BREAST CANCER PREDICTION USING RANDOM FOREST AND CNN MODELS

By

Davis Sebi

A DISSERTATION

Submitted to

The University of Liverpool

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

01 December 2023

ABSTRACT

BREAST CANCER PREDICTION USING RANDOM FOREST AND CNN MODELS

By

Davis Sebi

Breast cancer is a life-threatening disease arising predominantly from the aberrant proliferation of cells lining the breast's milk ducts. Despite being more prevalent in women, it can also affect men (Cutress, 2023). Prompt detection and treatment can significantly enhance breast cancer survival rates, necessitating improved early detection methodologies. Current detection approaches, such as diagnostic mammography, may be impeded by various challenges, including dense breast tissue in some women many images need to be taken and it is difficult for a large number of audiences (Rabiei et al., 2022). In recent years, machine learning has emerged as a robust tool in healthcare, offering significant potential for early disease prediction by deriving knowledge from data and uncovering hidden patterns. Previous studies have used different data types for breast cancer prediction, including mammographic characteristics, biopsy data, demographic risk variables, and genetic information.(Boeri et al., 2020; Rabiei et al., 2022).

In this project, the primary aim was to develop a robust and efficient machine-learning model to improve the speed and accuracy of breast cancer prediction. A range of Python libraries and tools were employed for data manipulation, visualization, and deep learning model development. Three diverse datasets were utilized: Mammographic Data for binary classification of benign and malignant masses, the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) for image classification, and histopathological data for another image classification task. The Mammographic Data underwent pre-processing to handle missing values, and four machine learning models (Random Forest, Decision Tree, K-Nearest Neighbors, and Naïve Bayes) were evaluated, with Random Forest emerging as the top performer with 82.25% accuracy. The CBIS-DDSM dataset was cleaned, and images were resized and normalized. The histopathological dataset underwent a similar treatment, and both image datasets were merged and expanded through data augmentation. Combining two different types of datasets helped my model to work for a large variety of breast cancer images because most of the breast cancer images were taken in both formats. For image classification, three convolutional neural networks (CNN) architectures (CNN, ResNet, DenseNet, and VGG) were examined, with CNN delivering superior performance with 89.14% accuracy. Both models were integrated into a user-friendly web application using Streamlit, allowing users to predict breast cancer conditions from mammographic data and images.

In summary, this project successfully developed a robust machine-learning system for breast cancer prediction, integrating diverse datasets and employing both traditional machine learning and deep learning models. The optimized Random Forest model for mammographic data and the CNN model for image data showcased exceptional predictive performance. The user-friendly web application further enhances accessibility, making this project a valuable contribution to the field of medical diagnostics and the quest for early breast cancer detection.

DECLARATION

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I confirm that I have not copied material from another source nor committed plagiarism nor commissioned all or part of the work (including unacceptable proof-reading) nor fabricated, falsified, or embellished data when completing the attached piece of work.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

Davis Sebi

“This dissertation contains material that is confidential and/or commercially sensitive. It is included here on the understanding that this will not be revealed to any person not involved in the assessment process”.

ACKNOWLEDGEMENTS

I commence by expressing my profound gratitude to the Divine for endowing me with the strength, wisdom, and unwavering determination to undertake and successfully complete this academic odyssey. I owe an immeasurable debt of gratitude to my parents, whose unflagging support, boundless love, and continuous encouragement have illuminated my path throughout this scholarly pursuit. Their unwavering faith in my potential and the sacrifices they made have served as the cornerstone of my perseverance. I extend my heartfelt appreciation to my esteemed supervisor, Dr. Fawada Qaiser, whose erudite guidance, mentorship, and profound expertise have been pivotal in shaping the trajectory of this project. Dr. Fawada's insightful feedback and unyielding belief in my capabilities have been instrumental in the realization of this work. To my dedicated friends and colleagues, who stood resolutely beside me during the challenging junctures of this scholarly voyage, providing words of motivation, encouragement, and lending a compassionate ear, I convey my heartfelt gratitude. Your camaraderie and unwavering support have been a wellspring of strength and resilience. Furthermore, I acknowledge with gratitude the University of Irvine, Machine Learning Repository, and Kaggle for their invaluable resources, which enriched the quality of this endeavor. I express my gratitude to Matjaz Zwitter and Milan Soklic for their creation of the Mammographic Mass Details dataset and to Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin for their pioneering work in developing the original CBIS-DDSM dataset, as well as to Abbas Cheddad for providing the simplified version of CBIS-DDSM utilized in this study. I also extend my appreciation to Paul Mooney for making the breast histopathological dataset available for this research. I am profoundly thankful to all the authors, researchers, and scholars whose seminal contributions formed the bedrock upon which this project was constructed. Your pioneering work in the field of breast cancer prediction has been a continual source of inspiration and guidance. In conclusion, this project's fruition would not have been conceivable without the steadfast support, encouragement, and unwavering belief of everyone mentioned in this acknowledgement and numerous others who may not have been explicitly named but have played a significant role, however modest, in my journey. I am deeply appreciative of your presence in my life and the indispensable roles you have fulfilled in this academic expedition.

**TABLE OF CONTENT**

Page

LIST OF TABLES vi

LIST OF FIGURES vii

Chapter 1. Introduction 1

Chapter 2. Aims and objectives 4

Chapter 3. Background and review of the literature 5

Chapter 4. Ethical Use of Data 9

Chapter 5. Design 10

Chapter 6. Implementation 21

Chapter 7. Evaluation 34

Chapter 8. Learning Points 35

Chapter 9. Professional Issues 36

Chapter 10. Conclusion And Future Work 37

REFERENCES CITED 39

Appendix 43

LIST OF TABLES

Page

Appendix Table 1: Project Log 45

LIST OF FIGURES

Page

Figure 1. Overall Project Design 10

Figure 2. Loaded Libraries for the Project 11

Figure 3. Loading mammographic mass dataset and its visualization. 12

Figure 4. Mammographic dataset after removing null values. 13

Figure 5. (left): Outliers in the mammographic dataset. Figure 6. (right): Mammographic dataset after removing outliers. 13

Figure 7. Values in Mammographic mass dataset after standard scale transformation 13

Figure 8. Saving of the model which has the best accuracy for the Mammographic mass dataset. 14

Figure 9. Importing The Mini CBIS DDSM and Histopathological dataset 15

Figure 10. Example of resized images as part of pre-processing 16

Figure 11. Example of Normalized images as part of pre-processing 16

Figure 12. Examples of Resized and Normalized Images as part of pre-processing. 17

Figure 13. Combining both Mini CBIS DDSM and Histopathological Dataset 18

Figure 14. Data Augmentation for Image Classification. 18

Figure 15. Implementation of callbacks in the models. 19

Figure 16. Web application created for Breast Cancer Prediction. 21

Figure 17. Working of Random Forest Model 22

Figure 18. Best hyperparameters and output of Random Forest Model 23

Figure 19. Working of Decision Tree Model 23

Figure 20. Best hyperparameters and output of Decision Tree 24

Figure 21. Working of KNN Model 24

Figure 22. Best hyperparameters and output of KNN Model 25

Figure 23. Working of Naïve Bayes Model 25

Figure 24. Output of Naïve Bayes Model 26

Figure 25. Comparison of Machine Learning models 26

Figure 26. Testing of the random Forest model. 27

Figure 27. Architecture of CNN Model 28

Figure 28. Confusion matrix of CNN Model 28

Figure 29. Architecture of ResNet Model 29

Figure 30. Confusion Matrix of ResNet Model 30

Figure 31. Architecture of DenseNet Model 31

Figure 32. Confusion matrix of DenseNet Model 31

Figure 33. Architecture of VGGNet Model 32

Figure 34. Confusion matrix of VGGNet 32

Figure 35. Comparison of deep learning models 33

Figure 36. Tested output of non-cancerous image in CNN model 33

Figure 37. Streamlit application testing with Mammographic data 34

Figure 38. Streamlit application testing with image data 34

Figure 39. Image showing how to access the web application 44

# Introduction

Breast cancer, a life-threatening disease affecting both women and men, primarily originates in the breast's milk duct cells, there are two types of stages in Breast cancer, Benign state in which the breast cancer has just started and the other one is a malignant case where the breast cancer has already spread to other tissues (UK, 2023). Globally, there are over 2.3 million new cases annually, with 685,000 women dying from it by 2020 but there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world’s most prevalent cancer. Breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life (Anon., 2023). Timely detection and treatment are vital for improved survival rates, prompting the need for better early detection methods. Current methods, like mammography, face challenges due to factors like dense breast tissue and the need for numerous images, making widespread adoption difficult(Rabiei et al., 2022). Machine learning in healthcare shows promise for early disease prediction by analyzing diverse data types such as mammography, biopsy, demographics, and genetics(Boeri et al., 2020; Rabiei et al., 2022).

The main steps required for building a machine learning model are Setting Up the Environment, Understanding the Dataset through Exploratory Data Analysis, Pre-processing the data, Splitting the Dataset for training and testing, Defining the Model, Model Evaluation, Fine-Tuning and Optimization, Making Predictions (ProjectPro, 2023)

The primary objective of this project was to develop a robust yet streamlined machine-learning model to enhance the speed and reliability of breast cancer prediction. To achieve this, three distinct datasets were employed: patient demographic and health data (Mammographic Data) (Elter, 2007), mammographic scans from the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) (C.D. Lekamlage, F. Afzal, E. Westerberg and A. Cheddad, 2020), and images derived from histopathological data, obtained through microscopic examination of tissue samples(Veeling, Bastiaan S and Linmans, Jasper and Winkens, Jim and Cohen, Taco and Welling, Max, 2018).

The Mammographic Data were employed for binary classification, distinguishing between benign and malignant masses, with a focus on evaluating the performance of a computer-aided detection system. This dataset comprises 516 instances of benign cases and 445 instances of malignant cases, and it involves the correlation of BI-RADS scores with sensitivities and specificities. It is worth noting that the BI-RADS assessment, despite being non-predictive, serves as a benchmark for comparison. The dataset encompasses predictive attributes such as patient age, mass shape, mass margin, and mass density, with the binary Severity attribute serving as the target variable (Elter, 2007). The pre-processing phase of the project revealed several missing values and errors in feature data, which were subsequently addressed by removing affected data points to ensure that the overall model accuracy remained unaffected otherwise predicted values can cause irregularity in the overall prediction. The dataset was then split into a 70:30 ratio for training and testing purposes so that there are enough values available for training and testing without causing overfitting or underfitting (Dewang, 2023). The images were standardized using a scalar.fit module from the scikit learn library and rescaled the images to 0,1 scale.

Four distinct machine learning models, namely Random Forest (Varsha Nemade1,\*,Vishal Fegade2, 2023), Decision Tree (Varsha Nemade1,\*,Vishal Fegade2, 2023), K-Nearest Neighbors (KNN) (Varsha Nemade1,\*,Vishal Fegade2, 2023), and Naïve Bayes (Varsha Nemade1,\*,Vishal Fegade2, 2023)were individually constructed and assessed. Each model was trained and tested separately, with performance evaluated based on key metrics including Accuracy, Precision, Recall, and F1 score (Varsha Nemade1,\*,Vishal Fegade2, 2023). Among these metrics, particular emphasis was placed on model accuracy as it represents the overall predictive capability. Through the process of hyperparameter tuning and exhaustive experimentation, the Random Forest model demonstrated high accuracy rate of 86.69% as compared to other models.

In this research, the CBIS DDSM dataset, a scaled-down version of the original dataset, was chosen due to its size and complexity, and the JPEG image format was preferred over PNG for its smaller file size and improved computational efficiency (C.D. Lekamlage, F. Afzal, E. Westerberg and A. Cheddad, 2020). Only breast images were selected for the image classification task, excluding images representing contour lines of cancer over a black background, which were deemed irrelevant. An accompanying Microsoft Excel dataset provided details about the images but after pre-processing I kept this dataset aside because the model, I am making is for image classification, but it can be used for later updates, and a thorough review and cleaning process was undertaken to ensure data alignment. Although some images were duplicated, they were retained so that it wouldn’t affect the sample count. The dataset encompassed three distinct image classes: Benign, Cancer, and Normal. Image resizing was performed to ensure uniformity and enhance compatibility with machine learning models by standardizing the height and width of the images to 200 pixels. Normalization was applied to scale the images to a range between 0 and 1, improving convergence and stability.

Additionally, the histopathological dataset (Veeling, Bastiaan S and Linmans, Jasper and Winkens, Jim and Cohen, Taco and Welling, Max, 2018) was considered, which featured two classes: Cancer and Non-Cancer images. This dataset required minimal pre-processing because it was already in its ideal form, and the resizing and normalization procedures were applied, mirroring the treatment of the CBIS DDSM dataset. The datasets were amalgamated into a unified dataset, which involved assigning labels as follows: 0 for Normal images in the CBIS DDSM dataset and Non-Cancer in the Histopathological dataset, 1 for Benign images in the CBIS DDSM dataset, and Cancer images in the Histopathological dataset, and 2 for Cancer images in the CBIS DDSM dataset, renamed as "Malignant”. A 70:30 split of the data into training and testing subsets was performed as like the mammographic image to reduce the overfitting and underfitting and make the values work perfectly for training and testing purposes, and data augmentation techniques were employed to expand the dataset by creating image variations through resizing, zooming, rotation, noise addition, and other transformations. Here I did a multi-dimension classification using three dimensions non-cancer(0), cancer(1) and malignant(2).

Three convolutional neural network (CNN) architectures, namely CNN (Ziquan Zhu, Shui-Hua Wang, Yu-Dong Zhang, 2023), ResNet (Ziquan Zhu, Shui-Hua Wang, Yu-Dong Zhang, 2023), DenseNet (Ziquan Zhu, Shui-Hua Wang, Yu-Dong Zhang, 2023)and VGG (Ziquan Zhu, Shui-Hua Wang, Yu-Dong Zhang, 2023), were selected for the study. An extensive hyperparameter tuning process was undertaken to identify optimal parameters for each model. Subsequently, the models were trained and evaluated for 25 epochs, and to expedite the training process, early stopping and learning rate reduction callbacks were introduced. The EarlyStopping callback is used to monitor the validation loss during training and stop the training process if the loss does not improve for a specified number of consecutive 5 epochs, with the option to restore the model weights to the best epoch. The ReduceLROnPlateau callback reduces the learning rate when the validation loss does not improve for a set of 5 epochs, allowing for finer adjustments during training to reach an optimal model. This approach facilitated model convergence and optimization. CNN exhibited superior performance among the models assessed, as determined by accuracy metrics and loss analysis the accuracy was 89.14%.

After identifying the best models, the models were saved, and both models were tested by importing values for the Random Forest model from mammographic data. Similarly, for the CNN model, the image data was imported for testing. The results were found to be correct when I tested the mammographic mass details it showed the correct classification of the state and similarly correct output was given for the image classification. To enhance user experience and facilitate ease of testing, a web application was developed using Streamlit. This application allowed users to load models and select the model they wished to test. For the Random Forest model, users could input various values, and upon running the classification, it would provide an output indicating whether the condition was benign or malignant. In the case of the CNN model, users could directly upload images from their local directory for classification, which determined whether the condition was benign, malignant, or non-cancer. The application accepted both JPG and PNG image formats.

Throughout this project, an array of Python libraries was employed to facilitate data manipulation, visualization, machine learning, and image processing. Notably, Pandas, NumPy, and Scikit-Learn played vital roles in data handling and pre-processing, while Matplotlib, Seaborn, and Plotly were instrumental in data visualization. Image-related tasks were addressed using OpenCV and the Python Imaging Library (PIL), and deep learning model development and evaluation were carried out using TensorFlow and Keras. These tools were complemented by confusion matrix analysis to assess model performance(A.L.Sayeth Saabith, T.Vinothraj, MMM.Fareez, 2020).

The scope of this project expanded to merge both the CBIS DDSM dataset and the Histopathological dataset. Furthermore, additional machine learning models, including Decision Tree, K-Nearest Neighbours (KNN), and Naïve Bayes, were incorporated for the mammographic data, while ResNet, DenseNet, and VGG models were introduced for the image dataset. A comparative analysis was conducted among these models to evaluate their respective performance. To enhance user-friendliness, a web application was developed using Streamlit. While further pre-processing options, such as imputation for missing values using methods like mean, median, mode, K-NN, or multivariate imputation by chained equations (MICE), were considered, they were ultimately deemed unnecessary for the project's objectives.

# Aims and objectives

* 1. **To apply a Machine Learning model to predict the risk of breast cancer in patients based on preliminary demographic and medical details in the Mammographic mass dataset.**
* To identify relevant patient data including demographic, lifestyle, laboratory, and mammographic data from the mammographic mass data set and preprocess it.
* To split, train, test, and validate Machine Learning models like Random Forest, Decision Tree, KNN and Naïve Bayes.
* Find the best machine learning model based on accuracy, precision, recall, f1 score and confusion matrix and save the model.
  1. **To further evaluate the breast cancer risk prediction using CBIS-DDSM and Histopathological images through the Deep Learning Model.**
* To pre-process The Complete Mini CBIS-DDSM image datasets and Histopathological dataset for the Deep Learning models like CNN, ResNet, DenseNet and VGGNet models
* To apply image pre-processing techniques, splitting, training, testing and validating deep learning models to detect breast cancer signs.
* To perform a comparative analysis find the best model between the deep learning models in terms of prediction accuracy, confusion matrix and accuracy and loss graph and save the model.
  1. **To create a user-friendly web app by taking the best models from both datasets.**
* To make sure the web app works properly without showing technical errors by accepting both png and jpg formatted images.
  1. **To evaluate the feasibility and scalability of these predictive models in a real-world healthcare system.**
* To examine the potential impact of these models in a clinical setting, considering factors such as early detection rates and treatment efficiency.
* To discuss the limitations, challenges, and potential improvements of the developed models.
* To ensure that all data manipulation, model training, and analysis processes adhere to relevant ethical guidelines and respect patient privacy.

In the project's initial phase, the intention was to utilize the original CBIS DDSM dataset; however, due to its considerable size and resource-intensive requirements, a decision was made to transition to the use of the complete Mini CBIS DDSM dataset. In the context of the mammographic mass dataset, there was a preliminary consideration of addressing missing values through techniques such as mean, median, mode imputation, K-Nearest Neighbors (K-NN), or multivariate imputation by chained equations (MICE). However, it was subsequently determined that these methods were superfluous for the project's objectives. This decision was based on the recognition that imputing missing values could potentially introduce inaccuracies into the dataset, thereby adversely affecting the overall model accuracy.

# Background and review of the literature

Breast cancer, a malignant condition, primarily arises from the glandular tissue(Lee et al., 2021) of the breast, originating in the lining cells which is known as the epithelium of the ducts that is 85% of cases or in other hand inside lobules is 15% of cases. The malignant development usually begins *in situ*, contained within the duct or lobule, often asymptomatic and presenting a low risk of metastasis. However, over time, the tumour can grow and spread, potentially leading to fatal outcomes (Wikipedia, 2023). According to the details provided by Cancer research UK modern medical science has categorised breast cancer into several types, including Ductal Carcinoma, Invasive Ductal Carcinoma, Inflammatory Breast Cancer, Lobular Carcinoma in situ, Invasive Lobular Carcinoma, and others with Invasive Ductal Carcinoma being the most prevalent (Anon., 2023). As a multifactorial disease, the condition results from a combination of genetic, environmental, and lifestyle factors. Recognised risk factors include age, family history, genetic mutations, reproductive history, and alcohol consumption etc. Symptoms often comprise a breast lump or thickening, changes in breast size, shape or appearance, alterations in skin texture, modifications in nipple appearance or surrounding skin known as areola, and abnormal nipple discharge (Clevelandclinic.org, 2022) .

By the year 1988 Breast Cancer Screening was already started and women aged between 47 – 73 years were randomly selected for the diagnosis, and this helped a lot to reduce the number of breast cancer mortality rate in UK(Al-Ajmi, n.d.). This is clear evidence that early prediction of breast Cancer can help the affected person to overcome from his condition. Diagnosis mammography was the technology used it assess abnormal breast cancer tissue in patients with subtle and inconspicuous malignancy signs. Approximately 50% of breast cancers were not detected in screenings of women with very dense breast tissue.(Rabiei et al., 2022) A mammogram is a simple picture of the breast, which doctors use to find signs of breast cancer. This is done by squishing the breast with a plate, then using low-energy X-rays to create a 2D picture. Normally, these pictures are taken from two angles: from the side and from above. If something suspicious, like a lump or abnormal area, is seen on the mammogram, more detailed images from different angles might be needed to check it out more closely.(Iranmakani et al., 2020). There are three ways of diagnosing breast cancer which are manual prediction checking the abnormalities, next by taking mammographic images by Xray and MRI Scanning and third way is through ultrasound scanning.(Iranmakani et al., 2020). There are two different studies which shows that 97% of diagnosis are done by using Xray images and Ultrasound scanning whereas 74% of diagnosis are only done by using mammography with manual examination.(Chae et al., 2013; Taori et al., 2013). So, from the above data it is evident that mammography is one of the important diagnosis technology used in modern medical world for breast cancer prediction.

Machine learning plays a pivotal role in enhancing the efficiency of systems as they repetitively engage in analogous tasks over time. This iterative process leads to continual improvement, marking a crucial facet of artificial intelligence. As a discipline, machine learning investigates the emulation of human learning activities through computational means. It delves into the mechanisms by which computers can assimilate new knowledge and skills, leveraging existing information. Implicit in the essence of machine learning is the notion that it constitutes a purpose-driven acquisition of knowledge. Internally, it manifests as a progression from the unknown to the known, contributing to the growth of knowledge within the system. Externally, machine learning is demonstrated through heightened system performance and adaptability, enabling the system to either accomplish previously unattainable tasks or enhance its task execution. When contextualized within the framework of data mining, machine learning can be categorized into four principal types: Classification, Regression Analysis, Association Rules, and Clustering. This classification facilitates a nuanced understanding of how machine-learning techniques can be strategically employed to extract valuable insights from data, contributing to the continual evolution and refinement of intelligent systems (Teng & Gong, 2018).

Deep learning, a subset of machine learning and artificial intelligence, is recognized as a foundational technology in the contemporary landscape of the Fourth Industrial Revolution. Positioned at the forefront of technological advancements, deep learning's roots in artificial neural networks have propelled it to the forefront of computing discourse. Its pervasive applications span diverse domains such as healthcare, visual recognition, text analytics, cybersecurity, and beyond, owing to its adeptness in learning from data. However, the development of an effective deep learning model remains a formidable challenge, primarily due to the dynamic nature and inherent variations within real-world problems and datasets. Characterized by the utilization of multi-layer neural networks for computation, deep learning embodies a learning paradigm that processes information through successive levels or stages. The term "Deep" in deep learning denotes the incorporation of multiple layers, illustrating the intricate stages through which data is processed to construct a data-driven model. The significance of deep learning intersects substantially with the realm of "Data Science," wherein the latter encompasses the comprehensive process of extracting meaning and insights from data within a specific problem domain. Deep learning methodologies play a pivotal role in data science, contributing to advanced analytics and fostering intelligent decision-making capabilities (Sarker, 2021).

The utilization of machine learning and data mining has revolutionized healthcare, particularly in the prediction of diseases such as breast cancer. These technologies draw insights from large datasets, unearthing hidden relationships and knowledge that significantly improve disease prediction and patient outcomes(Rabiei et al., 2022). Machine learning uses a range of information - such as age, ethnicity, mammogram images, biopsy results, and genetic data - to help predict disease outcomes(Fatima et al., 2020). It's becoming a big part of how we find and diagnose cancer. Traditionally, we've relied on doctor's check-ups, scans, and lab tests. But now, machine learning is stepping in, offering a way to predict what's not immediately obvious. This involves cleaning the data, picking out the key details that is called features, and using these features to classify, or distinguish between, harmless and harmful tumours. The ability to pick out these key details is especially important in telling apart benign that is not dangerous and malignant which is dangerous tumours.(Fatima et al., 2020). Data mining, which involves methods like grouping, trend analysis, and prediction, provides important information about breast cancer patients. Using this information and data from previous patients, we can predict different types of breast cancer more accurately. This increases the chances of a correct diagnosis and a successful treatment plan. There are many different computer algorithms we use for these predictions, including models inspired by the human brain that is Artificial Neural Network (ANN)(Syarif et al., n.d.), models that use a voting system called K-Nearest Neighbour (KNN)(Syarif et al., n.d.), models that use a decision-making tree like Decision Tree (DT)(Syarif et al., n.d.), Random Forest (RF)(Syarif et al., n.d.), and more. We also use methods that combine multiple models, such as Bagging(Syarif et al., n.d.), Boosting(Syarif et al., n.d.), and Stacking(Syarif et al., n.d.), to improve accuracy (Syarif et al., n.d.). Additionally, advanced techniques like Auto Encoders and CNN (Convolutional Neural Network), also contribute to our ability to predict cancer types and its state based on the image data.(Tran et al., 2021). In simple terms, these techniques help us use information from past patients to make more accurate predictions about future patients, which helps doctors diagnose and treat the disease more effectively. From mammographic images, abnormalities in breast tissues can be detected, enabling the assessment of breast cancer risk and identification of its stage as either benign or malignant. This approach provides a valuable preliminary insight into the specific type of breast cancer like Ductal Carcinoma, Invasive Ductal Carcinoma (Khamparia et al., 2021a; Yala et al., 2019).

As one of the most impactful deep learning algorithms, CNNs have significantly advanced medical image analysis through their multi-layered artificial neuron structure, extracting and passing complex features between layers for object recognition; this study specifically explores the convolutional features extracted from the deepest layer of a CNN model, capitalizing on the model's effective information representation that has been associated with highly accurate prediction results.(Ortega-Martorell et al., 2022). Various image classification models are utilized for breast cancer prediction, including traditional machine learning models such as Support Vector Machines (SVMs), Decision Trees, and Random Forests, as well as deep learning models like CNNs, and more advanced variants such as Residual Networks (ResNets)(He et al., n.d.), and DenseNets (Huang et al., n.d.). Each model has its strengths and unique applications, but CNNs have gained significant attention due to their remarkable performance on image-based tasks. The superior efficacy of CNNs can be attributed to their inherent architecture designed to learn spatial hierarchies of features directly automatically and adaptively from images, thereby capturing complex patterns more efficiently (Mishra, 2019). Furthermore, CNNs can handle larger datasets and image sizes, making them particularly well-suited for mammography or histopathology image analysis in breast cancer prediction. Their ability to extract high-level features from raw pixel data, without needing manual feature extraction, also contributes to their dominance in the field(Poojary & Pai, 2019). But when we come to other types of data set like neumerical dataset Random Forests often outperform other models due to their ability to handle high-dimensional, categorical, and missing data without requiring much pre-processing. They utilize an ensemble of decision trees, reducing the risk of overfitting and offering higher prediction accuracy by aggregating results through majority voting or averaging(Nemade & Fegade, 2023). Additionally, Random Forests inherently provide feature importance, aiding in interpretability and understanding of the underlying data relationships.(Nemade & Fegade, 2023).

The paper named Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques and Their Analysis(Fatima et al., 2020) gives the comparison of various models like Random Forest, Decision Tree, KNN, CNN, Naïve Bayes, Resnet, Densnet, VGG etc on different datasets like Mammograhic mass, CBIS DDSM, Histopathological, Wincosin etc with their accuracy. In the paper named Machine Learning Techniques for Breast Cancer Prediction (Nemade & Fegade, 2023) the authors were able to work with the models like Random Forest, Decision Tree, Naïve Bayes and KNN models with Breast Cancer data and able to achieve a result of 90% above accuracy. In the paper named Graph Temporal Ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis (Shi et al., 2020) The CNN model was able to achieve 89.5% accuracy. In the paper named A Deep Learning Approach for Breast Invasive Ductal Carcinoma Detection and Lymphoma Multi-Classification in Histological Images(Brancati et al., 2019) the Resnet model can showcase an accuracy of 87% using a histological image dataset. In the paper titled Analysis of Histopathological Images for Prediction of Breast Cancer Using Traditional Classifiers with Pre-Trained CNN (Gupta & Chawla, 2020) the VGG model has an accuracy of 78.6% and the Resnet Model has an accuracy of 86.4% accuracy. In the paper Breast Cancer Detection in Mammography Images: A CNN-Based Approach with Feature Selection(Zahra Jafari & Ebrahim Karami, 2023) they tried to make a prediction with CBIS DDSM dataset along with other MRI image breast dataset and in that Resnet Model had an accuracy of 84%. In the paper Diagnosis of Breast Cancer Based on Modern Mammography using Hybrid Transfer Learning (Khamparia et al., 2021b) the author was able to create a Resnet model with 85.1% accuracy and VGGnet with 83% accuracy and the dataset used here is CBIS DDSM dataset.

# Ethical Use of Data

4.1 Mammographic Mass Data

The Mammographic Mass Data was taken from the UC Irvine Machine Learning Repository and the data was licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license this allows for the sharing and adaptation of the datasets for any purpose, by providing the appropriate credit.

4.2 The Complete Mini CBIS DDSM Dataset

The Complete Mini Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) dataset on Kaggle is open and publicly available for research and educational purposes. It is a comprehensive, standardized dataset aimed to foster research in the field of mammographic image analysis. It is a refined and updated version of the original DDSM, one of the largest publicly available datasets for mammographic images. content on this site is licensed under a CC BY-ND 4.0 DEED.

4.3 Breast Histopathology Images

The Histopathological data is taken from Kaggle and it is open and publicly available for research and educational purposes. The content on this site is licensed under a CC0 1.0 Universal.

# Design

A diagram of a company

Description automatically generated with medium confidence Figure 1. Overall Project Design

5.1 IMPORTING THE NECESSARY LIBRARIES

In this Python script, various libraries are imported to facilitate data analysis, visualization, and machine-learning tasks. The `pandas` library is used for data manipulation and analysis, while `numpy` aids in numerical operations. Visualization is enhanced through `matplotlib`, `seaborn`, and `plotly.express` for creating plots and charts. The `os` library is employed for interacting with the operating system, and `cv2` and `PIL` are utilized for image processing tasks. Machine learning aspects are covered with `tensorflow` and `keras`, which are frameworks for building and training neural networks. Additionally, pre-processing functions from `sklearn` and utilities for image handling are included. The script also sets a random seed for reproducibility using the `random` library. Overall, these libraries collectively empower the script to handle diverse tasks ranging from data pre-processing to machine learning model evaluation.

A screen shot of a computer code

Description automatically generated

Figure 2. Loaded Libraries for the Project

5.2 PREDICTION OF BREAST CANCER USING MAMMOGRAPHIC MASS DATA

Breast cancer prediction is conducted using mammographic mass data, employing four distinct models: Random Forest, Decision Tree, KNN, and Naïve Bayes. The design process is explained step by step below.

5.2.1 Importing Mammographic Mass Dataset

The dataset is loaded into a Pandas Data Frame using the pd.read\_csv function. The read\_csv function parses the CSV file and creates a tabular representation of the data, storing it in the variable data.

A screenshot of a computer

Description automatically generated

Figure 3. Loading mammographic mass dataset and its visualization.

5.2.2 Data Pre-processing

The dataset under consideration manifests missing attribute values within the BI-RADS assessment, Age, Shape, Margin, and Density columns, while notably, the Severity attribute, serving as the target variable for prediction, exhibits completeness. Prior to the deployment of the dataset for modelling endeavours, essential editing procedures are imperative. The removal of rows containing missing values is executed to cultivate a dataset devoid of NaN values, thereby enhancing its integrity. This operation is effectuated through the utilization of the `dropna` method, serving as a robust mechanism for purging rows afflicted by undefined or absent attribute values. Furthermore, a critical facet of data preparation involves the conversion of designated columns to the float data type. This conversion holds significance, particularly for facilitating numerical operations and optimizing compatibility with machine learning algorithms predicated on numeric inputs.

In tandem with addressing missing values, the dataset undergoes a refinement process through the exclusion of rows wherein the 'Score' column assumes values equal to 55 or 0. This strategic omission of rows, characterized by specific 'Score' values, aligns with the overarching objective of data cleaning, ensuring the elimination of potential outliers or incongruent instances that might compromise analytical robustness.

Subsequently, an integral phase of data preprocessing unfolds with the standardization of features using the `StandardScaler` from the scikit-learn library. This transformative procedure imparts a standardized scale to the features, endowing them with a mean of 0 and a standard deviation of 1. Such normalisation is paramount for machine learning algorithms that reveal sensitivity to the scale of input features. Noteworthy is the utilisation of the 'Age' column as the target variable (`y`) during the scaling process, signifying its pivotal role in dictating the scaling parameters.

A screenshot of a black screen

Description automatically generated

Figure 4. Mammographic dataset after removing null values.

A diagram of a number of boxes

Description automatically generated with medium confidence A chart with lines and text

Description automatically generated with medium confidence

Figure 5. (left): Outliers in the mammographic dataset. Figure 6. (right): Mammographic dataset after removing outliers.

A screenshot of a computer screen

Description automatically generated

Figure 7. Values in Mammographic mass dataset after standard scale transformation

5.2.3 Splitting Data for Training and Splitting Purposes

I split the dataset in a ratio of 70:30 for the training and testing sets, respectively will help to perform the model well and also help to do the training and testing work perfectly without creating underfitting and overfitting issues.

5.2.4 Machine Learning Model Creation and Evaluation

Four distinct machine learning classifiers, namely Random Forest, Decision Tree, and K-Nearest Neighbours (KNN), along with the Gaussian Naive Bayes classifier, were systematically evaluated for their predictive efficacy on a given dataset. The Random Forest classifier underwent a meticulous hyperparameter tuning process through a GridSearchCV approach, optimizing parameters such as the number of estimators, maximum depth, and minimum samples for effective decision-making. The resulting model exhibited notable accuracy, as evidenced by the calculated accuracy score, classification report, and confusion matrix. Similarly, the Decision Tree classifier, equipped with hyperparameters such as criterion, maximum depth, and minimum samples, demonstrated commendable predictive performance following a rigorous tuning process. K-Nearest Neighbours, employing parameters like the number of neighbours, weights, and distance metric, showcased its effectiveness in proximity-based classification. Lastly, the Gaussian Naive Bayes classifier, a probabilistic model, demonstrated its inherent simplicity and efficiency in handling the dataset. Each model's performance was rigorously assessed through accuracy metrics, classification reports, and confusion matrices, providing a comprehensive understanding of their strengths and limitations in the context of the given dataset. More details about the model are provided in the implementation chapter.

5.2.5 Finding the Best Machine Learning Model and Saving the Model

A comparative analysis of four distinct machine learning models, namely Random Forest, Decision Tree, K-Nearest Neighbours (KNN), and Naive Bayes, was conducted to assess their predictive performance on the given dataset. The models were evaluated based on accuracy scores. To ensure the retention of this optimal model for future use, it was serialized and saved using the joblib library. This meticulous process of model selection and preservation is essential for reproducibility and deployment in real-world scenarios, contributing to the overarching objective of deploying robust and accurate machine-learning solutions in practical applications. More details regarding the best model evaluation are mentioned in the Implementation chapter.

A screen shot of a computer program

Description automatically generated

Figure 8. Saving of the model which has the best accuracy for the Mammographic mass dataset.

5.3 PREDICTION OF BREAST CANCER USING CBIS DDSM AND HISTOPATHOLOGICAL IMAGE DATASET

Breast cancer prediction is conducted using the complete mini CBIS DDSM and Histopathological dataset, employing four distinct models: CNN Model, ResNet Model, DenseNet Model, and VGGNet Model. The design process is explained step by step below.

5.3.1 Importing The Complete Mini CBIS DDSM Dataset and Histopathological Dataset

The Mini CBIS-DDSM dataset information, consisting of relevant metadata, was imported, and organized from an Excel file using the Pandas library in Python. The Mini CBIS-DDSM dataset was further stratified into distinct categories representing diverse medical conditions, namely Benign, Cancer, and Normal. Each category was meticulously curated by extracting relevant image file paths from the corresponding directories. For the Benign, Cancer, and Normal datasets, specific criteria were applied to filter out auxiliary files, such as those associated with masks, ensuring the focus on the primary image data.

Additionally, the Histopathological dataset was procured through the utilization of the glob library, which facilitated the retrieval of breast histopathological images in the PNG format. The dataset was sourced from a recursive search spanning various directories, capturing a diverse array of histopathological specimens. This collective assemblage of medical image datasets serves as a foundational resource for subsequent analytical endeavours, fostering a multidimensional exploration of both radiological and histopathological perspectives in the broader context of medical research and diagnosis.

A screen shot of a computer

Description automatically generated

Figure 9. Importing The Mini CBIS DDSM and Histopathological dataset

5.3.2 Pre-processing of Mini CBIS DDSM Dataset

In the preparatory stages of the comprehensive Mini CBIS DDSM dataset, several vital data pre-processing steps were meticulously executed to enhance the dataset's integrity and suitability for subsequent analytical endeavours. Initially, extraneous details pertaining to 'Tumour\_Contour' and 'Tumour\_Contour2' were expeditiously removed from the dataset, streamlining its structure for focused analysis. A judicious approach to handling missing data ensued, wherein the 'Age' column's NaN values were imputed with the mean age of the dataset, ensuring a coherent and complete representation of patient demographics. The metadata which I pre-processed kept aside because it is not useful for current image classification but later it can be used for detailed study. Subsequently, an imperative aspect of image processing unfolded through resizing the dataset's distinct categories (Benign, Cancer, and Normal) to uniform dimensions of 200x200 pixels. This standardization facilitates homogeneity in image dimensions, a prerequisite for many machine learning models. The consequential normalization of pixel values within the resized images further refined the dataset, ensuring uniformity in the pixel value range across all images. This normalization, achieved by scaling pixel values to the [0, 1] range, not only fosters computational efficiency but also establishes a standardized framework for subsequent model training. Collectively, these methodical data pre-processing stages represent a pivotal foundation, positioning the Mini CBIS DDSM dataset for optimal utilization in subsequent model development and medical image analysis endeavours within the realm of radiology and diagnostic research.

A collage of images

Description automatically generated

Figure 10. Example of resized images as part of pre-processing

A collage of images of a person's chest

Description automatically generated

Figure 11. Example of Normalized images as part of pre-processing

5.3.3 Pre-Processing of Histopathological Dataset

In the pre-processing phase of the Histopathological dataset, a meticulous stratification between cancerous and non-cancerous images was executed, delineating distinct categories for focused analysis. Employing a systematic classification based on file naming conventions, images designated as "class0.png" were identified as non-cancerous, while those labelled "class1.png" were classified as cancerous. This discerning categorization laid the foundation for targeted exploration and subsequent analytical endeavours within the realm of histopathological image analysis. To facilitate a comprehensive visual understanding, a custom display function was employed, enabling the depiction of the first five resized and normalized images from both cancerous and non-cancerous categories. Each displayed image underwent resizing to uniform dimensions of 200x200 pixels, contributing to standardized image dimensions conducive to machine learning model development. Moreover, the normalization process involved scaling pixel values to the [0, 1] range, promoting uniformity in pixel value ranges across all images. This methodical pre-processing of the Histopathological dataset establishes a robust foundation for subsequent model training and diagnostic research, providing a nuanced understanding of cellular structures within the context of cancer pathology.

A close-up of a cancer

Description automatically generated

Figure 12. Examples of Resized and Normalized Images as part of pre-processing.

5.3.4 Combing both Mini CBIS DDSM and Histopathological Dataset

In the integration of the Mini CBIS DDSM and Histopathological datasets, a meticulous amalgamation process was undertaken to create a cohesive and comprehensive dataset for subsequent model development. The normalized and resized images from the Mini CBIS DDSM dataset, representing benign, malignant, and normal instances, were seamlessly combined with the resized and normalized images from the Histopathological dataset, encompassing both cancerous and non-cancerous samples. A systematic approach was employed to assign categorical labels to each image, designating label '0' for non-cancerous instances, label '1' for cancerous instances, and label '2' for malignant cases from the Mini CBIS DDSM dataset. The amalgamated dataset was strategically shuffled to ensure randomness in the arrangement of instances, mitigating the potential for bias during subsequent model training. This concerted effort in dataset integration fosters a nuanced and diverse representation of medical image data, catering to a broad spectrum of pathologies. The resulting combined dataset stands poised for utilization in the development of machine learning models aimed at comprehensive diagnostic analyses, drawing insights from both radiological and histopathological perspectives within the realm of medical research and healthcare.

A computer screen shot of a program

Description automatically generated

Figure 13. Combining both Mini CBIS DDSM and Histopathological Dataset

5.3.5 Data Augmentation

In the context of machine learning and image classification tasks, data augmentation serves as a pivotal technique employed to enhance the diversity and robustness of the training dataset. The implemented data augmentation strategy, utilizing the ImageDataGenerator from the TensorFlow Keras library, introduces controlled variations to the original images during the training phase. This augmentation encompasses a suite of geometric transformations, including rotation, width and height shifts, shearing, zooming, and horizontal flipping. Each transformation is parameterized, allowing for a tailored adjustment of augmentation intensity. The rationale behind data augmentation is twofold: firstly, it amplifies the training dataset by generating augmented instances, mitigating the risk of overfitting and enhancing the model's ability to generalize to unseen data; secondly, it fosters resilience by exposing the model to a diverse array of augmented representations, thereby improving its adaptability to variations and distortions present in real-world data. The resultant data generators, created for both training and testing datasets, encapsulate the augmented images in batches, facilitating the seamless integration of these augmented instances into the training pipeline. This judicious application of data augmentation contributes to the overall robustness and generalization capability of the machine learning model, substantiating its efficacy in handling variations within the input data distribution.

A screen shot of a computer program

Description automatically generated

Figure 14. Data Augmentation for Image Classification.

5.3.6 Splitting the dataset for training and testing purposes

I split the dataset in a ratio of 70:30 for the training and testing sets, respectively will help to perform the model well and also help to do the training and testing work perfectly without creating underfitting and overfitting issues.

5.3.7 Introducing Early Stopping and ReduceLROnPlateau Call-backs

The implementation of the EarlyStopping callback in the context of neural network training represents a strategic mechanism to optimize model performance while mitigating the risk of overfitting. Configured to monitor the validation loss during training, the EarlyStopping callback intervenes when the validation loss ceases to improve over a specified number of epochs, known as the "patience" parameter. This adaptive halting of training aims to prevent the model from continuing to learn from noise or minor fluctuations in the data, ultimately safeguarding against overfitting and enhancing the model's generalization capability to unseen data. The parameter "min\_delta" defines the minimum change in the monitored quantity required to be considered an improvement, ensuring that only substantial improvements trigger the early stopping mechanism. The "restore\_best\_weights" parameter further refines this strategy by reverting the model to the epoch with the optimal validation loss, preserving the configuration that yielded superior performance. In essence, the EarlyStopping callback serves as a prudent mechanism for model regularization, striking a balance between achieving high performance on the training set and maintaining generalization efficacy on unseen data.

A screenshot of a computer program

Description automatically generated

Figure 15. Implementation of callbacks in the models.

5.3.8 Deep Learning Model creation and evaluation

In this study, several convolutional neural network (CNN) architectures were implemented and evaluated for the classification of medical images, specifically in the context of breast cancer diagnosis. The first model, a traditional CNN, comprised multiple convolutional layers with batch normalization, max-pooling, and global average pooling, followed by fully connected layers. This model demonstrated competitive performance with consideration for model accuracy and loss. Subsequently, a Residual Network (ResNet) architecture was employed, leveraging residual blocks to facilitate the training of deeper networks. The ResNet model exhibited notable performance improvements, showcasing the efficacy of residual connections in mitigating vanishing gradient issues during training. Following this, a DenseNet model, characterized by densely connected blocks, was implemented, resulting in a highly parameter-efficient architecture. The DenseNet model showcased compelling performance metrics, reinforcing the effectiveness of dense connectivity in feature extraction. Additionally, VGGNet, a model renowned for its simplicity, was employed, featuring multiple convolutional blocks with max-pooling and dense fully connected layers. While VGGNet demonstrated competitive accuracy, it exhibited a comparatively higher model complexity. Overall, this comprehensive investigation into various CNN architectures provides valuable insights into their respective strengths and weaknesses in the domain of medical image classification, informing considerations for future model selection in similar diagnostic tasks. More details about the model are provided in the implementation chapter.

5.3.9 Finding the best Deep Learning model and Saving the Model

This study encompasses a comparative analysis of four distinct deep learning convolutional neural network (CNN) architectures, specifically CNN, ResNet, DenseNet, and VGGNet. The primary focus of this investigation is the assessment of their predictive performance on the provided dataset, with the evaluation criterion centred on accuracy scores. A meticulous approach was adopted to ensure the retention of the optimal model for prospective utilization. The serialization of the chosen model is imperative for the sake of reproducibility and its seamless deployment in real-world scenarios. This methodical process of model selection and preservation holds significance in the context of contributing to the overarching objective of deploying machine-learning solutions that are both robust and accurate in practical applications. Further elaboration on the details pertinent to the evaluation of the best model can be found in the Implementation chapter.

5.4 WEB APPLICATION

This web application serves as a medical breast cancer classifier, integrating two distinct models: Convolutional Neural Network (CNN) and Random Forest which were the best models after the evaluation. Users can opt for either the CNN model for image-based predictions or the Random Forest model for predictions based on histopathological data. For image classification using the CNN model, users can upload a CBIS-DDSM image, and upon clicking the "Classify" button, the model provides predictions with class labels such as 'Non-Cancer,' 'Cancer,' or 'Malignant.' On the other hand, the Random Forest model enables predictions based on histopathological details, allowing users to input variables like BI-RADS assessment, patient's age, mass shape, margin, and density. The "Predict" button triggers the model, yielding predictions with class labels such as 'Negative' or 'Positive.' This interactive platform enhances accessibility to predictive models, catering to different user preferences and types of input data, thus contributing to the broader integration of machine learning in medical diagnostics.

The application employs a streamlined user interface with a main section for model selection and image/data upload and a sidebar for choosing between the CNN and Random Forest models. The CNN model is designed to handle CBIS-DDSM images through a user-friendly image upload mechanism. Simultaneously, the Random Forest model accommodates tabular input, allowing users to input relevant histopathological details for predictions. Such an integrated approach offers flexibility and adaptability, allowing medical practitioners to leverage diverse data types for more comprehensive breast cancer predictions. More details regarding the web application are provided in the implementation chapter.

A screenshot of a computer

Description automatically generated

Figure 16. Web application created for Breast Cancer Prediction.

# Implementation

6.1 MACHINE LEARNING MODEL CREATION FOR MAMMOGRAPHIC MASS DATA AND ITS IMPLEMENTATION

In the initial project phase, a classification model was developed to predict breast cancer stage (benign or malignant) based on mammographic data. Extensive dataset pre-processing was followed by training four models: Random Forest, Decision Tree, K-Nearest Neighbours (KNN), and Naïve Bayes. To streamline hyperparameter tuning, a grid search method was employed to find optimal parameters for the models. Notably, Naïve Bayes, as a probabilistic classification algorithm, wasn't subject to traditional hyperparameter tuning.

6.1.1 Random Forest Model

In the process of preparing the Random Forest (RF) model, a comprehensive hyperparameter tuning approach, namely GridSearchCV, was employed to systematically explore a range of hyperparameter combinations and identify the optimal configuration. The hyperparameter grid encompassed variations in the number of estimators (`n\_estimators`), the maximum depth of the trees (`max\_depth`), the minimum number of samples required to split an internal node (`min\_samples\_split`), and the minimum number of samples required to be at a leaf node (`min\_samples\_leaf`). The RandomForestClassifier from scikit-learn was instantiated with an initial configuration, including setting the random seed for reproducibility (`random\_state=20`), using the Gini impurity criterion (`criterion='gini'`), and leveraging parallel computing for efficient processing (`n\_jobs=-1`).

Subsequently, the GridSearchCV was executed with 5-fold cross-validation (`cv=5`), leveraging parallel processing (`n\_jobs=-1`) for efficiency, and utilizing accuracy as the scoring metric (`scoring='accuracy'`). This rigorous exploration resulted in the identification of the best-performing hyperparameters, which were then extracted for reference. The best Random Forest Classifier was instantiated with these optimal hyperparameters, and the model was fitted to the training data. Following this training phase, predictions were generated on the test data, and the model's performance was rigorously assessed. The accuracy score, classification report, and confusion matrix were computed, providing a comprehensive evaluation of the Random Forest model's predictive capabilities. This meticulous model preparation process adheres to best practices in machine learning, ensuring that the chosen model configuration is robust and performs optimally on the given dataset.

A computer screen shot of a program code

Description automatically generated

Figure 17. Working of Random Forest Model

A screenshot of a computer

Description automatically generated

Figure 18. Best hyperparameters and output of Random Forest Model

6.1.2 Decision Tree Model

In the process of model preparation using a Decision Tree Classifier, a rigorous hyperparameter tuning approach was adopted to optimize the model's performance. The hyperparameter space was systematically explored using the GridSearchCV technique, encompassing various configurations for essential parameters such as 'criterion,' 'max\_depth,' 'min\_samples\_split,' 'min\_samples\_leaf,' and 'max\_features.' This meticulous exploration involved a five-fold cross-validation strategy to robustly assess each configuration's generalization ability. The decision tree classifier was instantiated with a random state for reproducibility.

The best-performing model, denoted as `best\_classifier\_DT`, was identified based on the highest accuracy achieved during the grid search. The selected hyperparameters, termed `best\_params\_DT`, represent the optimal configuration for the decision tree model. Subsequently, the model was fitted to the training data using these refined hyperparameters. Predictions were then generated on the test dataset, and the accuracy score was computed to quantify the model's effectiveness. Furthermore, a comprehensive evaluation of the model's predictive capabilities was conducted through the generation of a classification report, which includes metrics such as precision, recall, and F1-score for each class. Additionally, a confusion matrix was constructed to visualize the model's performance in terms of true positive, true negative, false positive, and false negative classifications. This rigorous model preparation process adheres to best practices in machine learning, ensuring that the Decision Tree Classifier is finely tuned for optimal predictive performance on the given dataset.

A screen shot of a computer program

Description automatically generated

Figure 19. Working of Decision Tree Model

A screenshot of a computer

Description automatically generated

Figure 20. Best hyperparameters and output of Decision Tree

6.1.3 KNN Model

In the process of preparing the K-Nearest Neighbours (KNN) classifier, hyperparameter tuning is employed to enhance its predictive performance. The hyperparameters under consideration include the number of neighbours (`n\_neighbors`), the weighting scheme (`weights`), and the distance metric (`metric`). A grid search is conducted over a predefined parameter grid, exploring different combinations of these hyperparameters. The candidate models are evaluated through 5-fold cross-validation, and their performance is assessed using the accuracy metric. The best-performing model, determined by the highest accuracy, is selected as the final KNN classifier.

Subsequently, the optimal hyperparameters for the KNN model are identified based on the outcomes of the grid search. The best estimator, encapsulating these hyperparameters, is instantiated as `best\_classifier\_KNN`. This classifier is then fitted to the training data (`X\_train` and `Y\_train`). Following the training phase, predictions are generated on the test data (`X\_test`). The accuracy of the KNN classifier is quantified using the accuracy score, reflecting the proportion of correctly classified instances. Additionally, a comprehensive evaluation of the model's performance is provided through a classification report, detailing metrics such as precision, recall, and F1-score for each class. The confusion matrix further illuminates the distribution of predicted and actual class labels, offering insights into the model's ability to discern between positive and negative instances. Overall, this systematic approach ensures the robustness and efficacy of the KNN classifier for the given dataset.

A screen shot of a computer program

Description automatically generated

Figure 21. Working of KNN Model

A screenshot of a computer

Description automatically generated

Figure 22. Best hyperparameters and output of KNN Model

6.1.4 Naïve Bayes Model

In this study, a Gaussian Naive Bayes (NB) classifier was employed for the predictive modelling task. The scikit-learn library was utilized to implement the classifier, specifically the ` Gaussian Naive Bayes ` class. Prior to model fitting, the dataset was divided into training and test sets (`X\_train`, `Y\_train`, `X\_test`, and `Y\_test`), adhering to the standard practice of supervised machine learning. Subsequently, the Naive Bayes classifier was fitted to the training data using the `fit` method, enabling the model to learn the underlying patterns and relationships within the features and labels of the training set.

Following the model training, predictions were generated on the test data using the `predict` method, allowing for the assessment of the model's performance on previously unseen instances. The accuracy of the Gaussian Naive Bayes classifier was evaluated using the `accuracy\_score` metric, providing a quantitative measure of the model's correctness in predicting the target variable. Additionally, a comprehensive analysis of the model's performance was conducted through the generation of a classification report, which included metrics such as precision, recall, and F1-score for each class. Furthermore, a confusion matrix was constructed to visualize the count of true positive, true negative, false positive, and false negative predictions, offering valuable insights into the classifier's ability to correctly classify instances. Overall, this rigorous model preparation and evaluation process adheres to established methodologies in the field of machine learning and contributes to the interpretability and reliability of the Gaussian Naive Bayes classifier in the given context.

A screen shot of a computer code

Description automatically generated

Figure 23. Working of Naïve Bayes Model

A screenshot of a computer

Description automatically generated

Figure 24. Output of Naïve Bayes Model

6.2 FINDING THE BEST MODEL AMONG THE MACHINE LEARNING MODELS

A comparative analysis was conducted among the four distinct Machine learning models by means of a bar graph. The results of this comparative assessment distinctly indicated that the Random Forest model exhibited the highest accuracy, achieving a notable accuracy rate of 86.69%.

A graph showing different colored squares

Description automatically generated

Figure 25. Comparison of Machine Learning models

6.3 TESTING THE BEST MACHINE LEARNING MODEL

The Random Forest, identified as the optimal model, underwent initial testing using the reserved test dataset. During this evaluation, the model demonstrated efficacy by accurately predicting the target variable. The primary testing involved passing the test data through the trained Random Forest classifier, and the model successfully provided correct predictions. This preliminary assessment serves as a crucial step in validating the model's generalization performance on independent data, substantiating its suitability for real-world applications. The outcomes of this testing phase contribute valuable insights into the Random Forest's predictive capabilities and reinforce its standing as the selected model for subsequent stages of evaluation and deployment.

A computer screen shot of a program code

Description automatically generated

Figure 26. Testing of the random Forest model.

6.4 DEEP LEARNING MODEL CREATION FOR IMAGE DATA CLASSIFICATION

After the comprehensive pre-processing steps, the unified dataset, amalgamating both the complete mini CBIS DDSM and Histopathological datasets, was systematically partitioned for training and testing purposes. Given the nature of the task involving image classification, Convolutional Neural Network (CNN) architectures were employed, encompassing models such as CNN, ResNet, DenseNet, and VGGNet. Each of these models underwent rigorous evaluation to discern their predictive efficacy. This methodical approach ensures a robust assessment of the combined dataset's characteristics and facilitates the selection of an optimal deep-learning model for subsequent stages of analysis and deployment, substantiating the scientific rigour and applicability of the chosen frameworks.

6.4.1 CNN Model

The convolutional neural network (CNN) model is constructed for the purpose of multiclass classification, specifically targeting a dataset with three classes (0, 1, and 2). The model architecture is meticulously designed with considerations for feature extraction and hierarchical representation learning. The random seed is set to 42 to ensure the reproducibility of results. The input layer, denoted as 'input\_layer,' is defined with a shape of (50, 50, 3), reflecting the dimensions of the input images, where 3 corresponds to the colour channels (RGB). Subsequently, a series of convolutional layers, activation functions, batch normalization, and max-pooling operations are applied to capture relevant features in a spatial hierarchy. The convolutional layers, named 'conv1' through 'conv4,' each contribute to the extraction of increasingly complex features. Batch normalization is incorporated to enhance training stability and convergence, and max pooling is employed to downsample spatial dimensions and reduce computational complexity.

Following the convolutional layers, a global average pooling layer, denoted as 'global\_avg\_pool,' is applied to reduce the spatial dimensions to a single vector, facilitating global feature summarization. This global average pooling step contributes to spatial invariance and ensures that the model remains robust to variations in object position within the input images. Subsequently, fully connected layers are introduced, starting with 'dense1,' which consists of 128 neurons and employs the rectified linear unit (ReLU) activation function to introduce non-linearity. To prevent overfitting, a dropout layer with a rate of 0.3 is incorporated, strategically discarding a portion of neurons during training.

The output layer, designated 'output,' is tailored for multiclass classification and employs the SoftMax activation function. The model is compiled with the Adam optimizer, utilizing a learning rate of 0.001, and categorical cross-entropy is chosen as the loss function to optimize the network for multiclass classification scenarios. The chosen evaluation metric is accuracy. This meticulously configured CNN model, named 'CNN\_Model,' is poised for training on the designated dataset, demonstrating a thoughtful balance between feature extraction, spatial summarization, and regularization techniques for effective and robust multiclass classification. This model preparation, characterized by its architectural choices and hyperparameter configurations, serves as a foundational framework for subsequent training on specific datasets. It reflects a thoughtful consideration of the nuances in the given dataset and aims to strike a balance between model expressiveness and generalization ability.

A screenshot of a computer

Description automatically generated

Figure 27. Architecture of CNN Model

A graph with a number of blue and purple squares

Description automatically generated with medium confidence

Figure 28. Confusion matrix of CNN Model

6.4.2 ResNet Model

The ResNet (Residual Network) model prepared in this work follows a meticulously designed architecture tailored for image classification tasks. The implementation adheres to best practices in deep learning and employs the TensorFlow and Keras frameworks. To ensure reproducibility, a random seed (seed value: 42) is set at the beginning, contributing to consistent model initialization during training and evaluation. The model architecture begins with an input layer of shape (50, 50, 3), accommodating images of size 50x50 pixels with three colour channels. The initial Convolutional layer, employing a kernel size of (7, 7) and a stride of (2, 2), extracts hierarchical features from the input images. Following this, a Batch Normalization layer stabilizes the activations, and a Rectified Linear Unit (ReLU) activation function introduces non-linearity.

Subsequently, a MaxPooling layer with a kernel size of (3, 3) and a stride of (2, 2) is applied, reducing spatial dimensions, and enhancing computational efficiency. The core of the ResNet model consists of residual blocks, specifically designed to mitigate vanishing gradient issues associated with deep networks. In this implementation, three residual blocks are employed, each comprising two Convolutional layers with Batch Normalization and ReLU activation. The residual connection, facilitated by the Adding layer, bypasses the convolutional layers, fostering the flow of gradient information and facilitating smoother convergence during training.

The model concludes with a Global Average Pooling layer, which spatially averages the tensor dimensions, effectively compressing the spatial information into a single vector. Subsequent fully connected layers, including a Dense layer with 128 units and a ReLU activation function, introduce high-level abstractions. To prevent overfitting, a Dropout layer with a dropout rate of 0.3 is incorporated. The output layer, customized to the specific dataset, comprises three units with a SoftMax activation function, aligning with the multi-class classification objective. Finally, the model is compiled using the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss function, and accuracy as the evaluation metric. The resulting ResNet model is poised for effective training and subsequent deployment in image classification scenarios.

A screenshot of a computer program

Description automatically generated

Figure 29. Architecture of ResNet Model

A graph of a graph

Description automatically generated with medium confidence

Figure 30. Confusion Matrix of ResNet Model

6.4.3 DenseNet Model

The DenseNet model architecture is meticulously prepared to address classification tasks, specifically tailored for medical image analysis in this context. The model begins with an input layer of shape (50, 50, 3), indicating an image input with a resolution of 50x50 pixels and three color channels. The initial convolution layer, a pivotal component, employs a kernel size of (7, 7) with padding set to 'same' and utilizes the rectified linear unit (ReLU) activation function to introduce non-linearity. Following this, a batch normalization layer is applied, enhancing convergence and mitigating overfitting.

The model features two dense blocks, each characterized by a sequence of Batch Normalization, ReLU activation, 1x1 convolution, and 3x3 convolution layers. These dense blocks foster feature reuse and facilitate the extraction of intricate hierarchical features from the input image. Between the dense blocks, transition layers, comprising 1x1 convolution and average pooling, are strategically placed to reduce spatial dimensions and control model complexity. The final stage of the model incorporates a global average pooling layer, effectively summarizing the spatial information, followed by flattening to prepare for fully connected layers.

In the fully connected layers, a dense layer with 128 units and ReLU activation is introduced, promoting feature aggregation and abstraction. To mitigate overfitting, a dropout layer with a dropout rate of 0.3 is incorporated. The output layer, critical for binary classification tasks, comprises three units with a softmax activation function, aligning with the three classes present in the dataset. The model is compiled using the Adam optimizer with a learning rate of 0.001, binary cross-entropy serving as the loss function, and accuracy as the evaluation metric. This strategic architectural design, coupled with meticulous parameter tuning, positions the DenseNet model as a robust and effective tool for the classification of medical images in breast cancer diagnostics.

A screenshot of a computer program

Description automatically generated

Figure 31. Architecture of DenseNet Model

A graph of a graph with a number of blue and purple squares

Description automatically generated with medium confidence

Figure 32. Confusion matrix of DenseNet Model

6.4.4 VGGNet Model

The VGGNet model is a convolutional neural network (CNN) architecture designed for image classification tasks. In the presented code, the model is prepared using the TensorFlow and Keras framework. The model begins with an input layer defined by the shape of the input data, which in this case is a 3-channel image of size 50x50 pixels. The subsequent architecture follows the VGGNet convention, organized into convolutional blocks. The first convolutional block consists of two consecutive convolutional layers, each with 64 filters and a 3x3 kernel, followed by max-pooling with a 2x2 pool size and strides of 2. This process is repeated for additional convolutional blocks, gradually increasing the number of filters to capture more complex features. The last convolutional block is followed by max-pooling and serves as a feature extractor.

The flatten layer reshapes the output of the convolutional blocks into a one-dimensional vector, enabling the connection to fully connected layers. Two dense (fully connected) layers with 4096 units each and ReLU activation functions follow, incorporating non-linearities to capture intricate patterns in the data. Dropout layers are introduced with a dropout rate of 0.5 to mitigate overfitting. Finally, the output layer comprises three units (matching the number of classes in the binary classification task) with a softmax activation function, facilitating probability distribution interpretation. The model is compiled using the Adam optimizer, binary cross-entropy loss, and accuracy as the evaluation metric. This rigorous model preparation adheres to standard practices in deep learning for image classification, ensuring effective learning and generalization.

A screenshot of a computer

Description automatically generated

Figure 33. Architecture of VGGNet Model

A graph of a graph

Description automatically generated with medium confidence

Figure 34. Confusion matrix of VGGNet

6.5 FINDING THE BEST MODEL AMONG THE DEEP LEARNING MODELS

A comparative analysis was conducted among all the developed deep learning models like CNN, ResNet, DenseNet and VGGNet, and the results were visually represented using a bar graph. The findings of this comparative assessment indicated that the CNN model outperformed the others, boasting a notably **high accuracy rate of 89.14%.**

A graph showing different colored rectangular shapes

Description automatically generated

Figure 35. Comparison of deep learning models

6.6 TESTING THE BEST DEEP LEARNING MODEL

The CNN model, identified as the optimal model through prior evaluation, underwent initial testing using the designated test dataset. During this evaluation, the model demonstrated its efficacy by providing accurate predictions on the test data. The performance assessment involved comparing the model's predicted outcomes with the true labels, affirming the model's ability to correctly classify instances within the given dataset. This preliminary testing phase substantiates the efficacy of the CNN model and underscores its capability to yield accurate predictions in a controlled evaluation environment. This testing procedure aligns with established practices in machine learning evaluation and contributes to the empirical validation of the CNN model's predictive performance on the provided test dataset.

A close-up of a radiography

Description automatically generated

Figure 36. Tested output of non-cancerous image in CNN model

6.7 WORKING OF THE WEB APPLICATION

A user-friendly web application was developed utilizing Streamlit to facilitate the effortless classification of both image data for image classification and detailed information for the binary classification of mammographic data

6.7.1 Web Application Working For Brest Cancer Prediction with Mammographic Mass Details

The application accommodates predictions based on histopathological data using the Random Forest model. Users input key features such as BI-RADS assessment score, patient age, mass shape, margin, and density. These input features are processed, and the Random Forest model predicts the likelihood of a positive or negative diagnosis. The results are then communicated to the user, offering insights into the potential presence of breast cancer based on the provided clinical details.

A screenshot of a computer

Description automatically generated

Figure 37. Streamlit application testing with Mammographic data

6.7.2 Web Application Working For Brest Cancer Prediction with Image Dataset

For the CNN model, users can upload medical images, specifically CBIS-DDSM images, through the application's interface. The uploaded image undergoes pre-processing, including resizing, normalization, and conversion to RGB format. The pre-processed image is then fed into the trained CNN model, which predicts the likelihood of the presence of non-cancerous, cancerous, or malignant conditions. The CNN's predictions are presented to the user, contributing to an interactive and informative diagnostic experience.

A screenshot of a medical image

Description automatically generated

Figure 38. Streamlit application testing with image data

# Evaluation

7.1 Potential Impact of the Model

Timely detection is crucial in medical diagnostics, especially for breast cancer, allowing for early intervention and better outcomes. Convolutional Neural Networks (CNNs) play a vital role in analysing mammographic and histopathological images to detect subtle malignancy indicators. They provide precise classification of breast cancer cases (benign, malignant, or non-cancerous), reducing the need for invasive procedures and optimizing resource allocation. These models offer consistent and objective assessments, mitigating inter-observer variability and improving diagnostic reliability. Additionally, CNNs can identify features that may be challenging for human clinicians to discern among a large volume of images. They are effective even in cases with a significant proportion of fat tissue in a patient's breast, improving diagnostic accuracy. These models can also be used for self-administered preliminary tests through mammographic images, empowering individuals to take proactive steps in early detection, which is essential for timely treatment and improved patient outcomes.

7.2 Limitations and Challenges

The effectiveness of these models heavily relies on the quality and diversity of training data. Biased or unrepresentative datasets can introduce limitations and potential biases in real-world use. Addressing false positives and false negatives is another significant challenge, as false positives can lead to unnecessary interventions, and false negatives may result in missed diagnoses. Continual refinement and evaluation are crucial to minimise these errors. Ethical and regulatory considerations, including patient data privacy, informed consent, and compliance with healthcare regulations, need attention when deploying these models in clinical settings. Compliance with regulatory standards, such as FDA approval for medical devices, is essential for responsible clinical use.

# 

Overall, the machine learning models, including Random Forest, Decision Tree, KNN, and Naïve Bayes, were applied to predict breast cancer risk based on demographic and medical details from the Mammographic mass dataset. The models were thoroughly trained, tested, and validated, with Random Forest emerging as the best-performing model, achieving an accuracy of 86.69%. The deep learning models, CNN, ResNet, DenseNet, and VGGNet, were evaluated using CBIS-DDSM and Histopathological images. The CNN model exhibited outstanding performance with an accuracy of 89.14%, surpassing the other deep-learning models. A user-friendly web app was developed, seamlessly integrating the top-performing models from both datasets, allowing users to make predictions conveniently. The feasibility and scalability of these predictive models in a real-world healthcare system were explored, considering factors such as early detection rates and treatment efficiency. The study adhered to ethical guidelines and privacy standards, and potential limitations, challenges, and avenues for improvement were discussed.

# Learning Points

The developmental trajectory of the "Breast Cancer Prediction Using Random Forest and CNN Model" project has yielded profound learning outcomes across diverse domains, encompassing technical proficiencies, project management acumen, and insightful revelations in Breast Cancer Prediction. As the project culminates, several salient learning points have surfaced, markedly influencing the project's triumph, and charting a course for future initiatives. Foremost among these is a heightened technical acuity, involving an in-depth exploration of machine learning models (Random Forest, Decision Tree, KNN, and Naive Bayes), deep learning architectures (CNN, ResNet, DenseNet, and VGGNet), and web application development using Streamlit, coupled with rigorous data pre-processing. This immersive process of comprehension, implementation, and refinement has substantially augmented machine learning expertise, while the incorporation of Streamlit has broadened practical programming competencies.

A pivotal learning juncture underscores the paramount significance of meticulous data collection and pre-processing procedures. The acquisition of a multifaceted dataset and the judicious application of techniques such as data splitting and text cleaning emerged as indispensable facets, ensuring the efficacy of the predictive models. This experience emphatically underscores the pivotal role of high-quality data as the foundational bedrock of any machine learning undertaking. Additionally, the discernment of optimal models through the evaluation of multiple methodologies has unveiled valuable insights into the techniques and parameters conducive to model excellence.

A central tenet of this transformative journey is the recognition of the cardinal role played by user-centric design principles. Prioritizing user-friendliness, intuitive navigation, and lucid presentation of results has distinctly contributed to the project's overall usability and garnered user acceptance. These assimilated learning points stand poised to guide subsequent endeavours, promising the formulation of more resilient and efficacious solutions in the realm of Breast Cancer prediction.

# Professional Issues

The "Breast Cancer Prediction Using Random Forest and CNN Model" project aligns comprehensively with the British Computer Society (BCS) Code of Conduct, reflecting a commitment to ethical, professional, and socially responsible practices in the realm of computing. The BCS Code of Conduct emphasizes the paramount importance of promoting the public good, and the project's overarching aim to enhance breast cancer prediction directly contributes to the societal welfare. By leveraging machine learning and deep learning models, the project adheres to the BCS principle of striving for high standards of quality and professionalism in computing applications.

Moreover, the project underscores the BCS Code of Conduct's emphasis on competence and integrity in professional conduct. The meticulous attention given to technical proficiency in implementing diverse machine learning models and deep learning architectures aligns with the BCS's call for maintaining and updating professional competence. Furthermore, the rigorous adherence to ethical guidelines in data collection, model training, and analysis demonstrates the project's commitment to integrity, transparency, and accountability, reflecting the ethical considerations stipulated by the BCS Code of Conduct.

In relation to the BCS Code's focus on user-centric design and the societal impact of computing, the project prioritizes user-friendliness in its web application development, ensuring accessibility and clear presentation of results. By addressing the critical domain of breast cancer prediction, the project resonates with the BCS's commitment to utilizing computing expertise for societal benefit. Overall, the project embodies the values and principles articulated in the BCS Code of Conduct, emphasizing ethical, competent, and socially responsible practices within the field of computing.

# Conclusion And Future Work

10.1 CONCLUSION

In conclusion, the comprehensive evaluation of machine learning and deep learning models for breast cancer prediction revealed notable insights into their performance and characteristics. The Random Forest model, meticulously optimized through GridSearchCV, emerged as a robust performer, attaining a high accuracy of 86.69% and demonstrating a commendable balance between precision and recall. In the medical context, it showcased minimal false positives and false negatives, positioning it as a reliable choice for medical data classification. Additionally, the Decision Tree model exhibited solid accuracy and well-balanced precision and recall, further emphasizing its competence in breast cancer prediction. The K-Nearest Neighbors and Gaussian Naive Bayes models, though competitive, displayed distinct characteristics, with the former achieving reasonable accuracy and the latter demonstrating balanced metrics.

Moving to the realm of deep learning, the Convolutional Neural Network (CNN) model designed for multiclass medical image classification outshone others, achieving an impressive validation accuracy of 89.14%. Despite occasional challenges in distinguishing between benign and non-cancer cases, the CNN model showcased outstanding performance, learning intricate features from a vast dataset. The ResNet and DenseNet models also exhibited competitive accuracies, with the former demonstrating resilience to overfitting through residual connections. However, challenges in classifying certain malignant cases were evident, indicating areas for fine-tuning and addressing class imbalance. The VGGNet, while effective in distinguishing certain classes, faced hurdles in classifying malignant images.

In this evaluation, the CNN model emerged as the preferred choice among the four convolutional neural network models, boasting exceptional accuracy and overall effectiveness. Despite the strengths displayed by ResNet and DenseNet, the CNN model's simplicity, outstanding accuracy, and architectural considerations position it as the optimal model for image classification tasks in the context of breast cancer prediction. These findings provide valuable guidance for future endeavours, emphasizing the significance of model selection based on evaluation metrics and the specific nuances of the medical dataset under consideration.

10.2 FUTURE WORK

Looking forward, the project's achievements pave the way for several promising avenues in future research and development. Implementing advanced data augmentation techniques stands out as a potential enhancement to further diversify the dataset, contributing to improved generalization capabilities of the models. The exploration of hybrid models, combining Convolutional Neural Networks (CNNs) with traditional machine learning methods, presents an intriguing prospect to elevate classification accuracy by leveraging the strengths of both paradigms. Future efforts can focus on the development of explainability methods tailored to interpret the predictions of the models, fostering trust and understanding among healthcare professionals and stakeholders. Additionally, the incorporation of ensemble learning techniques such as Bagging and Boosting holds promise for refining the performance of the already established superior models, potentially achieving even greater predictive accuracy and robustness in breast cancer prediction applications. These avenues collectively underscore the potential for ongoing advancements and refinements in the domain of medical image classification for breast cancer prediction.

REFERENCES CITED

A.L.Sayeth Saabith, T.Vinothraj, MMM.Fareez, 2020. POPULAR PYTHON LIBRARIES AND THEIR APPLICATION DOMAINS. *International Journal of Advance Engineering and Research Development,* 7(11), pp. 18 - 26.

Anon., 2022. *Clevelandclinic.org.* [Online]   
Available at: https://my.clevelandclinic.org/health/diseases/3986-breast-cancer

Anon., 2023. *Cancer Research UK.* [Online]   
Available at: https://www.cancerresearchuk.org/about-cancer/breast-cancer

Anon., 2023. *How to Build a Machine Learning Model ?.* [Online]   
Available at: https://www.projectpro.io/article/how-to-build-a-machine-learning-model/914  
[Accessed 12 October 2023].

Anon., 2023. *Wikipedia.* [Online]   
Available at: https://en.wikipedia.org/wiki/Breast\_cancer

Anon., 2023. *World Health Organisation.* [Online]   
Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer  
[Accessed 12 July 2023].

C.D. Lekamlage, F. Afzal, E. Westerberg and A. Cheddad, 2020. *Mini-DDSM: Mammography-based Automatic Age Estimation.* Japan, 3rd International Conference on Digital Medicine and Image Processing (DMIP 2020).

Cutress, P. R., 2023. *Cancer Research UK.* [Online]   
Available at: https://www.cancerresearchuk.org/about-cancer/breast-cancer  
[Accessed 18 April 2023].

Dewang, N., 2023. *ML | Underfitting and Overfitting.* [Online]   
Available at: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/  
[Accessed 31 August 2023].

Elter, M., 2007. *UCI Machine Learning Repository..* [Online]   
Available at: https://archive.ics.uci.edu/dataset/161/mammographic+mass  
[Accessed 2007].

Mishra, P., 2019. *Medium.* [Online]   
Available at: https://medium.datadriveninvestor.com/why-are-convolutional-neural-networks-good-for-image-classification-146ec6e865e8

ProjectPro, 2023. *How to Build a Machine Learning Model.* [Online]   
Available at: https://www.projectpro.io/article/how-to-build-a-machine-learning-model/914  
[Accessed 12 October 2023].

UK, C. R., 2023. *Breast Cancer.* [Online]   
Available at: https://www.cancerresearchuk.org/about-cancer/breast-cancer

Varsha Nemade1,\*,Vishal Fegade2, 2023. Machine Learning Techniques for Breast Cancer Prediction. *ScienceDirect,* Issue 218, p. 1314–1320.

Veeling, Bastiaan S and Linmans, Jasper and Winkens, Jim and Cohen, Taco and Welling, Max, 2018. Rotation Equivariant {CNNs} for Digital Pathology. *The Journal of the American Medical Association,* 22(318), pp. 2199 - 2210.

Yvonne, E., 2023. *Breast Cancer.* [Online]   
Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Ziquan Zhu, Shui-Hua Wang, Yu-Dong Zhang, 2023. A Survey of Convolutional Neural Network in Breast Cancer. *National Library of Medicine,* 3(136), pp. 2127 - 2172.

Al-Ajmi, K. (n.d.). *Development of breast cancer risk prediction models using the UK Biobank dataset*.

Boeri, C., Chiappa, C., Galli, F., De Berardinis, V., Bardelli, L., Carcano, G., & Rovera, F. (2020). Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer Medicine*, *9*(9), 3234–3243. https://doi.org/10.1002/cam4.2811

Brancati, N., De Pietro, G., Frucci, M., & Riccio, D. (2019). A Deep Learning Approach for Breast Invasive Ductal Carcinoma Detection and Lymphoma Multi-Classification in Histological Images. *IEEE Access*, *7*, 44709–44720. https://doi.org/10.1109/ACCESS.2019.2908724

Chae, E. Y., Kim, H. H., Cha, J. H., Shin, H. J., & Kim, H. (2013). Evaluation of Screening Whole-Breast Sonography as a Supplemental Tool in Conjunction With Mammography in Women With Dense Breasts. *Journal of Ultrasound in Medicine*, *32*(9), 1573–1578. https://doi.org/10.7863/ultra.32.9.1573

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. In *IEEE Access* (Vol. 8, pp. 150360–150376). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2020.3016715

Gupta, K., & Chawla, N. (2020). Analysis of Histopathological Images for Prediction of Breast Cancer Using Traditional Classifiers with Pre-Trained CNN. *Procedia Computer Science*, *167*, 878–889. https://doi.org/10.1016/j.procs.2020.03.427

He, K., Zhang, X., Ren, S., & Sun, J. (n.d.). *Deep Residual Learning for Image Recognition*. http://image-net.org/challenges/LSVRC/2015/

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (n.d.). *Densely Connected Convolutional Networks*. https://github.com/liuzhuang13/DenseNet.

Iranmakani, S., Mortezazadeh, T., Sajadian, F., Ghaziani, M. F., Ghafari, A., Khezerloo, D., & Musa, A. E. (2020). A review of various modalities in breast imaging: technical aspects and clinical outcomes. *Egyptian Journal of Radiology and Nuclear Medicine*, *51*(1), 57. https://doi.org/10.1186/s43055-020-00175-5

Khamparia, A., Bharati, S., Podder, P., Gupta, D., Khanna, A., Phung, T. K., & Thanh, D. N. H. (2021a). Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. *Multidimensional Systems and Signal Processing*, *32*(2), 747–765. https://doi.org/10.1007/s11045-020-00756-7

Khamparia, A., Bharati, S., Podder, P., Gupta, D., Khanna, A., Phung, T. K., & Thanh, D. N. H. (2021b). Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. *Multidimensional Systems and Signal Processing*, *32*(2), 747–765. https://doi.org/10.1007/s11045-020-00756-7

Lee, S. H., Ryu, H.-S., Jang, M., Yi, A., Ha, S. M., Kim, S.-Y., Chang, J. M., Cho, N., & Moon, W. K. (2021). Glandular Tissue Component and Breast Cancer Risk in Mammographically Dense Breasts at Screening Breast US. *Radiology*, *301*(1), 57–65. https://doi.org/10.1148/radiol.2021210367

Nemade, V., & Fegade, V. (2023). Machine Learning Techniques for Breast Cancer Prediction. *Procedia Computer Science*, *218*, 1314–1320. https://doi.org/10.1016/j.procs.2023.01.110

Ortega-Martorell, S., Riley, P., Olier, I., Raidou, R. G., Casana-Eslava, R., Rea, M., Shen, L., Lisboa, P. J. G., & Palmieri, C. (2022). Breast cancer patient characterisation and visualisation using deep learning and fisher information networks. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-17894-6

Poojary, R., & Pai, A. (2019). Comparative Study of Model Optimization Techniques in Fine-Tuned CNN Models. *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 1–4. https://doi.org/10.1109/ICECTA48151.2019.8959681

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022). Prediction of Breast Cancer using Machine Learning Approaches. *Journal of Biomedical Physics and Engineering*, *12*(3), 297–308. https://doi.org/10.31661/jbpe.v0i0.2109-1403

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Shi, X., Su, H., Xing, F., Liang, Y., Qu, G., & Yang, L. (2020). Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. *Medical Image Analysis*, *60*, 101624. https://doi.org/10.1016/j.media.2019.101624

Syarif, I., Zaluska, E., Prugel-Bennett, A., & Wills, G. (n.d.). *Application of Bagging, Boosting and Stacking to Intrusion Detection*.

Taori, K., Dhakate, S., Rathod, J., Hatgaonkar, A., Disawal, A., Wavare, P., Bakare, V., & Puri, R. P. (2013). Evaluation of Breast Masses Using Mammography and Sonography as First Line Investigations. *Open Journal of Medical Imaging*, *03*(01), 40–49. https://doi.org/10.4236/ojmi.2013.31006

Teng, X., & Gong, Y. (2018). Research on Application of Machine Learning in Data Mining. *IOP Conference Series: Materials Science and Engineering*, *392*(6), 062202. https://doi.org/10.1088/1757-899X/392/6/062202

Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. In *Genome Medicine* (Vol. 13, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13073-021-00968-x

Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*, *292*(1), 60–66. https://doi.org/10.1148/radiol.2019182716

Zahra Jafari, & Ebrahim Karami. (2023). Breast Cancer Detection in Mammography Images: A CNN-Based Approach with Feature Selection. *Preprints*.

# Appendix

###### Installation and User Guide

This appendix furnishes a concise guide detailing the installation and utilization procedures for the "Breast Cancer Classifier" web application.

SOFTWARE REQUIREMENTS

Prior to software installation, it is imperative to verify that your system satisfies the specified prerequisites:

Python Version: Ensure that your system runs Python version 3.7 or later.

Dependencies: Streamlit, CV2, Numpy, PIL, Tensorflow, Joblib, Pandas

INSTALLATION AND RUNNING STEPS

Following are procedures that need to be followed in order to install and run the web application

Open Terminal

Navigate to the directory where the "app.py" file is located, using the cd command. For instance: **cd /path/to/directory/containing/app.py**

Activate Virtual Environment and Install the Required Packages

If the application is developed within a virtual environment, activate the environment using the appropriate command. For example, using Conda: **conda activate your\_virtual\_environment**

Ensure that all necessary packages are installed. If using a virtual environment, this step may not be required if the environment already includes the necessary dependencies. If not, install the packages listed in the "requirements.txt" file like streamlit: **pip install -r requirements.txt**

Run the Streamlit App

Execute the following command to run the Streamlit web application: **streamlit run app.py**

Access the Application

Once the application is successfully launched, a local server address (typically starting with "http://localhost") will be provided in the terminal. Open a web browser and navigate to this address to access and interact with the Streamlit web application.

A screen shot of a computer

Description automatically generated

Figure 39. Image showing how to access the web application

INSTRUCTION FOR USING THE APPLICATION

**1 Select Model:**

Use the sidebar to choose between "CNN (Image)" or "Random Forest (Details)" models.

**2 Upload Image / Enter Data:**

For "CNN (Image)," upload a CBIS-DDSM image (in jpg, jpeg, PNG, or png format).

For "Random Forest (Details)," input histopathological data using the sliders and dropdowns provided.

**3 Classify / Predict:**

Click the "Classify" button for image-based predictions or "Predict" for details-based predictions.

**4 View Results:**

The application will display the uploaded image (if applicable) and provide the model's prediction.

Ensure your uploaded image follows the supported formats, and input valid histopathological data for accurate predictions. Explore both models for a comprehensive understanding of breast cancer predictions based on different input modalities.

###### PROJECT LOG

This appendix provides a chronological log of important project milestones, events, and

activities throughout the development of the project.

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Timeline** | **Activity** |
| Initial Project Meeting | 13 June 2023 | * Discussion Regarding the Topic and its scope |
| Project Initiation | 31 July 2023 – 10 August 2023 | * Project officially initiated. * Key Literature review and Background reading started. * Initiated the Proposal design |
| Data Collection and Proposal development | 10 August 2023 – 1 September 2023 | * Identified data from different sources. * Selected Mammographic Mass data, The complete Mini DBIS DDSM and Histopathological data for final work. * Completed the project proposal |
| Model Development | 1 September 2023 – 15 October 2023 | * Successfully completed data pre-processing * Finalised the models and created. * Fine-tuned the model for best prediction and output. |
| Web application Development | 15 October 2023 – 20 October 2023 | * Created the web application using Streamlit |
| Testing the model and Web application | 20 October 2023 – 30 Octobber 2023 | * Both models along with the web application were tested to find out the working of the model. * Required error correction were done for the best working of the application. |
| Finalizing the Dissertation | 30 October 2023 – 3 November 2023 | * Preparation of Project Repot and Presentation |
| Feedback Meeting |  |  |
| Dissertation Submission | 3 November 2023 -1 December 2023 | * Final Submission of Dissertation * Project Officially Concluded |

Appendix Table 1: Project Log

NB: There were frequent meetings with the supervisor twice or thrice every month to discuss the progression of the Project.