAP) [11] are further employed to explain our results. The model incorporates unique risk factors like breastfeeding duration, exercise, tobacco use, and socioeconomic indicators, alongside conventional ones such as age, family history, and BMI. Results demonstrate a significant improvement in accuracy, with the Random Forest algorithm achieving an Area Under the Curve (AUC) of 1 and accuracy of 99% during internal validation.

A literature review highlights gaps in current breast cancer risk models, showing the need for a more global approach in populations that are not predominantly Caucasian.

The results improve breast cancer early detection for Cuban women. The work highlights the importance of taking into consideration local data into global models thus extending the need for machine learning in public health. Future research will focus on other regions of the world like Africa, utilizing broader features, and improving data collection for machine learning in health care. This report shows a significant step in early detection of breast cancer in less privileged regions which leads to lower mortality in societies that need improved public health.

***Index Terms*—Machine Learning, Breast Cancer, Risk Estimation, Random Forest, Cuban Women.**

I. Introduction

Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide, with a particularly high impact in developing nations like Cuba [14]. Limited healthcare resources and late detection contribute to significantly higher mortality rates compared to developed countries. Traditional risk prediction models, such as the Gail model [13], while effective in Western predominantly Caucasian populations, often fail to address the unique demographic, cultural, and socioeconomic factors that influence breast cancer risk in Cuban women. This results in inadequate predictions and delayed interventions, increasing the public health burden.

Artificial Intelligence (AI), particularly Machine Learning (ML), offers a promising solution to this problem. By analyzing large datasets, identifying patterns, and dynamically integrating new variables, ML algorithms can create population-specific predictive models. These models can incorporate a wide range of risk factors, including age, BMI, family history, lifestyle habits (e.g., smoking, alcohol consumption), and breastfeeding history, to provide personalized risk assessments. The use of AI in breast cancer prediction is particularly important in Cuba, where tailored, cost-effective, and scalable solutions are essential to improving early detection rates.

This study leverages a dataset of 1,697 Cuban women to develop a localized breast cancer risk prediction model using ML techniques. Algorithms like Decision Trees (DT), Support Vector Machine (SVM) and Random Forest (RF) are validated using quantitative and qualitative variables, achieving more accurate predictions compared to traditional models.

This research addresses critical gaps in existing prediction tools, offering an adaptable framework that enhances early detection efforts while considering the unique characteristics of underserved populations. The results contribute to the growing body of work advocating for AI-driven, population-specific healthcare solutions.

II. Background and Motivation

Breast cancer remains the leading cause of cancer mortality among women worldwide, with particularly high mortality rates in developing nations such as Cuba due to late detection and limited access to healthcare services [7]. Traditional models such as the Gail model [13] have limitations in specific populations, often underestimating or overestimating risk for non-Caucasian groups.

The need for an adjusted breast cancer prediction model for Cuban women influenced this research. Various machine learning (ML) models were employed to find the optimal solution. The output is a machine learning-based risk estimation model designed to predict individual breast cancer risks using factors specific to this population.

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to learn and make decisions or predictions without being explicitly programmed to perform specific tasks [17]. Instead, machine learning models identify patterns in data and use these patterns to make informed decisions. We have four classes of ML models, supervised, semi-supervised, unsupervised and reinforcement learning models. In this project, we shall use three supervised ML models to analyse our data; decision trees (DT), support vector machine (SVM) and random forest (RF).

Support Vector Machine (SVM): A machine learning algorithm that finds the optimal boundary (or hyperplane) to separate data points into different classes by maximizing the margin between them.

Decision Trees: A tree-structured model that splits data into branches based on feature values, making decisions at each node until it reaches an outcome or prediction.

Random Forest: An ensemble learning method that combines multiple decision trees to improve prediction accuracy and prevent overfitting by averaging (for regression) or voting (for classification).

After running our models, we used Explainable Artificial intelligence (XAI) to give us further insight into how our selected model works. **Explainable Artificial intelligence (XAI)** is a set of processes and methods that allows humans to better understand and trust the results and outputs generated by machine learning algorithms [1]. This is crucial in order to ensure transparency, trust, and accountability in ML systems. This in turn gives users a reason to easily adopt a model.

III. Literature Review

Research on breast cancer prediction models has evolved significantly over the years, and various models have been developed to estimate individual risk. One of the most widely known models is the Gail model [13], developed in 1989. It uses risk factors such as age, family history, and reproductive history to predict breast cancer risk, and has been calibrated for several populations, including African Americans, Asians, and Hispanics. However, its application in developing countries, especially for Cuban women, has shown poor performance due to cultural and socioeconomic differences.

A. Traditional Models

1) Gail Model: [13]: Initially developed for white women of US birth, the Gail model was later modified to estimate risk for other racial groups. It considers age, reproductive history, and family history but has limitations in accurately predicting breast cancer in non-Caucasian populations due to its inability to integrate localized risk factors such as body mass index (BMI), obesity, diet, or hormonal therapies [2].

2) Barlow Model: [4]: An extension of the Gail model, the Barlow model uses additional factors such as breast density and hormone replacement therapy. It has performed well in the US population, but like the Gail model, its predictive power diminishes for Cuban women, highlighting a gap in accurate risk assessment for this demographic.

B. Machine Learning in Breast Cancer Prediction

Recent advancements in machine learning (ML) have led to the development of more adaptive and personalized models for the assessment of cancer risk. Studies such as Gao et al. [8] and Ang et al. [3] incorporated genetic markers into risk models, increasing accuracy but also adding complexity and cost. For developing countries like Cuba, these approaches are often unaffordable due to the lack of access to genetic testing and the high costs associated with such advanced diagnostics.

More recent studies, such as those by Valencia-Moreno et al. [15] and Li et al. [9], demonstrate how ML models using demographic and medical data can outperform traditional models by incorporating additional risk factors, such as stress levels, obesity, tobacco, and alcohol consumption. ML models like Random Forests, Support Vector Machines (SVM), and Gradient Boosted Trees (GBT) have shown improved performance in populations where traditional models fail to capture specific risk factors.

C. Gaps in the Literature

Despite advancements in prediction models, a significant gap exists in terms of their applicability to non-Western populations. Studies consistently show that models like Gail, even after calibration for Hispanic populations, fail to accurately predict breast cancer risk in Cuban women due to missing variables related to diet, physical activity, and social conditions specific to the Cuban demographic. For instance, a study by Colmenares et al. [5] on Venezuelan population demonstrated that traditional models underestimated risks for women as the models showed a low precision by only identifying 41% of the cases.

D. Contribution of ML Models

Machine learning models offer several advantages over traditional models, particularly in their ability to dynamically update and integrate new risk factors. For instance, studies have demonstrated that ML models improve cancer risk prediction [12] by incorporating non-linear relationships between variables. The literature also highlights the importance of feature selection techniques like forward selection and correlation analysis, which allow ML models to automatically identify the most relevant predictors of breast cancer risk for a specific population.

E. The Need for Population-Specific Models

The literature supports the argument that breast cancer risk prediction models must be population-specific to be effective [6]. The use of generalized models leads to either overestimation or underestimation of risk, which can result in inappropriate medical advice or unnecessary tests. As seen in Cuban women, traditional models that do not account for local risk factors related to socioeconomic status, diet, and healthcare accessibility perform poorly.

F. Conclusion

While traditional models have served as a starting point for breast cancer risk prediction, the literature reveals a growing need for more adaptive and localized models, particularly in underserved populations. Machine learning offers a path forward by addressing the limitations of traditional models, allowing for the incorporation of a wider range of variables that reflect local demographics and risk factors.

IV. Methodology

The problem being investigated is the inaccuracy of traditional breast cancer risk prediction models when applied to Cuban women due to a lack of consideration for local risk factors. This study seeks to improve early detection by developing a machine learning (ML) based risk prediction model tailored to Cuban women. The significance lies in providing an affordable and accurate tool for healthcare systems in resource-limited settings.

The research adopts a binary classification approach using machine learning to model breast cancer risk. Data is preprocessed, wrangled, and used to train models like Random Forest, Decision Trees, and Support Vector Machines.

The ML evaluation framework uses performance metrics such as accuracy, Area Under the Curve (AUC), and interpretability, which allow for comprehensive evaluation of the model's effectiveness and usability in clinical settings.

Detailed analysis and interpretation of results is undertaken to ensure a robust performance evaluation. Further more, LIME and SHAP are used to understand the results more effectively.

A. Data Description

The dataset used in this study consists of 1,697 breast cancer cases from Cuban women, gathered from the Hospital Universitario Clínico-Quirúrgico Comandante Manuel Fajardo between 2016 and 2019. It includes both quantitative features (e.g., age, BMI, number of pregnancies) and qualitative features (e.g., family history of cancer, smoking status, breastfeeding duration).

1) Real-life Problem Addressed: The dataset is crucial for developing a localized breast cancer risk prediction model that can provide early detection tools for Cuban women. Traditional models often fail due to cultural and lifestyle differences; this dataset offers the specific variables necessary to create a more accurate, population-tailored model.

2) Feature Selection Considerations: Several factors were considered before selecting this dataset for the ML project:

- Relevance to Local Population: It includes risk factors that are specific to Cuban women, which helps in improving model accuracy and relevance.
- Data Availability: The dataset provides comprehensive demographic and medical data, enabling the construction of a robust predictive model.
- Quality and Consistency: The data collected from a single hospital ensures uniformity, reducing variability that could arise from multi-center data.

3) Important Features: Key features in the dataset that the AI model uses include:

1) **Age:** A critical factor for breast cancer risk prediction.
2) **Weight:** Patients weight at screening. A known risk factor for cancer.
3) **Family History of Breast Cancer (nrelbc):** First-degree relatives with breast cancer increase an individual's risk.
4) **Breastfeeding Duration:** This feature captures the influence of reproductive health.
5) **Tobacco Use and Alcohol Consumption:** Lifestyle choices that have been linked to increased cancer risk.
6) **Number of Biopsies:** Past biopsies may indicate prior issues with abnormal cell growth, affecting future risk.
7) **Age of menarche (menarche)**
8) **Age of menopause (menopause)**
9) **Age at first successful delivery (agefirst)**
10) **Number of children born alive (children)**
11) **Hyperplasia:** An increase in the number of cells in an organ or tissue
12) **Race of the patient:** White, Black and Mixed.
13) **Weekly physical activity (exercise):** Weekly physical activity - Quantitative (0 for no exercise at all - 7 for daily exercise)
14) **Invasive micropapillary carcinoma (imc):**

By leveraging these features, the ML model can predict the risk of breast cancer, leading to better early diagnosis and timely interventions.

B. Data Preparation and Exploratory Data Analysis

The data was analysed and no duplicates were found.

- **Total Cases:** 1,697
- **Total Positives:** (Those who had cancer) 1,160
- **Total Negatives:** (Those who did not have cancer) 537

1) Outliers: Random forests are robust to outliers as they split data based on ranks rather than absolute values. Outliers have less impact on the model's overall performance since each decision tree considers only a subset of features and instances.

2) Missing values: Bellow were the missing values as seen in figure 1

- biopsies: 1
- year: 537
- imc: 7
- weight: 10
- allergies: 276
- histologicalclass: 537
- birads: 537

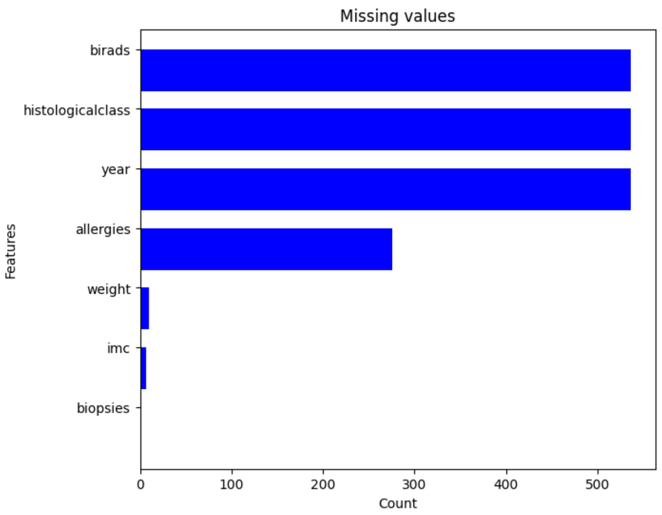

Fig. 1. Missing values.

We have 537 cases where patients didn't have cancer. Which corresponds with the value for year of diagnosis, histological class and birads. These columns therefore won't be needed in our data analysis because it is clear that they are present with a dianosis and absent without it.

Id is also not needed in our predictions, it is just used as a counter.

Emotional, Allergies and Depressive is also irrelevant in our analysis.

This leaves us with only; biopsies, imc, and weight.

Only one biopsie is missing, we can fill it with the 50% percentile value.

Only 7 imc are missing, we can fill it with the mean.

Only 10 weights are missing, we can fill it with the mean.

3) Label Encoding: All data must be numeric.

Variables with a binary choice of yes and no should be encoded with 1 or 0 respectively.

- cancer mapping: 'No': 0, 'Yes': 1
- hyperplasia mapping: 'No': 0, 'Yes': 1
- alcohol mapping: 'No': 0, 'Yes': 1
- tobacco mapping: 'No': 0, 'Yes': 1
- race mapping: 'Black': 0, 'Mixed': 1, 'White': 2

Cleaning menopause: Menopause is represented in months, which is usually numeric but where a patient has not reached menopause, a 'no' is inserted. Therefore, where it says no replace with 0.

Cleaning exercise: Exercise is represented by the number of days a patient exercises per week. If they do not exercise, a 'no' is inserted. Therefore, where it says no replace with 0. This works for us as exercise is the number of days in a week, so 0 is as good as No. For a patient that exercises daily, the word 'Diary' is used. Therefore, where it says Diary replace with 7. This means that someone exercises all days of the week.

Cleaning children, where it says 5+ replace with 6. We have no way of knowing if a patient has more than 6 children. This can skew the data, but on analysis it was found that only 2 patients had this entry which was fine.

Cleaning agefirst which is the age a patient first had a child at. Where a patient has not given birth, 'no' was inserted. Therefore, where it says no replace with 0.This can be missleading as the patient does not have a child at 0, but it can be treated as a special case.

Cleaning breastfeeding, where a patient has never breast fed 'no' was inserted. Therefore, where it says No replace with 0. It is measured in months and 0 months should work fine.

We also, strip out the months part so that we remain with only numeric values.

Cleaning nrelbc (Number of first-degree relatives with breast cancer). We can see that the data in this column is split by a forward slash:

- Mother/Sister
- Mother
- No

therefore each time we encounter this we count another family member. Where the patient had no first-degree relative with cancer, a 'no' was inserted. Therefore, where it says no replace with 0.

All object datatypes were also converted to numerics.

C. Anaylsis

At this point all our data is clean and ready for analysis, the first analysis we carried out was generate a heat map as seen in figure 2

There is a significant correlation between biopsies and cancer. We also see strong correlation between menopause and age, which is to be expected and does not influence our model.

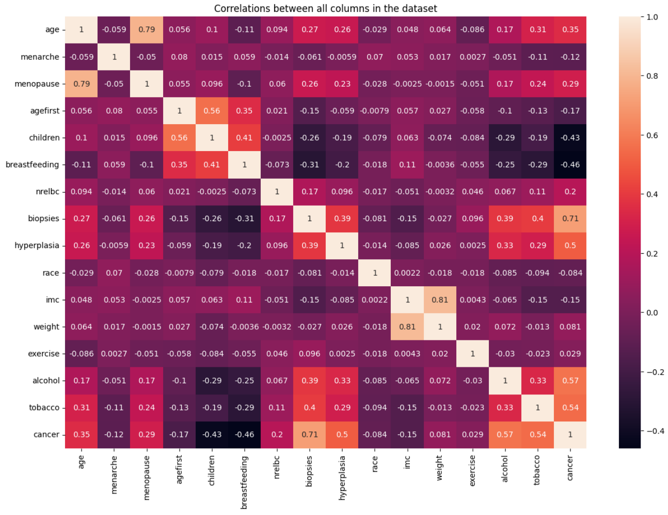

Fig. 2. Heat Map.

D. ML model selection and optimization.

In this paper, three models were used to analyse the data; Decision Trees, Support Vector Machine and Random Forest.

1) Decision Trees: A Decision Tree is a supervised learning model that splits data into branches to make decisions based on feature values. It's highly interpretable and easy to visualize.

Parameters:

- Criterion: Determines the function to measure the quality of a split (e.g., Gini Impurity or Entropy). We used Gini Impurity

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

- Max Depth: Limits the maximum depth of the tree to prevent overfitting.
- Random State: Controls the randomness of the estimator.
- Min Samples Split: The minimum number of samples required to split an internal node.
- Min Samples Leaf: The minimum number of samples required to be at a leaf node.

Hyperparameters:

- Max Depth, 10 was the optimal value.
- Min Samples Split, 2 was the optimal value.
- Min Samples Leaf, 2 was the optimal value.

2) Support Vector Machine: Support Vector Machine (SVM) is a supervised learning algorithm that finds the hyperplane that best separates data into different classes. **Parameters:**

- C (Regularization Parameter): Controls the trade-off between a smooth decision boundary and classifying training points correctly..
- Kernel: The function used to map input data into a higher-dimensional space (e.g., linear, polynomial, RBF)..
- Gamma: Controls the influence of a single training example on the decision boundary.

Hyperparameters:

- C (Regularization Parameter), best fit was 0.1.
- Kernel, best fit was linear.
- Gamma, best fit was 0.1.

3) Random Forest: Random Forest is an ensemble method that builds multiple decision trees and aggregates their results to improve accuracy and reduce overfitting.

Parameters:

- n_estimators: The number of trees in the forest.
- Criterion: Determines the function to measure the quality of a split (e.g., Gini Impurity or Entropy). We used Gini Impurity

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

- Random State: Controls the randomness of the estimator. This was set at 42
- Max Depth: Limits the maximum depth of the tree to prevent overfitting.

Hyperparameters:

- n_estimators, best fit was 69.
- Max Depth, best fit was 13.

Each model was tuned using cross-validation to ensure the best-performing hyperparameters and to avoid overfitting, ensuring the models generalize well to new, unseen data.

E. Machine Learning model selection Accountability.

This is the concept that a machine learning model and its output can be explained in a way that "makes sense" to a human being at an acceptable level [10] .

We are going to try two popular methods used for explainability; Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to explain our chosen model.

1) LIME: An attempt was made to use LIME to show the sample effect of features at 5 instances. LIME however wasn't the best at explaining our results. It was observed that each change in our values drastically affected the LIME effect of each feature.

LIME relies on sampling and random perturbations, leading to inconsistent results across runs [1]. Explanations for the same input can vary significantly.

We therefore move to using SHAP.

2) SHAP: We are going to focus more on SHAP since our data set is small enough. SHAP gives us more insight than LIME, as LIME is more localized.

From figure 3 we observe that $E[f(x)] = 0.322$ gives the average predicted probability of having cancer. $f(x) = 1$ is the predicted probability for this particular entry.

The SHAP values are all the values in between. For example, the biopsies has increased the predicted probability by 0.4, breast feeding by 0.1 etc

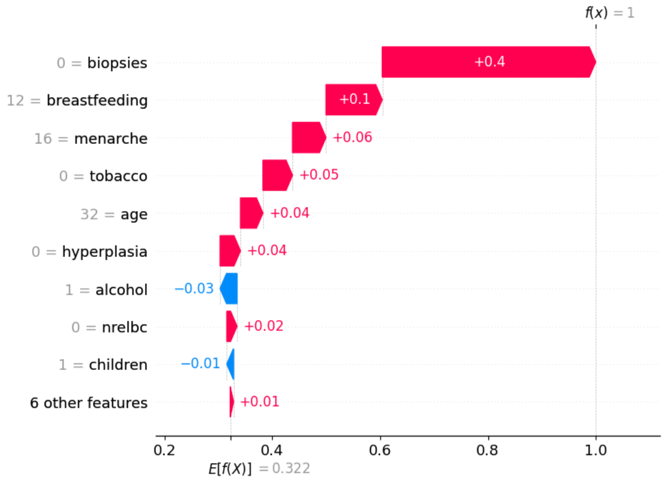

Fig. 3. SHAP Waterfall.

For each feature, we calculate the mean SHAP value across all observations. Specifically, we take the mean of the absolute values as we do not want positive and negative values to offset each other. In the end, we have the bar plot below. There is one bar for each feature. For example, we can see that biopsies had the largest mean SHAP value.

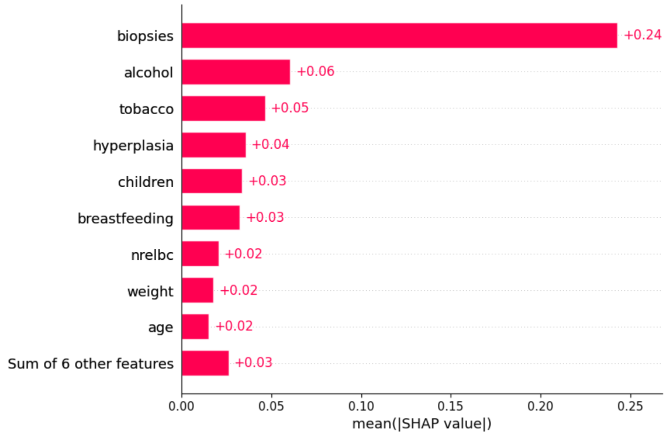

Fig. 4. Mean SHAP values.

V. Results and Discussion

A. Evaluation metrics

Analysis of the results obtained from the machine learning algorithm was done using the bellow specified evaluation metrics. These values were rounded off to 4 decimal places. Random Forest had the highest accuracy of all the three models we run, this is therefore the model we shall use for our analysis. Figure 5, shows a sample of random forests.

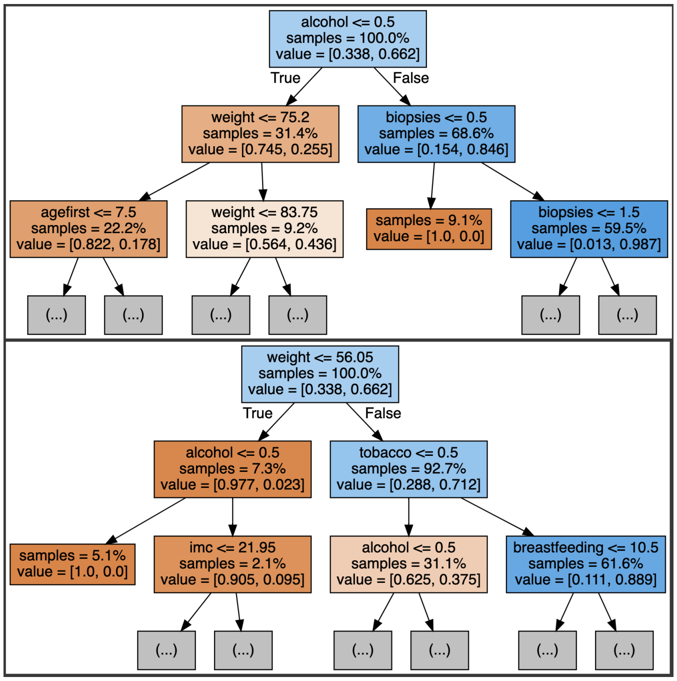

Fig. 5. Sample Random Forest.

Metrics:

- Accuracy of Random Forest: 99.0196%
- Precision of Random Forest: 99.0333%
- Recall of Random Forest: 99.0196%
- F1-Score of Random Forest: 99.0152%

Analyzing the ROC curve for random forest showed that the AUC was 1 as shown in figure 6

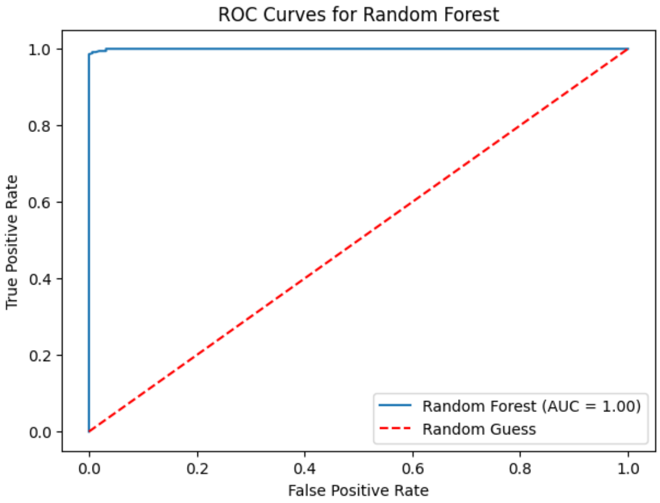

Fig. 6. ROC Curves for Random Forest Model.

Analyzing the confusion matrix for random forest showed that we had zero false negatives as seen in figure 7. A false negative would be dangerous because it would mean that our model predicted that someone did not have breast cancer when they actually have it, therefore this is a good thing. We have 5 false positives, this is acceptable because doctors can carry out further investigations to find the information they want.

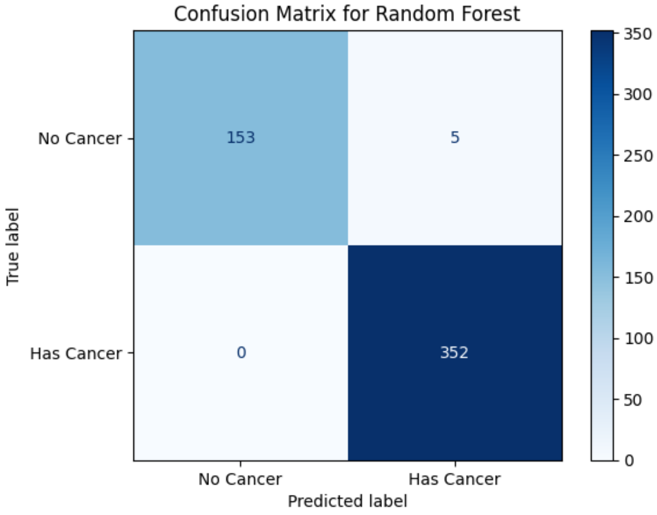

Fig. 7. Confusion Matrix for Random Forest Model.

From figure 8, we observe that biopsies were the most important feature by a huge margin. However, this doesn't mean that other factors weren't important. In fact when biopsies are eliminated, the rest of the features are rather relative to each other.

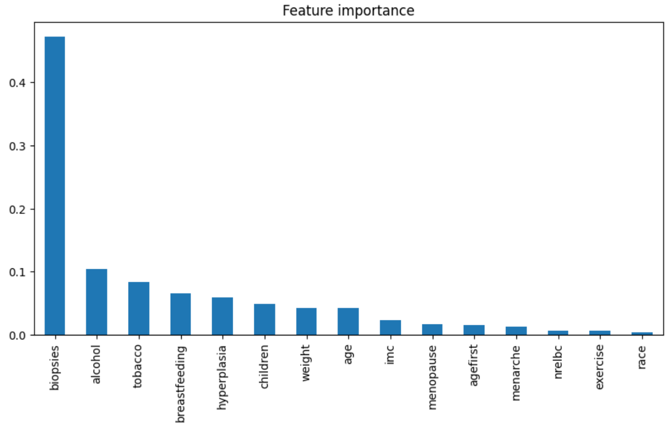

Fig. 8. Feature Importance from Random Forest Model.

VI. Conclusion and Future Work

This study introduced a machine learning-based model to address the shortcomings of traditional breast cancer risk prediction methods for Cuban women. By leveraging locally relevant risk factors such as BMI, breastfeeding history, family history, and lifestyle habits, the model tailored its predictions to the specific demographic and cultural needs of the population. The Random Forest algorithm emerged as the most effective, achieving an accuracy of 99% and an AUC of 1. These results demonstrate the potential of ML-driven approaches to significantly enhance early detection efforts, reduce mortality, and support healthcare systems in resource-limited settings. Furthermore, the integration of interpretability as an evaluation criterion ensures that the model's outputs are accessible and actionable for healthcare providers.

The findings highlight the importance of population-specific solutions and validate the use of machine learning to address pressing public health challenges. This model not only improves accuracy but also offers a cost-effective and scalable alternative to traditional methods, making it applicable to other Hispanic and underserved populations.

To build on the results of this study, future work will focus on:

1) **Broader Data Collection:** Expanding the dataset to include data from additional hospitals and regions world wide to enhance the model's generalizability and robustness.
2) **Feature Integration:** Incorporating additional risk factors such as breast density, occupational exposure, and detailed dietary habits to further refine predictions.
3) **ML Optimization:** Investigating more advanced AI techniques, such as ensemble learning and neural networks, for further performance gains while maintaining interpretability.

This work lays the foundation for a scalable, data-driven solution that has the potential to transform breast cancer detection and prevention in underserved regions.

VII. Dataset and Python Source Code

For the python source code used in this project see Link to Notebook

For the dataset used in this project see Link to Dataset

For this projectś presentation see Link to PPT slides

References

[1] Shamim Ahmed, M Shamim Kaiser, Mohammad Shahadat Hossain, and Karl Andersson. A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions. *IEEE Access*, 2024.

[2] Richard Allman, Yi Mu, Gillian S Dite, Erika Spaeth, John L Hopper, and Bernard A Rosner. Validation of a breast cancer risk prediction model based on the key risk factors: family history, mammographic density and polygenic risk. *Breast Cancer Research and Treatment*, 198(2):335–347, 2023.

[3] Boon Hong Ang, Weang Kee Ho, Eldarina Wijaya, Pui Yoke Kwan, Pei Sze Ng, Sook Yee Yoon, Siti Norhidayu Hasan, Joanna MC Lim, Tiara Hassan, Mei-Chee Tai, et al. Predicting the likelihood of carrying a brca1 or brca2 mutation in asian patients with breast cancer. *Journal of Clinical Oncology*, 40(14):1542–1551, 2022.

[4] William E Barlow, Emily White, Rachel Ballard-Barbash, Pamela M Vacek, Linda Titus-Ernstoff, Patricia A Carney, Jeffrey A Tice, Diana SM Buist, Berta M Geller, Robert Rosenberg, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, 98(17):1204–1214, 2006.

[5] Josepmilly del Valle Peña Colmenares, Carmen Cristina García, Yazmin José Velásquez Velásquez, Leider Arelis Campos Pino, Álvaro Gómez Rodríguez, Wladimir José Villegas Rodríguez, David José González Vargas, and Douglas José Angulo Herrera. Is using the gail model to calculate the risk of breast cancer in the venezuelan population justified? *ecancermedicalscience*, 17, 2023.

[6] Run Fan, Yufan Chen, Sarah Nechuta, Hui Cai, Kai Gu, Liang Shi, Pingping Bao, Yu Shyr, Xiao-Ou Shu, and Fei Ye. Prediction models for breast cancer prognosis among asian women. *Cancer*, 127(11):1758–1769, 2021.

[7] Laura Fejerman, Amelie G Ramirez, Anna María Nápoles, Scarlett Lin Gomez, and Mariana C Stern. Cancer epidemiology in hispanic populations: what have we learned and where do we need to make progress? *Cancer Epidemiology, Biomarkers & Prevention*, 31(5):932–941, 2022.

[8] Chi Gao, Eric C Polley, Steven N Hart, Hongyan Huang, Chunling Hu, Rohan Gnanaolivu, Jenna Lilyquist, Nicholas J Boddicker, Jie Na, Christine B Ambrosone, et al. Risk of breast cancer among carriers of pathogenic variants in breast cancer predisposition genes varies by polygenic risk score. *Journal of Clinical Oncology*, 39(23):2564–2573, 2021.

[9] Jiaxin Li, Zijun Zhou, Jianyu Dong, Ying Fu, Yuan Li, Ze Luan, and Xin Peng. Predicting breast cancer 5-year survival using machine learning: A systematic review. *PloS one*, 16(4):e0250370, 2021.

[10] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[11] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pages 4593–4603, 2022.

[12] Reza Rabiei, Seyed Mohammad Ayyoubzadeh, Solmaz Sohrabei, Marzieh Esmaeili, and Alireza Atashi. Prediction of breast cancer using machine learning approaches. *Journal of biomedical physics & engineering*, 12(3):297, 2022.

[13] Donna Spiegelman, Graham A Colditz, David Hunter, and Ellen Hertzmark. Validation of the gail et al. model for predicting individual breast cancer risk. *JNCI: Journal of the National Cancer Institute*, 86(8):600–607, 1994.

[14] Kathryn P Trayes and Sarah EH Cokenakes. Breast cancer treatment. *American family physician*, 104(2):171–178, 2021.

[15] José Manuel Valencia-Moreno, Everardo Gutiérrez López, José Felipe Ramírez Pérez, Juan Pedro Febles Rodríguez, and Omar Álvarez Xochihua. Exploring breast cancer prediction for cuban women. In *Information Technology and Systems: Proceedings of ICITS 2020*, pages 480–489. Springer, 2020.

[16] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for lime: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022.

[17] Zhi-Hua Zhou. *Machine learning*. Springer nature, 2021.