

**SEMESTER ONE 2024/2025 ACADEMIC YEAR 1**

**SCHOOL COMPUTING AND IMFORMATICS TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE**

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

**MCS 7103**

**MACHINE LEARNING**

**ASSIGNMENT ONE**

**Eugene Munyaneza**

**2024/HD05/21936U**

**2400721936**

Table of Contents

[Breast cancer risk factors in Cuban women 3](#_Toc180680269)

[Introduction 3](#_Toc180680270)

[Background of the Project 3](#_Toc180680271)

[Key terms include: 3](#_Toc180680272)

[Literature Review 3](#_Toc180680273)

[Traditional Models: 3](#_Toc180680274)

[Machine Learning in Breast Cancer Prediction: 4](#_Toc180680275)

[Gaps in the Literature: 4](#_Toc180680276)

[Contribution of ML Models: 4](#_Toc180680277)

[The Need for Population-Specific Models: 4](#_Toc180680278)

[Conclusion: 5](#_Toc180680279)

[Gaps and Significance of Work 5](#_Toc180680280)

[Project Scope 5](#_Toc180680281)

[Problem Statement 5](#_Toc180680282)

[Research Questions 5](#_Toc180680283)

[Methodology 5](#_Toc180680284)

[Methodology Breakdown 6](#_Toc180680285)

[Results Presentation 6](#_Toc180680286)

[Conclusion and Future Work 6](#_Toc180680287)

[References 7](#_Toc180680288)

# Breast cancer risk factors in Cuban women

## Introduction

This report presents an analysis of breast cancer risk factors in Cuban women, employing machine learning algorithms to improve early detection and treatment planning. The authors, including José Manuel Valencia-Moreno and colleagues from multiple institutions in Mexico, Cuba, and Russia, focus on personalized risk assessment models to address limitations in traditional prediction tools (Valencia-Moreno, 2024).

## Background of the Project

Breast cancer remains the leading cause of cancer mortality among women globally, with particularly high mortality rates in developing nations like Cuba due to late detection and limited access to healthcare services (Pinheiro, 2017). Traditional models such as the Gail model (Gail, 2000) have limitations in specific populations, often underestimating or overestimating risk for non-Caucasian groups. The need for an adjusted breast cancer prediction model tailored to Cuban women influenced this research. The output is a machine learning-based risk estimation model, designed to predict individual breast cancer risks using factors specific to this population.

### Key terms include:

1. Machine Learning: Algorithms used for predicting breast cancer risk.
2. Risk Factors: Demographic and health-related variables influencing cancer risk (age, biopsies, obesity, etc.).
3. AUC (Area Under Curve): A metric to evaluate model performance.

Expected outputs involve a model that improves the accuracy of predicting breast cancer risk among Cuban women, particularly through the successful application of algorithms like Random Forests and Decision Trees.

## Literature Review

Research on breast cancer prediction models has evolved significantly over the years, with various models being developed to estimate individual risk. One of the most widely known models is the Gail model (Gail, 2000), 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 flawed performance due to cultural and socioeconomic differences.

### Traditional Models:

* **Gail Model** (Gail, 2000): Initially developed for US-born white women, 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 (Spiegelman, 1994).
* **Barlow Model** (Barlow, 2006): 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.

### Machine Learning in Breast Cancer Prediction:

Recent advancements in machine learning (ML) have led to the development of more adaptive and personalized models for cancer risk assessment. Studies such as Parmigiani et al. (Parmigiani, Berry, & Aguilar, 1998) and Jonker et al. (Van, Jacobi, & Bock, 2005) 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. (2020) and Rajendran et al. (2020), 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.

### 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 Banegas et al. (2017) on US Hispanics demonstrated that traditional models underestimated risks for women born outside the US.

### 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 by 15-25% 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.

### The Need for Population-Specific Models:

The literature supports the argument that breast cancer risk prediction models must be population-specific to be effective. 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.

### 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.

## Gaps and Significance of Work

The research addresses gaps in the application of breast cancer prediction models to the Cuban population, where current global models like Gail and IBIS either underperform or overlook important local risk factors such as lifestyle, diet, and socioeconomic status. This model is significant for Cuban women’s health due to its ability to tailor prediction models for more accurate early detection and preventative measures.

## Project Scope

The project aims to develop a predictive model using machine learning algorithms to estimate breast cancer risk in Cuban women, incorporating local data. This work contributes to health informatics by providing a practical tool that enhances early detection and can be generalized to other Hispanic populations.

## Problem Statement

While existing breast cancer prediction models are well-calibrated for US populations, they fail to adequately estimate risks in Cuban women due to differing risk factors. This project seeks to develop a machine learning model specifically for the Cuban population, ensuring a more accurate and practical tool for clinicians.

## Research Questions

1. What machine learning algorithms are most effective in predicting breast cancer risk in Cuban women?
2. How do local risk factors, not accounted for in traditional models, impact breast cancer prediction accuracy?
3. Can this model be generalized to other Hispanic populations?

## Methodology

The research adopts a binary classification approach using machine learning to model breast cancer risk. Data is pre-processed, balanced, and used to train models like Random Forest, Decision Tree, and Support Vector Machines. Stratified sampling ensures data representativeness, and cross-validation ensures robust performance evaluation.

### Methodology Breakdown

1. Data Collection: The dataset comprises 1697 cases of Cuban women aged 20–90, with 23 risk factors.
2. EDA (Exploratory Data Analysis): Data cleaning, normalization, and variable encoding to manage missing values and outliers.
3. Data Engineering: Feature selection techniques such as forward selection and correlation analysis were used.
4. Modeling: Algorithms trained on a training-validation split (60%-40%) using metrics like Area Under Curve (AUC) of a Receiver-operation characteristic (ROC) curve and accuracy to measure performance.

### Results Presentation

EDA: Found significant predictors like age, BMI, and familial cancer history.

Modeling: Random Forest outperformed other models, with an AUC of 0.990 and training accuracy of 0.99. Other algorithms like Decision Trees showed strong but slightly lower performance.

## Conclusion and Future Work

This model demonstrates high accuracy in predicting breast cancer risk for Cuban women, outperforming traditional models. Future work involves refining the model by integrating additional risk factors like lifestyle habits and external validations across broader datasets. Expansion to mobile applications for practical use in clinical settings is also recommended.

# References

Barlow, W. E.-B.-E. (2006). Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, 1204-1214.

Gail, M. H. (2000). Gail model and breast cancer. *The Lancet*, 1017.

Parmigiani, G., Berry, D., & Aguilar, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *American journal of human genetics*, 145–158.

Pinheiro, P. S. (2017). Cancer mortality in Hispanic ethnic groups. *Cancer epidemiology, biomarkers and prevention*, 376-382.

Spiegelman, D. a. (1994). Validation of the Gail et al. model for predicting individual breast cancer risk. *JNCI: Journal of the National Cancer Institute*, 600-607.

Valencia-Moreno, J. M.-F.-L.-S.-R. (2024). Breast cancer risk estimation with intelligent algorithms and risk factors for Cuban women. *Computers in Biology and Medicine*, 108818.

Van, A., Jacobi, C. a., & Bock, C. a. (2005). Remarks on the BOADICEA model of genetic susceptibility to breast and ovarian cancer. *British journal of cancer*, 1336-1337.