A COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR EARLY BREAST CANCER DETECTION AND PROGNOSIS

In partial fulfillment of

FATHERS LOCKE AND STILLER RESEARCH AWARDS (LSRA)

MINI THESIS

LSRA CYCLE 5

Submitted to the

Department of Research

ST. XAVIER'S COLLEGE

Maitighar, Kathmandu, Nepal

By

RASHU SHRESTHA

Department of Computer Science

ST. XAVIER'S COLLEGE

KATHMANDU, NEPAL

ST. XAVIER'S COLLEGE
MAITIGHAR, KATHMANDU, NEPAL
Post Box : 7437
Contact: 4221365, 4244636
Email: ktm@sxc.edu.np
Website: www.sxc.edu.np

सेन्ट जेभियर्स कलेज
माईतीघर, काठमाडौं, नेपाल
पो.ब.नं. : ७४३७
फोन : ४२२१३६५, ४२४४६३६
ईमेल : ktm@sxc.edu.np
वेबसाइट : www.sxc.edu.np

CERTIFICATE OF APPROVAL

The head of research hereby makes known that ***Rashu Shrestha, B.Sc. 3rd Year, Computer Science Department, 019BSCIT026*** has been conferred the **LOCKE AND STILLER RESEARCH AWARD (LSRA) CYCLE 5 for her** research work embodied in this mini-thesis entitled **A comparative analysis of machine learning techniques for early breast cancer detection and prognosis** under the supervision and guidance of ***Er. Rajan Karmacharya from Computer Science Department*** for the full period prescribed by the LSRA-research office of St. Xavier's College, Maitighar, Kathmandu, Nepal.

Department of Computer Science
Place: St. Xavier's College, Kathmandu
Date:

Mr. Niraj Nakarmi
Research Head

ST. XAVIER'S COLLEGE
MAITIGHAR, KATHMANDU, NEPAL
Post Box : 7437
Contact: 4221365, 4244636
Email: ktm@sxc.edu.np
Website: www.sxc.edu.np

सेन्ट जेभियर्स कलेज
माईतीघर, काठमाडौं, नेपाल
पो.ब.नं. : ७४३७
फोन : ४२२१३६५, ४२४४६३६
ईमेल : ktm@sxc.edu.np
वेबसाइट : www.sxc.edu.np

RECOMMENDATION FOR APPROVAL

This is to certify that ***Rashu Shrestha, B.Sc. CSIT 3rd year, Department of Computer Science, 019BSCIT026*** has carried out the **FATHERS LOCKE AND STILLER RESEARCH AWARD (LSRA)** research work embodied in this mini-thesis under the supervision and guidance of ***Er. Rajan Karmacharya, Department of Computer Science*** for the full period prescribed by the LSRA-research office of St. Xavier's College, Maitighar, Kathmandu, Nepal.

We approve this mini-thesis entitled ***"A comparative analysis of machine learning techniques for early breast cancer detection and prognosis"*** for submission for the **FATHERS LOCKE AND STILLER RESEARCH AWARD (LSRA) CYCLE 5.**

Department of Computer Science
Place: St. Xavier's College, Kathmandu
Date:

Mr. Ganesh Yogi
Head of Department
St. Xavier's College, Kathmandu

Er. Sarjan Shrestha
LSRA- Department Research coordinator
St. Xavier's College, Kathmandu

ST. XAVIER'S COLLEGE
MAITIGHAR, KATHMANDU, NEPAL
Post Box : 7437
Contact: 4221365, 4244636
Email: ktm@sxc.edu.np
Website: www.sxc.edu.np

सेन्ट जेभियर्स कलेज
माईतीघर, काठमाडौं, नेपाल
पो.ब.नं. : ७४३७
फोन : ४२२१३६५, ४२४४६३६
ईमेल : ktm@sxc.edu.np
वेबसाइट : www.sxc.edu.np

CERTIFICATE OF RECOMMENDATION

This is to certify that this mini-thesis entitled **"*A comparative analysis of machine learning techniques for early breast cancer detection and prognosis*"** submitted by ***Rashu Shrestha, B.Sc. 3rd Year, Computer Science Department, 019BSCIT026*** for the **FATHERS LOCKE AND STILLER RESEARCH AWARD (LSRA) CYCLE 5,** is a record of research work done by the candidate from ***(July, 2022)*** to ***(March, 2023)*** at the Department of ***Computer Science,*** St. Xavier's College, Maitighar, Kathmandu, Nepal, under my supervision and guidance and has not previously been submitted for any degree, diploma or project work.

I recommend this mini-thesis for submission.

Department of Computer Science
Place: St. Xavier's College
Date:

Er. Rajan Karmacharya
LSRA Advisor

DECLARATION

I, ***Rashu Shrestha, B.Sc. 3rd Year, Computer Science Department, 019BSCIT026*** hereby declare that the work presented in the mini thesis entitled **"*A comparative analysis machine learning techniques for early breast cancer detection and prognosis*"** is a record of independent research work done by me from ***(July, 2022)*** to ***(March, 2023)*** at ***Computer Science Department,*** St. Xavier's College, Maitighar, Kathmandu, Nepal, under the supervision and guidance of ***Er. Rajan Karmacharya***, St. Xavier's College, Maitighar, Kathmandu, Nepal. This work or part of it has not previously been submitted for any degree, diploma or project work.

I submit this mini-thesis for the **LOCKE AND STILLER RESEARCH AWARD (LSRA) cycle 5.**

Department of Computer Science
Place: St. Xavier's College, Kathmandu
Date:

Rashu Shrestha
Candidate

ACKNOWLEDGEMENTS

I would want to take this chance to express my sincerest gratitude to everyone who helped me throughout my research and supported me to devise this mini-thesis.

It brings me great pleasure to express my sincere gratitude and heartfelt appreciation to my highly respected and esteemed supervisor, **Er. Rajan Karmacharya, Chief Technology Officer, Lecturer, Department of Computer Science**, for his valuable direction, inspiration, and assistance in completing this research. I truly thank him for his helpful recommendations and cooperative attitude during this entire project. I would also like to express our gratitude towards **Mr. Ganesh Yogi**, **Head of Department** for his constant support.

I also want to thank **Er. Sarjan Shrestha, Research Coordinator, Lecturer, Department of Computer Science and Mr. Niraj Nakarmi, Head of Research** for helping me with the research and documentation. I thank my lecturers for their unwavering support and direction. Additionally, I would like to express our sincere gratitude to **Dr. Bikash Bhaila, Surgeon, Bhaktapur Cancer Hospital** for his support and guidance. In the end, I would want to extend my sincere gratitude to all of my friends, my research companions and everyone who supported and assisted me in some way during this project.

RASHU SHRESTHA
DEPARTMENT OF COMPUTER SCIENCE

ABSTRACT

Breast cancer is the leading factor causing cancer-related deaths among women worldwide. Early detection is crucial for effective treatment and lower fatality rates. Recent advances in machine learning have shown promising results in predicting breast cancer. This study uses the Breast Cancer Wisconsin (Diagnostic) dataset for training and the Bhaktapur Cancer Hospital dataset for testing to compare the performance of three well-known classification algorithms: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). The data is preprocessed by scaling and partitioning. The evaluation of each algorithm's performance on the testing dataset involves measuring its accuracy, precision, recall, and F1-score metrics followed by training on the training dataset. The results reveal that SVM and Random Forest perform better than Logistic Regression in terms of accuracy and F1-score. The study demonstrates that machine learning algorithms, especially SVM and Random Forest, can accurately predict breast cancer, which can aid in early detection and improve patient outcomes. These models can assist doctors in identifying breast cancer more precisely, leading to fewer unnecessary biopsies and better patient outcomes. The findings of this research can be applied in clinical settings for early detection and treatment of breast cancer, with significant implications for improving breast cancer diagnosis and patient outcomes.

Keywords: *Breast cancer, Machine learning, Classification algorithms, Support Vector Machine, Logistic Regression, Random Forest, Breast Cancer Wisconsin dataset, Clinical settings, Biopsies.*

TABLE OF CONTENTS

LISTS OF TABLES

LISTS OF FIGURES

ABBERVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area under the ROC Curve
BC	Breast Cancer
BCCD	Breast Cancer Coimbra Dataset
BDT	Bagged Decision Tree
CAD	Computer Aided Diagnostics
DCIS	Ductal Carcinoma in Situ
DL	Deep Learning
DT	Decision Trees
ELM	Extreme Learning Machines
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GLM	General Linear Model
IBC	Inflammatory Breast Cancer
IDC	Invasive Ductal Carcinoma
K-NN	K-Nearest Neighbors
LBC	Lobular Breast Cancer
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MBC	Mucinous Breast Cancer
ML	Machine Learning
MTBC	Mixed Tumors Breast Cancer
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RF	Random Forest
RFC	Random Forest Classifier
ROC	Receiver Operating Characteristic
SEER	Surveillance, Epidemiology, and End Results
SVM	Support Vector Machine

TN	True Negative
TP	True Positive
TPR	True Positive Rate
WAUCE	Weighted Area under the Receiver Operating Characteristic Curve Ensemble
WBCD	Wisconsin Breast Cancer Diagnostic

CHAPTER 1
INTRODUCTION

1.1 Background

In developing nations, breast cancer is a serious issue within terms of public health. According to the American Cancer Society (Giri et al., 2018), the United States (US) will see about 1,735,350 new cases of cancer and 609,640 cancer-related fatalities in 2018, including 266,120 new instances of invasive breast cancer. In Nepalese women, breast cancer ranks as the second most prevalent cancer.

While breast cancer is a significant burden on the Nepalese healthcare system, little is known about the number of women who have this disease. Approximately 1,700 new instances of breast cancer were detected in Nepal in 2012, with an age-standardized rate (ASR) of 13.7 new cases per 100,000 men, while 870 women died, with an ASR of 7.2 deaths per 100,000 women, according to GLOBOCAN, 2012 (GLOBOCAN, 2012).

In the world, 25% of all cancers in women are breast cancer, making it a serious health concern. Breast cancer is the most prevalent cancer among women worldwide, according to the World Health Organization, with an estimated 2.3 million new cases expected to be diagnosed in 2020 alone. Effective treatment and higher survival rates for breast cancer depend on early detection and a precise diagnosis. Mammography, ultrasound, biopsy, and physical examination have historically been used to diagnose breast cancer (“Cancer Facts & Figures”, 2022).

1.1.1 Breast Cancer

Breast cancer originated through malignant tumors, when the growth of the cell got out of control (Lu et al., 2018). Breast cancer is brought on by the abnormal proliferation of numerous fatty and fibrous breast tissues. Breast cancer develops from malignant tumors that develops when cells expand uncontrollably (Zhang et al., 2021). A lump or tumor is often the outcome of this disease which is a collection of conditions in which cells in the breast tissue of a person alter and divide out of

control. The milk glands (lobules) or the tubes (ducts) connecting the milk glands to the nipple are where most breast cancers start (“Cancer Facts & Figures”, 2022). Different stages of cancer are caused by the spread of cancer cells throughout tumors. There are different types of breast cancer which occur when affected cells and tissues spread throughout the body.

1.2 Machine Learning

Medical imaging research now heavily relies on machine learning (ML). Over time, ML techniques have changed from manually seeded inputs to automatically initialized models. The science of machine learning (ML) has advanced, and as a result, computer-aided diagnostic (CAD) systems have become more intelligent and self-sufficient. Deep feature learning and representations-based automated approaches are becoming more and more prevalent. Recent developments in deep learning (DL), also known as ML with deeper and extensive representation methodologies, have significantly improved the diagnostics capabilities of CAD systems (Gardezi et al., 2019).

For the detection and prediction of breast cancer, machine learning algorithms have emerged as a promising option. These algorithms can examine huge, complex datasets to find patterns and make highly precise predictions about what will happen. Early detection, a reduction of pointless biopsies, and better patient outcomes can all be attributed to the application of machine learning algorithms in breast cancer prediction.

Due to their capacity for handling sizable and complicated datasets, Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) have been among the machine learning algorithms most frequently utilized in the prediction of breast cancer.

This paper presents supervised learning algorithms for predicting breast cancer. Through the methodology presented in paper, it has compared three different approaches of prediction: Support Vector Machine, Logistic Regression and Random Forest Classification.

1.3 Hospital Data

Hospital data is the most crucial part of the research. Data of a patient who has done mammography is collected to train and test the model. Data consisting of both positive and negative results are included. Data of 10 real valued attributes computed for each nuclei is purposed to be used to train the model. The mean, standard error and worst or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 attributes.

i. radius (mean of distances from center to points on the perimeter)
ii. texture (standard deviation of gray-scale values)
iii. perimeter
iv. area
v. smoothness (local variation in radius lengths)
vi. compactness
vii. concavity (severity of concave portions of the contour)
viii. concave points (number of concave portions of the contour)
ix. symmetry
x. fractal dimension

1.4 Problem Statement

Breast cancer is the most prominent cause of death in the context of underdeveloped countries that lack awareness of cancer related issues. Some of the problems of breast cancer according to the current evidence are:

1. Delays in cancer care in low- and middle-income countries are caused by access issues and poor quality of care. In these nations, where it is most necessary for the development of cost-effective public policies that strengthen health systems to combat this costly and deadly disease, research on specific access barriers and deficiencies in the quality of care for the early diagnosis and treatment of breast cancer is essentially nonexistent.
2. These mammography tests are both time consuming, risky and hectic which directly or indirectly impact on the patient’s psychology.
3. This tests upsurges the patient's time in the hospital and tests are very expensive, it increases the cost of diagnosis by a significant amount.

4. In the context of Nepal, where healthcare and high speed transportation is not available everywhere, this can be a situation of death for many individuals.

1.5 Research Objective

The main objectives of this research is to build a model that can be able:

1. To compare the performance of three machine learning algorithms: Support Vector Machine, Random Forest, and Logistic Regression and suggest the algorithm with highest accuracy for breast cancer prediction.

2. To help health professional minimize the false negative cases in breast cancer prediction by building a system to predict breast cancer using machine learning algorithms.

1.6 Scope of the research

Although research has been done to predict and cure the breast cancer in developed countries. In underdeveloped country, like Nepal still breast cancer is main cause of female mortality rate. Only few people who have access to good education and are aware are aware of the breast cancer. Other women in the underdeveloped places are more likely to find about the breast cancer at the last stage causing breast cancer to have high mortality rate. Using machine learning techniques and comparing them in order to find the best and efficient algorithm for the early breast cancer detection even in the developing country is the aim of this paper.

The prognosis for breast cancer can be predicted using machine learning (ML), which enables us to identify relationships between prognostic markers. Research on new algorithms and doing comparative study to test their accuracy by considering the factors of underdeveloped countries is to be carried out in this paper. This will allow us to develop more efficient method for breast cancer detection in Nepal. This research study focuses specifically on implementation of machine learning algorithms in early breast cancer detection. It can help health professionals in minimizing the false negative cases, which is common because there are very few highly skilled health professionals and high volume of cases. It can be used in hospital or in clinics

as an expert system assisting doctors and radiologist for detecting early stages of breast cancer which will be more effective.

CHAPTER 2
LITERATURE REVIEW

2.1 Breast Cancer and its types

Breast cancer, which can be classed as benign or malignant, is the most prevalent invasive cancer in women and the second leading cause of cancer death in females. In recent years, researchers' interest in breast cancer research and prevention has increased (Li, 2018). Breast cancer develops from malignant tumors that develop when cells expand uncontrollably (Lu et al., 2018). Invasive Ductal Carcinoma (Pervez & Khan, 2007), Lobular Breast Cancer (Masciari et al., 2007), Mucinous Breast Cancer (Gradilone et al., 2011), Inflammatory Breast Cancer (Robertson et al., 2010), Mixed Tumors Breast Cancer (Lee et al., 2017) and Ductal Carcinoma in Situ (Hou et al., 2020) are the different types of cancers prevalent in women.

Ductal Carcinoma in Situ (DCIS) is a type of breast cancer that occurs when abnormal cells spread outside the breast. It is also known as the non-invasive cancer (Hou et al., 2020). The second type is Invasive Ductal Carcinoma (IDC) also known as infiltrative ductal carcinoma which occurs when the abnormal cells of breast spread over all the breast tissues and IDC cancer is usually found in men (Pervez & Khan, 2007). Mixed Tumors Breast Cancer (MTBC) is the third type of breast cancer, also known as invasive mammary breast cancer. Abnormal duct cells and lobular cells cause such kinds of cancer (Lee et al., 2017).

The fourth type of cancer is Lobular Breast Cancer (LBC) (Masciari et al., 2007) which occurs inside the lobule. It increases the chances of other invasive cancers. Mucinous Breast Cancer (MBC) (Gradilone et al., 2011)is the fifth type that occurs because of invasive ductal cells, also known as colloid breast cancer. It occurs when the abnormal tissues spread around the duct. Inflammatory Breast Cancer (IBC) is the last type that causes swelling and reddening of breast (Robertson et al., 2010).

2.2 Situation of Breast Cancer in Nepal

Breast cancers are typically found in advanced stages in underdeveloped nations like Nepal and patients may not receive proper therapy, pain management, or palliative care. Socioeconomic divisions and a lack of financial resources are barriers to Nepal's initiatives to prevent breast cancer.

In Nepal, Bhaktapur Cancer Hospital has recorded a significant increase in the number of breast cancer cases over the past two years. According to data collected in 2020, Bhaktapur Cancer Hospital alone had a total of 260 breast cancer patients, which included 1 male and 259 females. This number increased the following year, with a total of 405 breast cancer patients, including 9 males and 396 females (Bhaktapur Cancer Hospital, 2023).

Breast cancer is the most common type of cancer among women worldwide, and early detection and treatment are crucial for improving survival rates. The increase in breast cancer cases may be due to various factors such as lifestyle changes, environmental factors, and genetic predisposition. People should be aware of the risk factors for breast cancer and take precautions to lower their risk, especially women. Mammograms, clinical breast exams, and routine self-examinations of the breast are important in the early detection techniques.

The chance of developing breast cancer can also be decreased by living a healthy lifestyle that includes regular exercise, a balanced diet, moderate alcohol intake, and quitting smoking.

2.3 Supervised Machine Learning Algorithms

Machine learning (ML) is a subset of artificial intelligence (AI) that has emerged as an effective technique for dealing complicated datasets. It gives computers the ability to learn from the data and extract insightful information that may not be simple for humans to understand (Great Learning Team, 2023). Machine learning has grown in popularity as a result of the number of datasets available, and numerous sectors have begun to use it to glean insightful information from their data. Making it possible for

machines to learn from data, spot patterns, and take actions depending on the knowledge available is one of the fundamental goals of machine learning.

Supervised Machine Learning algorithm has demonstrated promising result in disease prediction using health data in recent years. These algorithms can determine risk factors and forecast the likelihood of illness onset by evaluating vast volumes of health data. Better health outcomes for individuals and communities can result from early disease detection and prevention (Linardatos et al., 2020).

Algorithms for supervised machine learning consist of a function with X as an input and y as an output, where y = f (X). The algorithm's objective is to map a function to the provided data that fits it the best. Based on the features gathered, the dataset for supervised learning always has the desired value (Nasteski, 2017). The value that the algorithm is supposed to anticipate is known as the target value. There are two different categories of supervised algorithms:

a) **Classification**

   The problem has an output value in a category such as Yes, No or Doing, Halted, Not doing which is from a set of value.

b) **Regression**

   The problem where output value is real value such as mass, length, quantity etc.

2.3.1. Logistic Regression

In machine learning, the popular supervised learning technique known as logistic regression is used to address categorization issues. It is a statistical model that examines the link between one or more independent variables and a dependent variable (outcome).The chance of the event occurring serves as the dependent variable in logistic regression, which might be binary or categorical. Any input value is converted to a value between 0 and 1 using the logistic function used in logistic regression to model the probability of the result.

The maximum likelihood estimation technique is used by the logistic regression algorithm to calculate the parameters of the model. The likelihood of the outcome for brand-new, untested data can be predicted using the model once it has been trained on

the training data. Binary classification problems (where the outcome is either 0 or 1) and multi-class classification problems can both be solved using logistic regression (where the outcome can take more than two values). Several binary logistic regression models are trained in the case of multi-class classification, and the model with the highest probability is selected as the projected outcome (Jason Brownlee, 2016).

2.3.2 Support Vector Machine

A powerful supervised learning method in machine learning, is used to solve classification and regression issues. Based on the properties of the data points, a binary linear classifier divides them into various classes. The SVM algorithm locates the hyper plane in the feature space that best divides the data points of various classes. The margin between the classes is maximized by selecting the hyper plane in a way that maximizes the distance between the hyper plane and the closest data points from each class. The term "support vectors" refers to these data points.

A kernel method that translates the input data to a higher dimensional feature space where it is more likely to be linearly separable allows SVM to handle non-linearly separable data as well. Commonly used kernels include linear, polynomial, and radial basis functions (RBF). SVM can handle high-dimensional data, efficiency in handling small sample numbers, and robustness against over fitting (Mustafa Abdullah & Mohsin Abdulazeez, 2021).

2.3.3 Random Forest Classification

An effective supervised learning approach in machine learning for classification issues is called random forest classification. It uses an ensemble learning technique to merge different decision trees to build a more reliable and precise model. A huge number of decision trees are built using various randomly selected subsets of the training data and characteristics in a Random Forest Classification technique. For each decision tree, the algorithm chooses the attributes that offer the optimal split. Combining all the decision trees' predictions, either by majority vote or weighted average, yields the final forecast.

The Random Forest Classification can handle high-dimensional data, the ability to deal with missing values and outliers, and the capacity to reveal the relative relevance of the features employed in the classification (Biau & Scornet, 2016).

2.4 Related Works

A number of researchers have explored the feasibility of predicting early breast cancer by examining data mining, machine learning, and hybrid mining systems as potential learning systems. Most research papers in machine learning use existing algorithms, but some refine and customize them to better suit their dataset and objectives.

Recent developments in deep learning (DL), also known as Machine Learning (ML) with deeper and extensive representation methodologies, have significantly improved the diagnostics capabilities of Computer Aided Design (CAD) systems. A study conducted in 2019 evaluated traditional machine learning methods and found that it had limited application in clinical analysis. However, the authors suggested that deep learning methods had promising potential for use in computer-aided diagnosis systems, which could improve the accuracy of breast cancer diagnosis (Gardezi et al., 2019).

A study was conducted in 2018 that involved 500 women residing in the Kathmandu Valley, and the authors found that only 3.4% of them received mammography every two years, while 7.2% underwent clinical breast exams annually, and 14.4% performed breast self-examinations monthly. The research showed that women with high subjective norms, high perceived behavioral control, and a positive outlook were more inclined to use all three screening methods (Bhandari et al., 2021).

Using blood analysis data that included age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistance, and MCP1, a 2018 study wanted to identify breast cancer using various machine learning techniques, including Artificial Neural Network (ANN), Extreme Learning Machines (ELM), K-Nearest Neighbor (K-NN), and SVM. According to the studies, ELM had an accuracy rate of 80%, which was higher than ANN's 79.4304%, k-NN’s 77.5%, and SVM's 73.5% (Celik et al., 2018).

A study was conducted where the authors reviewed 14 research papers on breast cancer to gain knowledge on its types, symptoms, and causes in 2020. The study explored various machine learning techniques, including ensemble and deep learning, and provided detailed explanations of the algorithms used for breast cancer prediction (Fatima et al., 2020).

In 2018, a study examined how to detect breast cancer using different machine learning algorithms, including Random Forest, Naive Bayes, SVM, and k-NN. The results showed that SVM was the most successful algorithm, with a 97.9% accuracy rate (Khourdifi & Bahaj, 2018).

A study conducted on predicting breast cancer risk by combining demographic risk variables with a set of interrelated genetic variants using machine learning in 2020. It was found that by using interacting genetic and familial history features, it was able to predict breast cancer risk with a higher mean average precision (77.78) compared to using only familial history features (74.19) or combination of interacting genetic variants (73.65). Similarly, using interacting genetic and oestrogen metabolism also resulted in a higher mean average precision (78.00) compared to a system based solely on oestrogen metabolism features (72.57) (Behravan et al., 2020).

In 2020, a study comparing five supervised machine learning techniques for breast cancer prediction, namely Support Vector Machine, K-Nearest Neighbors, Random Forest, Artificial Neural Networks, and Logistic Regression was conducted. The authors found that ANNs had the highest accuracy, precision, and F1 score of 98.57%, 97.82%, and 0.9890, respectively, while SVM had an accuracy of 97.14%, precision of 95.65%, and F1 score of 0.9777 (Islam et al., 2020).

In 2019, a study presented a hybrid method for diagnosing breast cancer by first utilizing Linear Discriminant Analysis (LDA) to reduce the dimensionality of features, and then using the resulting reduced feature dataset to apply Support Vector Machine. In this study, it was found that the proposed approach had an accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07%, and area under the receiver operating characteristic curve of 0.9994 (Omondiagbe et al., 2019) .

In 2020, a study used Decision Tree Classifier and Logistic Regression for breast cancer prediction and compared their accuracies. The authors found that Decision Tree Classifier was the best-suited algorithm for prediction, as it had a higher pinpoint prediction accuracy compared to Logistic Regression (Sengar et al., 2020).

A study examines the performance of five different classification models, including Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), and Logistics Regression (LR), and uses these models to classify two different datasets related to breast cancer: the Wisconsin Breast Cancer Database and the Breast Cancer Coimbra Dataset (WBCD). The random forest model can outperform and adapt better than the other four techniques, according to comparative experiment research (Li, 2018).

A research team that carried out a study on the automated diagnosis of breast cancer using machine learning techniques in 2019. The purpose of this work was to use genetic programming and machine learning approaches to create a system that could accurately distinguish between benign and malignant breast cancers. The results of the study showed that genetic programming may efficiently integrate feature preprocessing methods and classifier algorithms to automatically select the best model (Dhahri et al., 2019).

According to a 2018 study, clinical traits such age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, and MCP-1 might be used to predict whether or not breast cancer will be present. Using various combinations of predictor variables, the study used a variety of machine learning algorithms, including logistic regression, random forests, and support vector machines. Using the Monte Carlo Cross-Validation technique, the models' sensitivity, specificity, and AUC were assessed, and 95% confidence intervals were generated. Using glucose, resistin, age, and BMI as predictors, the study discovered that support vector machine models had high sensitivity ranging between 82 and 88%, specificity ranging between 85 and 90%, and an AUC with a 95% confidence interval of [0.87, 0.91]. The study concludes that these predictors and support vector machine models can accurately predict the development of breast cancer in women (Patrício et al., 2018).

A study was carried out in 2019 to predict breast cancer utilizing the most recent advancements in CAD systems and associated methods. Their major goal was to use machine learning algorithms trained on pertinent data to assess whether or not a person has breast cancer (Yarabarla et al., 2019).

Using several machine learning model parameters, a study in 2020 was conducted to predict breast cancer. K-Nearest Neighborhood, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine with Radial Basis Function kernel were six supervised machine learning methods employed in the study. Moreover, Adam Gradient Descent Learning, which combines the advantages of the adaptive gradient approach with root mean square propagation, was used for deep learning. To increase accuracy both within the model and when comparing models, each model was changed using unique hyper parameters. The deep learning model achieved an accuracy rate of 98.24% while producing the least amount of loss (Gupta & Garg, 2020).

Using Optimal Ensemble Learning, a computer-aided diagnosis system for predicting breast cancer recurrence was developed in 2017. In order to pick discriminating factors from the clinic pathologic characteristics of 579 breast cancer patients, the researchers used statistical feature selection approaches. Particle Swarm Optimization (PSO) was used as an input for the ensemble learning classification system to sharpen the features (Bagged Decision Tree: BDT). Age at diagnosis, tumor size, lymph node involvement ratio, number of affected axillary lymph nodes, and expression of the progesterone receptor, hormone therapy use, and kind of surgery were the features that were considered. During all cross-validation folds and the hold-out test fold, the system achieved a minimum sensitivity, specificity, precision, and accuracy of 77%, 93%, 95%, and 85%, respectively (Mohebian et al., 2017).

A study in 2018 used an ensemble algorithm based on support vector machines to diagnose breast cancer (SVM). The Weighted Area under the Receiver Operating Characteristic Curve Ensemble (WAUCE) technique, which integrates twelve different SVMs, was used in the study to increase the accuracy of breast cancer diagnosis. The findings demonstrated that the WAUCE model beat two commonly

used ensemble models (adaptive boosting and bagging classification tree) and five other ensemble processes in terms of accuracy and variance for the diagnosis of breast cancer. The suggested WAUCE model outperformed the best single SVM model on the SEER dataset by 33.34% and decreased variance by 97.89% (Wang et al., 2018).

Using descriptions of the nuclei in breast tumors, a study conducted in 2019 showed how well machine learning techniques work for diagnosing cancer. Single-layer artificial neural networks, Support Vector Machines (SVMs) with radial basis functions, and regularized General Linear Model (GLMs) regression were utilized as the three prediction models. For categorizing cell nuclei, the trained algorithms showed high specificity (.99), sensitivity (.99), and accuracy (.94–.96). (.85-.94). The SVM method produced the highest area under the curve (.96) and accuracy (.96). (.97). The prediction performance was marginally enhanced when integrated into a voting ensemble (accuracy =.97, sensitivity =.99, specificity =.95) (Sidey-Gibbons & Sidey-Gibbons, 2019).

In 2017, a study was done on the risk factors and prevention of breast cancer. This study found breast cancer stem cells, elucidated the mechanisms and causes of tumor medication resistance, and identified a number of breast cancer-related genes and emphasizes the accessibility of recently established chemoprevention and biological prevention strategies, which raise the standard of living for breast cancer patients. The findings shed light on the pathogenesis of breast cancer, as well as its linked genes, risk factors, and recently proposed preventative strategies. Even though the improvements are small, they still show success in the ongoing fight against breast cancer (Sun et al., 2017).

Deep learning was used to predict the Estrogen Receptor (ER) status in breast cancer data by a research team in 2017. In their study they showed the benefits of employing deep learning to categorize ER state based on metabolomics which achieved the highest prediction accuracy with an AUC of 0.93, Furthermore, the deep learning method allowed for a greater comprehension of the biology of the disease (Alakwaa et al., 2018).

A study on the prevalence of breast cancer in Nepal was carried out in 2018. The results demonstrated that in underdeveloped nations like Nepal, where women might not receive proper therapy, pain management, or palliative care, breast cancer is frequently discovered at an advanced stage. Socioeconomic differences and a lack of funding make it difficult to prevent breast cancer in Nepal. The study gives a comprehensive review of Nepal's screening and treatment options, as well as the prevalence of breast cancer there. The study also emphasizes Nepali women's current awareness of breast cancer and related prevention strategies (Giri et al., 2018).

CHAPTER 3
RESEARCH METHODOLOGY

The methodology used in given research for breast cancer prediction and prognosis is given as below.

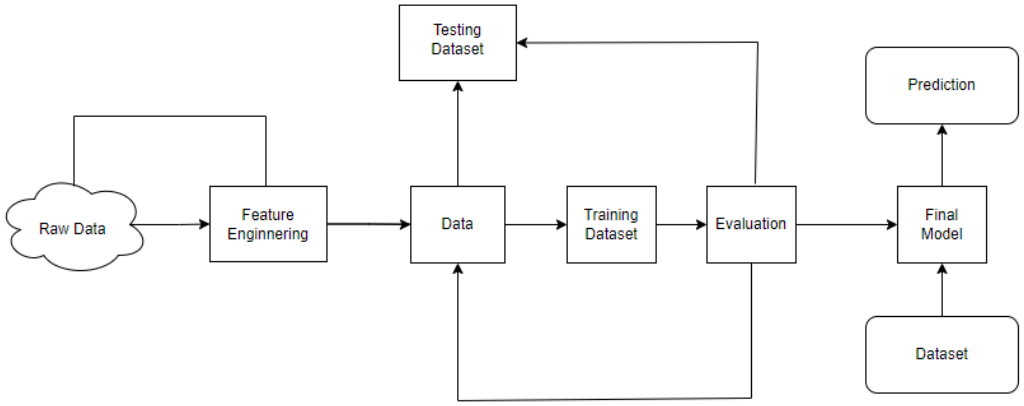


Figure 1: Methodology Used

3.1 Machine Learning Algorithms

The development of analytical models is automated by a data analysis method known as machine learning. It is an area of artificial intelligence based on the idea that machines are capable of learning from data, identifying patterns, and making decisions with minimal human input (Machine Learning: What It Is and Why It Matters., 2023). Machine learning is used to develop trends that assist computers in understanding data and rendering fact-based decisions. In the upcoming years, particularly in 2023 and 2024, this technology is probably going to spread. According to sources, 35% of businesses claim to be employing AI in their operations (Sarvesh, 2023).

3.2 Data Collection

The data collection and processing involves analyzing past papers and collecting data from various patients. Batch learning, a method that divides training sets into smaller groups, is proposed due to the diverse data features. The Wisconsin Breast Cancer Diagnostic (WBCD) dataset is a widely used dataset for breast cancer prognosis

studies. There are 569 samples total in the collection, 212 of which are cancerous, and 357 of which are benign (UCI Machine Learning, n.d.). The proposed machine learning algorithms will be employed on WBCD dataset and test the model on the data collected from Bhaktapur Cancer Hospital.

The collected data has 31 columns with 30 attributes and 1 target variable. They are:

Table 1: List of Data Attributes

Diagnosis/ Target Variable	worst smoothness
mean radius	worst compactness
mean texture	worst concave points
mean perimeter	worst symmetry
mean area	worst fractal dimension
mean smoothness	worst radius
mean compactness	worst perimeter
mean concavity	smoothness error
mean concave points	compactness error
mean symmetry	concavity error
mean fractal dimension	concave points error
radius error	symmetry error
texture error	fractal dimension error
perimeter error	worst texture
area error	worst area
worst concavity	

3.3 Data Distribution

The dataset is divided into malignant (cancerous) and benign (non-cancerous) attributes as the target variable.

3.3.1 Malignant and Benign

The training dataset and testing dataset is divided into malignant (0) and benign (1). The dataset malignant and benign is taken as the target variable and output shown after the testing the model and also predicts the accuracy of the model. The training

dataset consists of 357 benign tumors and 212 malignant tumors. The testing dataset consists of 76 benign tumors and 51 malignant tumors.

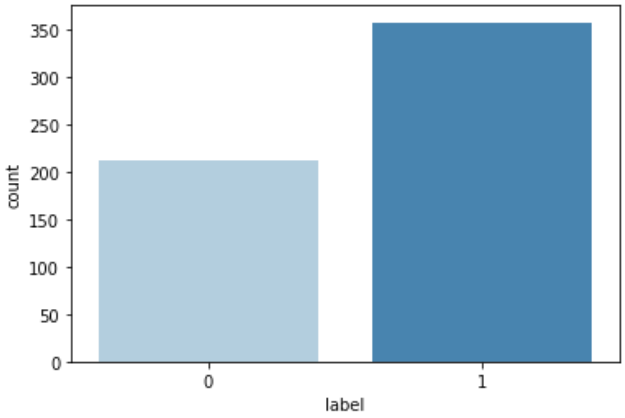


Figure 2: Training Dataset from WBCD

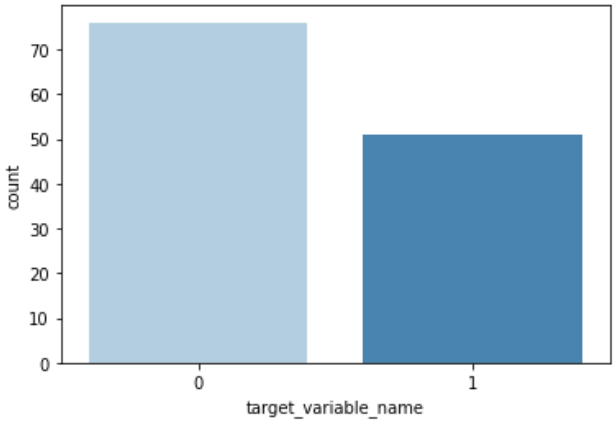


Figure 3: Testing Dataset from data collection

3.4 Feature Engineering

The gathered raw data is put through feature engineering, which includes feature extraction and feature selection. By extracting a distinct collection of attributes from the input, feature selection lowers the complexity of the machine learning model and expedites training. Additionally, it addresses the over-fitting issue and raises the model's accuracy. Feature extraction is used to create new features from the original ones after the feature selection process, producing a summarized dataset with a less number of features (Sarker, 2021). The dataset is divided into features and target in which features contains the attributes and the label contains the target variables. The following table shows the mean of the malignant and benign dataset for particular attribute.

label	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture
0	17.462830	21.604906	115.365377	978.376415	0.102898	0.145188	0.160775	0.087990	0.192909	0.062680	...	21.134811	29.318208
1	12.146524	17.914762	78.075406	462.790196	0.092478	0.080085	0.046058	0.025717	0.174186	0.062867	...	13.379801	23.515070

2 rows × 30 columns

Figure 4: Data labeling for training dataset

target_variable_name	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius
0	16.451842	21.910921	108.664868	865.009211	0.104135	0.144650	0.152717	0.081679	0.196263	0.063670	...	19.902368
1	11.549627	17.883529	74.157843	420.611765	0.095868	0.080788	0.048362	0.025884	0.175853	0.064201	...	12.741902

2 rows × 30 columns

Figure 5: Data labeling for testing dataset

3.5 Data Cleaning and Splitting

Before feeding the raw data from hospitals and medical facilities into a machine learning algorithm, it is necessary to clean the data. This involves digitizing paper-based data and sorting through the features to include essential ones and eliminate undesirable ones from the domain. The cleansed dataset is then split into a training set and then the testing dataset is done similarly and formulated a test set to test the model's accuracy. The machine model is trained using the training set, and its accuracy is evaluated using the test set.

3.5.1 Training Dataset

The training dataset was taken from Breast Cancer Wisconsin (Diagnostic) that contains 569 training dataset among which 357 are benign (1) and 212 are malignant (0) tumors as shown in the graph below:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	26.50	98.87	567.7
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	26.40	166.10	2027.0
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	38.25	155.00	1731.0
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	34.12	126.70	1124.0
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	39.42	184.60	1821.0
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	30.37	59.16	268.6

Figure 6: Feature Training Dataset

```
print(Y_train)
```

```
0      0
1      0
2      0
3      0
4      0
      ..
564    0
565    0
566    0
567    0
568    1
Name: label, Length: 569, dtype: int64
```

Figure 7: Target Training Dataset

3.5.2 Testing Dataset

The testing dataset was collected from Bhaktapur Cancer Hospital that contains 127 testing dataset among which are 76 malignant (0) and 51 are benign (1) tumors as shown in the graph below:

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	14.54	27.54	96.73	658.8	0.11390	0.15950	0.16390	0.07364	0.2303
1	14.68	20.13	94.74	684.5	0.09867	0.07200	0.07395	0.05259	0.1586
2	16.13	20.68	108.10	798.8	0.11700	0.20220	0.17220	0.10280	0.2164
3	19.81	22.15	130.00	1260.0	0.09831	0.10270	0.14790	0.09498	0.1582
4	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885
...	...	...	...	...	...	...	...	...	...
122	13.03	18.42	82.61	523.8	0.08983	0.03766	0.02562	0.02923	0.1467
123	14.99	25.20	95.54	698.8	0.09387	0.05131	0.02398	0.02899	0.1565
124	13.48	20.82	88.40	559.2	0.10160	0.12550	0.10630	0.05439	0.1720
125	13.44	21.58	86.18	563.0	0.08162	0.06031	0.03110	0.02031	0.1784
126	10.95	21.35	71.90	371.1	0.12270	0.12180	0.10440	0.05669	0.1895

Figure 8: Feature Testing Dataset

```
print(Y_test)
```

```
0      0
1      0
2      0
3      0
4      1
      ..
122    1
123    0
124    0
125    0
126    0
Name: target_variable_name, Length: 127, dtype: int64
```

Figure 9: Target Testing Dataset

3.6 Model Training

After cleaning the data, it is now prepared to be input into a machine model. To determine the best performing model, various machine learning models will be trained. During this stage, the sorted features from the training data are given weights to aid the model in predicting outcomes for unobserved data. The primary objective is to select the appropriate feature weight that minimizes overall error in the training set and generalizes it effectively. This stage addresses the issues of over fitting and under fitting in the model.

Machine Learning algorithms used in this papers are:

3.6.1 Logistic Regression

It is a statistical model that examines the link between one or more independent variables and a dependent variable (outcome). Any input value is converted to a value between 0 and 1 using the logistic function used in logistic regression to model the probability of the result. The data is filtered, preprocessed, grouped according to the target variable and then model is created and the training dataset is fitted into the model and the accuracy of the model is predicted using the testing dataset.

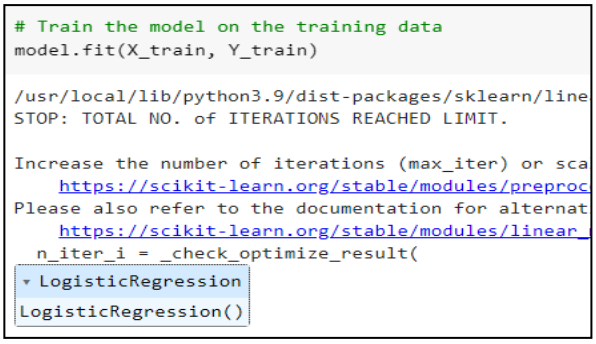


Figure 10: Logistic Regression

3.6.2 Support Vector Machine

The SVM algorithm locates the hyper plane in the feature space that best divides the data points of various classes and works on the basis of hyper planes. The preprocessed data is fed into the SVM linear kernel model and the fitted using the training dataset.

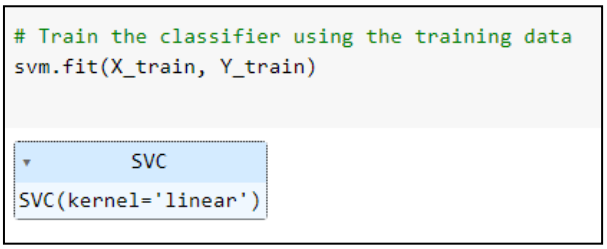


Figure 11: Support Vector Machine

3.6.3 Random Forest Classification

It uses an ensemble learning technique to merge different decision trees to build a more reliable and precise model. In RF model, the preprocessed data is classified and training dataset is fitted. The classification is done when n_estimators=1 and random_state=0.

```
# Train the model on the training data
rfc.fit(X_train, Y_train)
```

```
RandomForestClassifier
RandomForestClassifier(n_estimators=1, random_state=0)
```

Figure 12: Random Forest Classifiaction

3.7 Model Testing

Once the model has been trained on a set of training examples, it is moved on to the next step where it is evaluated using a test set that was not used in the training phase. The precision metric calculates the percentage of all positive forecasts that are true positive predictions, or cases that are actually positive. Recall is a metric that counts the percentage of positive instances that were really predicted to be positive. The F1-score, which is a harmonic mean of precision and recall, is helpful when evaluating models when both precision and recall are crucial. The evaluation involves calculating the accuracy, precision, and F1-measure of the model.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-Measure = 2×Precision×RecallPrecision+Recall

Where

True positive (TP): Prediction is +ve

True negative (TN): Prediction is -ve

False positive (FP): Prediction is +ve

False negative (FN): Prediction is –ve (Gaël Varoquaux, 2023).

3.8 Plotting Function

A graphing tool uses several examples from the train set, test set, and curves for easy comparison between these various approaches. In this paper, machine learning libraries NumPy as np, pandas as pd, seaborn as sns, matplotlib.pyplot as plt, then

from sklearn metrics import classification report, roc curve, and auc to plot the necessary curves. Using a consistent interface, the Sklearn offers a variety of effective methods for statistical modeling and machine learning, including classification, regression, clustering, and dimensionality reduction.

3.9 Deployment

Once the data is been collected and processed, the training dataset is fed into a preprocessing system, and the results are passed on to the classifier model. The model is currently in the learning phase. Following this, the test dataset is input, and the performance of the output is assessed. Standard metrics such as accuracy can be used to evaluate the performance.

CHAPTER 4
RESULTS

The study compares three algorithms: logistic regression, SVM, and random forest for breast cancer prediction based on performance metrics such as accuracy, sensitivity, specificity, F1 score, and ROC curve analysis. The dataset includes patient age, tumor size, tumor type, and malignancy indicators such as cell size and shape, tumor radius, and concavity. Preprocessing involved managing missing values and feature scaling to standardize the data. The models can predict if the cancer is malignant or benign based on the input data on the mentioned attributes.

After evaluating the performance of a binary classifier on a dataset consisting of 127 instances, a classification report is generated. This report provides metrics for precision, recall, and F1-score for both classes (Benign and Malignant) and also includes the relative support, which refers to the number of instances in each class.

Three prediction models were used to analyze breast cancer data: Logistic Regression, Support Vector Machine (SVM), and Random Forest. The SVM model had the highest average accuracy, while the Logistic Regression model had the lowest. The SVM model's ability to handle high-dimensional data and its nonlinear nature may have contributed to its better performance. Random Forest has slightly lower accuracy than SVM and may be prone to over fitting. Further findings and result are discussed as below:

4.1 Logistic Regression

Using Logistic Regression as a prediction model, the average accuracy was less than that of the SVM and Random Forest. This may be due to the linear nature of logistic regression, inability to capture complex relationships, and probable constraints in the dataset utilized for analysis are responsible for its poorer accuracy in breast cancer prediction compared to SVM and random forest.

```
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(Y_train, X_train_prediction)
print(f"Accuracy on training data is {training_data_accuracy*100:.2f} %")

Accuracy on training data is 94.73 %


# Calculate the accuracy of the model/ testing data
accuracy = accuracy_score(Y_test, predictions)
print(f"Accuracy of the model is {accuracy*100:.2f}%")

Accuracy of the model is 91.34%
```

Figure 13: Logistic Regression Accuracy

The figure below shows the prediction whether the dataset input is either malignant or benign for LR model. This helps us to predict the early stage breast cancer just by providing the attribute to the given model. Some of the sample data for validation of the LR model is represented as:

```
input_data = (17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4

# change the input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the numpy array as we are predicting for one datapoint
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
  print('The Breast cancer is Malignant')

else:
  print('The Breast Cancer is Benign')

[0]
The Breast cancer is Malignant
```

Figure 14: Prediction of Breast Cancer using LR (Malignant)

```
input_data = (11.52,18.75,73.34,409,0.09524,0.05473,0.03036,0.02278,0.192,0.05907,0.3249,0.9591,2.183,
# change the input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the numpy array as we are predicting for one datapoint
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
  print('The Breast cancer is Malignant')

else:
  print('The Breast Cancer is Benign')
```

```
[1]
The Breast Cancer is Benign
```

Figure 15: Prediction of Breast Cancer using LR (Benign)

4.2 Support Vector Machine

Using Support Vector Machine as a prediction model, the average accuracy was higher than that of Logistic Regression and Random Forest. SVM is non-linear, it can handle high-dimensional data, and it may have benefits in the dataset being analyzed maybe reasons for predicting breast cancer more accurately than logistic regression and random forest.

```
# Calculate the accuracy score of the model for training data
train_accuracy = accuracy_score(Y_train, Y_train_pred)
print(f"Training Accuracy: {train_accuracy*100:.2f}%")Z
```

```
Training Accuracy: 98.77%
```

```
# Calculate the accuracy score of the model for testing data
test_accuracy = accuracy_score(Y_test, Y_test_pred)
print(f"Testing Accuracy: {test_accuracy*100:.2f}%")
```

```
Testing Accuracy: 97.64%
```

Figure 16: Support Vector Machine Accuracy

The figure below shows the prediction whether the dataset input is either malignant or benign for SVM model. This helps us to predict the early stage breast cancer just by providing the attribute to the given model. Some of the sample data for validation of the SVM model are represented as:

```
input_data = [17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4,0.006399,

# Preprocess the input data
input_data = np.array(input_data).reshape(1, -1)
input_data = scaler.transform(input_data)

# Make a prediction using the input data
label = svm.predict(input_data)[0]

# Print the predicted label
if label == 0:
    print('The Breast Cancer is Malignant')
else:
    print('The Breast Cancer is Benign')
```

```
The Breast Cancer is Malignant
```

Figure 17: Prediction of Breast Cancer using SVM (Malignant)

```
input_data = [11.52,18.75,73.34,409,0.09524,0.05473,0.03036,0.02278,0.192,0.05907,0.3249,0.9591,2.183,23.47,
# Preprocess the input data
input_data = np.array(input_data).reshape(1, -1)
input_data = scaler.transform(input_data)

# Make a prediction using the input data
label = svm.predict(input_data)[0]

# Print the predicted label
if label == 0:
    print('The Breast Cancer is Malignant')
else:
    print('The Breast Cancer is Benign')
```

```
The Breast Cancer is Benign
```

Figure 18: Prediction of Breast Cancer using SVM (Benign)

4.3 Random Forest Classification

Using Random Forest as a prediction model, the average accuracy was slightly less than that of SVM. This maybe be caused as Random Forest is more prone to the over fitting, if the number of features are larger than that of the number of samples.

```
# Calculate the accuracy score of the model for training data
training_accuracy_rfc = accuracy_score(Y_train, Y_train_predict)
print(f"Training Accuracy: {training_accuracy_rfc*100:.2f}%")

Training Accuracy: 96.31%


# Calculate the accuracy score of the model/ testing accuracy
testing_accuracy_rfc = accuracy_score(Y_test, Y_test_predict)
print(f"Testing Accuracy: {testing_accuracy_rfc*100:.2f}%")

Testing Accuracy: 96.06%
```

Figure 19:Random Forest Classifer Accuracy

The figure below shows the prediction whether the dataset input is either malignant or benign for RFC model. This helps us to predict the early stage breast cancer just by providing the attribute to the given model. Some of the sample data for validation of the RFC model are represented as:

```
input_data = np.array([17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4,
# Reshape the input data to a 2D array with a single row and the same number of columns as the training data
input_data_2D = input_data.reshape(1, -1)

# Predict the class label for the input data using the trained random forest classifier
prediction = rfc.predict(input_data_2D)

# Print the predicted class label
if prediction == 0:
    print("The input data is classified as Malignant.")
else:
    print("The input data is classified as Benign.")

The input data is classified as Malignant.
```

Figure 20: Prediction of Breast Cancer using RFC(Malignant)

```
input_data = np.array([13.08, 15.71, 85.63, 520.0, 0.1075, 0.127, 0.04568, 0.0311, 0.1967, 0.06811, 0.1852, 0.7477,

# Reshape the input data to a 2D array with a single row and the same number of columns as the training data
input_data_2D = input_data.reshape(1, -1)

# Predict the class label for the input data using the trained random forest classifier
prediction = rfc.predict(input_data_2D)

# Print the predicted class label
if prediction == 0:
    print("The input data is classified as malignant.")
else:
    print("The input data is classified as benign.")

The input data is classified as benign.
```

Figure 21: Prediction of Breast Cancer using RFC (Benign)

4.4 Comparative Analysis

A comparison was conducted using the same dataset to train and test three algorithms. The accuracy varied among them, with SVM achieving the highest accuracy of 98.77% for the testing dataset and 97.64% for the training dataset. Similarly, Random Forest achieved an accuracy of 96.06% for the testing dataset and 96.31% for the training dataset. In contrast, Logistic Regression had the lowest accuracy, with a testing accuracy of 91.34% and a training accuracy of 94.73%.

Table 2: The classification table using Logistic Regression, Support Vector Machine and Random Forest

	Precision			**Recall**			**F1-score**			**Support**		
	LR	**SVM**	**RF**	**LR**	**SVM**	**RF**	**LR**	**SVM**	**RF**	**LR**	**SVM**	**RF**
Malignant	0.99	1.00	0.99	0.87	0.96	0.95	0.92	0.98	0.97	76	76	76
Benign	0.83	0.94	0.93	0.98	1.00	0.98	0.90	0.97	0.95	51	51	51
Accuracy							0.91	0.98	0.96	127	127	127
Macro Average	0.91	0.97	0.96	0.92	0.98	0.96	0.91	0.98	0.96	127	127	127
Weighted Average	0.92	0.98	0.96	0.91	0.98	0.96	0.91	0.98	0.96	127	127	127

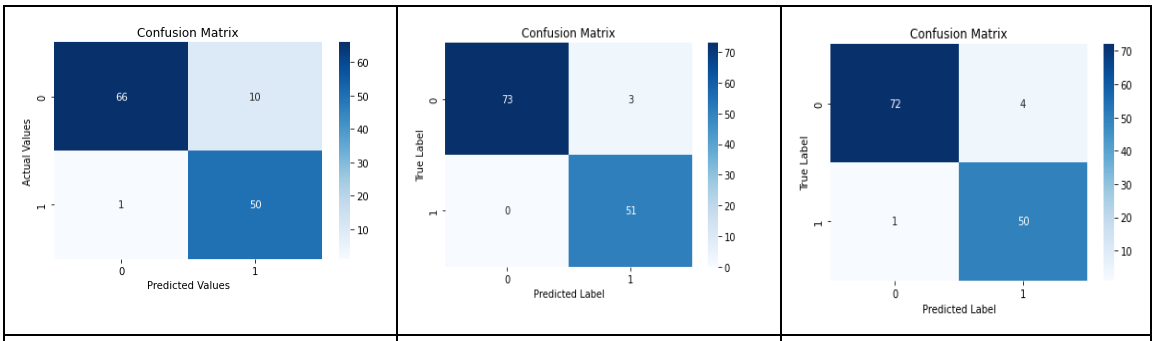
This paper has a multi-class metrics classification i.e; the cancer prediction system classifies whether the cancer is benign or malignant. For malignant there is one set of data in which LR receives precision of 0.99 for malignant that means that the patients that 99% of patients predicted to have cancer actually had the disease and for benign 83% patients actually had benign breast cancer.

The ratio of positively anticipated occurrences for each class that were successfully predicted out of all positively predicted examples is known as the precision. On the other hand, recall describes the proportion of positive instances properly predicted for each class out of all actual positive instances. The F1-score incorporates a weighted average of precision and recall to provide a fair evaluation of the model's accuracy for each class. The Support displays how many instances there are of each class. The

Macro Average calculates the mean of the scores for each class to give a general sense of the model's accuracy. By computing the average of the scores for each class, weighted by the number of examples in each class, the Weighted Average provides a general indicator of the model's accuracy and accounts for the distribution of cases in each class (Teemu Kanstrén, 2020).

4.4.1 Confusion matrix

A table that shows how well a machine learning model performs in a classification test is called a confusion matrix. It is a useful tool for locating potential error sources and learning about the model's advantages and disadvantages. A number of assessment measures, such as accuracy, precision, recall, and F1-score, can be derived by looking at the matrix to assess the model's efficacy. This paper might be useful for utilizing the confusion matrix to calculate these evaluation criteria. The matrix divides the results into four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) (Ting, 2017). It does this by comparing the predicted labels of a model to the actual labels (FN).



Logistic Regression	Support Vector Machine	Random Forest

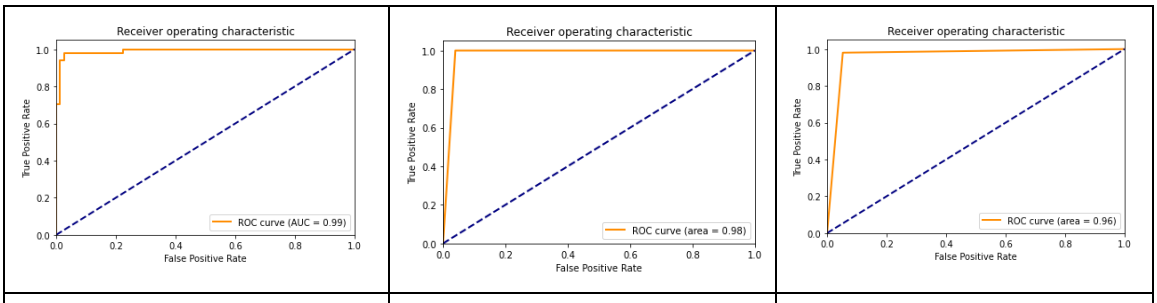
Figure 22: Confusion Matrix showing the true value and predicted value

For Logistic Regression, the confusion matrix classifies that the model Predicted 66 instances as positive that were actually positive, 10 instances as positive but they were actually negative, 50 instances as negative as they were actually negative and 1 instance and negative but it was actually positive. For Support Vector Machine, the confusion matrix classifies that the model predicted 73 positive instances that were actually positive, 3 instances as positive but they were negative, 51 negative that were actually negative and none of the negative instances were positive. For Random

Forest, the confusion matrix classifies that the mode predicted 72 instances as positive that were actually positive, 4 instances as positive that were actually negative, 50 instances as negative and they were actually negative and 1 instance as negative but it was actually positive.

4.4.2 ROC Curve

The performance of a binary classifier is shown visually using the Receiver Operating Characteristic (ROC) curve at various classification levels. This method is frequently used to assess how accurately a classifier's predictions are made in machine learning, statistics, and medical diagnosis. Plotting the True Positive Rate (TPR) vs the False Positive Rate (FPR) for various categorization criteria results in the ROC curve (Majnik & Bosnić, 2013). TPR, which also refers to sensitivity or recall, quantifies the percentage of real positives that the model properly detected. FPR estimates the percentage of real negatives that the model misclassified as positives. A good predictive model should have high TPR and low FPR at all thresholds resulting in a curve closer to the upper- left corner of the plot.



Logistic Regression	Support Vector Machine	Random Forest

Figure 23: ROC Curve showing the true and false positive rate

The area under the ROC curve (AUC) is used as a metric for evaluating the overall performance of the model. In the models we presented in this paper, the Logistic Regression has an AUC of 0.99, Support Vector Machine has an AUC of 0.98 and Random Forest has an AUC of 0.96.

4.4.3 Heat maps

A heat map represent two-dimensional tables of numbers as shades of colors. It graphically represents data using different colors to indicate the intensity of the values

in the matrix. This can be used to identify the clusters of features that are highly correlated with each other and also used for identifying a group of features that provide higher predictive power. By analyzing these diagrams we can gain knowledge regarding the underlying patterns and relationships between the features and breast cancer outcomes to improve breast cancer prediction models (Gehlenborg & Wong, 2012).

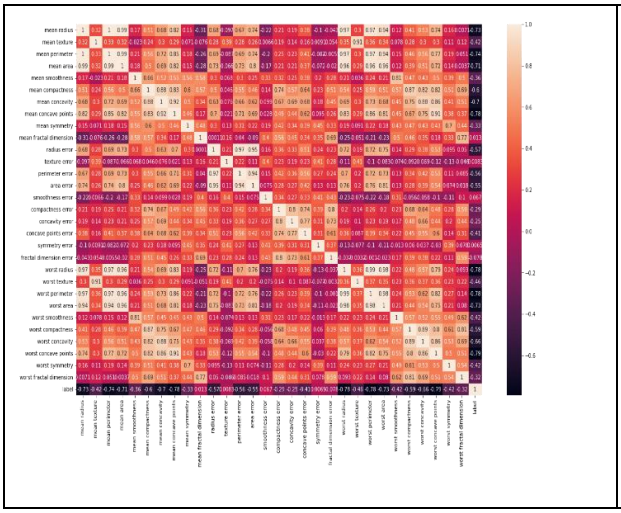


Figure 24: Heat map for training dataset

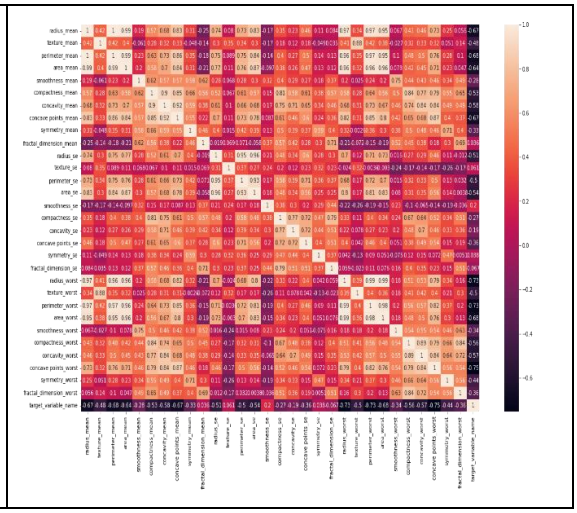


Figure 25: Heat map for testing dataset

In Figure 24, the heat map visualizes the relationship between the different features in the dataset taken during the training of the model (such as mean radius, concavity, compactness of the lymph nodes etc.) and the likelihood of breast cancer occurrence and recurrence. In Figure 25, the heat map visualizes the relationship between the different features in the dataset taken during the testing of the model and the likelihood of breast cancer occurrence and recurrence.

This can be used to identify the clusters of features that are highly correlated with each other and also used for identifying group of features that provide higher predictive power. By analyzing these diagrams we can gain knowledge regarding the underlying patterns and relationships between features and breast cancer outcomes to improve breast cancer prediction models.

4.4.4 Histograms

Histogram, a graphical representation of the distribution of a dataset, is used to visualize the distribution features in the dataset and their relationship with the cancer outcomes. In a histogram, a preprocessed dataset is taken and then the dataset is grouped into bins and the number of patients that lie under the bin is counted. These are helpful exploratory data visualizations for identifying outliers, skew, bimodality, and other distributional shape properties as well as for comparing subgroups in the data (Nuzzo, 2019).

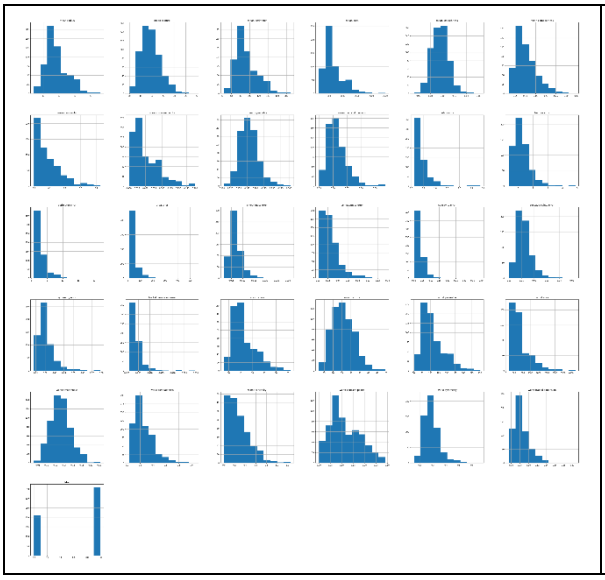


Figure 26: Histogram for training dataset

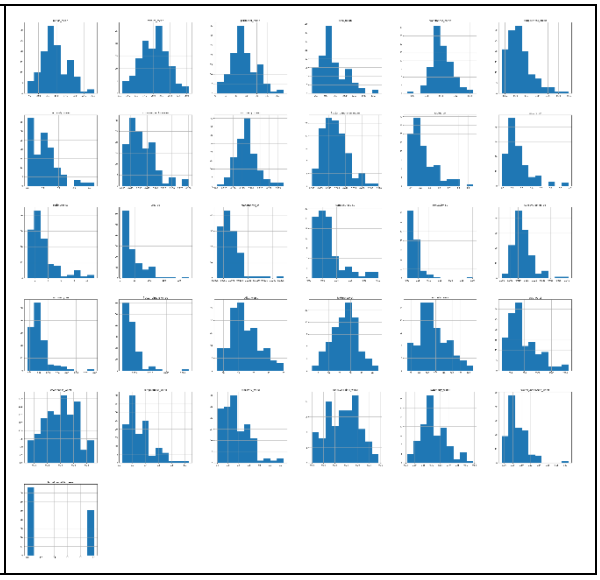


Figure 27: Histogram for testing dataset

The X- axis and Y-axis represents the bin range and the frequency of patients with that feature. This provides insights into the underlying patterns in the data and helps to identify the potential risk factors for breast cancer. In Figure 26 and 27 the histogram for the training and testing dataset is shown respectively that is used for predicting breast cancer.

CHAPTER 5
DISCUSSION

The result demonstrated by that the proposed research is beneficial for predicting early breast cancer since it assists healthcare professionals in reducing the number of false-negative breast cancer cases. Due to the fact that they have been trained on the machine and there is less chance of human error, this system is significantly more accurate, simpler, and quicker.

The optimal method with the highest accuracy for predicting breast cancer is discovered by comparing the accuracy of machine-level algorithms like Support Vector Machine, Random Forest, and Logistic Regression. The problem here is that in order to detect if there are malignant or non-cancerous lump nodes, the supplied model needs to be trained with the mentioned 31 data attributes. The model will not fit the dataset if any of the attributes are missing, and it may also contain errors. Another issue with the accuracy of the models is that there is not enough training data. SVM has the highest accuracy of the models, at 98.77%, but it still is unable to predict accurate results since these models are affected by varied data and data attributes that cause over fitting and under fitting of the data. The expanded dataset may yield various results for certain attributes.

As per the study conducted on 2018 regarding detection breast cancer using different machine learning algorithms. The results showed that SVM was the most successful algorithm (Khourdifi & Bahaj, 2018). Similarly, a study was done to show the performance of five different classification models, including Decision Tree, Random Forest, Support Vector Machine, Neural Network and Logistics Regression that used these models to classify two different datasets related to breast cancer showed random forest model outperforming and adapting better than the other four techniques (Li, 2018). These both papers showed that SVM had better accuracy in comparison to other algorithms. Hence, this paper follows similar approach as previous research.

CHAPTER 6
CONCLUSION

According to the result of the above experiment, breast cancer detection using SVM, Logistic Regression and Random Forest has somewhat similar accuracy and is purely dependent upon the size of the dataset and the attributes of the dataset. The total experimental analysis on the same dataset using three different supervised machine level algorithms SVM, Logistic Regression and Random Forest has an average accuracy of 98.77%, 91.34% and 96.06% respectively.

The method of predicting breast cancer is complicated and requires looking at a variety of variables, such as a woman's age, family history, genetic markers, and lifestyle choices. There are numerous techniques and tests that can assist identify women who may be at increased risk for breast cancer, even though there is no foolproof way to predict with confidence whether a woman will develop the disease. However due to the advancement in technology and the evolving machine learning methods we can predict breast cancer early using machine learning to some extent. Women should monitor their breast health carefully and have frequent tests, such as mammograms, clinical breast exams, and self-exams. Early identification and treatment can significantly increase a woman's chances of surviving breast cancer, if it does occur.

The findings of this research can be used to help health professional minimize the false negative cases in breast cancer detection that is much easier and faster comparative to previous detection that includes numerous mammography to be done and health professionals to look after each of result. This research is not only limited for breast cancer prediction by applying different attribute other diseases can also be detected using these algorithms.

CHAPTER 7
RECOMMENDATIONS

These algorithms struggled to process the dataset containing attributes of various sizes. Several deep learning algorithms may be employed as a consequence for more precise and accurate prediction. For various results, the datasets can be expanded and new attributes added. It was challenging to comprehend how machine learning algorithms make their predictions which led to lack of assurance of correct predictions. These algorithms are not the most effective ones for detection because there are many alternative algorithms for prediction that yield results that are far more precise. For the dataset to be properly fitted and predicted by these algorithms, as well as for the required mammography findings for those set of qualities, a certain set of attributes must be provided. The line graph had many variations and a rough diagram due to the huge number of attributes and low number of training and testing datasets, respectively. Overall, deep learning neural networks can be used to predict breast cancer by using mammography images, which can help to enhance the algorithm.

References

Alakwaa, F. M., Chaudhary, K., & Garmire, L. X. (2018). Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data. *Journal of Proteome Research*, *17*(1), 337–347. https://doi.org/10.1021/acs.jproteome.7b00595

Behravan, H., Hartikainen, J. M., Tengström, M., Kosma, V., & Mannermaa, A. (2020). Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Scientific Reports*, *10*(1), 11044. https://doi.org/10.1038/s41598-020-66907-9

Bhandari, D., Shibanuma, A., Kiriya, J., Hirachan, S., Ong, K. I. C., & Jimba, M. (2021). Factors associated with breast cancer screening intention in Kathmandu Valley, Nepal. *PloS One*, *16*(1), e0245856. https://doi.org/10.1371/journal.pone.0245856

Bhaktapur Cancer Hospital. (2023). Retrieved from https://bch.org.np/

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Cancer Facts & Figures (2022). *American Cancer Society*. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2022.html

Celik, Y., Sabanci, K., Durdu, A., & Aslan, M. F. (2018). Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. *International Journal of Intelligent Systems and Applications in Engineering*, *6*(4), 289–293. https://doi.org/10.18201/ijisae.2018648455

Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. *Journal of Healthcare Engineering*, *2019*, 1–11. https://doi.org/10.1155/2019/4253641

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. https://doi.org/10.1109/ACCESS.2020.3016715

Gaël Varoquaux, O. C. (2023). Evaluating machine learning models and their diagnostic value. *Springer*.

Gardezi, S. J. S., Elazab, A., Lei, B., & Wang, T. (2019). Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review. *Journal of Medical Internet Research*, *21*(7), e14464. https://doi.org/10.2196/14464

Giri, M., Giri, M., Thapa, R. J., Upreti, B., & Pariyar, B. (2018b). Breast Cancer in Nepal: Current status and future directions. *Biomedical Reports*, *8*(4), 325–329. https://doi.org/10.3892/br.2018.1057

Gradilone, A., Naso, G., Raimondi, C., Cortesi, E., Gandini, O., Vincenzi, B., Saltarelli, R., Chiapparino, E., Spremberg, F., Cristofanilli, M., Frati, L., Aglianò, A. M., & Gazzaniga, P. (2011). Circulating tumor cells (CTCs) in metastatic breast cancer (MBC): prognosis, drug resistance and phenotypic characterization. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, *22*(1), 86–92. https://doi.org/10.1093/annonc/mdq323

Great Learning Team. (2023, February 7). *What is Machine Learning?* . Great Learning.

GLOBOCAN, (2012). *Estimated Cancer Incidence, Mortality and Prevalence Worldwide 2012*. http://globocan.iarc.fr/Pages/fact_sheets_population.aspx

Gupta, P., & Garg, S. (2020). Breast Cancer Prediction using varying Parameters of Machine Learning Models. *Procedia Computer Science*, *171*, 593–601. https://doi.org/10.1016/j.procs.2020.04.064

Hou, R., Mazurowski, M. A., Grimm, L. J., Marks, J. R., King, L. M., Maley, C. C., Hwang, E.-S. S., & Lo, J. Y. (2020). Prediction of Upstaged Ductal Carcinoma in Situ Using Forced Labeling and Domain Adaptation. *IEEE Transactions on Bio-Medical Engineering*, *67*(6), 1565–1572. https://doi.org/10.1109/TBME.2019.2940195

Islam, Md. M., Haque, Md. R., Iqbal, H., Hasan, Md. M., Hasan, M., & Kabir, M. N. (2020). Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*, *1*(5), 290. https://doi.org/10.1007/s42979-020-00305-w

Jason Brownlee. (2016). Logistic Regression for Machine Learning. *Machine Learning Mystery*.

Khourdifi, Y., & Bahaj, M. (2018). Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification. *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 1–5. https://doi.org/10.1109/ICECOCS.2018.8610632

Lee, B., Kim, K., Choi, J. Y., Suh, D. H., No, J. H., Lee, H.-Y., Eom, K.-Y., Kim, H., Hwang, S. Il, Lee, H. J., & Kim, Y. B. (2017). Efficacy of the multidisciplinary tumor board conference in gynecologic oncology: A prospective study. *Medicine*, *96*(48), e8089. https://doi.org/10.1097/MD.0000000000008089

Li, Y. (2018). Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction. *Applied and Computational Mathematics*, *7*(4), 212. https://doi.org/10.11648/j.acm.20180704.15

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lu, Y., Li, J.-Y., Su, Y.-T., & Liu, A.-A. (2018). A Review of Breast Cancer Detection in Medical Images. *2018 IEEE Visual Communications and Image Processing (VCIP)*, 1–4. https://doi.org/10.1109/VCIP.2018.8698732

Machine learning: What it is and why it matters. (2023, January 18). *SAS Institute*.

Majnik, M., & Bosnić, Z. (2013). ROC analysis of classifiers in machine learning: A survey. *Intelligent Data Analysis*, *17*(3), 531–558. https://doi.org/10.3233/IDA-130592

Masciari, S., Larsson, N., Senz, J., Boyd, N., Kaurah, P., Kandel, M. J., Harris, L. N., Pinheiro, H. C., Troussard, A., Miron, P., Tung, N., Oliveira, C., Collins, L., Schnitt, S., Garber, J. E., & Huntsman, D. (2007). Germline E-cadherin mutations in familial lobular breast cancer. *Journal of Medical Genetics*, *44*(11), 726–731. https://doi.org/10.1136/jmg.2007.051268

Mohebian, M. R., Marateb, H. R., Mansourian, M., Mañanas, M. A., & Mokarian, F. (2017). A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning.

Computational and Structural Biotechnology Journal, *15*, 75–85. https://doi.org/10.1016/j.csbj.2016.11.004

Mustafa Abdullah, D., & Mohsin Abdulazeez, A. (2021). Machine Learning Applications based on SVM Classification A Review. *Qubahan Academic Journal*, *1*(2), 81–90. https://doi.org/10.48161/qaj.v1n2a50

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Nuzzo, R. L. (2019). Histograms: A Useful Data Analysis Visualization. *PM&R*, *11*(3), 309–312. https://doi.org/10.1002/pmrj.12145

Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine Learning Classification Techniques for Breast Cancer Diagnosis. *IOP Conference Series: Materials Science and Engineering*, *495*, 012033. https://doi.org/10.1088/1757-899X/495/1/012033

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, *18*(1), 29. https://doi.org/10.1186/s12885-017-3877-1

Pervez, S., & Khan, H. (2007). Infiltrating ductal carcinoma breast with central necrosis closely mimicking ductal carcinoma in situ (comedo type): a case series. *Journal of Medical Case Reports*, *1*(1), 83. https://doi.org/10.1186/1752-1947-1-83

Robertson, F. M., Bondy, M., Yang, W., Yamauchi, H., Wiggins, S., Kamrudin, S., Krishnamurthy, S., Le-Petross, H., Bidaut, L., Player, A. N., Barsky, S. H., Woodward, W. A., Buchholz, T., Lucci, A., Ueno, N. T., & Cristofanilli, M. (2010). Inflammatory breast cancer: the disease, the biology, the treatment. *CA: A Cancer Journal for Clinicians*, *60*(6), 351–375. https://doi.org/10.3322/caac.20082

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sarvesh. (2023, January 18). Machine Learning Trends 2023. *Emeritus*.

Sengar, P. P., Gaikwad, M. J., & Nagdive, A. S. (2020). Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 796–801. https://doi.org/10.1109/ICSSIT48917.2020.9214267

Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1), 64. https://doi.org/10.1186/s12874-019-0681-4

Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., & Zhu, H.-P. (2017). Risk Factors and Preventions of Breast Cancer. *International Journal of Biological Sciences*, *13*(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

UCI Machine Learning. (n.d.). *Breast Cancer Wisconsin (Diagnostic) Data Set*. Kaggle.

Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*, *267*(2), 687–699. https://doi.org/10.1016/j.ejor.2017.12.001

Teemu Kanstrén (2020). A Look at Precision, Recall, and F1-Score. *Towards Data Science*

Ting, K. M. (2017). Confusion Matrix. In Encyclopedia of Machine Learning and Data Mining (pp. 260–260). *Springer US*. https://doi.org/10.1007/978-1-4899-7687-1_50

Yarabarla, M. S., Ravi, L. K., & Sivasangari, A. (2019). Breast Cancer Prediction via Machine Learning. *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 121–124. https://doi.org/10.1109/ICOEI.2019.8862533

Zhang, Y.-N., Xia, K.-R., Li, C.-Y., Wei, B.-L., & Zhang, B. (2021). Review of Breast Cancer Pathologigcal Image Processing. *BioMed Research International, 2021*, 1994764. https://doi.org/10.1155/2021/1994764