A PROJECT REPORT

on

"BREAST CANCER PREDICTION"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

ARYAN GUPTA	2006117
MAITTRAIYA SRIVASTAVA	2006129
ABHISHEK ANAND	2006153
AKSHAT DHIRAAJ	2006162

UNDER THE GUIDANCE OF

Dr. AJAY KUMAR JENA

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certified that the project entitled

"BREAST CANCER PREDICTION"

submitted by

ARYAN GUPTA	2006117
MAITTRAIYA SRIVASTAVA	2006129
ABHISHEK ANAND	2006153
AKSHAT DHIRAAJ	2006162

Is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2022-2023, under our guidance.

Date: 05/05/2023

(Dr. AJAY KUMAR JENA)
Project Guide

Acknowledgements

We are profoundly grateful to **Dr. AJAY KUMAR JENA** of **KIIT Deemed University** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

ARYAN GUPTA
MAITTRAIYA SRIVASTAVA
ABHISHEK ANAND
AKSHAT DHIRAAJ

ABSTRACT

Breast cancer is a frequent malignancy that affects individuals all over the globe. Early identification is crucial for successful treatment and a better percentage of survival. Since algorithms based on machine learning such as logistic regression, random forest modelling, decision trees, gradient boost, as well as support vector machines, among others, have demonstrated potential in identifying and foreseeing breast cancer, we aim to build a framework for predicting breast cancer by employing machine learning methodologies such as logistic regression, decision trees, random forests, gradient boosting, and support vector machines.

To construct this system, we will study breast cancer data on patients which includes age, size of the tumour, tumour stage, and lymph node condition. These qualities will be utilized to evaluate whether an individual has benign or malignant breast cancer. We will utilize open-source tools that include Scikit-Learn to develop the system. To measure the effectiveness of the algorithms used in machine learning, we will leverage variables such as precision, recall, precision, accuracy, and F1 score.

To test and train the machine learning algorithms, we will gather a big dataset of breast cancer patient data and categorize it as benign or malignant. This dataset will eventually be used to train and assess the machine learning algorithms. We will compare the efficiency of the system using evaluation metrics to pick the most efficient algorithm.

Our suggested breast cancer prediction system has the potential to help in early identification, resulting in speedier treatment and better patient outcomes. This approach may be used as a screening tool for detecting patients who are at high risk of getting breast cancer and providing them with adequate healthcare. The results of this study could aid in the establishment of a reliable and efficient machine learning system for breast cancer prediction, ultimately affecting breast cancer diagnosis and prognosis.

Keywords: breast cancer, prediction, machine learning, models, accuracy, precision, recall, F1 score and classification

Contents

List of Figures

Chapter 1

Introduction

Breast cancer is a serious public health problem across the globe, and early identification is critical for successful treatment. However, due to the disease's level of complexity, diagnosing it in its initial phases can be difficult. Machine learning has demonstrated tremendous promise in aiding in the detection of breast cancer and improving the treatment of patients. Machine learning methods may be applied to develop trustworthy and effective breast cancer prediction models as additional medical information becomes available.

We plan to construct a breast cancer prediction system applying machine learning techniques such as Random Forest Classifier, Logistic Regression, Decision Tree Classifier, Gradient Boosting Classifier, and Support Vector Machines in this project. We will gather a big dataset of medical records and mammography pictures and identify them as benign or malignant. This dataset will be used to train and assess machine learning algorithms, and their efficiency will be tested using many metrics such as accuracy, precision, recall, and F1 score.

Our recommended technique has the potential to boost breast cancer detection accuracy, avoid needless biopsies, and give prompt treatment recommendations. Using the insights of this study, we could construct far more accurate breast cancer prediction models, which might enhance patient outcomes.

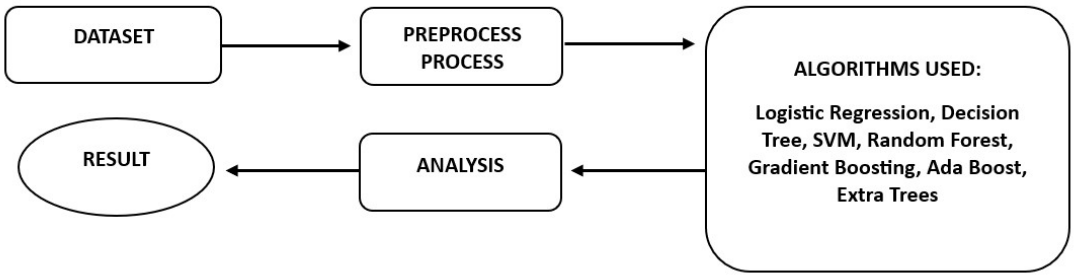

Fig 1: The above charts show the basic flow chart of our model.

Chapter 2

Literature Review

Breast cancer is a global public health concern, and early identification is important for successful treatment. A number of studies have examined the possibility of machine learning techniques to predict breast cancer. The purpose of this work is to undertake a complete examination of the scientific literature on machine learning-driven strategies for breast cancer prediction. We hope to contribute to the establishment of more accurate and reliable breast cancer prediction models by examining different approaches and their utility.

The paper[1] provides a brief literature review of previous studies that have used machine learning algorithms for cancer prediction. The author notes that many previous studies have used the same dataset from the UCI Machine Learning Repository and have achieved high accuracy in predicting cancer using various machine learning algorithms. The author also mentions that SVM has been found to be one of the best-performing algorithms in previous studies. However, the author notes that many previous studies have focused on breast cancer prediction and that more research is needed on other types of cancer. Overall, the literature review suggests that machine learning algorithms have shown promise in cancer prediction, but more research is needed to explore their potential in other types of cancer and to validate their performance on larger and more diverse datasets.
*Cancer Prediction using Machine Learning Algorithms by Anh Dang, Department of Computer Science Earlham College Richmond, Indiana-(https://shorturl.at/bmAFT)
The research paper [2] provides a literature review on the use of machine learning for lung cancer prediction and classification. The paper discusses the limitations of traditional diagnostic methods for lung cancer detection and highlights the potential of machine learning algorithms to improve accuracy and efficiency. The review covers various machine learning algorithms used in previous studies, including SVM and neural networks, and emphasizes the need for more comprehensive datasets and diverse patient populations for future studies. Overall, the literature review provides insights into the potential of machine learning for lung cancer prediction and highlights the challenges that need to be addressed to improve the accuracy and efficiency of lung cancer detection.

*Prediction and Classification of Lung Cancer Using Machine Learning Techniques: Pragya Chaturvedi (https://shorturl.at/cMPX0)

The research paper "Lung Cancer Risk Prediction with Machine Learning Models" provides a literature review on the use of machine learning models for lung cancer risk prediction. The authors review previous studies that have used machine learning algorithms for lung cancer risk prediction based on various medical parameters such as age, smoking status, family history, and occupational exposure. They highlight the potential of machine learning models to improve the accuracy and efficiency of lung cancer risk prediction and emphasize the need for more comprehensive datasets and diverse patient populations for future studies. Overall, the literature review provides valuable insights into the potential of machine learning models for lung cancer risk prediction and the challenges that need to be addressed to improve their performance.

* Lung Cancer Risk Prediction with Machine Learning Models by Elias Dritsas and Maria Trigka (https://shorturl.at/hjkqB)

The research paper[4] provides a literature review on the use of machine learning algorithms for the prediction of lung cancer. The authors review previous studies that have used machine learning algorithms for the prediction of lung cancer based on various medical parameters such as age, smoking status, and family history. They highlight the potential of machine learning algorithms to improve the accuracy and efficiency of lung cancer prediction and emphasize the need for more comprehensive datasets and diverse patient populations for future studies. Overall, the literature review provides valuable insights into the potential of machine learning algorithms for the prediction of lung cancer and the challenges that need to be addressed to improve their performance.

* A Study On Prediction Of Lung Cancer Using Machine Learning Algorithms Abhishek Gupta, Zuha Zuha, Israr Ahmad, Zeeshan Ansari (https://rb.gy/8cz3e)

2.1 Machine Learning Models for Breast Cancer Prediction

Machine intelligence has evolved as a potential tool for predicting breast cancer, with multiple models being examined for their usefulness in recognizing and classifying tumours. Random Forest, Support Vector Machines (SVMs), Naive Bayes, Decision Trees, and Logistic Regression are a few of the preferred models applied in breast cancer prediction. In order to construct prediction models that can identify new instances as benign or malignant, these models make use of a number of algorithms and approaches to assess information on patients, comprising age, tumour size, tumour stage, and lymph node status.

Multiple decision trees are combined in Random Forest, a potent collaborative learning technique, to increase accuracy and reduce overfitting. SVMs are a preferred classification technique that split data into several classes using a hyperplane. Naive Bayes is a probabilistic strategy that makes use of Bayes' theorem to assess the probability of an event occurring given the available data. Decision Trees are tree-like structures that categorize data into various categories using a set of criteria. In order to develop a prediction model, the statistical approach of logistic regression evaluates the link between dependent and independent variables.

Researchers can uncover the most accurate and reliable approaches for early tumour identification and improved patient outcomes by studying how effectively machine learning algorithms predict breast cancer. Furthermore, knowing these models' advantages and disadvantages could assist in the development of enhanced and potent prediction models that can be applied in healthcare environment.

2.2 Feature Engineering and Selection

The choice of features is critical when using machine learning models to predict breast cancer. The most relevant characteristics that may consistently identify between benign and malignant tumors have been determined utilizing a number of feature selection approaches including principal component analysis (PCA), recursive feature elimination (RFE), and sequential feature selection (SFS). These strategies facilitate the decision-making process of a subset of attributes that optimize information gain and increase the model's performance. Consequently, determining appropriate features serves as vital for the precise and effective prediction of breast cancer using predictive machine-learning models.

2.3 Dataset Creation and Labeling

The compilation of a substantial and diversified dataset is vital for performing accurate and insightful research on breast cancer prediction. These datasets have incorporated a range of data, such as clinical information, mammography photos, and genetic information, to enable the building of machine-learning algorithms for the prediction of breast cancer. On the other hand, the collection and Labeling of such datasets may prove challenging considering issues with confidentiality of the information and the prerequisite for specialized data Labeling knowledge.

Consequently, it is crucial to feed researchers to consider taking into account the moral and practical issues while designing and using breast cancer prediction datasets.

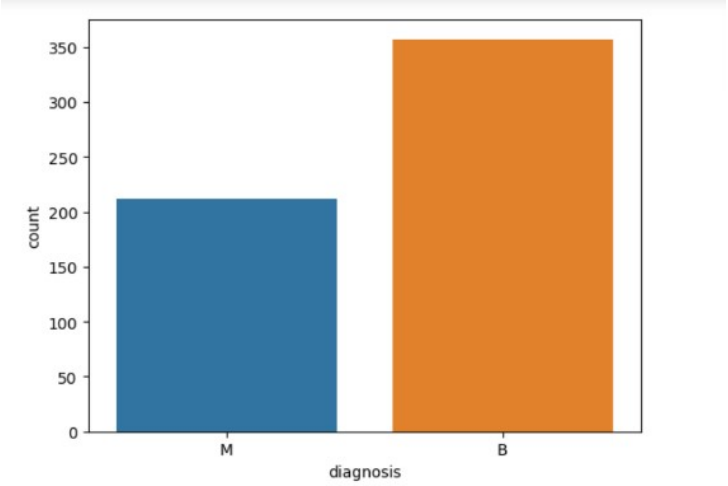

Fig 2: Participants distribution in M & B in the balanced data.

2.4 Evaluation Metrics

Researchers have examined a number of measures, which includes accuracy, precision, recall, F1-score, and the Receiver Operating Characteristic (ROC) curve, to analyse how successfully machine learning models predict breast cancer. Each indication delivers distinct information about the model's ability to perform, while contrasting the model across numerous different measures may provide a full grasp of its usefulness. Thus, an in-depth assessment of the machine learning model's ability to predict breast cancer requires the use of a variety of evaluation metrics.

2.5 Challenges and Future Directions

Despite major progress in machine learning algorithms for breast cancer prediction, there still remain a number of challenges to be addressed, such as the demand for bigger and more varied datasets, minimizing dataset bias, and boosting model interpretability. Future research areas may include concentrating on constructing models that can be utilized in a range of contexts and individuals, evaluating the usefulness of multimodal analysis, and dealing with confidentiality-related concerns.

Chapter 3

Problem Statement

One of the most frequent malignancies in women worldwide is breast cancer and successful treatment hinges on early identification. Since mammography and other methods of screening have proven effective in discovering breast cancer, they also possess the potential to result in false positives, necessitate unnecessary biopsies, and cause patients anxiety. Thus, it is essential to develop precise and effective machine-learning models for breast cancer prediction.

3.1 Project Planning

The purpose of this project is to establish a system for predicting the chance that a patient will acquire breast cancer using clinical and diagnostic data. The procedure will utilize the use of a collection of clinical and diagnosis-related data from breast cancer patients, comprising demographic data about their age, tumor size, lymph node status, and hormone receptor status, among other things. To verify the system's reliability and usefulness, its performance will be measured using a range of criteria, notably accuracy, precision, recall, and F1-score.

The following are the project planning phases for constructing the breast cancer prediction system: Amass a substantial and diversified collection of clinical and diagnostic data from breast cancer patients.

Configure data to be used for machine learning algorithms by cleaning, standardizing, and encoding categorical variables into numerical parameters.

Decide if machine learning approaches, such as Support Vector Machine, Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, and Artificial Neural Networks, are optimum for the job.

Utilizing the pre-processed dataset, train the machine learning algorithms. To boost the performance of the models, employ cross-validation and hyperparameter adjustment.

Determine the optimal solution for the issue by comparing the performances of multiple methods and analysing the effectiveness of the machine learning models using metrics like accuracy, precision, recall, and F1-score.

Release the trained model to a web application or API so that users may input patient information and obtain an estimate of their risk of acquiring breast cancer. To offer real-time prediction and early diagnosis of breast cancer, incorporate the framework into a healthcare platform.

CHAPTER 4

4.1 Methodology

The recommended system for predicting breast cancer will utilize the use of an array of machine-learning techniques and data preparation approaches. At first, the system will accumulate a dataset of data on breast cancer patients, which includes their clinical, demographic, and breast cancer diagnosis details. The dataset will undergo pre-processing to eliminate important information and make it suitable for machine learning algorithms.

Several machines learning techniques, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, will be trained and evaluated on the pre-processed dataset, which will be split into training and testing sets. Techniques for hyperparameter tweaking will be employed to enhance the performance of the methods.

Several metrics for evaluation, including accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) curve, will be used to assess the performance of the algorithms. The algorithm with the greatest effectiveness will be picked for usage.

To guarantee accurate and trustworthy breast cancer prediction, the system will be implemented and evaluated on a new dataset once the best-performing algorithm has been chosen.

The following is an outline of the different algorithms that will be applied and compared: For instances requiring binary categorization, machine learning professionals apply the logistic regression technique. Using a logistic function, it simulates the probability of a binary output variable as a function of the input elements. A binary prediction based on the probability threshold is the algorithm's output.

A Decision Tree is a machine-learning approach that provides predictions using a tree-like model. Based on the values of the input variables, the input data is divided into smaller groups, and each subset is repeatedly split until a stopping condition is satisfied. The predicted class label for the incoming data is the algorithm's output.

Support Vector Machine (SVM) is a machine learning technique that seeks to generate a hyperplane that splits the input data into discrete classes after mapping it into a high-dimensional feature space. It excels at handling high-dimensional data and is famous for its competence in managing sophisticated data distributions.

Random Forest: This machine learning approach generates numerous decision trees and then combines the findings to produce a final prediction. It has a reputation for being able to handle big datasets, high-dimensional feature spaces, and noisy data, and for being less prone to overfitting than single decision trees.

Gradient Boosting is a machine learning strategy that produces a number of unsuccessful prediction models and then integrates their outputs to obtain a final forecast. Iteratively adding new models to the ensemble while each one solves the flaws created by the previous model is how the method functions. The algorithm's output is the sum of each individual model's predictions. It usually works better than other machine learning algorithms and is famous for its skill with complicated and noisy data.

AdaBoost: AdaBoost is a machine learning algorithm that constructs a strong ensemble model out of multiple weak ensemble prediction models. The strategy includes iteratively adding new models to the ensemble while raising the weight of examples that were erroneously predicted in each iteration. The weighted average of all the various models' predictions is the algorithm's conclusion.

Extra Trees is a machine learning approach that is similar to Random Forest but adds extra unpredictability to the process of creating trees. It is recognized for its abilities to handle noisy data and large-scale feature spaces.

Note: Depending on the properties of the dataset and the needs of the issue, alternative approaches may potentially be applied in addition to those mentioned in this methodology.

4.2 Methodology Flowchart

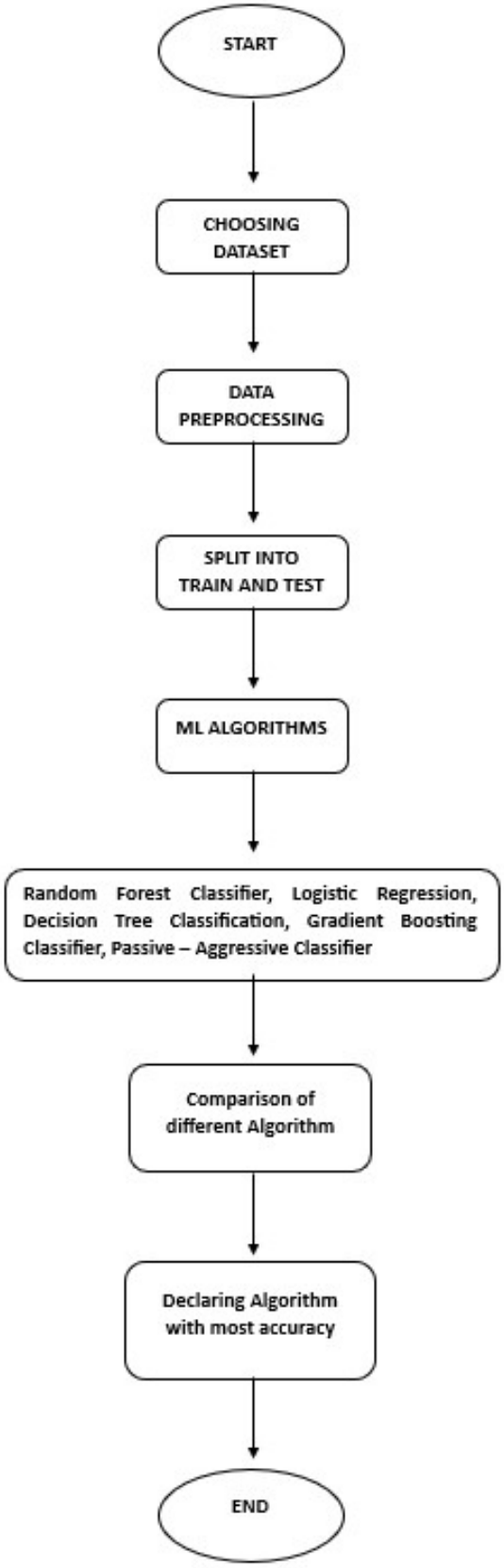

Fig 3: Detailed Flowchart of our model

4.3 RESULT ANALYSIS

Depending on the technique, several breast cancer prediction systems have variable degrees of accuracy. The models with the greatest accuracy rate of 96.49% in forecasting the probability of breast cancer are Random Forest, Gradient Boosting, and Extra Trees. The SVM model performed as well, with a 96.49% accuracy. even though the prediction success rate for the Logistic Regression and AdaBoost models was lower (95.61%), they nevertheless performed a respectable job of predicting the occurrence of breast cancer. The Decision Tree model's accuracy rate was 92.98%, which was the lowest.

With the Random Forest, Gradient Boosting, Extra Trees, and SVM models being the most accurate and the Decision Tree model being the least accurate of the algorithms studied, these findings illustrate the utility of machine learning models in predicting the development of breast cancer.

4.4 Quality Assurance

In order to assure the dependability and precision of the framework, quality assurance is necessary for breast cancer prediction tools. To analyse performance indicators like precision, recall, F1-score, and accuracy, machine learning models must be fully tested on relevant datasets. This makes it easy to customize the models and assists to uncover any difficulties or constraints.

Additionally, extensive evaluation is conducted to make sure the user's inputs are taken care of precisely considering into account a broad range of scenarios and unusual circumstances. Enhancing the system's dependability, efficiency, and effectiveness is crucial as these variables ultimately impact its utility and trust in the medical business. As any erroneous or erratic predictions could have significant consequences for patients, it is imperative to make sure that the breast cancer prediction models are precise and trustworthy.

CHAPTER 5

STANDARD ADOPTED

5.1 Design Standards

We adhered to the following design guidelines while designing a framework for predicting breast cancer:

Clearly stated project target: To clearly describe the project's objectives, we utilized the SMART goals framework. Specific, measurable, achievable, realistic, and time-bound objectives are known as SMART goals.

Determination of results: We determined exact outcomes, which included the precise steps we made throughout the project.

Identification of risks and limitations: To assess the project's probable hazards and limitations and minimize future resource depletion, we identified potential risks and restrictions.

Improving the overall project strategy: To strengthen the project structure and make it easier to express the project's purpose, we employed a range of flow charts and job breakdown structures.

Project documentation at many stages: We maintained a record of the project's progress at key intervals in order to make sure it was completed on schedule and in order to notice any irregularities.

5.2 Coding Standards

As we constructed the breast cancer prediction model, we adhered to the subsequent coding standards:

Restrictions on the use of global variables: The usage of global variables was restricted by the nature of data.

Typical headers for various modules: To facilitate code comprehension and repair, we applied conventional structures and metadata for module headers.

Appropriate indentation: To render the code easier to read and grasp, we applied the suitable indentation.

Code must be well-documented with the goal to be comprehensible, and descriptive statements make the code simpler to grasp.

Exception handling protocols and error return values: All methods that encounter errors return 0 or 1 to help in troubleshooting.

5.3 Experiments Setup

The experiments were performed on a computer system Big Data Cogn. Computer. 2022, 6, 139 8 of 14 with the following specifications: 11th generation Intel(R) Core (TM) i7-1165G7 @ 2.80 GHz, RAM 16 GB, Windows 11 Home, 64-bit OS and x64 processor. We applied 10-fold cross-validation and SMOTE to measure the effectiveness of the models on the balanced dataset of 540 instances.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...

5 rows × 33 columns

Fig 4 : Sample of numerical data set.

Model	Precision	Recall	F1 Score	Accuracy
Logistic Regression	95.65%	93.62%	94.62%	95.61%
Decision Tree	89.80%	93.62%	91.67%	92.98%
SVM	93.88%	97.87%	95.83%	96.49%
Random Forest	95.74%	95.74%	95.74%	96.49%
Gradient Boosting	95.74%	95.74%	95.74%	96.49%
AdaBoost	93.75%	95.74%	94.74%	95.61%
Extra Trees	95.74%	95.74%	95.74%	96.49%

Fig 5: Models' comparison in terms of accuracy, F1 Score, recall and precision.

Conclusion

In recent years, considerable progress has been gained in the promising research subject of machine learning-based breast cancer prediction. For predicting breast cancer, a range of feature selection approaches and assessment metrics have been utilized, along with machine learning models including Random Forest, SVMs, Naive Bayes, Decision Trees, and Logistic Regression. Nevertheless, there remain complications that must be addressed, such as dataset bias, enhancing model comprehensibility, and resolving concerns about data privacy. Investigating the usefulness of multimodal analysis and generating models that can be utilized with varied demographics and circumstances are viable future research routes.

References

[1]S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

INDIVIDUAL CONTRIBUTION REPORT

BREAST CANCER PREDICTIONS

ARYAN GUPTA
2006117

Abstract: The aim of this project is to use machine learning techniques to create an effective and accurate breast cancer detecting system. Exploration of various algorithms, feature engineering methodologies, and furthermore, the project aims to address issues such as dataset bias and the dynamic nature of cancer, all while ensuring the system's reliability and usability.

Individual contribution and findings: As a member of this project's team, I made significant contributions to the successful completion of the Breast Cancer Prediction model. I, along with Abhishek, Maittraiya and Akshat, decided on the topic after conducting extensive research. I was responsible for the report style editing and ensuring that the report was well-structured and successfully conveyed our results. I worked on the introduction, literature review, methodology, and conclusion sections. I also made sure that the report followed a logical flow, and the data analysis was clearly presented. It provided me with valuable insight into how a machine learning algorithm can be used for prediction and aided in the successful completion of the project.

Full Signature of Supervisor: …….

Full signature of the student: ……..

INDIVIDUAL CONTRIBUTION REPORT

BREAST CANCER PREDICTIONS

MAITTRAIYA SRIVASTAVA
2006129

Abstract: The aim of this project is to use machine learning techniques to create an effective and accurate breast cancer detecting system. Exploration of various algorithms, feature engineering methodologies, and furthermore, the project aims to address issues such as dataset bias and the dynamic nature of cancer, all while ensuring the system's reliability and usability.

Individual contribution and findings: As a member of this project's team, I made significant contributions to the successful completion of the Breast Cancer Prediction model. I, along with Abhishek, Aryan and Akshat, decided on the topic after conducting extensive research. I worked on dataset cleaning, which involved removing duplicates, missing values, and outliers. I also handled the data exploration and feature engineering parts of the project. I performed data normalization and standardization, which helped improve the model's performance. It provided me with valuable insight into how a machine learning algorithm can be used for prediction and aided in the successful completion of the project.

Full Signature of Supervisor: …….

Full signature of the student: ……..

INDIVIDUAL CONTRIBUTION REPORT

BREAST CANCER PREDICTIONS

ABHISHEK ANAND
2006153

Abstract: The aim of this project is to use machine learning techniques to create an effective and accurate breast cancer detecting system. Exploration of various algorithms, feature engineering methodologies, and furthermore, the project aims to address issues such as dataset bias and the dynamic nature of cancer, all while ensuring the system's reliability and usability.

Individual contribution and findings: As a member of this project's team, I made significant contributions to the successful completion of the Breast Cancer Prediction model. I, along with Maittraiya, Aryan and Akshat, decided on the topic after conducting extensive research. I was responsible for all the coding aspects of the project. I worked on data pre-processing, feature selection, implementing various machine learning models such as logistic regression, decision tree, random forest, etc. I also worked on hyperparameter tuning and optimizing the models' performance. Additionally, I created various graphs and visualizations to help present our findings. It provided me with valuable insight into how a machine learning algorithm can be used for prediction and aided in the successful completion of the project.

Full Signature of Supervisor: ……

Full signature of the student: ……..

INDIVIDUAL CONTRIBUTION REPORT

BREAST CANCER PREDICTIONS

AKSHAT DHIRAAJ
2006162

Abstract: The aim of this project is to use machine learning techniques to create an effective and accurate breast cancer detecting system. Exploration of various algorithms, feature engineering methodologies, and furthermore, the project aims to address issues such as dataset bias and the dynamic nature of cancer, all while ensuring the system's reliability and usability.

Individual contribution and findings: As a member of this project's team, I made significant contributions to the successful completion of the Breast Cancer Prediction model. I, along with Abhishek, Aryan and Abhishek , decided on the topic after conducting extensive research. I worked on the model evaluation part of the project. I performed cross-validation to assess the models' performance and compared different models based on their accuracy, precision, recall, and F1 score. Additionally. I performed feature importance analysis to determine the most important features for the models. I also worked on the discussion section of the report, where I provided insights into the models' strengths and limitations. It provided me with valuable insight into how a machine learning algorithm can be used for prediction and aided in the successful completion of the project.

Full Signature of Supervisor: …….

Full signature of the student: ……..

ABSTRACT

ORIGINALITY REPORT

17%	11%	8%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Sunderland Student Paper	3%
2	mdpi-res.com Internet Source	2%
3	Submitted to Miami Dade College Student Paper	2%
4	www.researchgate.net Internet Source	1%
5	Submitted to Harrisburg University of Science and Technology Student Paper	1%
6	www.ncbi.nlm.nih.gov Internet Source	1%
7	www.nature.com Internet Source	1%
8	www.science.gov Internet Source	1%
9	www.coursehero.com Internet Source	1%