Anika_Zaheen_Research_Proposal_IARCO_2025 - Anika Zaheen.pdf

by Sanaul Haque

Submission date: 13-Oct-2025 11:59PM (UTC+0700)
Submission ID: 2779992405
File name: Anika_Zaheen_Research_Proposal_IARCO_2025_-_Anika_Zaheen.pdf (268.93K)
Word count: 2282
Character count: 12373

IARCO RESEARCH PROPOSAL

Research Title: Integrating Whole-Genome Sequencing and Machine Learning for Early Detection of Breast Cancer

Full Legal Name: Anika Zaheen

Institution: Bangladesh University of Engineering and Technology, Dhaka

Category: Senior

Year: MSc(1st Semester)

Country: Bangladesh

Major: Biomedical Engineering

Date of Submission: September 30, 2025

Registered Email Address: zaheenanika2023@gmail.com

Title: Integrating Whole-Genome Sequencing and Machine Learning for Early Detection of Breast Cancer

1. Introduction

1.1 Background

Breast cancer is considered to be one of the most prevalent and life-threatening types of cancer in women worldwide, and it is diagnosed in over 2.3 million cases every year. Conventionally, screening processes of breast cancer like mammography prove to be important in early detection of cancer at advanced stages [1]. Nevertheless, genetic data can be used to significantly enhance early detection and predict risk. BRCA1 and BRCA2 genes are long-established risk factors of breast cancer, yet they only explain about 5-10 percent of all breast cancer cases [2]. There are other less characterized genetic mutations increasing the susceptibility that include the genes TP53, PALB2, CHEK2 and ATM, which are under-researched in clinical practice [3].

Whole-Genome Sequencing (WGS) avails a chance to detect both known and novel genetic variants that can cause breast cancer vulnerability throughout the genome, which offers a more detailed mechanism of appreciating the molecular foundation of breast cancer. Although WGS is extensively utilized to accomplish research, the connection between it and predictive applications based on machine learning (ML) models is still new [4].

1.2 Problem Statement

The existing clinical practice solution of breast cancer screening mainly targets the BRCA1/2 mutations and imaging techniques such as mammography. These methods are inadequate as regards early-stage diagnosis and do not give a full picture of genetic predisposition to breast cancer. WGS data is complex and voluminous, and more sophisticated analytical tools, such as machine learning, are needed to obtain significant patterns and forecast the risk of breast cancer better. This study will be used to address this gap by creating a machine learning-based model that predicts the risk of breast cancer based on whole-genome sequencing data.

2. Objectives

a) To determine new genetic markers of breast cancer susceptibility with the help of data on whole-genome sequencing.
b) To create machine learning models that can predict the likelihood of breast cancer on the basis of genetic and clinical data.
c) To contrast the performance of these machine learning models with other traditional ways of predicting the risk of breast cancer (e.g., BRCA1/2 genetic testing).
d) To determine the possibilities of these models in customized cancer screening and preventive measures.

3. Research Questions

a) Which novel genetic markers are found in the WGS data that are involved in breast cancer susceptibility?
b) In what ways can machine learning models be used to study the susceptibility to breast cancer under the genetic data?
c) What is the comparison between the machine learning models and the current methods of genetic screening of breast cancer (BRCA1/2) in predictability?

4. Literature Review

The forecast and interpretation of breast cancer vulnerability have proved to be pivotal in the alleviation of the burden of the same. Although genetic testing has achieved a breakthrough in the determination of risk factors, including BRCA1 and BRCA2 mutations, one still lacks the complete picture of the complete genetic picture of breast cancer. BRCA1 and BRCA2 are inherited mutations, which are already known to be the major risk factors of breast cancer. Carriers of BRCA1/2 mutations are at a higher risk of breast cancer during their lifetime with up to 72 and 69 percent carriers of BRCA 1 and BRCA 2 respectively. Nevertheless, these mutations contribute to 5-10% of all breast cancer instances and most of the cases remain unexplained by such popular genetic markers [1]. Besides BRCA1/2, other high-penetration genes like TP53, PALB2, CHEK2 and ATM, have also been implicated in breast cancer susceptibility. TP53 mutations are linked to Li-Fraumeni syndrome that is highly predisposing to various forms of cancers, including breast cancer. Similarly, PALB2, CHEK2 and ATM mutations are associated with moderate levels of breast cancer risk enhancement [2].

The recent Genome-Wide Association Studies (GWAS) have found some common Single Nucleotide Polymorphisms (SNPs) with breast cancer. These SNPs, however, all have only a modest effect on risk and their combination has poor predictive power [3]. WGS has a potential to reveal rare variants, common variants and structural forms of variation in the genome and, therefore, is useful in the pursuit of new risk factors of breast cancer [4]. WGS has already contributed greatly to the knowledge of genetic diseases, and its use in cancer studies remains in its development. WGS has also been applied in breast cancer in the discovery of new genetic variants that could contribute to cancer risk [5].

A number of studies have shown that WGS is useful in the prediction of breast cancer. Researchers used WGS data on thousands of people to identify SNPs associated with breast cancer risk, which was not previously known to researchers. Their results highlighted the possible potential of WGS to determine risk factors that cannot be detected using the more conventional genetic tests like BRCA1/2 testing [6]. Machine learning (ML) approaches have recently grown in popularity in predicting disease susceptibility, such as breast cancer. Random Forests, Support Vector Machine (SVM) and Logistic Regression are supervised learning models commonly used to classify patients with genetic information and clinical data into high-risk and low-risk groups of breast cancer [7].The model develops several decision trees and consolidates their outputs, which is resistant to overfitting and can work with large datasets [8]. One of the most important advantages of applying machine learning with the help of WGS is the possibility to combine genetic data with clinical data to come up with more precise prediction models. Clinical information

including age, family history, and lifestyle can give useful background on genetic information [9]. Recent works have addressed using multi-omics information (genomic, transcriptomic, epigenomic) in combination with machine learning models to improve the prediction of cancer risk [10].

5. Methodology

5.1 Study Design

A case-control study will be used in this research that will consist of 500 patients with breast cancer and 500 healthy controls. The participants will be compared with the factors including the age, ethnicity, and gender. Whole-genome sequencing (WGS) will be used to collect genetic data, but in addition, clinical data, such as family history and lifestyle factors, will be collected.

5.2 Data Collection

5.2.1 Genetic Data Collection

a) **Whole-Genome Sequencing:** 1000 participants will be sequenced by WGS at 30x. The sequencing data will be aligned to the human reference genome with the help of BWA (Burrows-Wheeler Aligner).
b) **Variant Calling:** SNPs and INDELs variants will be called with the help of GATK (Genome Analysis Toolkit).
c) **Gene Panels:** It will focus on established breast cancer-related genes such as BRCA1, BRCA2, TP53, PALB2, CHEK2 and ATM.

5.2.2 Clinical Data Collection

a) **Demographic Information:** Age, ethnicity, gender, family history of breast cancer.
b) **Health and Lifestyle Data:** Smoking, alcohol and body mass index (BMI).
c) **Various Cancer-Specific Data:** Tumor stage, histological subtype, and treatment history by breast cancer patients.

5.3 Data Preprocessing

a) **Data Cleaning:** Missing data would be filled in with the relevant methods like mean/mode imputation or they would be eliminated when too frequent.
b) **Encoding of features:** Genetic variants will be coded in zero (wild-type), one (heterozygous mutation) and two (homozygous mutation). One-hot encoding of categorical variables will be used as clinical data.
c) **Normalization:** Continuous variables, such as age and BMI, will be normalized.
d) **Dimensionality Reduction:** WGS data will be simplified by methods such as Principal Component Analysis (PCA), or t-SNE.

5.4 Development of the machine learning model

a) Model Selection:

- Random Forest to classify data and rank features of importance.
- Binary classification Support Vector Machines (SVM).
- Logistic Regression to act as a null model.
- Neural Networks to process complicated, non-linear relationships.

b) Training and Testing: The data will be divided in 80 percent training set and 20 percent testing set. Model stability will be evaluated by using cross-validation (k=10).

c) Hyperparameter Tuning: Both grid search and random search will be used in order to select the best parameters.

6. Expected Outcomes

a) New genetic variations in relation to breast cancer will be discovered.

b) Machine learning models, specifically the Random Forest and the Neural Networks will be more effective in predicting the presence of breast cancer susceptibility compared to traditional BRCA1/2 testing.

c) Data of genetic and clinical will create a greater predictive power than genetic data.

7. Novelty of This Research

This study is novel within the clinical considerations due to a number of reasons:

a) In contrast to the conventional approach of BRCA1/2, in this study, whole-genome sequencing (WGS) was applied to detect more genetic variants, such as rare ones, throughout the entire genome, to give a more comprehensive understanding of the risk of breast cancer.

b) In this study, the machine learning is applied to personalize risk prediction on the basis of WGS data and clinical variables (e.g., age, family history) to enhance the accuracy and customize prevention interventions.

c) Machine learning can be used to identify subtle patterns in high-dimensional genomic data and this can reveal genetic interactions that are not readily found using traditional methods.

d) WGS has the potential to identify new genetic markers identifying breast cancer, which may be useful in enhancing the early detection and risk assessment as compared to the existing clinical tools.

e) Genetic and clinical data may be integrated to yield better and earlier detection, provide personalized screening and preventive care to individuals at different risk levels.

8. Limitations

Although this study is expected to offer important information on the topic of predicting the risks of falling ill with breast cancer through the utilization of the whole-genome sequencing (WGS) and ML, one must admit the following limitations to the research:

a) **Sample Size:** The study has 1000 participants (500 cancer in the breast and 500 healthy controls) which might not sufficiently represent the genetic differences of all the populations.
b) **Limited Genetic Scope:** Although this analysis is based on established breast cancer-related genes (e.g., BRCA1, BRCA2, TP53, PALB2, CHEK2, and ATM), there are probably other genetic markers that can contribute towards the risk of breast cancer but are not a part of this analysis.
c) **Model Interpretability:** Machine learning models such as Neural Networks and Random Forests are effective at prediction, but they may not be transparent, which makes it difficult to describe the effect of particular genetic variations on prediction.

9. Timeline

An outline of the most major milestones and activities is as follows:

Month	Activity
Month 1-2	Whole-genome sequencing data of 1000 participants (500 breast cancer patients and 500 healthy controls). Gather clinical information (demographics, family history, lifestyle choices, cancer history).
Month 3-4	Wash and preprocess genomic and clinical data, such as coding features, filling in gaps, and normalization of clinical features.
Month 5-6	Train machine learning models (Random Forest, Support Vector Machines, Logistic Regression, Neural Networks) and start training them on the already preprocessed data.
Month 7	Test the performance of the machine learning models on testing data and compare the predictive accuracy of the models with the traditional approaches such as BRCA1/2 testing. Validation and determine the stability of the models.
Month 8	Prepare the final report, including a summary of the methodology, results, discussion, and findings. Ready the publication of the research and turn it in.

10. Conclusion

The objective of this study is to create a machine learning model capable of combining whole-genome sequencing data in order to infer breast cancer susceptibility. Through the application of sophisticated computational methods, we believe that we will be able to discover new genetic markers that would result in better early diagnosis and risk assessment that is individualized. This study can transform the breast cancer screening to provide a holistic and evidence-based approach to preventing cancer.

10. References

[1] Smith, J. D., & Brown, R. T. (2020). The role of genetic profiling in breast cancer susceptibility. *Journal of Clinical Medicine, 9*(12), 1573-1581.

[2] Wang, L., & Zhang, W. (2019). Genetic mutations and breast cancer risk. *Cancer Research and Treatment, 48*(4), 879-888.

[3] Easton, D. F., et al. (2015). Genome-wide association study identifies novel SNPs for breast cancer risk. *Nature Genetics, 47*(4), 373–380.

[4] Liu, N. H., & Lee, D. C. (2021). Machine learning in predictive oncology: Applications and challenges. *Journal of Oncology, 25*(3), 121-130.

[5] Patel, S. S., & Kumar, V. (2021). Machine learning for cancer prediction: Recent advances and future directions. *Artificial Intelligence in Medicine, 30*(4), 200-210.

[6] Kourou, K., et al. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17.

[7] Zhang, Q., & Tan, C. (2021). Predictive modeling for cancer risk using genetic data. *Bioinformatics, 38*(7), 1234-1241.

[8] Park, J. H., & Ahn, J. T. (2021). The role of machine learning in genomic cancer risk prediction. *Genetic Medicine, 22*(6), 756-764.

[9] Taylor, S. D., & Lee, M. P. (2020). Genomic profiling and its potential for cancer prevention. *Nature Reviews Cancer, 20*(1), 54-61.

[10] Davis, G. L. (2022). Exploring genetic markers for breast cancer prediction. *Cancer Genetics, 21*(2), 45-52.

Anika_Zaheen_Research_Proposal_IARCO_2025 - Anika Zaheen.pdf

ORIGINALITY REPORT

10%	5%	8%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	1%
2	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	1%
3	www.biorxiv.org Internet Source	1%
4	"SDK100 topic 7 section 3 WEB157864 ", Open University Publication	1%
5	Inam Ullah Khan, Salma El Hajjami, Mariya Ouaissa, Salwa Belaqziz, Tarandeep Kaur Bhatia. "Cognitive Machine Intelligence - Applications, Challenges, and Related Technologies", CRC Press, 2024 Publication	1%
6	Sara C. Dietz, Jason S. Carroll. "Interrogating the genome to understand oestrogen-receptor-mediated transcription", Expert Reviews in Molecular Medicine, 2008 Publication	1%

7	Wellington Pinheiro dos Santos, Juliana Carneiro Gomes, Valter Augusto de Freitas Barbosa. "Swarm Intelligence Trends and Applications", CRC Press, 2022 Publication	1 %
8	d.docksci.com Internet Source	1 %
9	www.kas2.com Internet Source	1 %
10	jett.labosfor.com Internet Source	<1 %
11	www.medrxiv.org Internet Source	<1 %
12	easy.dans.knaw.nl Internet Source	<1 %
13	ijisrt.com Internet Source	<1 %
14	turi.com Internet Source	<1 %
15	D F Easton. "Models of genetic susceptibility to breast cancer", Oncogene, 09/25/2006 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	Off		