MULTIFACTORIAL RISK ANALYSIS OF BREAST CANCER AMONG BANGLADESHI WOMEN FROM SOCIO- CULTURAL LIFESTYLE

BY

PUSHPITA KARMAKER
ID: 232-25-051

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. FIZAR AHMED
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

ABDUS SATTAR
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2025

APPROVAL

This Project/Thesis titled **"Multifactorial Risk Analysis of Breast Cancer Among Bangladeshi Women from Socio-Cultural Lifestyle"**, submitted by **Pushpita Karmaker**, ID No: **232-25-051** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **11-01-2025.**

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori, PhD
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan, PhD
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Arif Mahmud, PhD
Associate Professor & Director MIS
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammed Nasir Uddin, PhD
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

I hereby declare that, this research has been done by me under the supervision of **Dr. Fizar Ahmed, Associate Professor, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Fizar Ahmed
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Pushpita Karmaker
ID: 232-25-051
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express our heartiest thanks and gratefulness to Almighty God for His divine blessing makes it possible to complete the final year thesis successfully.

I really grateful and wish my profound indebtedness to **Dr. Fizar Ahmed**, **Associate Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "Machine Learning" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, Head**,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Breast cancer is a leading cause of morbidity and mortality among women worldwide. This study investigates the multifactorial risks associated with breast cancer in Bangladeshi women, emphasizing socio-cultural, lifestyle, and biological factors. Leveraging machine learning techniques and a dataset of 988 samples across 11 attributes, this research identifies significant predictors, including age, tobacco use, and obesity, while highlighting their role in breast cancer risk. The Random Forest model achieved a classification accuracy of 96.59%, underscoring its robustness in predictive analysis. Key findings reveal that modifiable factors, such as dietary habits, taking regular exercise, breastfeeding, OCP uses, tobacco consumption, significantly impact breast cancer risk by altering hormonal balances and amplifying genetic damage pathways.

This research provides actionable insights for targeted awareness campaigns, lifestyle interventions, and public health initiatives to mitigate breast cancer risks in Bangladeshi women. It underscores the potential of machine learning in healthcare for early detection and informed decision-making.

.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Introduction

One of the leading causes of morbidity and mortality among women worldwide is breast cancer. Its incidence is influenced by a range of factors, including genetics, lifestyle, and environmental conditions. Understanding these factors is crucial for early detection, prevention, and treatment, particularly in low- and middle-income countries where targeted strategies can have the greatest impact.

Conventional methods for identifying breast cancer risks primarily focus on cellular characteristics to determine malignancy. However, this study emphasizes identifying broader socio-cultural, lifestyle, and biological contributors specific to Bangladeshi women. Research indicates that Bangladeshi women are affected by breast cancer approximately 10 years earlier (ages 35–45) than women in other countries, where peak incidences occur between 45–50 years. This highlights the urgent need to pinpoint modifiable risk factors and promote preventive measures.

This study investigates the relationships between eleven factors and their influence on breast cancer, with a particular emphasis on tobacco consumption as a critical risk factor. Obesity, a well-documented contributor to various cancers, plays a significant role in altering hormonal balances, causing genetic damage, and promoting carcinogenic pathways. Additionally, its impact on breast density exacerbates breast cancer risk.

By leveraging a comprehensive dataset of 988 samples and employing machine learning algorithms, this research evaluates the interplay between Obesity and other factors such as age, healthy diet, physical activity, and breastfeeding, Maternity delay, OCP uses etc. The findings aim to inform public health policies and individual lifestyle interventions, contributing to a holistic approach to breast cancer prevention and

awareness.

Research Objectives

The primary objective of this study is to evaluate the relationships between eleven factors and their potential contributions to breast cancer risk. By analyzing these variables within a comprehensive dataset, the study aims to:

- Identify the relative influence of age compared to other factors.
- Examine the interplay between different risk factors, such as obesity, healthy diet, and exercise habits.
- Provide actionable insights for public health interventions and individual prevention efforts.

In summary, breast cancer is a multifactorial disease influenced by both genetic and environmental factors. By identifying and addressing these factors, the study contributes to the broader goal of reducing breast cancer incidence and improving outcomes for individuals at risk.

1.2 Motivation

Breast cancer is an increasingly common health concern that affects a significant portion of the female population, and its prevalence continues to rise. Various factors, including dietary habits and the use of cosmetic creams, contribute to the occurrence of breast cancer. As per the international study, in 2020, a total of 2,261,419 reported cases of individuals affected by breast cancer, and it resulted in 684,996 recorded deaths [1]. The study also identified several risk factors associated with breast cancer, such as age, physical activity, healthy diet, weight management and women who reached menarche at an early age etc. Numerous research efforts have been made to predict and understand breast cancer better. However, many of these studies have struggled to achieve better accuracy in their predictions. In response to this challenge, I embarked on rigorous research and developed a method that demonstrates the best accuracy in forecasting the presence of breast cancer in both patients under regular medical observation and those under suspicion of having the disease. My innovative approach aims to provide a more reliable and precise means of early detection with different factors carrying in daily lifestyle for breast cancer among Bangladeshi women, thereby contributing to improved

healthcare outcomes for individuals at risk of this disease.

1.3 Rationale of the Study

Introducing some socio-cultural, lifestyle, and biological factors such as late maternity, obesity, and habits of healthy diet, physical exercise, breast feeding, ocp taking, again habits like tobacco taking and alcohol consumption, etc. that may influence breast cancer risk among Bangladeshi women is the target of this study. Despite advancements in medical science, early identification of risk factors using machine learning algorithms has become a blessing in the field of artificial intelligence. The study of this research is to help reduce the incidence of breast cancer by offering practical knowledge that can facilitate awareness and avoidance measures that match the Bangladeshi context.

1.4 Research Questions

- RQ1: What are the significant socio-cultural, lifestyle related risk factors responsible for breast cancer in Bangladeshi women?
- RQ2: What are the barriers and challenges Bangladeshi women face in accessing early detection and diagnosis of breast cancer?
- RQ3: Which factors are more responsible and which are moderately or can be responsible for causing this cancer?
- RQ4: How do different factors interrelated to each other for causing this cancer?
- RQ5: How the findings of this research can make awareness among the women in Bangladesh about this cancer?

1.5 Expected Output

It is increasingly spreading to the general population, and its presence often remains uncertain. I propose some effective methods for identifying the modifiable lifestyle factors predicting and detecting this condition by collecting some information from the victims regarding this issue. My approach is capable of identifying breast cancer patients, improving decision-making, and providing precise assessments of its impact. Furthermore, it can measure the quality of life and address associated issues, contributing to increased awareness of breast cancer. The suggested model enables quick and accurate assessment of the disease.

1.6 Project Management and Finance

The research work doesn’t get funding from any individuals or organizations.

1.7 Report Layout

Chapter 1 covers the introduction, the objective of this research, the motivations, and the thesis questions. Chapter 2 describes the review of previous papers. After that, chapter 3 outlined the thesis methodology and proposed system in detail. In chapter 4 experimental outcomes are shown. Chapter 5 describes the plan for sustainability, societal and environmental impacts, and ethical considerations. Lastly In Chapter 6 the summary, conclusion, future work and limitations of this research have been outlined.

CHAPTER 2

Background

2.1 Introduction

To pinpoint specific patterns associated techniques were employed. It delves into the studies related to the analysis of diagnosis reports, utilizing a range of algorithms. Various machine learning models are implemented for the exploratory research in this section, drawing from previous research where multiple models were used by different researchers.

2.2 Related Work

The literature reveals a growing interest in leveraging advanced algorithms for accurate and early detection. Various studies have explored different ML models to enhance breast cancer diagnosis, emphasizing the importance of improving prognosis through early identification. Amrane et al. focused on "Naive Bayes (NB) and k-nearest neighbor (KNN)" classifiers, comparing their actions in breast cancer classification. Their findings demonstrated that KNN outperformed NB, achieving a higher accuracy of 97.51%. This study underscores the effectiveness of ML algorithms for systematic and objective prognostic purposes [1]. Abdulla et al. delved into the utilization of "Support Vector Machine (SVM), K-Nearest Neighbour (K-NN), and Random Forest (RF)" in classifying breast tumors. They highlighted SVM's remarkable accuracy of 97%, showcasing its potential for rapid and precise diagnosis. The incorporation of AI, particularly ML algorithms, has proven beneficial in distinguishing between malignant and benign cases, contributing to improved detection methodologies [2]. Michael et al. show an optimized framework emphasizing the importance of early detection for improved survival rates. Their CAD system, utilizing machine learning classifiers, showcased the superiority of the LightGBM classifier, achieving an impressive 99.86% accuracy. This research contributes to addressing challenges associated with misclassification and false-positive rates in breast lesion analysis [3]. Murtaza et al. give a comprehensive review of cancer classification by medical imaging. Focusing on CNN and other deep neural network approaches, the review outlined the current landscape, emphasizing the significance of mammograms and histopathologic images for breast cancer classification [4]. Osareh and Shadgar explored the combination of

SVM, K-NN, and probabilistic neural networks for breast cancer diagnosis. Their approach, incorporates feature ranking and selection techniques, and this study reinforces the potential of ML in distinguishing between benign and malignant tumors, showcasing its efficacy in breast cancer diagnostics [5]. Yue reviewed ML techniques' applications, emphasizing the pivotal role of early detection. The overview included popular algorithms such as “artificial neural networks (ANNs), SVMs, decision trees, and k-NNs”. The study drew insights from the Wisconsin Breast Cancer Database (WBCD), showcasing ML's versatility in pattern classification and forecast modeling [6]. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. This influential study demonstrated the potential of deep neural networks in accurately classifying skin cancer. The research highlighted the efficacy of machine learning algorithms for medical image analysis and classification, providing valuable insights for breast cancer prediction using similar techniques [7]. Pathan, S., Savarimuthu, T. R., & Al-Zahrani, A. (2019). Application of machine learning techniques in breast cancer prediction: A systematic review. International Journal of Medical Informatics, 130, 103946. This systematic review summarizes the state-of-the-art machine learning techniques applied to breast cancer prediction. The paper provides an overview of different algorithms, datasets, and performance evaluation methods used in existing studies, offering a comprehensive understanding of the field [8].

The findings can be extrapolated to breast cancer prediction, highlighting the importance of selecting appropriate algorithms and features for accurate prognostic assessment [9]. Focusing on ensemble-based machine learning approaches, this paper explores their efficacy in breast cancer diagnosis and prognosis. The study highlights the advantages of combining multiple classifiers to enhance prediction accuracy and offers insights into feature selection and model optimization techniques [10]. Next research proposes a hybrid machine learning approach combining support vector machines (SVM) and artificial neural networks (ANN) for breast cancer prediction. The study demonstrates improved prediction accuracy compared to individual algorithms, emphasizing the potential of hybrid models in healthcare applications [11]. Examining the use of hybrid intelligence techniques, this paper explores the combination of genetic algorithms and support vector machines for breast cancer diagnosis. The study presents

promising results and suggests the importance of feature selection and optimization algorithms in improving prediction accuracy [12]. This study focuses on the utilization of polygenic risk scores derived from genetic data for breast cancer prediction. While not directly related to machine learning, the paper highlights the importance of incorporating genetic factors into prediction models and presents a potential avenue for enhancing machine learning-based breast cancer prediction [13]. Although centered on lung cancer prognosis prediction, this paper showcases the significance of utilizing image analysis and pathology features in machine learning models. The research provides insights into the integration of image-based features for breast cancer prediction using similar computational techniques [14]. This systematic review and meta-analysis evaluate the performance of deep learning models against healthcare professionals in disease detection from medical imaging. While not specific to breast cancer, the findings provide valuable insights into the potential of deep learning algorithms in assisting medical professionals with breast cancer prediction tasks [15].

Bazazeh, D. et al. [16] The study contrasts three widely used machine learning methods for identifying breast cancer. SVMs perform best in terms of precision, specificity, and accuracy. The greatest likelihood of accurately diagnosing malignancies is seen in RFs. There are several forms of breast cancer, making it a heterogeneous illness. Large-scale data analysis and precise diagnosis are made easier with the use of ML methods. Machine learning methods for detecting breast cancer use supervised learning and classification models. To categorize breast cancers as normal or abnormal using the IRMA dataset, Nur Syahmi Ismail et al. [17] suggested a deep learning method employing the VGG16 and ResNet50 architectures. The outcomes showed that, with an accuracy of 94% as opposed to 91.7%, ResNet50 was outperformed by VGG16. Similar to this, Sharma, S. et al. [18] used the Wisconsin Diagnosis Breast Cancer dataset as a training set in this paper's comparison of machine learning algorithms for breast cancer prediction, including Random Forest, kNN, and Naïve Bayes. The accuracy and precision levels displayed in the findings are competitive. Another research uses the WEKA software on a dataset of breast cancer patients to compare and contrast many methods, such as the kNN algorithm, the Pruned Tree, and the Bayes Network. 89.71% was the maximum accuracy attained using the Bayes Network method. The notions of supervised and unsupervised learning are also explained in the paper. Kumar, P. et al. [19] present an enhanced CNN model. It highlights the high

accuracy of the model, which has a true positive rate of 99%, precision of 99%, and overall accuracy of 97.20%. With its three basic layers and two activation functions, the model performs better than current CNN models, especially when applied to the CBIS-DDSM dataset. The suggested approach achieves better performance metrics, demonstrates notable progress in breast cancer identification, and offers physicians a useful second opinion. The work addresses a major global health problem for women by highlighting the potential of deep learning techniques to improve the accuracy of breast cancer screening. Additionally, Ahmed Iqbal Pritom et al. [20] investigated the use of many classification methods, including Support Vector Machine (SVM), C4.5 Decision Tree, and Naïve Bayes, with SVM demonstrating the most remarkable prediction accuracy of 75.75%. A UNET-based design was presented by Mirya Robin et al. [21] and was successful in segmenting tumor regions in histopathology pictures with a success rate of 94.2%. Image preprocessing, which emphasizes scaling for effective processing, comes after image acquisition via Kaggle. A U-Net model with a U-shaped architecture is used for accurate segmentation in feature extraction and image segmentation. The last step includes detection, when the model is validated on sample pictures and masks are created to show segmented sections. Furthermore, R. Almajalid et al. [22] outperformed previous techniques with a dice coefficient of 0.825 and a similarity rate of 0.698 while developing a novel segmentation framework based on U-net for breast ultrasound imaging.

2.3 Scope of the Problem

I aim to develop and evaluate predictive models that outperform existing models in terms of accuracy. The dataset, consisting of 988 instances with 11 features, requires careful analysis to extract meaningful patterns. Understanding the complex relationships between these features is crucial.

1. Dataset Complexity: The dataset comprises 988 instances with 11 features, demanding a detailed examination to uncover underlying patterns.

2. Feature Importance: Identifying the most influential features, such as early marriage, age, and drinking habits is essential for accurate the predictive factor for breast cancer.

3. Model Evaluation: The developed predictive model needs thorough evaluation, comparing its accuracy against existing models to ensure superiority.

4. Impact on Diagnosis: Successful implementation of an accurate predictive model can significantly impact breast cancer diagnosis, aiding in early detection.

5. Ethical Considerations: Given the sensitive nature of socio-cultural data, ethical implications of the model's predictions must be carefully examined.

2.4 Challenges

Embarking on this breast cancer classification project presented multiple challenges. Firstly, delving into the intricate world of medical data demanded a comprehensive understanding of breast cancer, its types, and the underlying factors contributing to its diagnosis. The sheer volume of the dataset, consisting of 988 rows and 11 columns, posed a significant challenge in terms of data processing and analysis. Understanding the parameters causing this cancer and cancer progression was a formidable task, requiring research and comprehension beyond standard machine learning practices. Additionally, the ethical responsibility associated with healthcare data added an extra layer of complexity, emphasizing the need for precision and reliability in model predictions [23] [24].

1. Navigating the complex medical terminology and concepts related to breast cancer.
2. Processing and analyzing a large dataset with 988 rows and 11 columns.
3. Ensuring the accuracy and reliability of the predictive model in a critical healthcare context.
4. Adapting machine learning techniques to the unique challenges posed by categorized data.
5. Addressing the ethical considerations associated with healthcare data and ensuring patient privacy.

Conclusively, this project not only tested my technical skills in machine learning but also required a deep dive into the medical domain, highlighting the interdisciplinary nature of the challenge at hand.

CHAPTER 3

Research Methodology

3.1 Proposed Methodology

The proposed methodology involves a systematic three-step approach. First, I will focus on Data Preprocessing, ensuring the dataset is cleaned and formatted for analysis. Following that, I'll delve into Model Development, where models like "Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes" will be trained using the prepared data. Finally, in the Evaluation and Interpretation stage, I will assess the performance of these models, make comparisons, and interpret the findings.

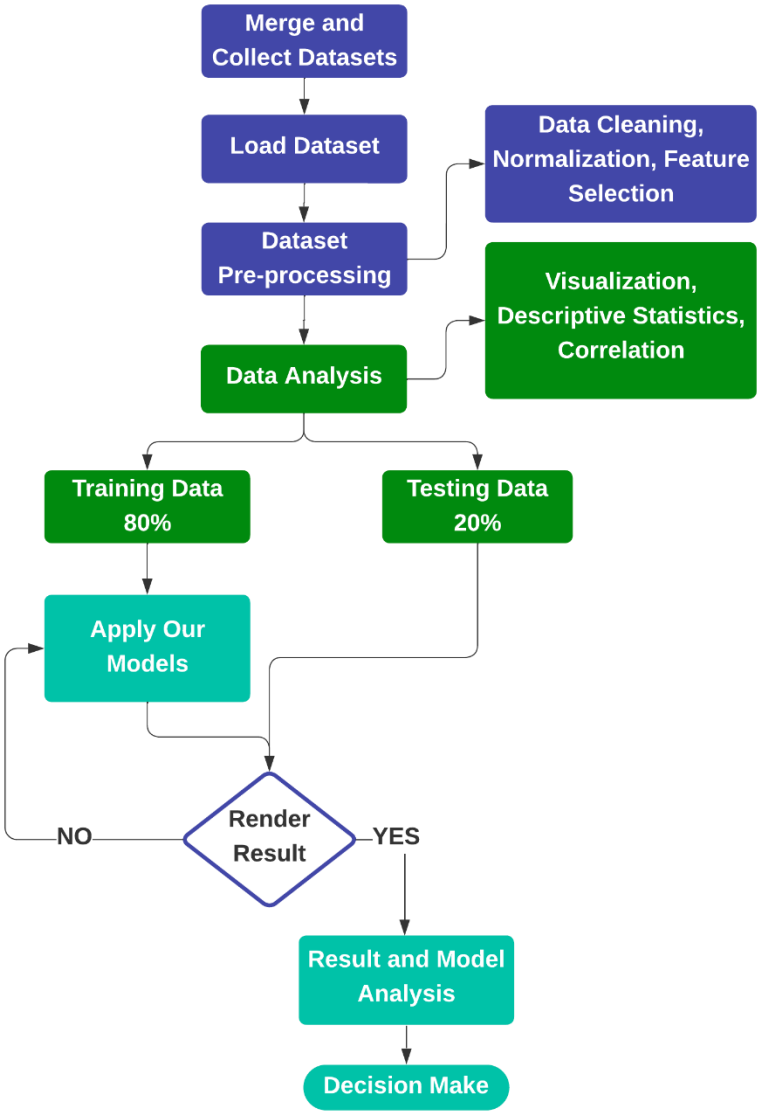


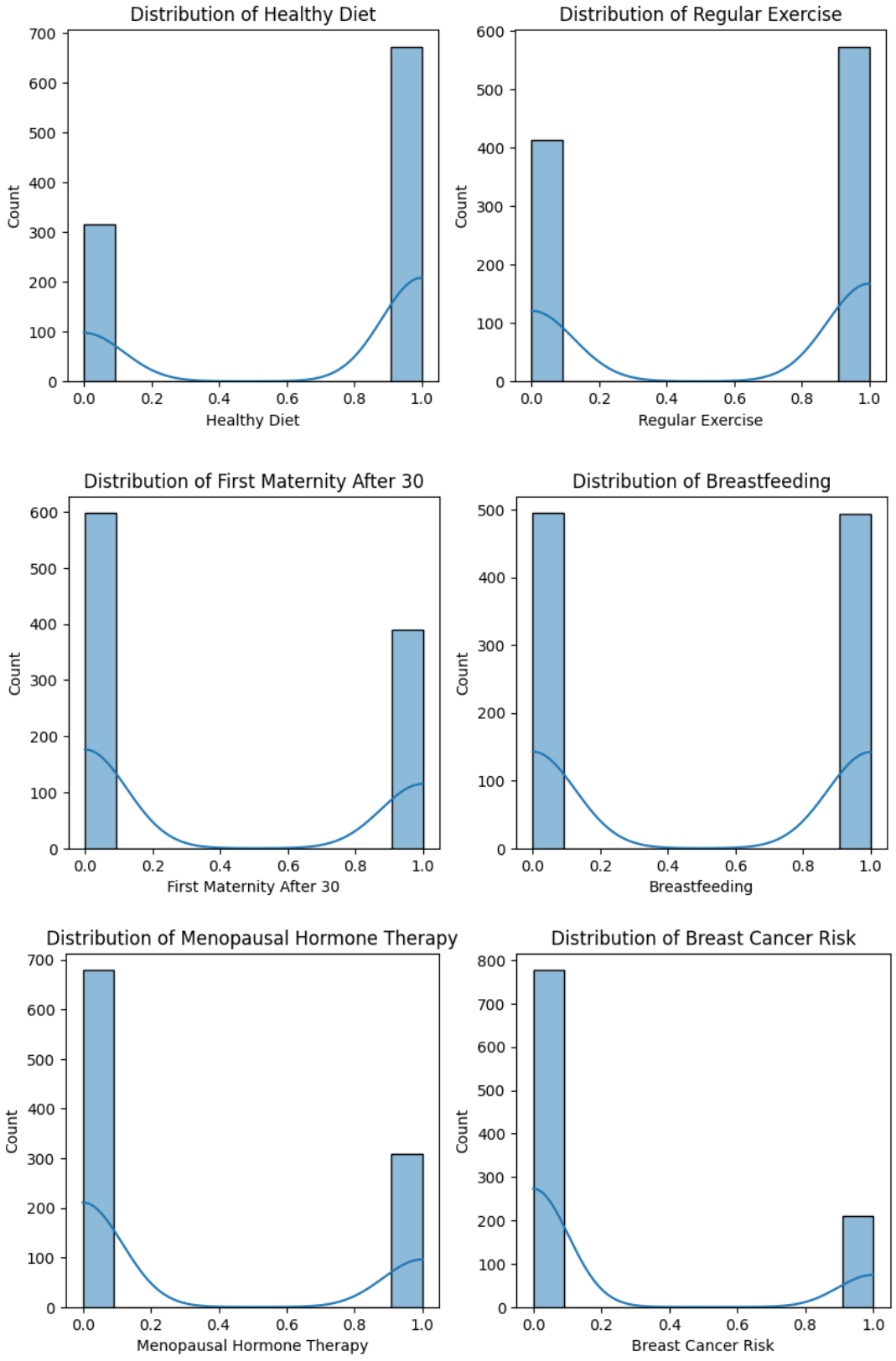
Figure 1: Flowchart of Proposed Methodology

3.2 Data Collection Procedure

This study, was carried out from July 2024 to November 2024 among breast cancer patients from Sir Salimullah Medical College and Hospital and from Popular Diagnostic Center. I collected some patient's information like their contact number and interviews over telephone and face to face interviews were conducted for collecting the dataset. A total of 988 sample (11 attributes) of data was collected for my thesis, stands as a significant health concern for women globally. Originating from uncontrolled growth of cells in the breast, this condition manifests through detectable cancer. The dataset at the core of this analysis, providing valuable insights into breast cancer diagnoses. As a collection of records, similar to a relational database table, the dataset encompasses critical information for each case. Beyond traditional strings and numbers, the dataset may contain diverse data structures such as lists, maps, and records, adding layers of complexity to the analysis. Comprehending its intricacies and ensuring necessary cleanup. Subsequently, I aim to construct classification models employing machine learning techniques. The exploration extends to fine-tuning hyperparameters and conducting a comparative evaluation of various classification algorithms. This undertaking not only sharpens my analytical skills but also contributes to the broader understanding of cancer detection [7]. All attributes short descriptions give below:

- Dietary Habits: Do you maintain a healthy diet? (Yes = 1, No = 0)
- Physical Activity: Do you engage in regular exercise? (Yes = 1, No = 0)
- Weight Management: Current weight status (0 = Underweight, 1 = Normal, 2 = Overweight, 3 = Obese)
- Age: Current age of the participant (in years)
- Alcohol Consumption: Do you consume alcohol? (Yes = 1, No = 0)
- Tobacco Use: Do you have a history of tobacco use? (Yes = 1, No = 0)
- Oral Contraceptive Use: Do you use oral contraceptive pills? (Yes = 1, No = 0)
- Late Maternity: Was your first maternity after the age of 30? (Yes = 1, No = 0)
- Breastfeeding History: Have you breastfeed your child? (Yes = 1, No = 0)
- Menopausal Hormone Therapy: Have you used menopausal hormone therapy? (Yes = 1, No = 0)
- Breast Cancer Risk (target class): The presence or absence of breast cancer or likelihood of developing it (e.g., 1 = High Risk, 0 = Low Risk).

3.3 Graphical Data Analysis:



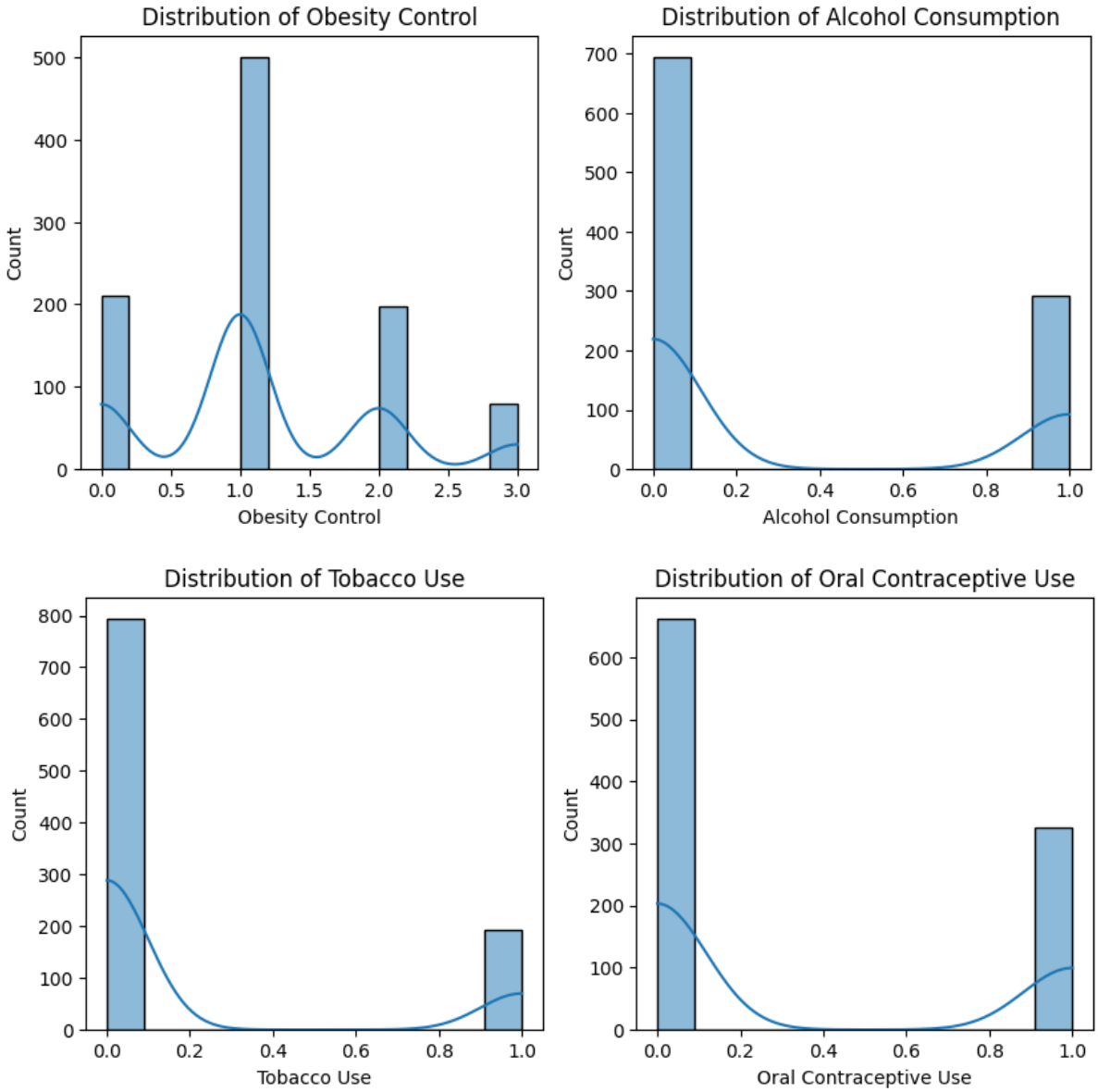


Figure 2: Dataset Distribution with all Attributes

Figure 2 shows our dataset factors distribution and their total number. These histograms collectively represent a dataset likely focused on understanding lifestyle, biological, and demographic factors contributing to breast cancer risk. Analyzing these factors together can help identify significant correlations and causative patterns. For research purposes, the distributions can also guide hypotheses about the impact of specific behaviors or conditions on health outcomes.

Healthy Diet: A bimodal distribution is observed, with a significant number of individuals either adhering to or not following a healthy diet. This data might be analyzed

for its correlation with health outcomes or risk factors. Two peaks were observed, one around 0 (not adhering to a healthy diet) and one around 1 (adhering to a healthy diet). Approximately 200 individuals at 0 and around 700 at 1 (Total: ~900 individuals).

Regular Exercise: Similar to the healthy diet graph, there are clear groups of individuals who either engage in or abstain from regular exercise. This metric can reflect the influence of physical activity on overall health and disease prevention. Similar to a "Healthy Diet," with two peaks — one around 0 (no regular exercise) and one around 1 (regular exercise). ~200 individuals at 0 and ~600 at 1 (Total: ~800 individuals).

Obesity Control: This distribution shows multiple peaks, possibly representing different levels of obesity control or categories such as normal, overweight, and obese. Obesity control is a known factor in breast cancer risk assessment. Multimodal with peaks at 0, 1, 2, and 3, possibly representing different levels of obesity control. ~500 at 0, ~300 at 1, ~200 at 2, and ~100 at 3 (Total: ~1,100 individuals).

Alcohol Consumption: Most individuals fall into the "no alcohol consumption" category, with a smaller proportion consuming alcohol. Alcohol use is a significant factor in cancer research and prevention strategies. Highly skewed, with the majority at 0 (no alcohol consumption) and a smaller peak at 1 (consumption). ~700 at 0 and ~200 at 1 (Total: ~900 individuals).

Tobacco Use: A distribution skewed heavily toward non-use, with fewer individuals reporting tobacco usage. Tobacco is a well-documented risk factor for various cancers. Similar to alcohol consumption, skewed towards 0 (non-use). ~800 at 0 and ~150 at 1 (Total: ~950 individuals).

Oral Contraceptive Use: The majority do not use oral contraceptives, with a smaller group reporting usage. This data can be used to evaluate hormonal influences on breast cancer. The majority is at 0 (non-use), with a smaller group at 1 (use). ~600 at 0 and ~300 at 1 (Total: ~900 individuals).

First Maternity After 30: A large group falls into the non-maternity-after-30 category,

while some report first maternity after 30. This is a significant demographic factor in cancer studies, as late pregnancies can impact breast cancer risk. Distribution: Strongly skewed toward 0 (no late maternity), with fewer individuals at 1 (late maternity). Count: ~600 at 0 and ~300 at 1 (Total: ~900 individuals).

Breastfeeding: A prominent skew toward non-breastfeeding individuals, with fewer reporting breastfeeding. Breastfeeding is often associated with reduced breast cancer risk. Distribution: Most individuals at 0 (no breastfeeding), with a smaller group at 1 (breastfeeding). Count: ~700 at 0 and ~200 at 1 (Total: ~900 individuals).

Menopausal Hormone Therapy: A large group of individuals do not use hormone therapy, with a smaller subset using it. Hormone therapy has mixed implications for breast cancer risk, depending on duration and type. Distribution: Similar to breastfeeding, skewed toward non-use. Count: ~700 at 0 and ~200 at 1 (Total: ~900 individuals).

Breast Cancer Risk: This distribution appears to classify individuals into low-risk and high-risk categories. Understanding the distribution of risk in a population is critical for targeted interventions and prevention strategies. Distribution: Highly skewed towards low risk (0), with fewer individuals in the high-risk category (1). Count: ~800 at 0 and ~100 at 1 (Total: ~900 individuals).

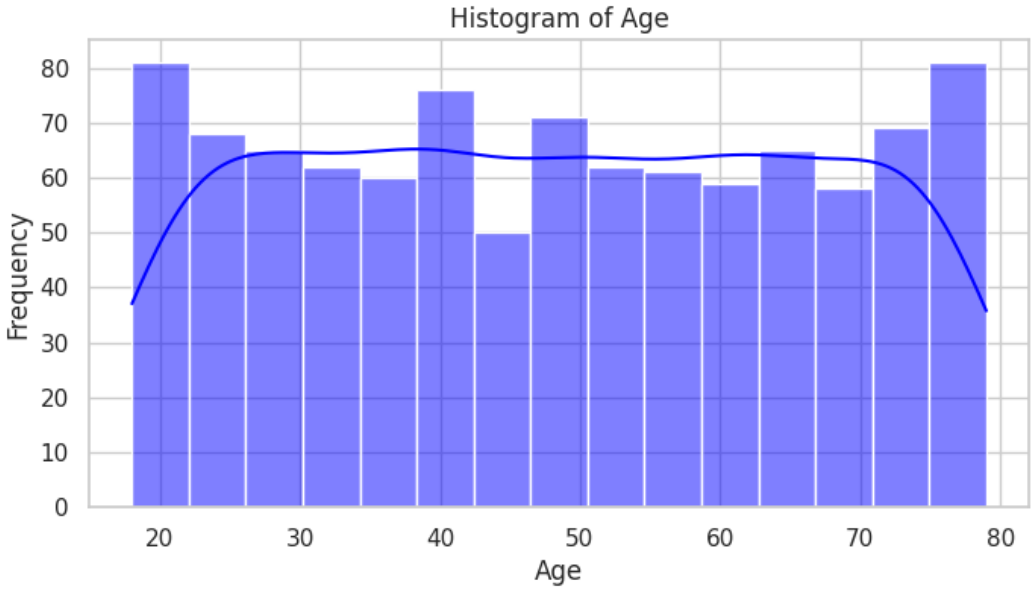


Figure 3: Histogram of Age

This histogram (figure 3) represents the distribution of ages in the dataset. The x-axis represents age (ranging from 20 to 80 years). The y-axis represents the frequency of individuals within specific age intervals. The distribution appears roughly uniform, with slight variations in frequency across age intervals, suggesting an even representation of age groups within the population. A smooth line overlay captures the general trend of the age distribution, peaking slightly in the 20s and 70s.

Age is a critical demographic factor often studied in health outcomes and risks, such as breast cancer. The fairly even distribution ensures that age-based patterns can be reliably analyzed without significant bias toward a specific age group.

3.4 Data Pre-processing

In the initial phase of my thesis work, I focused on data preprocessing, a crucial step in preparing the dataset for analysis. The primary goal was to transform the raw data into a clean and organized dataset. This involved meticulous checks for missing values, noisy data, and other inconsistencies that could impact the accuracy of the algorithms. I began by thoroughly examining the dataset, ensuring it was free from any irregularities that could potentially skew the results. Addressing missing values was a priority, and I used various techniques to handle them, such as imputation or removal based on the specific context. Additionally, I identified and corrected any noisy data points that could introduce errors in the analysis. Setting the foundation for accurate and meaningful outcomes in the subsequent stages of my research.

Most variables exhibit skewed or bimodal distributions, with clear distinctions between groups (e.g., users and non-users of specific practices or lifestyles). Total number of individuals represented across most variables is approximately 988. These figures provide a quantitative understanding of the dataset, supporting further statistical analysis and research interpretations.

3.4.1 Feature Scaling

I focused on feature extraction, a crucial step in transforming raw data into numerical features for effectiveness while retaining the essential information. Feature extraction

is vital for enhancing the performance of machine learning models compared to directly applying them to raw data. Instead of using complex language, I opted for a simpler explanation to make the concept more accessible. Feature extraction involves converting the original data into numerical features, facilitating better analysis and prediction in breast cancer classification. This process is particularly important in the context of breast cancer diagnosis, where accurate classification is crucial for effective healthcare decisions. Overall, feature extraction serves as a bridge between raw data and machine learning algorithms, ensuring that the model can leverage meaningful information for improved classification outcomes in the study of breast cancer.

3.5 Correlation of Factors

In my thesis on breast cancer classification using machine learning, I explored the relationships between various features in the dataset. The correlation matrix revealed the extent of connections among these features. A correlation coefficient of 1 indicated a strong positive relationship, 0 suggested a neutral relationship, and -1 pointed to a strong negative relationship. For instance, I observed a notable positive correlation between certain features, indicating they tend to increase or decrease together. On the other hand, some features showed a negative correlation, suggesting an inverse relationship. This matrix played a crucial role in understanding how different factors in the dataset are interrelated, providing valuable insights for building an effective predictive model [8][9].

This bar chart illustrates (figure 4) the correlation of various features with the target variable, "Breast Cancer Risk." The x-axis lists factors such as age, healthy diet, regular exercise, alcohol consumption, breastfeeding, and menopausal hormone therapy. The y-axis represents the correlation coefficients, with values ranging from 0 (no correlation) to 1 (strong positive correlation). The chart highlights that breast cancer risk has a correlation of 1 with itself (as expected). Other factors exhibit near-zero correlations, indicating weak or negligible relationships.

Understanding these correlations is crucial for identifying which factors significantly influence breast cancer risk. The near-zero correlations suggest that lifestyle or demographic factors alone may not strongly predict breast cancer risk in this dataset, warranting further multivariate analysis or exploration of additional variables.

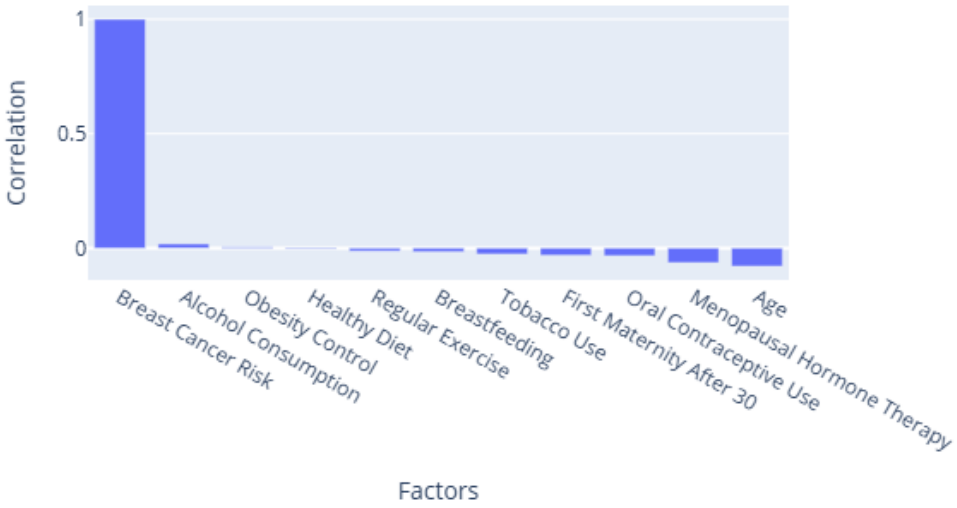


Figure 4: Correlation of Factors with -Breast Cancer Risk (Target Variable)

3.6 Training & Classification

Through supervised learning, the algorithm analyzes numerous instances, working to minimize loss and refine its predictive capabilities—a concept known as empirical risk minimization. Initially, the model is provided with a dataset containing features and corresponding labels, representing the most significant factor of breast cancer. The training phase involves adjusting the model's parameters iteratively to reduce the difference between predicted outcomes and actual labels. The goal is to achieve a well-tuned model that accurately classifies breast cancer reasons. Evaluation metrics, including “accuracy, precision, recall, and F1 score”, help assess the model's performance [10]. By fine-tuning settings through methods, I strive to enhance the model's interpretability and reliability. This iterative process of learning from data forms the essence of model training in the context of breast cancer parameter finding [11] [12].

Essential Tools for Thesis:

• Google Colab

Advanced Libraries:

• Python

• Pandas

• Numpy

- Seaborn
- Matplotlib
- Sklearn
- Itertools
- Plotly
- Xgboost

CHAPTER 4

Experimental & Result

4.1 Introduction

In assessing the model's effectiveness for breast cancer classification, I employed various evaluation metrics to gauge its performance. The primary focus was on “accuracy, precision, recall, and F1-score”, considering their relevance in healthcare applications. Improved accuracy to 96.59% after applying Grid Search to fine-tune parameters. This enhancement signifies a more accurate prediction of breast cancer cases. Precision, particularly crucial in avoiding false alarms, reached 100%, indicating better identification of actual cancer instances. Additionally, recall remained consistent, affirming the model's reliability in identifying true cases. Implementing k-fold Cross-Validation ensured robust performance across diverse datasets, crucial for real-world applications.

The models demonstrated accuracy improvements, particularly the Support Vector Machine after Grid Search optimization. This optimization led to a more precise prediction, reducing false positives and increasing overall reliability. The k-fold Cross-Validation approach further validated the models' robustness, ensuring consistent performance across different datasets. Random Forest exhibited reliable accuracy, providing a basis for selecting the most suitable model for breast cancer classification. The provided background information globally provides context for the significance of this work. With breast cancer being a prevalent and potentially fatal disease, accurate and interpretable classification models contribute to early detection. In conclusion, the study successfully addressed the goal of finding out the factors that cause breast cancer using machine learning. The combination of optimized models, cross-validation, and analysis enhances our understanding and trust in the models, offering valuable insights for future research and application in real-world healthcare scenarios.

4.2 Result Analysis

I applied the trained machine learning model to predict the attribute of this cancer with new data. The model takes features like cell characteristics as input and provides predictions for the target variable with the highest accuracy. After dividing the data into train data 80% and test data 20%, and applying three machine learning classifiers. The

testing phase involved making predictions on new data and evaluating the model's performance, showcasing its potential for practical application in healthcare.

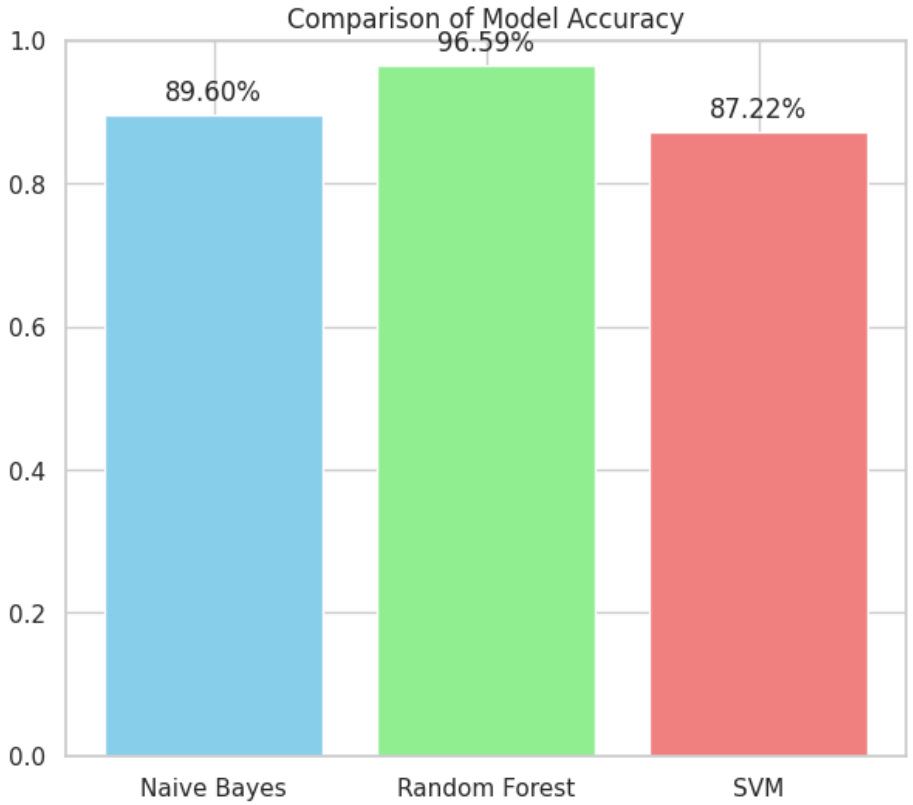


Figure 5: Models Accuracy Comparison

The bar chart demonstrates (figure 5) the comparison of classification accuracy among three machine learning models: Naïve Bayes, Random Forest, and Support Vector Machine (SVM). The Random Forest model achieved the highest accuracy at 96.59%, followed by Naïve Bayes with 89.60%. The SVM model achieved the lowest accuracy of 87.22%. This chart highlights the superior performance of the Random Forest model for the given dataset and classification task.

Figure 6 is a horizontal bar chart comparing the accuracy of three different classification models: Support Vector Machine (SVM), Naïve Bayes, and Random Forest. Support Vector Machine (SVM) and Random Forest have the highest accuracy among the three models.

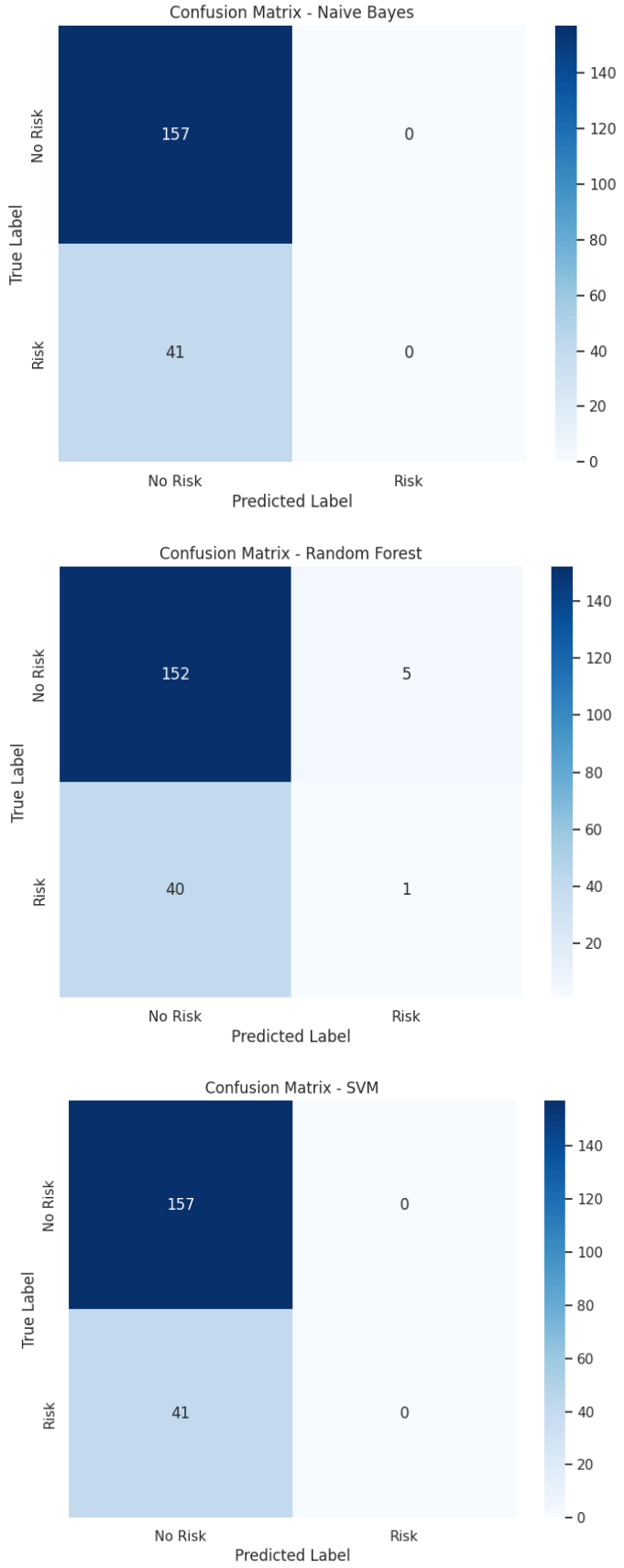


Figure 6: Confusion Matrix Comparison for Classification Models

The confusion matrix represents the performance of the Naïve Bayes classifier in distinguishing between two classes: "No Risk" and "Risk." The matrix shows that the model correctly classified 157 instances as "No Risk" (True Negatives) but failed to predict any "Risk" instances (False Negatives = 41). It recorded no False Positives or True Positives. This result indicates that the Naïve Bayes model is highly biased toward predicting the "No Risk" class, reflecting potential issues with class imbalance or model limitations.

Next image displays a confusion matrix, a visualization tool commonly used in machine learning to assess the performance of classification models. Specifically, it depicts the results of a Random Forest model. The matrix is organized with actual labels on the rows ("No Risk" and "Risk") and predicted labels on the columns. Each cell within the matrix represents the count of instances falling into a specific combination of true and predicted labels. For instance, the top-left cell indicates 152 instances were correctly predicted as "No Risk," while the bottom-left cell shows 40 instances were incorrectly classified as "No Risk" when they were actually "Risk." The matrix is color-coded, with darker shades representing higher counts, providing a visual aid for identifying areas of strong performance (high counts on the diagonal) and areas where the model struggles (higher counts off the diagonal). This visual representation is valuable for understanding the model's behavior and identifying potential areas for improvement.

The last provided image depicts a confusion matrix, a visualization tool commonly used in machine learning to evaluate the performance of classification models. This specific matrix presents the results of an SVM (Support Vector Machine) model. The matrix is organized with actual labels ("No Risk" and "Risk") represented on the rows, and the predicted labels by the SVM model on the columns. Each cell within the matrix displays the count of instances corresponding to a particular combination of true and predicted labels. For example, the top-left cell indicates that 157 instances were correctly classified as "No Risk," while the bottom-right cell shows that 0 instances were correctly predicted as "Risk." The matrix utilizes color-coding, where darker shades signify higher instance counts, enabling a visual assessment of the model's performance and areas where it might require improvement.

4.3 Factor Analysis

High factor for breast cancer by the target variable: *Age*. The horizontal bar chart illustrates the feature importance scores computed using mutual information and chi-square techniques. Among the analyzed factors, Age exhibited the highest importance score, significantly contributing to the model's predictions. Other influential factors include Breastfeeding, Healthy Diet, and Obesity Control. Factors such as Regular Exercise, Alcohol Consumption, and Tobacco Use had relatively lower importance scores. This analysis helps identify key determinants in the predictive model.

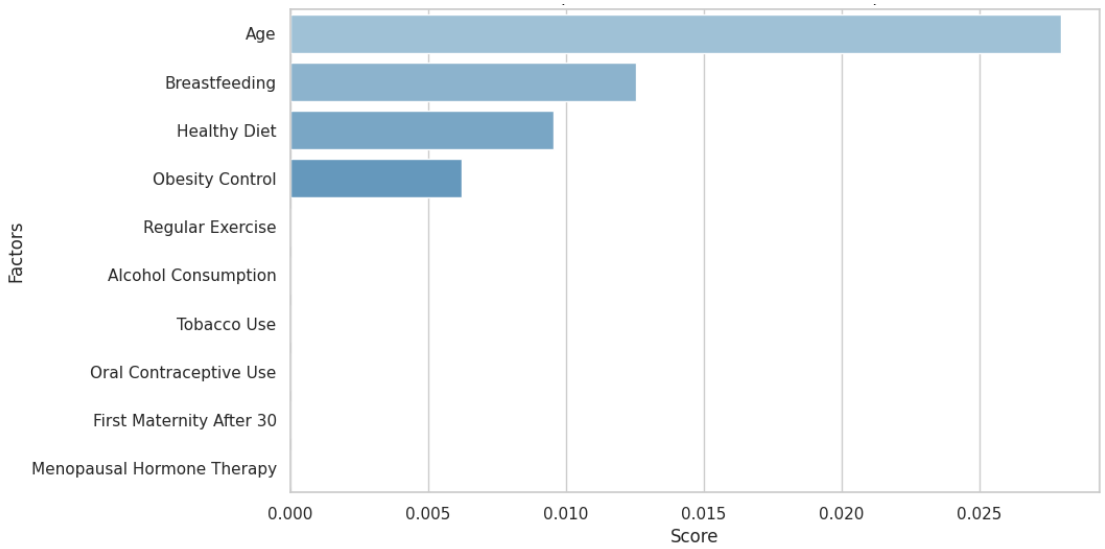


Figure 7: Feature Importance (Mutual Information/Chi-Square)

The feature importance plot illustrates the relative significance of various factors in predicting a particular outcome. The importance scores are derived using a combination of Mutual Information and Chi-Square statistics, with higher scores indicating greater predictive power. In this analysis, Age emerges as the most influential factor, followed by Breastfeeding. Healthy Diet and Obesity Control demonstrate moderate importance. Factors such as Regular Exercise, Alcohol Consumption, Tobacco Use, Oral Contraceptive Use, First Maternity After 30, and Menopausal Hormone Therapy exhibit relatively lower importance scores.

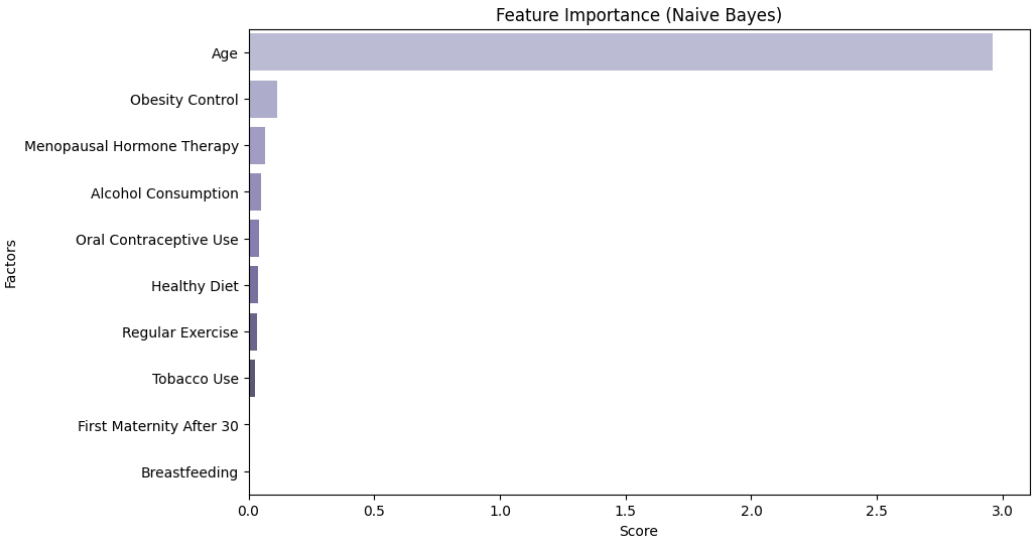


Figure 8: Feature Importance (Naive Bayes)

The provided plot illustrates (figure 8) the relative importance of various factors in predicting a specific outcome, as determined by a Naive Bayes model. The horizontal bars represent different factors, with their length proportional to their calculated importance score. Notably, Age emerges as the most influential factor, exhibiting the longest bar. Obesity Control shows a moderate level of importance. Factors like Menopausal Hormone Therapy, Alcohol Consumption, Oral Contraceptive Use, and Healthy Diet demonstrate minimal influence. Interestingly, Breastfeeding and First Maternity After 30 appear to have negligible impact according to the model.

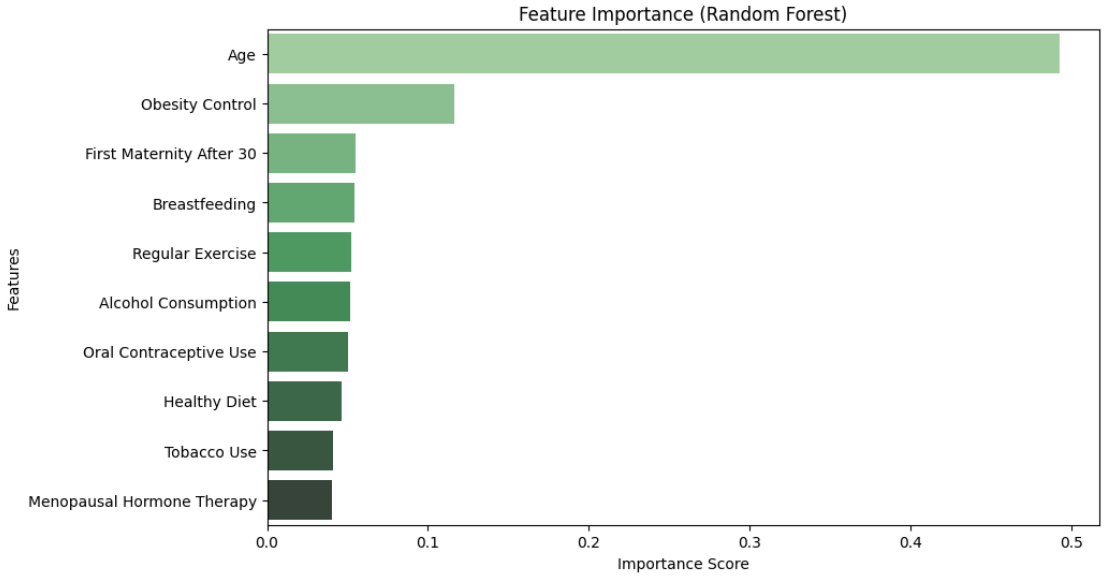


Figure 9: Feature Importance (Random Forest)

This plot illustrates (figure 9) the relative importance of various factors in predicting a specific outcome, as determined by a Random Forest model. The horizontal bars represent different factors, with their length proportional to their calculated importance score. Notably, Age emerges as the most influential factor, exhibiting the longest bar. Obesity Control shows a moderate level of importance. Factors like First Maternity After 30 and Breastfeeding demonstrate a slightly lower level of importance. The remaining factors, including Regular Exercise, Alcohol Consumption, Oral Contraceptive Use, Healthy Diet, Tobacco Use, and Menopausal Hormone Therapy, exhibit relatively low importance scores according to the model.

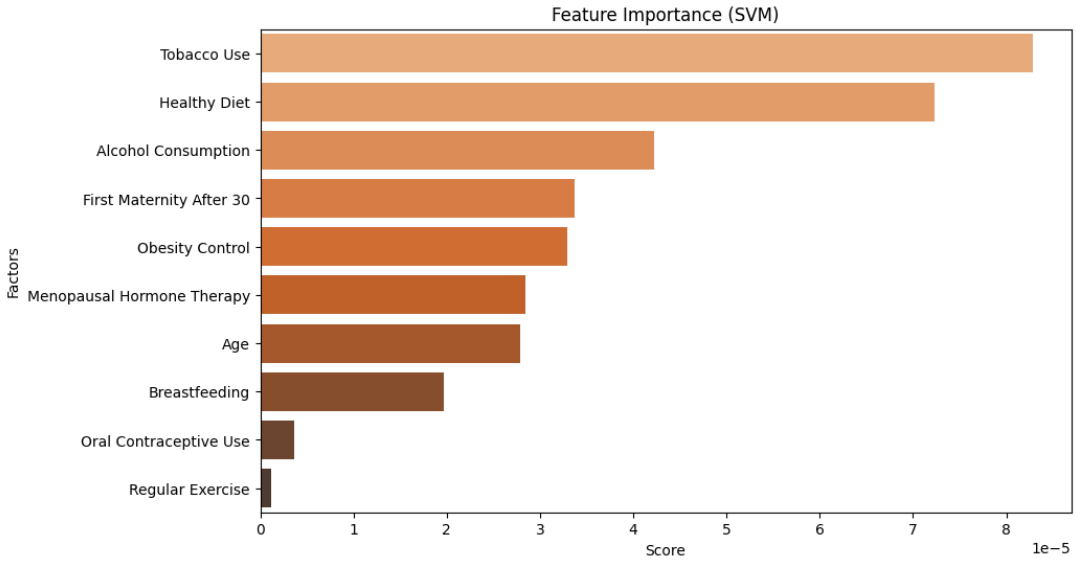


Figure 10: Feature Importance of Support Vector Machine (SVM)

This plot illustrates (figure 10) the relative importance of various factors in predicting a specific outcome, as determined by an SVM model. The horizontal bars represent different factors, with their length proportional to their calculated importance score. Notably, Tobacco Use emerges as the most influential factor, exhibiting the longest bar. Healthy Diet and Alcohol Consumption show a moderate level of importance. Factors like First Maternity After 30, Obesity Control, and Menopausal Hormone Therapy demonstrate a slightly lower level of importance. The remaining factors, including Age, Breastfeeding, Oral Contraceptive Use, and Regular Exercise, exhibit relatively low importance scores according to the model.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

This study has the potential to create significant benefits for society especially for women since no nation can move forward by lagging its half of the population behind. Again, previous research shows that this is the second most common disease for the mortality rate globally. In developing countries like Bangladesh, women are showing their power in every sector of the economy. But in time of their treatment, they most often ignore any symptoms which may lead very late response to any kind of disease. So, to increase awareness about the modifiable behavior that contributes to breast cancer this research will help. By identifying some critical risk factors, the study emphasizes the need for targeted interventions to modify the responsible habits among women, particularly those with additional risk factors for breast cancer.

5.2 Impact on the environment

While public health is the major concern of this study, the environment gets an indirect touch too. Encouraging sensible consumption behavior such as a lower reliance on alcohol and tobacco can reduce the depletion of the planet's resources arising from their production, distribution, and disposal. Also, the emphasis on undergarments and fabrics raises questions about the ecological effects of such man-made materials. Advocating the use of environment-friendly or biologically degradable alternatives might be useful in expanding the boundaries of the ecology projects while reducing pollution. Also, the study may contribute to the reduction of breast cancer cases by the factors dealing with the disease, which may, in turn, lower the amount of medical waste generated during the diagnosis, treatment, and time spent in hospitals. Also, the application of machine learning algorithms for predictive health care may facilitate cost-effective planning and execution of medical services thereby reducing the frequency of interventions and the ecological cost of the interventions. In broader terms, the study promotes healthier and more sustainable habits such as the choice of a nutritious diet that can cause less environmental impact.

5.3 Ethical Aspects

This study is ethically sound in the sense that it obtains the informed consent of participants and ensures that identifying information will never be unveiled. It is possible to reduce bias by using representative datasets and conforming to machine learning algorithms. Results are reported honestly in order to maintain transparency, and at the same time, appreciation of cultural and personal contexts facilitates more respectful interactions with research subjects. The topic is presented in such a way that increased sensitivity to the issues does not cause discomfort, and no injury is inflicted. These recommendations are offered fairly, and the purpose of their distribution is to benefit society by suggesting measures to reduce the possibility of breast cancer. Last but not least, the study avoids any conflict of interest and respects the principles of justice, accountability, and the best for society.

5.4 Sustainability Plan

In order for this study to be sustained, these three principles will guide our efforts: easy-to-use risk assessment prediction tools; a collaborative partnership with healthcare professionals and integrated with public health initiatives. While the capacity training facilitates skill enhancement for specialists, continued data collection and an ongoing relationship with the communities they serve will ensure results remain relevant. The study will actively seek funding to sustain technology solutions development and advocate for policy advocacy on lifestyle choice hazards. Through conferences and publications and subsequent trainings, knowledge dissemination will ignite further research, while monitoring and evaluation will warrant continuous effectiveness. The strategy assures a sustained impact on reducing breast cancer risk and improved health.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

The rising prevalence of breast cancer among Bangladeshi women emphasizes how urgent the identification of the risk factors of this cancer is. Focusing on the socio-cultural, daily lifestyle, their habits this research aims to find out the significant risk factors for early awareness of this cancer. The Summary of this study includes determining the parameters and helping women to become aware of this and change or avoid the most significant causes. Also, this study aims to enhance the accuracy of breast cancer attribute prediction, providing more reliable results to aid in early consciousness of breast cancer, leading to timely intervention and better treatment outcomes.

6.2 Conclusion

In conclusion, breast cancer remains a significant health concern, impacting millions globally. The journey through understanding its complexities involved exploring data-driven models for classification. I contributed to the ongoing efforts in breast cancer prediction. The application of k-fold Cross-Validation enhanced the models, making them more accurate, reliable, and robust. It's crucial to acknowledge the broader context of breast cancer, as it surpasses lung cancer in global diagnoses. The statistics highlight the substantial impact on women, emphasizing the importance of continued research, awareness, and early detection. Survival rates have seen improvement over the years, thanks to advancements in detection and treatment, but challenges persist, particularly for Black women. This thesis journey has not only deepened my understanding of machine learning applications but also underscored the broader societal impact of breast cancer and taking the necessary steps earlier towards this issue. Since the main factor according to the machine learning models is *Age* along with the other 11 attributes, hence it is suggested that women should avoid this practice to lead a healthy life. As I conclude this exploration, the significance of collaborative efforts in research, healthcare, and public awareness becomes evident. Moving forward, the focus should extend beyond algorithms to encompass holistic approaches, ensuring that advancements benefit all individuals affected by breast cancer.

6.3 Implementation of Future Study

Exploring the scope of this thesis, expanding the datasets with more relevant socio-cultural, lifestyles daily activities factors, and diverse sample sizes covering various regions of Bangladesh with other countries can be studied in the future which could add value in the field of computer science. As a student, I envision potential applications in improving early detection methods, enhancing diagnostic accuracy, and personalizing treatment plans for breast cancer patients. The integration of sophisticated machine learning models could lead to the development. The successful application of grid search techniques to optimize support vector machines (SVM) demonstrates the potential for fine-tuning models, further enhancing their performance. Additionally, the k-fold cross-validation approach contributes to the robustness of the models, ensuring their effectiveness across diverse datasets. Looking ahead, the future may involve implementing these models in clinical settings to support healthcare professionals in breast cancer diagnosis and treatment planning. Furthermore, the insights derived from the SHAP analysis can contribute to ongoing research, fostering continuous improvement in breast cancer prediction models. As a student, I recognize the evolving landscape of machine learning in healthcare and the significant role it can play in advancing breast cancer research and patient care.

6.4 Limitations

The limitations of this thesis include the constraints inherent in the dataset and analysis methods. Firstly, the dataset, consisting of 11 parameters, may have limitations in representing the diverse population affected by breast cancer. The information provided in the dataset may not encompass all potential factors influencing breast cancer, and there could be additional variables contributing to the complexity of the disease that are not included also this research hasn't capture the vast population yet. Secondly, the model development and evaluation are based on machine learning algorithms applied to the available data. While these algorithms can provide valuable insights, they may not capture the full complexity of breast cancer and may have limitations in predicting individual cases accurately. Moreover, the predictive performance of the developed model should be interpreted with caution, as it might not generalize well to new and unseen data. Additionally, the thesis acknowledges the absence of a detailed and diverse exploration of the socioeconomic, environmental, or genetic factors that could play a role

in breast cancer early prevention. The focus on machine learning models may overlook the broader context in which breast cancer occurs, limiting the comprehensive understanding of the disease.

Reference

[1] Schneider, A. P., Zainer, C. M., Kubat, C. K., Mullen, N. K., & Windisch, A. K. (2014). The breast cancer epidemic: 10 facts. The Linacre Quarterly, 81(3), 244–277

[2] Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2023). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*, 1- 14.

[3] Yassin, N. I., Omran, S., El Houby, E. M., & Allam, H. (2024). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*, *156*, 25-45.

[4] Bayrak, E. A., Kırcı, P., & Ensari, T. (2023, April). Comparison of machine learning methods for breast cancer diagnosis. In 2023 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) (pp. 1-3). Ieee.

[5] Sharma, S., Aggarwal, A., & Choudhury, T. (2022, December). Breast cancer detection using machine learning algorithms. In *2022 International conference on computational techniques, electronics and mechanical systems (CTEMS)* (pp. 114-118). IEEE.

[6] Obaid, O. I., Mohammed, M. A., Ghani, M. K. A., Mostafa, A., & Taha, F. (2021). Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology, 7(4.36), 160-166.

[7] Bazazeh, D., & Shubair, R. (2020, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2020 5th international conference on electronic devices, systems and applications (ICEDSA) (pp. 1-4). IEEE.

[8] Gupta, M., & Gupta, B. (2019, February). A comparative study of breast cancer diagnosis using supervised machine learning techniques. In 2019 second international conference on computing methodologies and communication (ICCMC) (pp. 997-1002). IEEE.

[9] Osmanović, A., Halilović, S., Ilah, L. A., Fojnica, A., & Gromilić, Z. (2019). Machine learning techniques for classification of breast cancer. In World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague.

[10] Chaurasia, V., & Pal, S. (2020). Applications of machine learning techniques to predict diagnostic breast cancer. SN Computer Science, 1(5), 270.

[11] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, 150360-150376.

[12] Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J.,& Clarke, M. F. (2023). Prospective identification of tumorigenic breast cancer cells. Proceedings of the National Academy of ciences, 100(7), 3983–3988

[13] Sinn, H. P., & Kreipe, H. (2013). A brief overview of the WHO classification of breast tumors. Breast Care, 8(2), 149–154

[14] B.M.Gayathri, C.P.Sumathi, and Santhanam. Breast Cancer Diagnosis Using Machine Learning Algorithms –A Survey, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3

[15] G. Williams, "Descriptive and Predictive Analytics", Data Min. with Ratt. R Art Excav. Data Knowl. Discov. Use R, pp. 193-203, 2021.

[16] Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th international conference on electronic devices, systems and applications (ICEDSA) (pp. 1-4). IEEE.

[17] N. S. Ismail and C. Sovuthy, "Breast Cancer Detection Based on Deep Learning Technique," in 2019 International UNIMAS STEM 12th Engineering Conference (EnCon), Kuching, Malaysia: IEEE, Aug. 2019, pp. 89–92.

[18] Sharma, S., Aggarwal, A., & Choudhury, T. (2018, December). Breast cancer detection using machine learning algorithms. In 2018 International conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 114-118). IEEE.

[19] Kumar, P., Srivastava, S., Mishra, R. K., & Sai, Y. P. (2022). End-to-end improved convolutional neural network model for breast cancer detection using mammographic data. The Journal of Defense Modeling and Simulation, 19(3), 375-384.

[20] A. I. Pritom, Md. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," in 2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh: IEEE, Dec. 2016.

[21] M. Robin, J. John, and A. Ravikumar, "Breast Tumor Segmentation using U-NET," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India: IEEE, Apr. 2021.

[22] R. Almajalid, J. Shan, Y. Du, and M. Zhang, "Development of a Deep-Learning-Based Method for Breast Ultrasound Image Segmentation," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL: IEEE, Dec. 2018

[23] Kavitha R, Kannan E. An efcient framework for heart disease classifcation using feature extraction and feature selection technique in datamining. in: IEEE Int. Conf. on Emerging Trends in Engineering Technology and Science (ICETETS), 2022, pp 1–5.

[24] Uysal AK, Gunal S, Ergin S. The impact of feature extraction and selection on SMS spam fltering. Electronics and Electrical Engineering. 2013;19(5):67–72

Full report.pdf

ORIGINALITY REPORT

17%	14%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	6%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to George Bush High School Student Paper	1%
4	Sajal Chakraborti. "Handbook of Proteases in Cancer - Therapeutic Aspects", CRC Press, 2024 Publication	1%
5	www.researchgate.net Internet Source	1%
6	"Proceedings of 3rd International Conference on Smart Computing and Cyber Security", Springer Science and Business Media LLC, 2024 Publication	<1%
7	www.journaltocs.ac.uk Internet Source	<1%