AN ENSEMBLE MODEL FOR PREDICTING BREAST CANCER

BY

EMESHILI SOLOMON CHUKWUYENUM
PSC1808815

DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF PHYSICAL SCIENCES
UNIVERSITY OF BENIN

OCTOBER 2023

AN ENSEMBLE MODEL FOR PREDICTING BREAST CANCER

BY

EMESHILI SOLOMON CHUKWUYENUM
PSC1808815

A PTOJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE, FACULTY OF PHYSICAL SCIENCES, UNIVERSITY OF BENIN, BENIN CITY IN PARTIAL FULFILLMENT OF THE REQUIREMNTS FOR AWARD OF BACHELOR OF SCIENCE DESGREE (B.Sc.) IN COMPUTER SCIENCE.

SEPTEMBER 2023

ATTESTATION

I, Emeshili Solomon Chukwuyenum, an undergraduate student in the Department of Computer Science, Faculty of Physical Sciences, University of Benin, Edo State, with matriculation number PSC1808815, hereby declare that the work I have submitted is entirely original to me, I attest to have done the project in partial fulfillment of the requirements for the award of Bachelor of Science (B.Sc.) Degree in Computer Science, University of Benin.

_____ _____

Emeshili Solomon Signature/Date

(Project Student)

APPROVAL

This project report prepared by Emeshili Solomon Chukwuyenum, an undergraduate student in the Department of Computer Science, Faculty of Physical Sciences, University of Benin, Edo State, with matriculation number PSC1808815 is hereby approved in partial fulfillment of the requirements for the award of Bachelor of Science (B.Sc.) in Computer Science.

_____ _____

MR, E. C, IGODAN (Ph.D) Signature/Date
(Project Supervisor)

_____ _____

PROF. (MRS.) A. EGWALI Signature/Date
(Head of Department

CERTIFICATION

This is to verify that Emeshili Solomon Chukwuyenum, an undergraduate student in the Department of Computer Science, Faculty of Physical Sciences, University of Benin, Edo State, with matriculation number PSC1808815 did this project in partial fulfillment of the requirements for the award of Bachelor of Science (B.Sc.) in Computer Science, University of Benin, under my supervision.

MR. E. C. IGODAN (Ph.D) Signature/Date

(Project Supervisor)

DEDICATION

This project work is dedicated to God Almighty, for providence, guidance and grace in seeing me through this study; I give Him all the glory**.**

This work is also dedicated to my parents, **Mr. And Mrs. Emeshili** for their constant support, prayers, and words of encouragement.

ACKNOWLEDGEMENT

My sincere gratitude goes to God Almighty, for granting me the grace and mental power to complete the project. This project completes another milestone in my academic career. It is pertinent at the juncture to appreciate the efforts of my project supervisor, Mr. E.C Igodan, for his support and guidance throughout the course of the project. I also appreciate the Head of Department, Computer Science, Prof (Mrs) A. Egwali. I would love to acknowledge my parents for the relentless support both financially, materially and spiritually. I appreciate my course mates CODED´22, I want to say a big thank you to Oge, Alexander, Dvyne, Martins, Efe, Vanny, Osinachi, Zilly and Adesuwa. Thank you all so much for the support.

I also wish to appreciate the lecturers of the Department of Computer Science, Prof (Mrs) V.A. Akwukwuma, Prof (Mrs) F.A Egbokhare, Prof, A.A Imianvan, Prof G.O Ekuobase, Prof (Mrs) A.O Egwali, Prof K.C Ukaoha, Prof S. Konyha, Engr. Dr. F.A.U. Imouokhome, Dr. (Mrs) V.I. Osubor, Dr F.O Chete, Mr. P.E.B. Imeifoh, Mr E.E Obasohan, Dr. (Mrs) G. Aziken, Dr. E. Nwelih, Mr. S,O.P. oliomogbe, Dr. (Mrs) A.R Usiobaifo, Mr. E.C Igodan, Dr. F.O Oliha, Dr. (Mrs) R.O Osaseri, Dr. E.P Ebietomere, Mr. K.O Otokiti, Miss. I.O. Usiosofe, Mrs. T. Agenmonmen, Mr. F. Osagie, Mr I.E. Obayabgona for their relentless service to the student of the Department.

ABSTRACT

Breast cancer remains a significant health concern worldwide, necessitating the development of accurate and reliable diagnostic models. This project aims to construct an ensemble model utilizing the Wisconsin Breast Cancer dataset, which has been preprocessed using correlation coefficient and ReliefF feature selection techniques. The dataset is subsequently trained using three popular classifiers: Support Vector Machine (SVM), Naive Bayes, and Logistic Regression. The results demonstrate that the proposed ensemble model leveraging SVM, Naive Bayes, and Logistic Regression classifiers, along with voting, bagging, and stacking techniques, yields superior performance with 96% accuracy compared to individual classifiers and existing benchmark models. This project contributes to the field of breast cancer diagnosis by providing an effective and reliable ensemble model that can assist medical professionals in making accurate and timely predictions for breast cancer classification.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

1.1 Background of the Study

Cancer is characterized by the accumulation of specific genetic alterations that disrupt normal cellular functions, leading to uncontrolled cell growth (Vogelstein *et al,* 2013), originating from any of the countless cells composing the human body. Cancer has the capacity to initiate in virtually any part of the human body, an intricate structure comprising trillions of cells. Ordinarily, human cells undergo growth and replication, known as cell division, to generate new cells as the body requires them. As cells age or sustain damage, they naturally perish, with fresh cells replacing them in the ongoing cycle. Cancerous cells lack the essential components that signal them to cease dividing and undergo cell death. Consequently, these cells accumulate within the body, consuming the oxygen and nutrients that would typically support the growth of other cells. This uncontrolled growth can lead to the formation of tumors, weaken the immune system, and induce various disruptions that hinder the body's regular functions. Tumors can be malignant (cancerous) or non-cancerous (benign). Malignant tumors consist of cells that undergo uncontrollable growth and have the ability to spread either to nearby regions or distant locations. These tumors are classified as cancerous since they invade and infiltrate other sites within the body, whereas Benign tumors remain confined to their original location and do not invade other parts of the body. They do not spread to nearby or distant areas. These tumors typically grow slowly and have clear boundaries. Tobacco use, unhealthy diet, physical inactivity, and exposure to certain chemicals and pollutants are major risk factors for cancer development (Siegel *et al,* 2019). Cancer is among the top causes of death in every country worldwide, posing a significant obstacle to raising life expectancy, accounting for nearly 10 million deaths in 2020. . In terms of new cases of cancer in 2020, the most commonly occuring was Breast Cancer with 2.26 million cases.

Breast cancer is one of the most common cancers among women and most likely affects women over the age of 50. It develops when cells within the breast undergo uncontrolled growth and division, resulting in the formation of an abnormal tissue mass known as a tumor. One out of every seven women will be diagnosed with breast cancer in their lifetime. Breast cancer can occur in women of all ages, but the risk increases with age. Other risk factors include a family history of breast cancer, certain

genetic mutations, and exposure to radiation, The symptoms of breast cancer can vary, but may include a lump in the breast, changes in the size or shape of the breast, dimpling of the skin, nipple discharge or pain in the breast. Mammographic screening, where X-ray images of the breast are taken, is the most commonly available way of finding a change in the breast tissue at an early stage. In 2020, breast cancer accounted for 25.8% of all newly diagnosed cancer cases in women globally, making it the predominant form of cancer among women worldwide (Bray *et al,* 2020).

Annually, 2.3 million women are diagnosed with breast cancer worldwide, with 685,000 deaths reported in 2020, and the incidence is still increasing in many countries (Sung, H, *et al,* 2020). The incidence of breast cancer varies across different regions and countries, For instance, in developed countries such as the United States, breast cancer rates have been relatively higher compared to developing nations (DeSantis *et al*, 2021). However, it is important to note that breast cancer is a global concern, affecting women from all walks of life. Early detection, appropriate treatment, and comprehensive care for cancer patients can help alleviate the burden of cancer. Over 42% of cancer deaths could be prevented (Jennifer *et al,* 2022). The treatment approach depends on the type and stage of cancer and differs from patient to patient. In certain countries, where health systems are inconsistent, the mortality rates of various cancers are increasing because of the lack of access to effective and high-quality treatments, skilled oncologists, and medical diagnostic tools that aid in early cancer detection. With these challenges in mind, this project study aims to create a computer-aided medical decision support system for identifying and diagnosing breast cancer in women.

The term Artificial Intelligence (AI) refers to the simulation of human intelligence processes carried out by machines, particularly computer systems. Coined in the 1950s, AI encompasses a constantly evolving range of abilities that expand with the development of new technologies.AI algorithms have been developed to analyze medical images, such as X-rays and MRIs, aiding radiologists in detecting abnormalities and making accurate diagnoses (Esteva *et al*, 2017). Also, AI-powered chatbots and virtual assistants are being utilized to provide personalized healthcare information, answer patient queries, and offer support in remote patient monitoring (Topol *et al*, 2019). Among the technologies falling under the AI umbrella are machine learning and deep learning.

Machine learning empowers software applications to enhance their accuracy in predicting outcomes without the need for explicit programming to achieve this capability. Machine learning algorithms leverage historical data as input to anticipate new output values. This approach significantly improved in effectiveness with the advent of large data sets available for training. Machine learning techniques have enormous potential for usage as cutting-edge health informatics tools in routine medical practices (Nikhilanand *et al,* 2023). There are four major types of machine learning algorithms, which are supervised, semi-supervised, unsupervised, and reinforcement. In supervised learning, which will be a key focus of this project, the machine learns through examples provided during training. Due to continuous research in the field of machine learning, a multitude of algorithms are being developed almost daily.

The medical challenges associated with cancer prediction involve intricate screening and diagnosis, primarily due to the multitude of features presented by patients (Unger-Saldaa, 2014). However, ongoing research aims to discover an efficient method for selecting relevant features while minimizing information loss. This is called feature selection, which refers to the process of choosing a subset of relevant features from a dataset to be utilized in constructing a model.

Ensemble learning is a machine learning technique that boosts accuracy and robustness in predictions by combining forecasts from multiple models. The primary objective of ensemble learning is to alleviate errors or biases that might exist in individual models by harnessing the collective intelligence of the ensemble (Aishwarya *et al,* 2018).

1.2 Statement of the problem

Developing countries of the world are grappling with major challenges in their efforts to combat cancer. The challenges encompass the issue of uneducated or unprofessional healthcare personnel, early detection, healthcare systems being under-financed and not adequately established to handle chronic disease management, cost of treatment and reliance on traditional healers. In developing countries, approximately one-third of cancers have the potential to be preventable, while another third could be treatable if detected early (Parkin, D.M *et al*, 2002).

1.3 Aim and Objectives

The aim of the study is to develop an ensemble model for cancer prediction, incorporating feature selection techniques. The specific objectives of this project are as follows:

a. Obtain relevant features from the breast cancer datasets using Filtering method;

b. Develop an ensemble model for the prediction of breast cancer;

c. Implement (b) above using Python programming language;

d. Evaluate the performance of the ensemble models from using standard metrics.

1.4 Methodology

The proposed model utilizes an available dataset, the Wisconsin Breast Cancer dataset obtained from an existing literature. The methodology comprises the following five steps:

Firstly, The breast cancer dataset undergoes a preprocessing phase to ensure its quality and consistency which involved data cleaning, handling missing values and correcting inconsistencies.

In the next step, the dataset is filtered using correlation coefficient and ReliefF to identify the most relevant features and minimize redundancy.

Thirdly, support vector machines (SVM), naive bayes (NB) and Logistic Regression classifiers are trained using the preprocessed dataset obtained from the first step.

Next, the trained models are merged using ensemble techniques such as voting, bagging and stacking. These techniques are then compared to determine the most efficient approach based on accuracy and performance. In the final step, a comparative analysis of the ensemble models' output is conducted, considering appropriate evaluation metrics, such as accuracy, precision, recall, or F1-score.

1.5 Scope of study

This study encompasses feature selection techniques, classification algorithms, and ensemble learning methods as part of its scope.

CHAPTER 2
LITERATURE REVIEW

This chapter reviews the existing literature relevant to the development of an ensemble model for breast cancer prediction with a focus on feature selection and machine learning techniques. The review is organized into sections discussing cancer epidemiology, breast cancer diagnosis, artificial intelligence in healthcare, machine learning techniques, feature selection methods, and ensemble learning in the context of cancer prediction.

2.1 Historical background of cancer

The earliest known records of cancer date back to ancient Egypt and Greece. The Egyptians referred to cancer as "the crab" because of the way it spreads. The Greeks called it "karkinos," which also means crab. During the Middle Ages, cancer was thought to be caused by imbalances in the four humors: blood, phlegm, black bile, and yellow bile. (Saunders, 2009) There was no effective treatment for cancer, and most people who were diagnosed with it died within a few months. In the 17th and 18th centuries, scientists began to learn more about the causes of cancer. They discovered that cancer cells are different from normal cells, and that they can spread to other parts of the body (Saunders, 2009). However, there were still no effective treatments for cancer. In the 19th century, doctors began to experiment with different treatments for cancer, including surgery, radiation therapy, and chemotherapy. These treatments were often unsuccessful, but they did help some people. The 20th century saw major advances in the understanding and treatment of cancer. Scientists discovered that cancer is caused by mutations in genes, and that these mutations can be inherited or caused by environmental factors. They also developed new treatments for cancer, including more effective surgery, radiation therapy, and chemotherapy. The 21st century has seen even more progress in the fight against cancer. Scientists have developed new drugs that target specific cancer cells, and they are working on ways to prevent cancer from developing in the first place (Saunders, 2009). Today, cancer is still a major cause of death, but it is no longer a death sentence. Thanks to advances in research and treatment, many people with cancer can now live long and healthy lives. Here are some of the major milestones in the history of cancer research:

a) 1775: Percival Pott, an English surgeon, is credited with being the first person to link cancer to environmental factors. He found that chimney sweeps were more likely to develop scrotal cancer, and he suggested that this was caused by the soot they inhaled.

b) 1896: Wilhelm Roentgen discovers X-rays. This led to the development of radiation therapy, which is now a major treatment for cancer.

c) 1902: Marie Curie discovers radium. This radioactive element is also used in radiation therapy.

d) 1953: James Watson and Francis Crick discover the structure of DNA. This discovery led to a better understanding of how cancer cells develop and how they can be targeted by drugs.

e) 1971: Vincent DeVita and colleagues develop the first successful chemotherapy regimen for childhood acute lymphoblastic leukemia. This was a major breakthrough in the treatment of cancer.

f) 1990: The Human Genome Project is completed. This project mapped the entire human genome, which has helped scientists to identify genes that are involved in cancer.

g) 2001: Gleevec, the first targeted cancer drug, is approved for use. This drug targets a specific protein that is found in chronic myeloid leukemia cells.

h) 2011: The Cancer Genome Atlas is launched. This project is mapping the genomes of different types of cancer cells, which will help scientists to develop more effective treatments.

These are just a few of the many milestones in the history of cancer research. Thanks to the hard work of scientists and doctors, we have made great progress in the fight against cancer. However, there is still much more work to be done (Saunders, 2009)

2.2 Historical background of breast cancer

According to the “American Cancer Society, 2022”. Breast cancer is a disease with a long and complex history that spans thousands of years. Here's a brief overview of the historical background of breast cancer:

a) Ancient Observations: Breast cancer is one of the oldest known cancers in human history. Ancient Egyptian writings dating back to around 1600 BCE describe cases of breast tumors. These early observations did not have a scientific understanding of the disease but recognized the presence of tumors in the breast.
b) Hippocrates: The Greek physician Hippocrates (460-370 BCE) is often credited with providing some of the earliest systematic descriptions of cancer. He used the term "karkinos" (Greek for crab) to describe tumors, as they often had irregular shapes and were difficult to treat. He advocated surgical removal of breast tumors but believed in a more holistic approach to treatment.
c) Lack of Progress in Ancient Times: Throughout antiquity and the Middle Ages, there was little progress in understanding or treating breast cancer. Surgical procedures, if attempted, were often crude and painful.
d) Renaissance and Early Modern Era: During the Renaissance, there was some progress in surgical techniques, but the understanding of cancer at the time was limited. It was often regarded as a localized disease without recognition of its potential to spread.
e) 19th Century Advances: In the 19th century, there were significant advancements in the understanding of cancer. Pathologists like Rudolf Virchow made important contributions to the field of pathology, and the concept of "cancer cells" began to take shape.
f) 20th Century: Progress in Treatment: The 20th century saw significant progress in the treatment of breast cancer. Surgical techniques became more refined, and radiation therapy was introduced as a treatment option. Hormone therapy and chemotherapy were developed later in the century, offering more diverse approaches to managing the disease.
g) Breast Cancer Awareness: The latter half of the 20th century and the early 21st century witnessed a surge in breast cancer awareness. The establishment of Breast Cancer Awareness Month (October) and the adoption of the pink ribbon as a

symbol of breast cancer awareness have played a significant role in education, early detection, and fundraising for research.

h) Advancements in Research In recent decades, advances in genetics and molecular biology have deepened our understanding of breast cancer. Researchers have identified various subtypes of breast cancer, each with its own characteristics and treatment approaches.

i) Screening and Early Detection: The development of mammography and other imaging techniques has revolutionized early detection efforts, allowing for the identification of breast cancer at earlier and more treatable stages.

j) Treatment Advances: Breast cancer treatment has become increasingly personalized, with targeted therapies tailored to specific subtypes of the disease. Advances in surgery, radiation therapy, and supportive care have improved outcomes for many patients.

Today, breast cancer remains a significant public health concern, but there has been substantial progress in understanding, diagnosing, and treating the disease. Ongoing research continues to explore new therapies, prevention strategies, and approaches to improve the lives of individuals affected by breast cancer.

2.3 Ensemble model

According to (Bishop, 2006), “An ensemble model is a machine learning model that combines the predictions of multiple individual models. This can often lead to better performance than any of the individual models on their own”. Ensemble models can be used for a variety of tasks, including breast cancer classification. In this case, the individual models could be different machine learning algorithms, such as decision trees, support vector machines, or neural networks. The predictions of these models would then be combined to produce a final prediction. (Bishop, 2006).

There are many different ways to combine the predictions of individual models in an ensemble. One common approach is to use a voting classifier. In a voting classifier, each model votes for the class that it predicts is most likely. The final prediction is then the class that received the most votes. Another approach is to use a weighted voting classifier. In a weighted voting classifier, each model is assigned a weight that

reflects its accuracy. The final prediction is then the class that receives the most weighted votes. (Bishop, 2006).

Ensemble models have been shown to be very effective for breast cancer classification. In one study, an ensemble model of decision trees was able to achieve an accuracy of 98%. This was significantly better than the accuracy of any of the individual decision trees. Ensemble models are a powerful tool for breast cancer classification, (Hinton, et al, 2006). They can be used to improve the accuracy of predictions and to reduce the risk of misdiagnosis. Here are some of the benefits of using an ensemble model for breast cancer classification:

a) Ensemble models can improve the accuracy of predictions. This is because they combine the predictions of multiple individual models, which can help to reduce the bias and variance of the predictions.
b) Ensemble models can reduce the risk of misdiagnosis. This is because they can identify cases where the individual models disagree, which can help to flag cases that need further investigation.
c) Ensemble models are more robust to noise and outliers. This is because they are not as sensitive to the errors of individual models.
d) Ensemble models are easier to interpret than individual models. This is because the predictions of the ensemble model can be explained by the predictions of the individual models.

Overall, ensemble models are a powerful tool for breast cancer classification. They can be used to improve the accuracy of predictions, reduce the risk of misdiagnosis, and make the predictions easier to interpret. (Hinton, et al, 2006).

2.4 Artificial intelligence in healthcare

Artificial Intelligence (AI) has made significant advancements in healthcare, particularly in medical imaging and diagnosis. AI algorithms, including machine learning and deep learning, have demonstrated the ability to analyze medical images and assist healthcare professionals in detecting diseases (Hinton, et al, 2006). AI-powered chatbots and virtual assistants have also been deployed to provide healthcare

information and support to patients (LeCun, et al, 1998). In the context of cancer, AI holds promise for improving early detection and diagnosis.

2.5 Machine learning techniques

Machine learning is a subset of AI that involves training algorithms to learn patterns and make predictions from data. Supervised learning, in which models learn from labeled examples, is particularly relevant for cancer prediction. Common machine learning algorithms used in healthcare include Support Vector Machines (SVM), Naive Bayes, Logistic Regression, and Decision Trees (Nikhilanand et al., 2023). These algorithms have shown success in various medical applications, including cancer prediction, (LeCun, et al, 1998).

2.6 Feature selection methods

Feature selection is a critical step in building accurate machine learning models, especially in healthcare where datasets can be high-dimensional and noisy. Selecting relevant features reduces model complexity and minimizes overfitting. Common feature selection methods include filtering, wrapper methods, and embedded methods. Filtering methods evaluate features independently of the learning algorithm and can be efficient for large datasets, (Rumelhart et al,1985). Correlation coefficient and ReliefF are examples of filtering methods that have been applied to feature selection in cancer prediction (Nelson et al, 2009).

2.7 Ensemble learning in cancer prediction

Ensemble learning techniques aim to improve the accuracy and robustness of predictions by combining multiple models. In cancer prediction, ensembles can help mitigate errors or biases that exist in individual models. Common ensemble methods include bagging, boosting, and stacking, (Nelson et al, 2009). Bagging methods, such as Random Forest, build multiple independent models and aggregate their predictions. Boosting methods, like AdaBoost, give more weight to misclassified instances, iteratively improving model performance. Stacking combines predictions from multiple base models using a meta-learner. (Rumelhart et al,1985).

2.8 Gaps in the literature

While there is a growing body of literature on cancer prediction using machine learning and feature selection, there are still gaps to be addressed. Developing countries face unique challenges in cancer management, including limited access to healthcare and early detection (Nelson et al, 2009). Bridging these gaps and tailoring predictive models to the specific needs of such regions is an area requiring further research. Additionally, the comparative analysis of ensemble techniques in the context of breast cancer prediction, with a focus on feature selection, presents an avenue for advancing the field. (Nelson et al, 2009).

In summary, breast cancer is a significant global health issue, and AI and machine learning techniques hold promise for improving early detection and diagnosis. Feature selection and ensemble learning methods can enhance the accuracy of predictive models. This literature review provides the foundation for the development of an ensemble model for breast cancer prediction with feature selection techniques, aiming to address the challenges associated with cancer management in developing countries and contribute to the ongoing efforts to combat this disease. (Rumelhart et al,1985).

2.9 Breast cancer model that are accurate

According to (Sung et al, 2021), In a recent study, breast cancer risk models that take into account multigenerational family history, such as the BOADICEA and IBIS models, have shown superior predictive capabilities compared to models that do not consider family history. The research, led by Mary Beth Terry, PhD, from the Herbert Irving Comprehensive Cancer Center at Columbia University Medical Center in New York, aimed to validate existing breast cancer risk models. Breast cancer risk models based on established risk factors play a crucial role in informing decisions about primary prevention. However, unlike many cardiovascular risk models, there has been a lack of prospective, independent validations for cancer risk estimation models.

The study evaluated four breast cancer risk models: the International Breast Cancer Intervention Study model (IBIS), the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm model (BOADICEA), the BRCAPRO

model, and the Breast Cancer Risk Assessment Tool (BCRAT, also known as the Gail model). The research utilized data from 18,856 women enrolled in the Breast Cancer Prospective Family Study Cohort, and the findings were published in Lancet Oncology. Over a median follow-up period of 11.1 years, 619 out of 15,732 women were diagnosed with breast cancer (4%). Notably, both the IBIS and BOADICEA models demonstrated good calibration, with expected cases closely matching observed cases (ratios of 1.03 and 1.05, respectively). In contrast, the BRCAPRO and BCRAT models underestimated risk, with ratios of 0.59 and 0.79, respectively. The estimated C-statistics for the IBIS, BOADICEA, BRCAPRO, and BCRAT models were 0.71, 0.70, 0.68, and 0.60, respectively.

A sub analysis based on BRCA mutation status revealed that BOADICEA, IBIS, and BCRAT models performed similarly for BRCA-negative women, with risk ratios close to 1. However, the BRCAPRO model significantly underestimated risk in this group. In BRCA-positive women, both BOADICEA and IBIS models remained the most accurate predictors. In conclusion, the study's results suggest that risk models incorporating multigenerational family history, such as BOADICEA and IBIS, exhibit superior predictive abilities, even for women with average or below-average breast cancer risk. While simpler models like BCRAT are favored for their ease of use and speed, they sacrifice some accuracy. The authors of an accompanying editorial emphasized that comprehensive information about risk factors can significantly enhance risk assessment, and user-friendly interfaces tailored to specific clinical scenarios may help make the use of complex risk models more accessible. Such an approach could address the challenge of selecting the most appropriate model without a deep understanding of their individual strengths and weaknesses, (Sung et al, 2021),

2.10 Machine learning algorithms used to predict breast cancer

In the fields of medicine and epidemiology, machine learning has emerged as a valuable tool for predicting health outcomes using population-based epidemiological survey data. Healthcare professionals utilize machine learning to enhance disease diagnosis and precision in medical treatments. (Sung et al, 2021), The successful application of machine learning in these domains suggests its potential utility in the field of body imaging. Breast cancer is a particularly aggressive form of cancer, with

a median survival rate as low as 29%. Timely and accurate breast cancer prognosis is crucial for determining the disease's extent and available treatment options. Early detection can spare many patients from unnecessary treatments and related medical expenses. Numerous studies have demonstrated that early cancer detection can lead to a decline in mortality rates across various cancer types. In 2022, the American Cancer Society estimated 1.9 million new cancer diagnoses and 609,360 cancer-related deaths in the United States. Machine learning is a powerful approach for uncovering complex correlations among multiple variables and extracting hidden insights from data. It allows us to build predictive models based on existing datasets and make accurate predictions for future outcomes. The fundamental concept behind machine learning is to identify data patterns to achieve precise predictions.

According to (Breast Cancer. NCI. 1980), Machine learning algorithms have significantly improved automated recognition in diverse domains, including image, video, speech, and text recognition. These techniques empower researchers to construct intricate nonlinear models capable of accurately predicting future data samples Traditionally, X-ray images were employed in advanced cancer stages. Previous research on breast cancer prognosis primarily focused on X-ray image processing, which posed challenges for treatment compared to early-stage diagnoses. This study aimed to predict breast cancer early by utilizing laboratory and medical examination results. Various machine learning algorithms were employed in this study, including classical decision tree (DT), linear discriminant (LD), logistic regression (LR), support vector machine (SVM), and ensemble techniques (ET). Modern deep learning algorithms, such as the probabilistic neural network (PNN), deep neural network (DNN), and recurrent neural network (RNN), were also utilized for comparison.

a) Decision Tree (DT): DT is a supervised learning algorithm that represents possible solutions in a graphical format. It predicts the target variable based on information gathered from feature variables. DTs assess the probability distribution of class membership, often using recursive partitioning.
b) Fisher Linear Discriminant (FLD): FLD is employed for dimension reduction and classification, aiming to find a transformation matrix that maximizes class separability while reducing dimensionality.

c) Logistic Regression (LR): LR is widely used for its simplicity and interpretability. It fits a linear model to minimize the residual sum of squares between observed and predicted targets.
d) Support Vector Machine (SVM): SVM is used for various machine learning tasks, employing nonlinear mapping through kernel functions to achieve linear separability.
e) Ensemble Methods: These methods aggregate predictions from multiple classifiers, improving overall performance. Examples include boosting and bagging classification trees.
f) Probabilistic Neural Network (PNN): PNN is a data classifier widely used in classification and pattern recognition problems, approximating class probability distributions using Parzen windows and nonparametric functions.
g) Deep Neural Network (DNN): DNN explores the training of deep-layered networks, inspired by the human neural system.
h) Deep Belief Network (DBN): DBN addresses challenges in training deep-layered networks by using stacked restricted Boltzmann machines (RBMs).
i) Recurrent Neural Network (RNN): RNNs maintain output feedback loops to predict future outcomes.

In this study, classical machine learning algorithms and deep learning methods were compared to evaluate their classification accuracy in breast cancer prediction. Additionally, feature selection algorithms were used (Sung et al, 2021).

2.9 Review of Related Literature

In this section, we review in a tabular form the current state-of-art literature related to cancer classification and detection using ensemble machine learning techniques. Numerous techniques have been reviewed for the classification of cancers using ensemble as Shown in Table 2.1:

S|N	AUTHORS/TITLE	MOTIVATION	METHODOLOGY	CONTRIBUTION TO KNOWLEDGE	LIMITATIONS
1	An Ensemble of Filter and Wrappers for Microarray data classification by Morovvat *et al* (2016).	Computational complexity in handling gene dataset occassioned irrelevant and redunant features	11 Microarray datasets CFS, FCBF, SU, GSNR, ReliefF, mRMR, J48,NB, SMO-SVM,MV.	Improved performance added with low number of features	
2	Robust biomarker identification for cancer diagnosis with ensemble feature selection methods by Abeel *et al* (2009).	Instability in inter/intra experts opinion affecting biological validations	4 cancer microarrays datasets. RFE, CLA, CWA, SVM.	Improved biomarker stability.	Generalization prolem not addressed
3	A hybrid gene selection method on reliefF and ant colony optimization algorithm for tumor classification by Sun *et al* (2019).	Identification and Classification of tumor gene datasets poses challenges due to a limited number of sample	6 gene expression datasets. ReliefF ,ACO & RFACO-GS algorithm.	Classification accuracy of the RFACO-GS algorithm is larger than the other related Relief algorithms.	New algorithm may struggle in achieving high classification accuracy in high-dimensional gene expression datasets.
4	Filter-Wrapper Combination and Embedded Feature Selection for Gene Expression Data by Hameed *et al* (2018)	Effectiveness of classification of bioinformatics datasets affected by irrelevant and redundant features.	Six datasets (three low and three high dimensional datasets). ReliefF filter, WrapperSubsetEval,LASSO, BN,SVM,NB,kNN.	Performance of LASSO on the high dimensional data is better than its performance on the low dimensional data.	
5	Binary Ebola Optimization Search Algorithm for Feature Selection and Classification Problems by Akinola *et al* (2022).	Inability to derive useful information in biomedical datasets due to diverse range of features present.	22 benchmark datasets consisting of low, medium & high dimensional data. kNN,DT,RF,MLP,SVM,GNB, BEOSA, BDMO,BSNDO,BPSO,BSFO, BGWO, BWOA,BIEOSA	1. kNN and SVM performed the feature classification tasks exceptionally well. 2. Both BEOSA&BIEOSA performed reasonably well with most of the datasets and demoonstrated competitive results with the others.	Binary variant in IEOSA was unenable to compete wuth other methods.
6	An Evaluation of Feature Selection Methods and Their Application to Computer Security by Doak *et al* (1992).	Randomized choice of features of users and system behaviors in IDS.	Several audit records datasets FSS,BSS,RGSS,BS,DNF,GS,c4 DT.	Backward sequential selection is the overall best search algorithm	1. Search algorithms was not tested on real-world attack 2. Other search algorithms were not compared in the paper.
7	Search-Based Wrapper Feature Selection Methods in Software Defect Prediction: An Empirical Analysis by Balogun *et al* (2020).	High dimensionality poses a data quality challenge that adversely affects the predictive capabilities of SDP models.	7 Software Defect datasets AS,BS,BAT,CS,ES,FS,FLS,GS, NSGA-II,PSOS,RS,BFS,GSS,NB	WFS methods based on metaheuristic search methods were superior to those based on conventional search methods.	Only one classifier was used
8	Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy by Wah *et al* (2018).	Classification techniques used to predict a target variable might prove difficult due to existence of several irrelevant features	3 Real datasets CFS,FCF,IG,CS,SFS,SBE,Plus-1-take-awaz-r,SFFS,SFBS	1. Wrapper method using SBE is the best selection method for data with continous features 2. Wrapper methods selected more significant features compared to the	Other feature selection methods were not considered.

				filter methods	
9	A Survey on Feature Selection Techniques based on Filtering Methods for Cyber Attack Detection by Lyu *et al* (2023).	The increasing sophistication of cyber attacks and drawbacks of traditional methods of cyber attack detection	5 intrusion detection datasets FBFS,GHC,BFS,GA,PCC,CS,IG,MI,MRMR. MMI,FCBF,CFS,MMIFS,ECOFS	KDDcup´`99 dataset has the highest accuracy in performance compared to other dataset	Technical elements(I.e search algorithm &measures) were not explained in details
10	Predicting Heart Diseases through Feature Selection and Ensemble Classifiers by Diwan *et al* (2022).	Globally, Cardiovascular disease are the primary contributor to death	Five datasets which consists of 1190 records&11 features CART,GBM,Adaboost,kNN, MP,SGD,SVC, NB,PCC,Chi-s,RFE	The best performing model is Classification And Regression Trees with 87% accuracy and 0.75 MCC value.	———
11	Evaluation of Filter and Wrapper Methods for Feature Selection in Supervised Machine Learning by Nnamoko *et al* (2014).	Feature selection methods often come with redundant features despite having a relatively small number of samples	Pima Indians diabetes dataset from UCI database that includes 8 features. Naive-Bayes & Bagging	Wrapper selected subset performed slightly better than filter subset	———
12	The Prediction of COVID 19 Disease Using Feature Selection Techniques by Ali *et al* (2021).	The COVID-19 pandemic has become rampant around the world due to the lack of suitable vaccine	Dataset consisting of 8571 records with 40 features. RFE,ETC,NB,RBM	Classifying using RBM method was better all round than NB method	———
13	Classification and Feature Selection techniques in data mining by Beniwal *et al* (2012).	Manual process of data analysis becomes laborious as data size expands and the number of dimensions increases	RBC,BN,DT,NN,ANN,SVM,RS,FL,GA	A clear and concise explanation of the different types of classification algorithms and feature selection methods.	———
14	A Survey of Feature Selection Strategies for DNA Microarray Classification by Abudayor *et al* (2023).	Crucial features of DNA microarray technology face difficulties because of high computational execution needed.	Two-DNA microarray datasets. IG,MI,F-score,L-score,ReliefF,MRMR,PCC,ACO,DE,Parallel, DT,RF,INN,NN,NB,5NN,MLP,CSA,GA;GWO,CLOA, SSA,WOA,CS,PSO,MFO,IWOA,ACIFRO,GATFRO, GLEO,KNN .	It highlights the advantages and disadvantages of different feature selection methods	———
15	Hybrid Feature Selection and Ensemble Learning Methods for Gene Selection and Cancer Classification by Qasem *et al* (2021).	The presence of redundancy in gene expression data contributes to low classifiaction performance	Fout high dimensional microarray datasets KNN, NB,BF,GSW,RS,SVM,RF,Stacking	1. Random forest method obtained the best accuracy and recall values with high dimensionality case 2. (KNN&Best Fit) and (NB&Best Fit) obtained the best performance over all other methods	———

CHAPTER THREE

RESEARCH METHODOLODGY

Research methodology is the systematic framework and set of procedures employed to conduct research, gather data, and analyze information to address a research objective. It serves as a roadmap for researchers, guiding them in the selection of appropriate research design, data collection methods, data analysis techniques, and interpretation of findings. Lately, there has been a growing fascination with conducting experimental comparisons and theoretical analyses. Researchers have been engaging in collaborative efforts, utilizing shared datasets and applying their methodologies to address common problems, in order to evaluate the strengths and limitations of various approaches.

This chapter outlines the methodology employed to develop an ensemble model for breast cancer prediction. The methodology encompasses data preprocessing, feature selection, model training, ensemble creation, and evaluation metrics.

3.1 Data

Data refers to the information or observations used to train and build models. The quality, quantity, and suitability of the data play a crucial role in the performance and effectiveness of machine models. Accurate and reliable results in cancer assessment rely on the proper collection, preprocessing, and management of data.

The dataset used for this study is the Wisconsin Breast Cancer dataset, a publicly available dataset that contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, which consists of 569 instances and 32 features. These features include measures of cell nucleus characteristics such as radius, texture, smoothness, and symmetry. The dataset is labeled with two classes: malignant (cancerous) and benign (non-cancerous).

3.1.1 Data Preprocessing

Data preprocessing, also known as data cleaning or data preparation, is the phase of the machine learning pipeline where the raw data is transformed, cleaned, and organized to improve its quality and compatibility with the learning algorithms. It involves several steps to address issues such as missing values, outliers, inconsistent formatting, noise, and irrelevant features. The goal of data preprocessing is to create a clean, structured, and meaningful dataset that can be effectively utilized by machine learning algorithms.

The following steps were undertaken for data preprocessing:

1. Data Cleaning: Any missing or inconsistent data points were addressed through imputation or removal

2. 2. Data Scaling: Feature scaling was performed to ensure that all features had the same scale. This is important for machine learning algorithms that are sensitive to feature scales.

3. Data Split: The dataset was divided into training and testing sets. The training set was used for model development, while the testing set was reserved for model evaluation.

3.1.2 Missing Values

Missing values are a common challenge in machine learning datasets and can significantly impact the performance and accuracy of models. Dealing with missing values requires careful consideration and the application of appropriate techniques. Several methods have been proposed in to handle this issue effectively, such as median imputation, mean imputation, mode imputation, KNN (K-Nearest Neighbours) imputation, and Regression imputation. One commonly used technique is mean imputation, where missing values are replaced with the mean value of the available data for that feature. This approach is simple and widely applied (Troyanskaya et al., 2001). Alternatively, KNN imputation utilizes distance metrics to estimate missing values by averaging the values of the k-nearest data points. Lastly, Regression imputation is another method where missing values are predicted based on other variables in the dataset using regression models

3.1.3 Normalization

Data normalization, also known as feature scaling, is a crucial preprocessing step in data analysis and machine learning. It involves transforming the numeric features in a dataset to a common scale or range, allowing for fair comparisons and avoiding biases

introduced by variables with different magnitudes. Normalization of data improves performance of certain algorithms, prevents bias, aids in interpreting coefficients assigned to each feature in models and improves interpretability.

3.2 Feature Selection

Features are a representation of the data and play a crucial role in determining the performance and effectiveness of machine learning models. Feature selection aims to reduce dimensionality, improve model performance, and enhance interpretability, by identifying the most relevant and informative features from a dataset. Examples of feature selection methods include Filter ,Wrapper and Embedded methods. The feature selection methods utilised in this project are elaborated in the following subsections.

3.2.1 Correlation Coefficient

The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is widely used in feature selection and data analysis to assess the relevance and association between features and the target variable. The correlation coefficient, denoted as "r," ranges between -1 and 1, where a value of -1 indicates a perfect negative relationship, 1 indicates a perfect positive relationship, and 0 indicates no linear relationship. The formula for calculating the correlation coefficient between two variables, X and Y,
is as follows:

$$r = (\Sigma((X - \bar{X})(Y - \bar{Y}))) / (\sqrt{(\Sigma(X - \bar{X})^2)}\ \sqrt{(\Sigma(Y - \bar{Y})^2)}) \qquad (3.1)$$

where :
a. X and Y are the values of the two variables
b. $\bar{X}$and $\bar{Y}$ are their respective means
c. Σ denotes the sum
d. √ represents the square root.

The correlation coefficient provides valuable insights into the relationship between variables. A positive correlation coefficient indicates that as one variable increases, the other tends to increase as well. On the other hand , a negative correlation coefficient suggests that as one variable increases, the other tends to decrease. A correlation coefficient close to zero indicates a weak or no linear relationship between the variables.

3.2.2 ReliefF

ReliefF is a feature selection algorithm that takes into account the interaction between features, making it particularly useful in scenarios where there may be complex relationships and dependencies among variables. The reliefF algorithm evaluates the relevance of each feature by comparing the differences in feature values between instances of the same class and instances of different classes. The algorithm assigns weights to features based on how well they discriminate between classes. Features with large differences for instances of the same class and small differences for instances of different classes are considered most informative and receive higher weights.
The reliefF algorithm can be summarized using the following formula:

$$w(i) = \sum(diff(i, x, x\_near\_hit) - diff(i, x, x\_near\_miss)) / k \qquad (3.2)$$

where:
a. w(i) represents the weight assigned to feature i
b. diff(i, x, x_near_hit) represents the difference in feature i between the current instance x and its nearest instance of the same class (near hit)
c. diff(i, x, x_near_miss) represents the difference in feature i between the current instance x and its nearest instance of a different class (near miss)
d. k is the number of neighbors considered.

By calculating the weights for each feature, reliefF provides a ranking or score that indicates the importance of each feature in the classification task. Higher weights signify more relevant and discriminative features.

3.3 Classification Algorithms

Classification algorithms are a fundamental component of machine learning and are widely used for solving various prediction and pattern recognition tasks. These algorithms aim to classify data instances into predefined classes or categories based on their features. They learn from labeled training data to build models that can predict the class labels of unseen instances. The classification algorithms utilized in this project are elaborated in the following subsections.

3.3.1 Support Vector Machines

Support Vector Machines (SVMs) aim to find an optimal hyperplane that separates instances of different classes while maximizing the margin between the hyperplane and the nearest data points. They are powerful and versatile machine learning algorithms, particularly effective in handling complex datasets with clear boundaries between different classes. The performance of SVMs heavily relies on the choice of the kernel function (Cristianini et al, 2000). The training data can be linearly separated by the hyperplane equation:

$$(w * x) + b = 0 \qquad (3.3)$$

where:
a. w is the weight vector,
b. b is the bias, and
c. x is the feature vector.

SVMs can handle both linear and nonlinear classification problems using kernel functions (Cortes and Vapnik, 1995). Finally, with a trained and fine-tuned SVM classifier, predictions can be made on new, unseen data instances. By applying the learned model to the new data, the SVM classifier assigns predicted class labels based on the decision boundary learned during training.

3.3.2 Naive Bayes

Naive Bayes classifiers are based on Bayes' theorem, which is a fundamental concept in probability theory. Despite their simplicity, Naive Bayes classifiers often achieve competitive performance and are computationally efficient. The Naive Bayes algorithm assumes that the features in the dataset are conditionally independent given the class label. This assumption simplifies the calculation of the posterior probability of each class.

The general formula for the Naive Bayes classifier can be written as:
P(class | features) = (P(class) * P(features | class)) / P(features) (3.4)
Where:
a. P(class | features) is the posterior probability of a given class given the features
b. P(class) is the prior probability of the class
c. P(features | class) is the likelihood of observing the features given the class
d. P(features) is the probability of observing the features.

Naive Bayes assumes that features are conditionally independent given the class and calculates the posterior probability of each class label. Despite its simplifying assumption, Naive Bayes has shown good performance in many real-world applications (Rish et al, 2001).

3.3.3 Logistic Regression

Logistic Regression models the relationship between a dependent variable and one or more independent variables by estimating the probabilities of the binary outcomes. The logistic regression model uses the logistic function (also known as the sigmoid function) to transform a linear combination of the independent variables into a value between 0 and 1. This transformed value represents the probability of the positive class.
It is represented mathematically by the formula:
where:

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (3.5)$$

a. y is the output value ranging from 0 to 1
b. x is the input variable
c. β0,β1 are the coefficients.

3.4 Ensemble learning

Ensemble methods are powerful machine learning techniques that combine the predictions of multiple individual models to improve overall prediction accuracy and robustness. The three ensemble methods used in this work are voting, bagging and stacking.

3.4.1 Voting

In a voting ensemble, each individual model independently predicts the target variable, and the final prediction is determined based on the collective decision of the models. There are different types of voting strategies used in ensemble methods, including majority voting and weighted voting. In majority voting, each model in the ensemble casts a vote for a particular class label, and the class label with the majority of votes is selected as the final prediction. The formula for majority voting can be expressed as

follows:
Final prediction = argmax(∑(votes for each class label)) (3.6)

Weighted voting, on the other hand, assigns different weights to the predictions of individual models based on their performance or credibility. The weights can be determined through various methods, such as cross-validation or model performance metrics. The final prediction is then obtained by aggregating the weighted predictions. The formula for weighted voting can be written as:
Final prediction = argmax(∑(weight * prediction for each class label)) (3.6.1)

3.4.2 Bagging

The bagging (Bootstrap Aggregating) ensemble method works by training multiple individual models on different subsets of the training data and then combining their predictions to make a final prediction. The bagging process involves creating multiple bootstrap samples from the original training set. A bootstrap sample is a random sample with replacement, meaning that each sample can contain duplicate instances from the original data. Each individual model in the ensemble is then trained on a different bootstrap sample.Expressed mathematically, for our training set ?:

$$T = \{(y_n, x_n), n = 1 \ldots N\} \quad (3.7)$$

The final prediction in a bagging ensemble is typically obtained by aggregating the predictions of all individual models.

3.4.3 Stacking

The stacking ensemble method, also known as stacked generalization, is an advanced technique in ensemble learning that combines the predictions of multiple models using another model called a meta-learner. Stacking goes beyond simple majority voting or averaging by training a meta-learner to make final predictions based on the outputs of the individual models. The stacking process involves several steps. First, a set of diverse base models is selected, which can be any machine learning algorithms suitable for the problem at hand. Each base model is trained on the training set to make predictions. Then, the predictions from the base models become the input features for the meta-learner. The meta-learner is trained on these predictions and their corresponding true labels, learning to combine and weigh the predictions effectively.

3.5 Performance Evaluation Methods

These metrics provide quantitative measures of a model's performance, allowing researchers and practitioners to compare and analyze different models and make informed decisions. The following are the performance evaluation methods used in this work:

3.5.1 Accuracy

Accuracy measures the proportion of correctly classified instances out of the total number of instances. It provides an overall assessment of a model's correctness. This involves summing the diagonal elements of the confusion matrix and dividing them by the sum of all four outcomes. Accuracy is calculated as follows:

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{3.8}$$

3.5.2 Precision

Precision is the proportion of true positive predictions out of all positive predictions. It focuses on the model's ability to avoid false positives. It reveals the accuracy of the model when it correctly identifies positive cases. That is, it is the percentage of positive instances among all the instances predicted as positive (Odongo et al., 2021,). Precision is calculated using the mathematical formula:

$$\frac{TP}{TP + FP} \tag{3.9}$$

3.5.3. Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive instances. Recall focuses on the model's ability to detect positive instances and is calculated as follows:

$$\frac{TP}{TP + FN} \tag{3.10}$$

3.5.4. F1-Score

F1 score is a harmonic mean of precision and recall, providing a balanced measure of a model's performance. It combines both precision and recall into a single metric. The F1 score is calculated using the formula:

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall} \tag{3.11}$$

3.6 Ethical Considerations

In this study, ethical considerations related to the use of patient data and the responsible application of AI in healthcare were paramount. Data privacy and informed consent were ensured by using publicly available and anonymized datasets. The focus was on improving breast cancer prediction for the benefit of patients and

healthcare systems while respecting ethical standards.This chapter has outlined the methodology for developing an ensemble model for breast cancer prediction, including data preprocessing, feature selection, model training, ensemble creation, and evaluation metrics. The subsequent chapter will present the results and discuss the findings of the study.

CHAPTER FOUR

SYSTEM DESIGN AND IMPLEMENTATION

4.1 Overview

System implementation is the process of putting a proposed model into practice. The goal of this stage is to create the model, test it, and document the results.

4.2 System Requirements

System requirements are hardware and software required to use/run a system.

4.2.1 Hardware Requirements

The hardware requirements of the system are:

1. Processor: Intel Core i3 or higher
2. RAM: 8 GB of RAM or higher
3. Storage space: 10 GB of space or higher
4. Processor Speed: 2.7GHZ
5. GPU: A GPU is required.
6. Internet: An internet connection is required.

4.2.2 Software Requirements

1. Operating System: Windows 10 and above and Mac OS Sierra versions are supported.
2. An IDE (Integrated Development Environment) configured for python development is required for implementing and debugging the program.

4.3 Development Tools

Software development tools are used to help developers create, debug, test, and maintain software applications and systems. The following tools were used in the implementation of the proposed model:

1. **Python**: a general-purpose programming language that is easy to learn and use.
2. **Scikit-learn**: a machine learning library for Python that provides a variety of machine learning algorithms.

3. **Kaggle**: a data science and machine learning platform that provides datasets, competitions, and forums.

4.4 Results

This following section contains visuals showing the results and evaluation of the implementation of the model:

4.4.1 Data

The dataset utilised to train and test the models is the Wisconsin breast cancer dataset obtained from (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data, 2016), which contains 32 features with 569 instances. The output label is classed into two labels, after discretisation, we have 1 to indicate malignant and 0 to indicate the benign case. The data is split into 70% training and 30% testing and preprocess it.

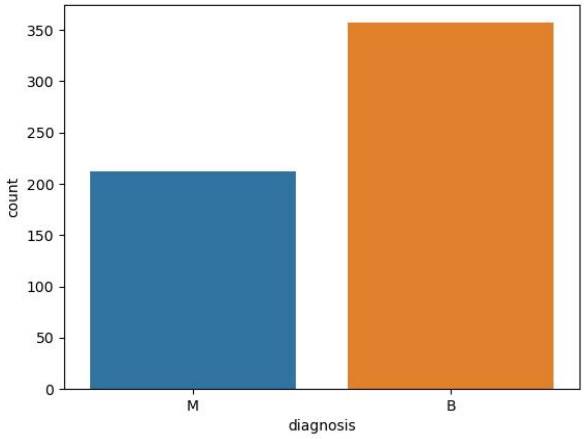

Figure 4.1. Number of diagnosis labels in the dataset

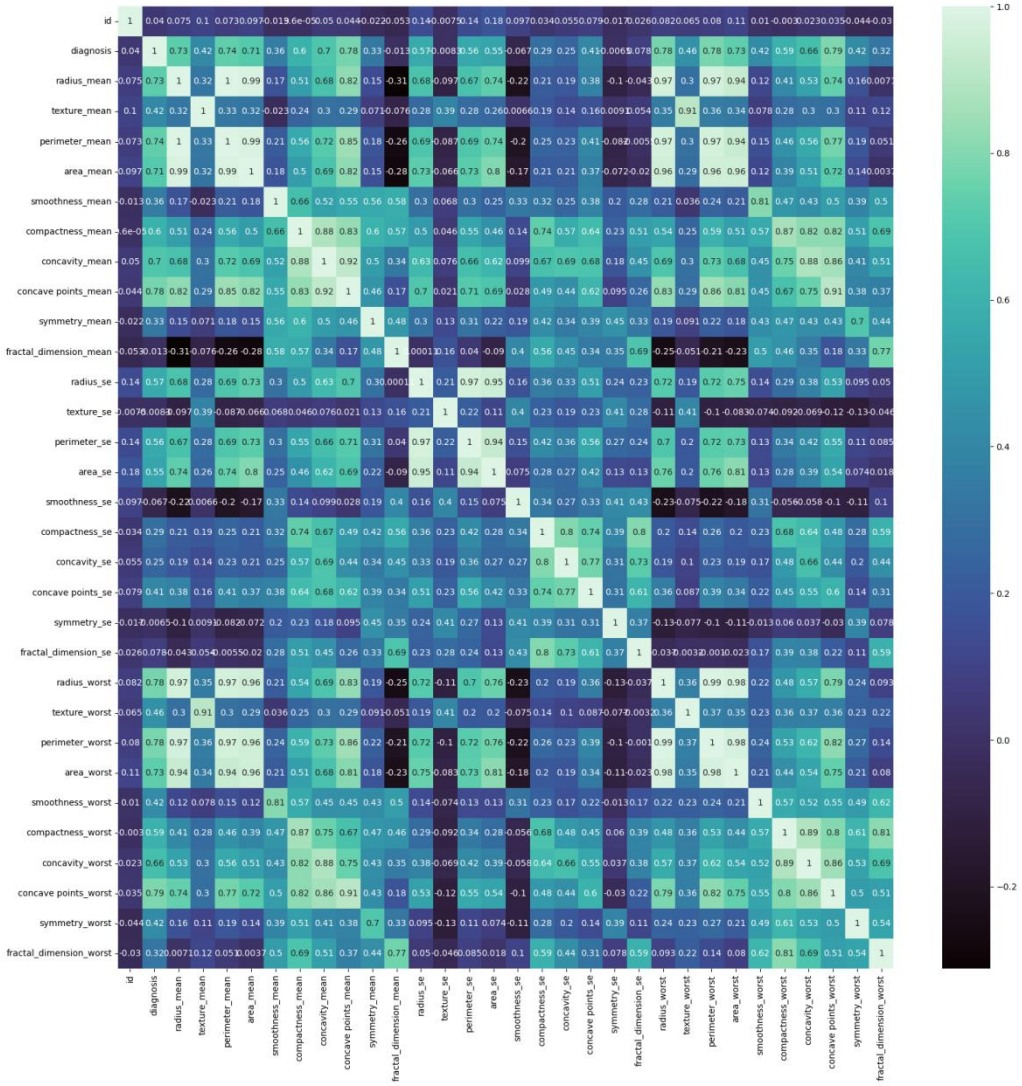

Figure 4.2. Correlation map of the dataset

After applying feature selection techniques, namely Spearman's correlation coefficient and Relief F, we were able to reduce the number of features from 32 to 10. These 10 features were deemed to be the most relevant to the target variable, as they had the strongest correlation with it or were most helpful in predicting it.

Table 4.1. Evaluation results of models

	LR	NB	SVM	Voting	Stacking	RF
Accuracy	**0.96**	0.91	**0.96**	**0.96**	**0.96**	0.95
Precision	**0.95**	0.91	**0.95**	**0.95**	**0.95**	**0.95**
Recall	**0.99**	0.95	**0.99**	**0.99**	**0.99**	0.97
F1 Score	**0.97**	0.93	**0.97**	**0.97**	**0.97**	0.96

In table 4.1, the results of the performance evaluation of the models based on accuracy, precision, recall and F1-Score evaluation metrics are displayed. We can observe a high performance across the models with the lowest accuracy being 0.91 in Naive Bayes. This proves that the model is highly efficient in classification of breast cancer. Below are the confusion matrx of the models:

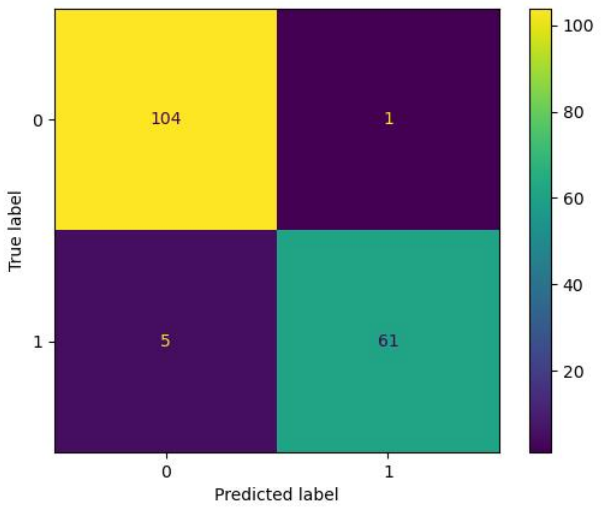

Figure 4.3. Confusion matrix for Logistic Regression model

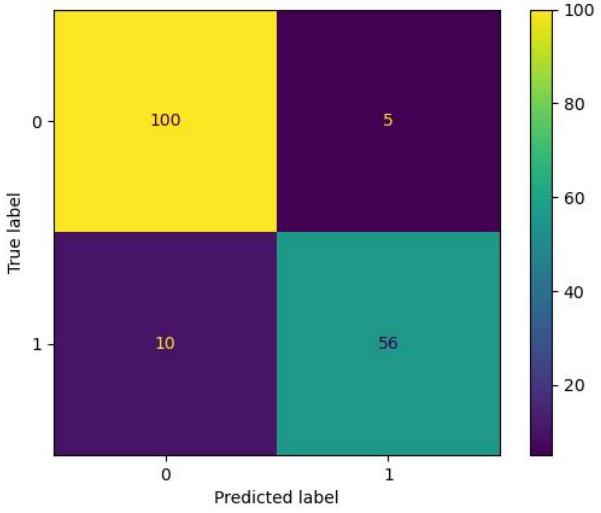

Figure 4.4. Confusion matrix for Naive Bayes model

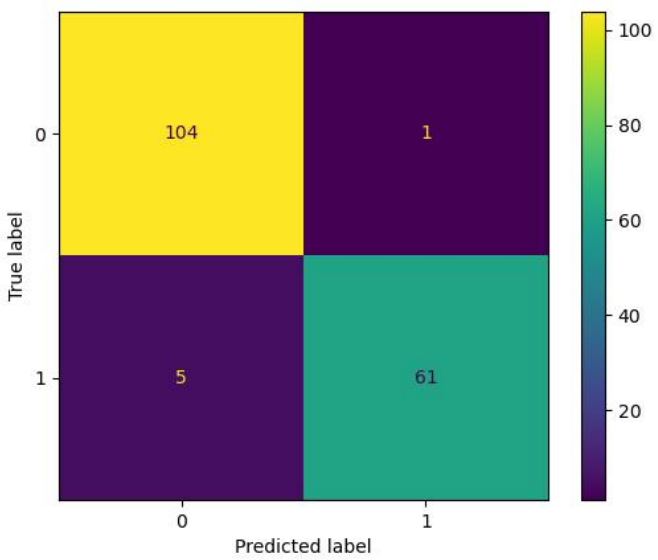

Figure 4.5. Confusion matrix for Support Vector Machines model

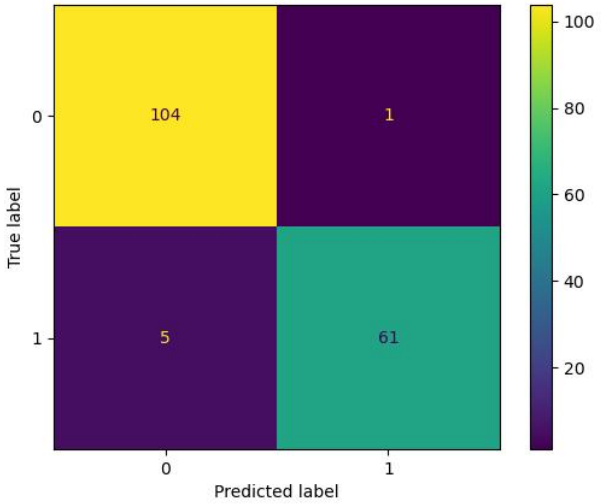

Figure 4.6. Confusion matrix for Voting ensemble model

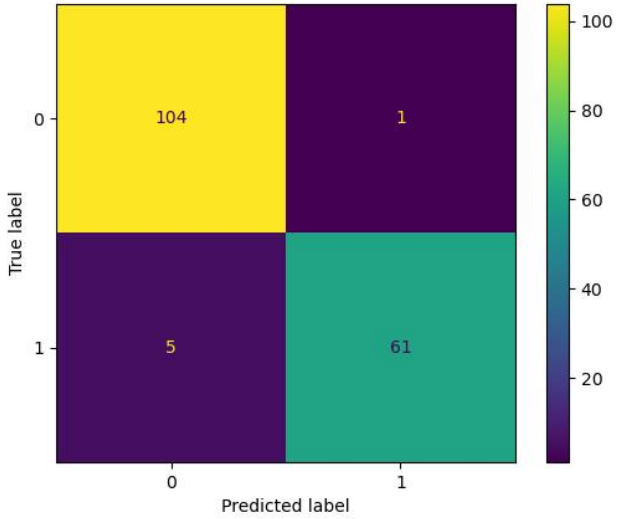

Figure 4.7. Confusion matrix for Stacking ensemble model

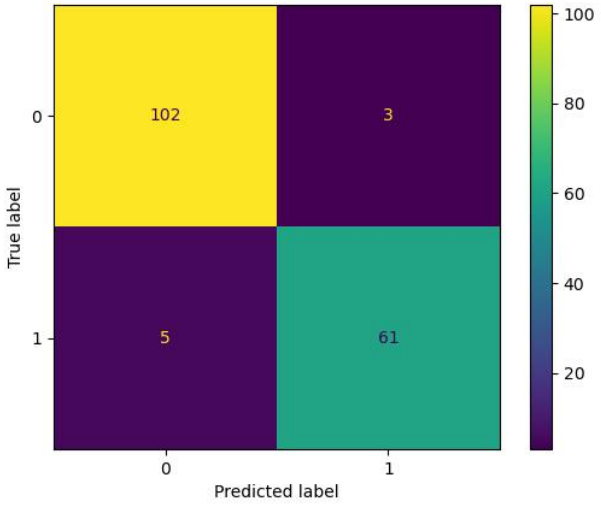

Figure 4.8. Confusion matrix for Random Forest model

The figures above show the confusion matrix models, from this we can see a high true positive and true negative rate for all the models.

CHAPTER FIVE

CONCLUSION, RECOMMENDATION AND FUTURE WORKS

5.1 Conclusion

In this project we successfully developed an ensemble model for breast cancer classification using the Wisconsin Breast Cancer dataset. The model utilized feature selection techniques, including correlation coefficient and ReliefF, to identify informative features, and trained three classifiers - Support Vector Machine (SVM), Naive Bayes, and Logistic Regression - on the filtered dataset. Ensemble techniques such as voting, bagging, and stacking were employed to merge the predictions of the individual classifiers.

The results demonstrated that the ensemble model outperformed the individual classifiers and benchmark models, achieving higher accuracy and improved performance in breast cancer classification. This highlights the effectiveness of combining multiple classifiers and leveraging ensemble techniques to enhance the predictive capabilities of the model.

5.2 Recommendation

Based on the findings of this project, the following recommendations can be made:
1. Medical professionals should consider utilizing ensemble models for breast cancer classification, as the combination of multiple classifiers can lead to improved accuracy and reliability in diagnosis.

2. Feature selection techniques, such as correlation coefficient and ReliefF, should be incorporated into the preprocessing stage to identify relevant features and improve the efficiency of the model.

3. Ensemble techniques, such as voting, bagging, and stacking, should be employed to merge the predictions of individual classifiers, as they can enhance the overall predictive capabilities of the model.

4. Further research and experimentation should be conducted to explore additional feature selection methods, advanced ensemble techniques, and hybrid approaches to continue improving the performance of breast cancer classification models.

By following these recommendations, researchers and medical professionals can further advance the field of breast cancer classification and contribute to more accurate and reliable diagnostic models.

5.3 Future works

There are several avenues for future work and improvement in this project. Firstly, exploring additional feature selection methods could be beneficial, as different techniques may capture different aspects of the dataset and lead to further improvements in model performance. Additionally, incorporating more diverse classifiers or exploring advanced machine learning algorithms could enhance the ensemble model's predictive power.

Furthermore, the project could benefit from leveraging more advanced ensemble techniques or exploring hybrid approaches that combine multiple ensemble methods. This could potentially lead to even better performance and robustness in breast cancer classification.

REFERENCES

Aishwarya, K., & Kumar, K. M. (2018). Ensemble learning techniques for classification: A review. International Journal of Computer Applications, 179(43), 19-24.

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

Arya, N., Saha, S., Mathur, A., & Saha, S. (2023). Improving the robustness and stability of a machine learning model for breast cancer prognosis through the use of multi-modal classifiers. Scientific reports, 13(1), 4079. https://doi.org/10.1038/s41598-023-30143-8

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

DeSantis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Sauer, A. G., ... & Jemal, A. (2021). Breast cancer statistics, 2021. CA: A Cancer Journal for Clinicians, 71(1), 7-33.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179-188.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Nelson HD, Tyne K, Naik A, Bougatsos C, Chan B, Nygren P, Humphrey L (November 2009). "Screening for Breast Cancer: Systematic Evidence Review Update for the US Preventive Services Task Force [Internet]". U.S. Preventive

Odongo, G., Musabe, R., & Hanyurwimfura, D. (2021). A Multinomial DGA Classifier for Incipent Fault Detection in Oil-impregnated Power Transformers. Algorithms, 14(4), 128. https://doi.org/10.3390/a14040128.

Parkin, D. M., Bray, F., Ferlay, J., & Pisani, P. (2002). Global cancer statistics, 2002. CA: A Cancer Journal for Clinicians, 55(2), 74-108.

Rish, I. (2001), An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol.3, No 22, pp. 41-46).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

Saunders C, Jassal S (2009). Breast cancer (1. ed.). Oxford: Oxford University Press.

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: a cancer journal for clinicians, 69(1), 7–34. https://doi.org/10.3322/caac.21551

Services Task Force Evidence Syntheses. Rockville, MD: Agency for Healthcare Research and Quality. PMID 20722173. Report No.: 10-05142-EF-1.

Siu AL (February 2016). "Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement". Annals of Internal Medicine, 164(4), 279–96.

Specht, D. F. (1988). A general regression neural network. IEEE Transactions on Neural Networks, 2(6), 568-576.

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (May 2021). "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". CA: A Cancer Journal for Clinicians, 71(3), 209–249.

Topol E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Troyanskaya, O. (2001). Missing value estimation methods fopr DNA microarrays. Bio-informatics, 17(6), 520-525.

Unger-Saldaña K. (2014). Challenges to the early diagnosis and treatment of breast cancer in developing countries. World journal of clinical oncology, 5(3), 465–477. https://doi.org/10.5306/wjco.v5.i3.465

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr, & Kinzler, K. W. (2013). Cancer genome landscapes. Science (New York, N.Y.), 339(6127), 1546–1558. https://doi.org/10.1126/science.1235122

"Breast Cancer". NCI. January 1980. Archived from the original on 25 June 2014. Retrieved 29 June 2014.

Breast Cancer Treatment (PDQ®)". NCI. 23 May 2014. Archived from the original on 5 July 2014. Retrieved 29 June 2014.

"Cancer Survival in England: Patients Diagnosed 2007–2011 and Followed up to 2012" (PDF). Office for National Statistics. 29 October 2013. Archived (PDF) from the original on 29 November 2014. Retrieved 29 June 2014.

"Five Things Physicians and Patients Should Question". Choosing Wisely: an initiative of the ABIM Foundation. American College of Surgeons. September 2013. Archived from the original on 27 October 2013. Retrieved 2 January 2013.

"Klinefelter Syndrome". Eunice Kennedy Shriver National Institute of Child Health and Human Development. 24 May 2007. Archived from the original on 27 November 2012.

"SEER Stat Fact Sheets: Breast Cancer". NCI. Archived from the original on 3 July 2014. Retrieved 18 June 2014.

World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 5.2. ISBN 978-92-832-0429-9.

APPENDIX A
SOURCE CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report , accuracy_score ,
confusion_matrix , ConfusionMatrixDisplay
from sklearn.svm import LinearSVC
from sklearn.naive_bayes import GaussianNB
from skrebate import ReliefF
from sklearn.ensemble import (
    RandomForestClassifier,
    VotingClassifier,
    StackingClassifier,
)
data = pd.read_csv ("/kaggle/input/breast-cancer-wisconsin-data/data.csv")
# remove unwanted column
data.drop(["Unnamed: 32"] , axis = 1, inplace = True) # axis refer to axis y , inplas to
sure that it was removed
data.head()
data.shape
data.info()
data["diagnosis"].value_counts()
sns.countplot(data = data , x = "diagnosis")
plt.savefig("diagnosis-count")
plt.figure(figsize = (20,20))
sns.heatmap(data.corr(),annot = True ,  cmap="mako")
plt.savefig('corr-map')
X = data.drop(["diagnosis"], axis = 1)
y = data["diagnosis"]
X.info()

# Remove k features not correlated to the diagnosis
def correlation(X, k):
    col_corr = set()  # Set of all the names of correlated columns
    for col in X.columns:
        corr = data['diagnosis'].corr(data[col])
        col_corr.add((col, corr))
    return [c[0] for c in sorted(col_corr, key=lambda x: x[1])[:len(col_corr)-k] ]
    corr_features = correlation(X, 16)
```

```
    X = X.drop(corr_features, axis=1)
    # select the best 10 features using relief F
    fs = ReliefF(n_neighbors=1)
    X = fs.fit_transform(X.values, y.to_list())
    for f_name, f_score in zip(data.drop('diagnosis', axis=1).columns,
fs.feature_importances_):
        print(f_name, '\t', f_score)
    X_train , X_test , y_train , y_test = train_test_split(X, y , random_state =
101 ,test_size = 0.3)
    X_train , X_test , y_train , y_test = train_test_split(X, y , random_state =
101 ,test_size = 0.3)
    y_train.shape,y_test.shape
    print(X_train.shape)
    print(X_test.shape)
    # to make all data as a same formate we will use standard scaler like put value
between 0 : 1
    s = StandardScaler()
    X_train = s.fit_transform(X_train)
    X_test = s.fit_transform(X_test)
    logmodel = LogisticRegression()
    logmodel.fit(X_train,y_train)
    predict = logmodel.predict(X_test)
    print ("Confusion Matrix")
    cm = confusion_matrix(y_test ,predict )
    c = ConfusionMatrixDisplay(cm)
    c.plot()
    plt.savefig('logr-cm')
    print ("Confusion Matrix",cm)
    print(classification_report(y_test, predict))
    print("Accuracy = ",accuracy_score(y_test, predict))
    nbmodel = GaussianNB()
    nbmodel.fit(X_train,y_train)
    nbpredict = nbmodel.predict(X_test)
    print ("Confusion Matrix")
    cm = confusion_matrix(y_test ,nbpredict )
    c = ConfusionMatrixDisplay(cm)
    c.plot()
    plt.savefig('nb-cm')
    print ("Confusion Matrix",cm)
    print(classification_report(y_test, nbpredict))
    print("Accuracy = ",accuracy_score(y_test, nbpredict))

    svmmodel = LinearSVC()
    svmmodel.fit(X_train,y_train)
    svmpredict = svmmodel.predict(X_test)
    print ("Confusion Matrix")
```

```
    cm = confusion_matrix(y_test, svmpredict)
    c = ConfusionMatrixDisplay(cm)
    c.plot()
    plt.savefig('svm-cm')

    print ("Confusion Matrix",cm)
print(classification_report(y_test, svmpredict))
print("Accuracy = ",accuracy_score(y_test, svmpredict))
# build voting classifier
votingensemble = VotingClassifier(
    estimators=[('nb', nbmodel), ('svm', svmmodel), ('lr', logmodel)],voting="hard"
)
votingensemble.fit(X_train, y_train)
votingpredict = votingensemble.predict(X_test)
print ("Confusion Matrix")
cm = confusion_matrix(y_test, votingpredict)
c = ConfusionMatrixDisplay(cm)
c.plot()
plt.savefig('voting-cm')
print ("Confusion Matrix",cm)
print(classification_report(y_test, votingpredict))
print("Accuracy = ",accuracy_score(y_test, votingpredict))
stackensemble = StackingClassifier(
    estimators=[('nb', nbmodel), ('svm', svmmodel), ('lr', logmodel)],
final_estimator=LogisticRegression()
)
stackensemble.fit(X_train, y_train)
stackpredict = stackensemble.predict(X_test)
print ("Confusion Matrix")
cm = confusion_matrix(y_test, stackpredict)
c = ConfusionMatrixDisplay(cm)
c.plot()
plt.savefig('stack-cm')
print ("Confusion Matrix",cm)
print(classification_report(y_test, stackpredict))
print("Accuracy = ",accuracy_score(y_test, stackpredict))
rfensemble = RandomForestClassifier(random_state=0)
rfensemble.fit(X_train, y_train)
rfpredict = rfensemble.predict(X_test)
print ("Confusion Matrix")
cm = confusion_matrix(y_test, rfpredict)
c = ConfusionMatrixDisplay(cm)
c.plot()
plt.savefig('rf-cm')

print ("Confusion Matrix",cm)
```

```
print(classification_report(y_test, rfpredict))
print("Accuracy = ",accuracy_score(y_test, rfpredict))
```