Breast Cancer Detection Using Deep Learning Algorithm

By

Waqar Younus

Submitted to

**The University of Roehampton**

In partial fulfilment of the requirements

for the degree of

**MASTER OF SCIENCE IN COMPUTING**

Abstract

## Breast cancer is a type of cancer that begins in the cells of the breast. It is the most common cancer among women worldwide, accounting for approximately 25% of all cancer cases in women. However, men can also develop breast cancer, though it is much less common. Breast cancer is a form of cancer that develops when cells start to proliferate uncontrollably. This study seeks to address the problem involved with predicting and controlling breast cancer in its earlier stages. People rarely perform self-tests, and this leads to detecting cancer in its later stages. This is particularly bad because upon discovery, the chances that the cancer has metastasized are high making the chances of curing it very low. Most breast cancer can be detected as a lump/mass on the breast, or through self-examination or mammography.

Signed (apply signature below)

**Declaration**

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

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

**Date:** Enter the date here

**Enter your name here**

Acknowledgements

Acknowledgements narrative

Table of Contents

[**1.** **Introduction** 7](#_Toc98426634)

[Research Question or Problem that will be Addressed 7](#_Toc98426635)

[Aims 7](#_Toc98426636)

[Objectives 7](#_Toc98426637)

[Legal, Social, Ethical and Professional Considerations 7](#_Toc98426638)

[Background 7](#_Toc98426639)

[Report overview 7](#_Toc98426640)

[**2.** **Literature or Technology Review** 8](#_Toc98426641)

[**3.** **Design or Methodology** 9](#_Toc98426642)

[**4.** **Implementation or Results** 10](#_Toc98426643)

[Evaluation 10](#_Toc98426644)

[Related Work 10](#_Toc98426645)

[**5.** **Conclusion** 11](#_Toc98426646)

[Reflection 11](#_Toc98426647)

[Future Work 11](#_Toc98426648)

[**6.** **References** 12](#_Toc98426649)

[**7.** **Appendices** 13](#_Toc98426650)

# **Introduction**

Breast cancer is a type of cancer that begins in the cells of the breast. It is the most common cancer among women worldwide, accounting for approximately 25% of all cancer cases in women. However, men can also develop breast cancer, though it is much less common. Breast cancer is a form of cancer that develops when cells start to proliferate uncontrollably. A tumour that develops from these cells can be felt as a bump or seen on an X-ray. The most frequent cancer among women globally is breast cancer, which can also affect men. Although though the precise causes of breast cancer are not entirely known, several risk factors have been established, including age, gender, family history of the disease, genetic abnormalities, oestrogen exposure, and some lifestyle choices including drinking alcohol and being overweight.

Key Facts about Breast Cancer:

* Risk Factors: Various factors can increase the risk of developing breast cancer, including age, family history of breast cancer, genetic mutations (e.g., BRCA1 and BRCA2 genes), early menstruation, late menopause, hormonal factors, obesity, alcohol consumption, and radiation exposure.
* Types of Breast Cancer: Breast cancer can be classified into several types, including ductal carcinoma in situ (DCIS), invasive ductal carcinoma, invasive lobular carcinoma, and triple-negative breast cancer, among others.
* Detection and Diagnosis: Regular breast self-exams and mammograms are important for early detection of breast cancer. If an abnormality is found, further diagnostic tests, such as biopsy, may be conducted to confirm the presence of cancer and determine its type and stage.
* Staging: Breast cancer is staged based on the size of the tumor, lymph node involvement, and whether it has spread to other parts of the body (metastasis). Staging helps guide treatment decisions.
* Treatment: Treatment for breast cancer depends on the stage and type of cancer. Common treatment options include surgery, radiation therapy, chemotherapy, hormone therapy, targeted therapy, and immunotherapy.
* Prognosis: Prognosis for breast cancer varies widely depending on the stage at diagnosis, tumor characteristics, and treatment response. Early detection and treatment offer better chances of successful outcomes.

The process of locating and defining areas of interest within a breast image that correlate to possible breast cancer, such as masses or calcifications, is known as breast cancer lesion segmentation. Accurate medical image lesion segmentation is crucial for medical image analysis since it can help with breast cancer detection, diagnosis, and therapy.

Breast lesion segmentation can be done in several ways, including manually by radiologists and through a variety of automated algorithms. Automated segmentation techniques frequently employ machine learning algorithms, such as convolutional neural networks (CNNs), to recognise regions of interest and discover patterns in breast pictures.

page4image48871440page4image48857712page4image48857920page4image48873104page4image48864992page4image48867072

**Figure 1.1 sample breast cancers**

In order to segment breast lesions using CNNs, one popular method entails training the network using a sizable dataset of breast cancer images with annotations that describe the position and size of any lesions present. By using the learnt patterns on the input data, the trained network may then be used to automatically segment lesions in new breast cancer images. Thresholding, region expanding, and active contours are other methods for breast lesion segmentation that are based on mathematical models and image processing algorithms. Although more manual tinkering and intervention may be needed than with machine learning- based solutions, these techniques may be useful in some situations.

Unlike machine learning models, mathematical models require less amount of data to perform segmentation. The drawback of most mathematical models are the fact that they require some amount of tinkering to find a combination of filters that solves particular image segmentation problems.

## 1.1 Research Question or Problem that will be Addressed

## The research questions to be answered are:

1. What is the percentage of belign and manignant sets in the dataset.
2. What ways can the mammogram images be best processed to achieve high detection accuracy.
3. How can the dataset be augmented to achieve variability in the dataset.
4. What parameters are particularly important in the prediction of breast cancer.
5. What combination of models can be used to achieve high detection rate.

## 1.2 Aims

## To identify the onset of breast cancer

1. Increase the accuracy of detection of breast cancer.
2. Incorporate image segmentation in the detection process.

## 1.3 Objectives

## Design a breast cancer detection model.

1. Perform feature selection on dataset.
2. Design a web interface to interact with the developed model.
3. Compare the accuracy of the developed model with other models in the problem domain.

## 1.4 Legal, Social, Ethical and Professional Considerations

## When conducting an academic project, it's essential to be mindful of various legal, social, ethical, and professional considerations to ensure the research is conducted responsibly and with integrity. Here are some key considerations for each category:

## Legal Considerations:

## Compliance: I ensured that the project complies with all applicable laws, regulations, and institutional policies. This includes data protection laws, research ethics guidelines, and intellectual property rights.

## Informed Consent: I obtained informed consent from research participants, ensuring they fully understand the purpose of the study, their rights, and the potential risks involved.

## Data Privacy: I safeguarded the confidentiality and privacy of personal data collected during the project. Anonymize or pseudonymize data when possible to protect participants' identities.

## Social Considerations:

## Social Impact: I considered the potential impact of your research on society, individuals, and communities. Strive to promote positive outcomes and avoid harm.

## Inclusivity: I was mindful of diversity and inclusion in your research, ensuring that the study represents and respects the perspectives of various social groups.

## Addressing Bias: I was aware of any potential biases in your research design, data collection, or interpretation of results. Take steps to mitigate bias and increase the study's objectivity.

## Ethical Considerations:

## Integrity: I conducted your research with honesty, transparency, and academic rigor. Avoid plagiarism and properly cite all sources.

## Conflicts of Interest: I disclosed any conflicts of interest that could influence the research outcomes or interpretation of results.

## Professional Considerations:

## Authorship and Collaboration: I clearly defined authorship roles and credit all contributors appropriately. Respect the intellectual property and contributions of others.

## Open Science: I am considering sharing my research data and findings openly and transparently, promoting reproducibility and collaboration.

## Peer Review: I seek feedback and constructive criticism from peers and mentors to improve the quality and validity of my research.

## By addressing these considerations, I ensured that my academic project is conducted ethically, responsibly, and in alignment with professional standards. It also enhances the credibility and impact of my research in the academic community and beyond.

## 1.5 Background

## This study seeks to address the problem involved with predicting and controlling cancer in its earlier stages. People rarely perform self-tests, and this leads to detecting cancer in its later stages. This is particularly bad because upon discovery, the chances that the cancer has metastasized are high making the chances of curing it very low. Most breast cancer can be detected as a lump/mass on the breast, or through self-examination or mammography.

## 1.6 Report overview

**Chapter 1**- Introduction; In this chapter, the background of the project is discussed in detail, the aims and objectives of the project are stated, and finally, legal, social, ethical, and professional considerations are also highlighted.

**Chapter 2**- Literature review; background research, theoretical approach. This chapter considers different literature by different authors related to the problem domain. Theoretical approach and background research are considered in this section of the project. The review of different technologies is also discussed.

**Chapter 3**- Design and methodology; technological review, methodological review. This chapter discusses the methodology taken to carry out the project and the technological stack that was adopted for the project.

**Chapter 4**- Implementation and result; the answers to the generated research questions in chapter one are evaluated here with discussions, and every step of the methodology to achieve the solution is also discussed.

**Chapter 5**- Conclusion; this part of the project provides a summary of the entire research. It discusses the answers to the research questions and provides a reflection of the result with recommendations and future directions.

In conclusion, a project plan for the project has been developed on Teamwork with a breakdown of the methodology and timeline for each stage of the methodology. The link to this plan is provided in Appendix 3.

# **Literature or Technology Review**

Breast cancer is a major global public health problem accounting for massive morbidity and significant mortality worldwide. Factors contributing to breast cancer mortality have been a topic of intense research and discussion in the scientific world. There is, however, a dearth of information on the incidence of breast cancer mortality in most resource-poor countries including Nigeria. Available data from most African workers on breast cancer focused on incidence, risk factors, and complications rather than mortality (Mohammed et al, 2017). It also affects one in eight women during their lives. It is the commonest site specific malignancy affecting women and the most common cause of cancer mortality in women worldwide. Breast cancer is a malignant (cancerous) growth that begins in the tissues of the breast. Cancer is a disease in which abnormal cells grow in an uncontrolled way (Siegel , Mille , Jemal , 2016). It is the most common cancer in women, but it can also appear in men. Breast cancer is now an epidemic, posing a serious threat to the health of women of all races globally. In Nigeria, cervical cancer was the commonest cause of cancer- related deaths among women for several decades but breast cancer is the leading cause of cancer related deaths among Nigerian women. This is not due to a reduction in cervical cancer but an increase in the incidence of breast cancer. Breast cancer is commonly seen in four stages that represents its progression. (Lee et al, 2010)

In stage I, the disease is confined entirely to the breast. The cancer usually start as a very tiny growth that cannot yet be felt but can be detected with imaging tests such as mammography and ultrasound. At this first stage, treatment is usually curative and more than 95% of those so detected will survive the disease beyond 5 years (Egenti, 2016). Stage II is a cancer that has involved lymph nodes in the armpit of the same side of the breast, while stage III disease is one that has involved the muscles under the breast. Stages II and III therefore require very aggressive treatment using different modalities to contain the spread of the disease. It is however difficult to cure a patient in stage IV because the disease has spread and may have involved other organs in the body such as the lungs, liver, bones, the brain or the spine (Egenti, 2016).

The five-year survival rate for breast cancer patients in the United States exceeds 85%, in Nigeria it is a dismal 10%. Breast cancer is responsible for about 16% of all cancer related deaths in Nigeria (Mohammed et al, 2017).

Breast cancer early detection is crucial for improving treatment outcomes and reducing mortality rates. Detecting breast cancer at an early stage allows for more effective treatment options, potentially less aggressive interventions, and a higher chance of successful recovery. Here are some key aspects of breast cancer early detection:

1. Breast Self-Examination (BSE):
   * BSE involves regularly examining one's breasts for changes in size, shape, or texture, as well as the presence of lumps or other abnormalities.
   * While BSE is not a definitive diagnostic tool, it encourages women to become familiar with their breasts, making it easier to detect any unusual changes and seek medical evaluation promptly.
2. Clinical Breast Examination (CBE):
   * CBE is performed by a healthcare professional, typically a doctor or nurse, during routine check-ups or in response to reported breast changes or concerns.
   * The healthcare professional examines the breasts and surrounding areas for any abnormalities or signs of breast cancer.
3. Mammography:
   * Mammography is the most common and effective screening tool for breast cancer.
   * It involves low-dose X-rays to create images of the breast tissue, allowing for the detection of tumors or abnormalities that may not be felt during a physical examination.
   * Mammography is recommended for women over a certain age (usually starting at 40 or 50) and can detect breast cancer at an early stage, even before symptoms are noticeable.
4. Breast Ultrasound and MRI:
   * In certain cases, breast ultrasound and MRI may be used in addition to mammography to provide further information about breast abnormalities.
   * These imaging techniques are often used to evaluate suspicious findings identified during a mammogram or physical examination.
5. Breast Cancer Risk Assessment:
   * Some women may have a higher risk of developing breast cancer due to family history or genetic factors.
   * Breast cancer risk assessment tools can help identify individuals at increased risk, allowing for more personalized and targeted screening and prevention strategies.
6. Education and Awareness:
   * Promoting breast cancer awareness and education is vital to encourage women to seek regular screenings and report any changes or concerns promptly.
   * Educational campaigns also aim to dispel misconceptions and encourage early detection and treatment-seeking behaviors.

It's important to remember that early detection is not a guarantee of breast cancer prevention, but it can significantly improve treatment outcomes. Women should discuss their individual risk factors and appropriate screening strategies with healthcare professionals to create a personalized breast health plan. Additionally, maintaining a healthy lifestyle, such as regular exercise, a balanced diet, and avoiding smoking and excessive alcohol consumption, may also contribute to breast cancer risk reduction.

Researchers have been investigating the use of machine learning (ML) and image filtering approaches to increase the accuracy of breast cancer diagnosis. There are various effective solutions in this problem domain that have achieved success in varying degrees. Convolutional neural networks (CNNs), a class of ML algorithm that can be trained to identify patterns in images, are one of these strategies. CNNs have been shown to be useful for detecting breast cancer in several trials, with some achieving accuracy rates of over 90%. For instance, Al-masni et al (2021) classified breast cancer images with 98% accuracy using a CNN-based method. In addition to using neural networks, other research has investigated the application of additional ML techniques, such as decision trees and support vector machines (SVMs), for the identification of breast cancer. For instance, El-Dahshan et al (2014) classified breast cancer images using an SVM-based method, and they did so with 96% accuracy.

The detection of breast cancer using image filtering methods has been a focus of current study. Image filtering is a method used to increase the clarity of medical images, which can help with diagnosis. One image filtering strategy is to smoothen an image using a method called Gaussian filtering, which improves edges and reduces noise. In several research, Gaussian filtering has been used to improve mammography pictures and raise the detection efficiency of breast cancer. Gaussian filtering, for instance, was employed in a study by Sadri et al. (2020) to improve mammography pictures and obtain a 90% accuracy in breast cancer diagnosis. Another strategy is to utilise image filtering methods like wavelet transformation and gaussian filtering to increase the quality of breast cancer pictures and diagnosis precision. In a research by Mirniaharikandehei et al. (2020), the accuracy of breast cancer diagnosis was significantly increased by enhancing mammography images using a mix of wavelet transformation and Gaussian filtering. Several research have investigated the use of other image filtering methods, such as edge detection and morphological procedures, for the diagnosis of breast cancer in addition to these methods. For instance, a research by Kamble et al. (2018) improved mammography images using a mix of edge detection and morphological processes, and they were able to diagnose breast cancer with a 95% accuracy rate.

Overall, the research points to the potential of ML and image filtering approaches to increase the precision of breast cancer identification. The results thus far are encouraging and lay the groundwork for additional study in this field, but further studies are required to confirm these findings and improve the performance of these techniques.

2.2 Technology Review

There are various technologies and tools used for data analysis, depending on the complexity of the data and the specific requirements of the analysis. Some of the common technologies and tools used for data analysis include:

1. Spreadsheet Software:
   * Spreadsheet software like Microsoft Excel or Google Sheets is widely used for simple data analysis and visualization. It is accessible to users with varying levels of technical expertise.
2. Statistical Software:
   * Statistical software packages like R and Python with libraries such as Pandas, NumPy, and SciPy are popular choices for data analysis. They offer powerful statistical functions and data manipulation capabilities.
3. Business Intelligence (BI) Tools:
   * BI tools like Tableau, Power BI, and QlikView are designed to create interactive and visually appealing data visualizations, dashboards, and reports.
4. Data Visualization Libraries:
   * Libraries like Matplotlib, Seaborn, and ggplot in Python and R enable the creation of static and interactive data visualizations, aiding in the exploration and communication of insights.
5. Data Mining Tools:
   * Data mining tools such as RapidMiner, KNIME, and Weka are used for exploring large datasets, discovering patterns, and building predictive models.
6. Machine Learning Frameworks:
   * Machine learning frameworks like scikit-learn in Python, TensorFlow, and PyTorch are used to develop and train machine learning models for various data analysis tasks.
7. Database Management Systems (DBMS):
   * DBMS such as MySQL, PostgreSQL, and Microsoft SQL Server are used to store, manage, and retrieve large datasets for analysis.
8. Big Data Tools:
   * For handling big data, technologies like Apache Hadoop, Apache Spark, and Apache Flink are used to process and analyze massive datasets in distributed computing environments.
9. Natural Language Processing (NLP) Tools:
   * NLP libraries like NLTK (Natural Language Toolkit) in Python and spaCy enable the analysis and understanding of textual data.
10. Cloud-Based Data Analysis Platforms:
    * Cloud-based platforms like Google Cloud Platform (GCP), Microsoft Azure, and Amazon Web Services (AWS) offer scalable and cost-effective solutions for data storage, processing, and analysis.
11. Data Integration and ETL (Extract, Transform, Load) Tools:
    * ETL tools like Apache NiFi, Talend, and Informatica facilitate the extraction, transformation, and loading of data from multiple sources for analysis.

The choice of technology depends on factors such as the nature and volume of data, the complexity of the analysis, the required level of interactivity and visualization, and the expertise of the data analysts and scientists involved in the project. Often, a combination of multiple tools and technologies is used to perform comprehensive data analysis and gain meaningful insights from the data.

# **Design or Methodology**

Breast cancer (BC) is one of the most common cancers among women worldwide, representing most new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem today.

Early discovery of breast cancer can help to improve patients’ chances of survival, as it will afford medical practitioners the opportunity of providing targeted care that are born out of reliable diagnosis. Further accurate classification of benign breast cancer can prevent patients undergoing unnecessary treatments. Thus, this research work is targeted at providing accurate diagnosis of breast cancer as either malignant or benign groups.

I propose a methodology that takes in the breast cancer dataset, perform necessary data pre-processing tasks on the dataset to remove and correct data inconsistencies to reduce dimensionality and avoid overfitting the model, divide the cleaned dataset into a training and testing dataset, use the training dataset to train the model and finally use the test dataset to validate the trained model. The model will be a combination of a bidirectional long short-term memory encoder and a two-dimensional convolutional neural network. The bidirectional long short-term memory will actively learn the dataset and the two-dimensional convolutional neural network will classify the dataset into the required class of either malignant or benign.

The proposed methodology is shown in figure 3.1 below.

The methodology is divided into three major parts namely:

1. Data collection and pre-processing
2. Model design, training and validation, and
3. Breast cancer detection and evaluation method

A diagram of a data processing process

Description automatically generated

**Figure 3.1 the proposed methodology**

**3.2 Data Collection**

This phase of the methodology involves the dataset used for the research work and how the dataset will be processed to make it ready for use in our analysis. Collecting and curating breast cancer image data is an essential step for developing and training machine learning models for tasks related to breast cancer detection, diagnosis, and classification. Invasive Ductal Carcinoma (IDC), also known as Infiltrating Ductal Carcinoma, is the most common type of breast cancer. It accounts for approximately 80% of all diagnosed breast cancers. IDC originates in the milk ducts (hence the term "ductal") of the breast but then invades surrounding breast tissues (hence "invasive"). IDC begins in the milk ducts of the breast when the cells lining the ducts undergo abnormal changes (mutation) that lead to uncontrolled growth. Over time, the cancer cells break through the duct walls and invade the surrounding breast tissue, potentially spreading to nearby lymph nodes and other parts of the body. IDC often presents as a lump or mass in the breast. Other possible signs and symptoms include changes in breast size or shape, nipple discharge, changes in the appearance of the breast skin (e.g., dimpling), and nipple inversion or changes.

IDC is typically diagnosed through a combination of methods, including mammography, ultrasound, biopsy, and sometimes MRI. A biopsy is essential to confirm the presence of cancer and to determine its characteristics, such as the grade and hormone receptor status. IDC tumors are usually graded based on their appearance under a microscope. The grade helps predict how aggressive the cancer is. Higher-grade tumors tend to grow and spread more quickly.

The dataset used for this project was collected by Paul Mooney. The initial dataset included 162 images of whole mount slides from Breast Cancer (BCa) specimens that had been scanned at 40x. 277,524 patches of size 50 × 50 were retrieved from it, with 198,738 IDCs being negative and 78,786 IDCs being positive. The format of the file name for each patch is u\_xX\_yY\_classC.png, for instance 10253\_idx5\_x1351\_y1101\_class0.png. Where u is the patient ID (10253\_idx5), X and Y are the x- and y-coordinates of the patch from which this image was cut, respectively, and C is the class, with 0 denoting non-IDC and 1 denoting IDC. Figure 3.2 shows sample of the collected mammogram images.

A collage of cells

Description automatically generated

**Figure 3.2 sample image dataset**

**3.3** **Data Pre-processing**

After collecting and storing the image data from the previous section, the next phase is to process the data. Image data pre-processing is a crucial step in preparing images for analysis, particularly in tasks like computer vision, image recognition, and deep learning. Proper pre-processing helps improve the quality and suitability of images for subsequent tasks. These tasks are discussed below:

Task 1: Read and Marge Images into Classification

The first task was to read in the dataset. The images are originally stored in folders with the patient ID and the class of the image. Zero was used to represent benign and one was used to represent malignant. Figure 3.3 shows the initial loading of the dataset to Jupyter notebook and sample folders in the dataset.

A screenshot of a computer screen

Description automatically generated

Figure 3.3 sample folders in the dataset

After loading the dataset, the content of the folders is read and marge into two folders: Benign and Malignant along with there classification. The function that does the merging is shown in figure 3.4.

A screenshot of a computer program

Description automatically generated

Figure 3.4 image folder merging function

After performing the merging and classification, the new dataset created from this process is then loaded as a data-frame for processing. These is shown in figure 3.5.

A screenshot of a computer

Description automatically generated

Figure 3.5 new dataset.

Task 2: Resizing and Scaling

Resizing and scaling are common operations in computer graphics and image processing, used to change the dimensions or size of an image or object while maintaining its aspect ratio or adjusting it as needed. Resizing involves changing the dimensions (width and height) of an image or object while preserving its original aspect ratio or not. It's often used to make an image fit within a specific space or to create different versions of an image for various purposes. To make the image usable by the machine learning algorithm, the image data was resized and scaled. Promotional scaling was employed to perform this task. This method maintains the aspect ratio of the image, ensuring that it doesn't get distorted. You specify either the width or the height, and the other dimension is adjusted accordingly. The image was resized so that the total area of the image is 224 x 224. Figure 3.6 shows a sample of the resized image.

A close-up of a microscope

Description automatically generated

Figure 3.6 resized image data.

Task 3: Image Filtering

Image filtering is a crucial pre-processing step in many machine learning tasks involving computer vision, image analysis, and deep learning. It helps enhance the quality of images, extract meaningful features, and reduce noise, making the images more suitable for subsequent processing and analysis. Random noise filter was applied to the dataset images to make them better for the machine learning algorithm that is used for this project. A random noise image filter, also known as a noise generator or noise filter, is a tool used to add random noise to an image. This noise can introduce a level of randomness or imperfection to an otherwise clean or smooth image. Noise is often used in various fields, including image processing, computer graphics, and machine learning, to simulate real-world imperfections or enhance the robustness of algorithms. For this project, the gaussian filter was used to add random noise to the images. A Gaussian noise image filter, often referred to as Gaussian noise, adds random noise to an image by introducing pixel-level variations following a Gaussian (normal) distribution. Gaussian noise is a type of noise that is commonly used for simulating various real-world noise sources in images, such as sensor noise in photography or electronic noise in imaging devices.

The Gaussian noise is generated based on a normal distribution with a mean (μ) of 0 (zero) and a standard deviation (σ) that determines the spread or the amount of noise. The mathematical formula for generating Gaussian noise is as follows:

For each pixel or data point, you generate a random value (N) from a normal (Gaussian) distribution:

N∼N(0,σ2)*N*∼N(0,*σ*2)

Where:

* N*N* is the generated random value following a Gaussian distribution.
* NN represents the normal (Gaussian) distribution.
* 00 is the mean of the distribution (μ), which is typically set to zero for Gaussian noise.
* σ2*σ*2 is the variance of the distribution, and σ*σ* is the standard deviation. The standard deviation (σ*σ*) controls the spread or the intensity of the noise. Larger values of σ*σ* result in more intense noise.

The gaussian noise was imported from the sci-kit learn image filter library. The output is shown in figure 3.7.

A blurry image of a pink surface

Description automatically generatedA pink and purple pixelated square

Description automatically generated

Figure 3.7 gaussian and random noise filter application.

**3.4 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a critical initial step in understanding and gaining insights from image data. While traditional EDA methods for tabular data may not directly apply to images, there are several techniques and approaches you can use to explore image data effectively.

One way is to show the distribution of the image. The distribution of the IDC present in the images is shown in figure 3.8.

A graph with pink lines

Description automatically generated with medium confidence

Figure 3.8 distribution of IDC in the dataset.

The next one is to find out how many patches are there per patient. This is show in figure 3.8.

A graph of patches

Description automatically generated

Figure 3.8 distribution of patches per patient.

After the above distribution, the next was to discover the count of positive IDC (malignant) and negative IDC (benign) cases there is in the dataset.

A graph with blue and orange squares

Description automatically generated

Figure 3.9 count of IDC in dataset.

Finally, we take a closer look at the shape of the patches and their distribution in each mammogram using Binary objective visualization for each tissue slice. This is shown in figure 3.10.

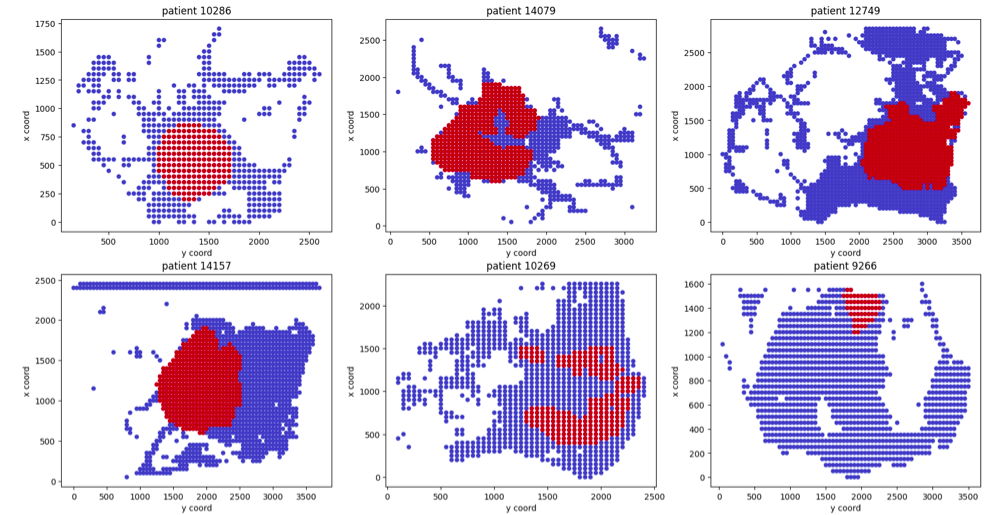


Figure 3.10 patches distribution using binary objective visualization.

**3.5 Model Design**

After collecting and cleaning the dataset, the next phase in the methodology is the design of the model. Designing a machine learning model is a crucial step in building predictive or analytical systems. It involves selecting the appropriate algorithms, architecture, and parameters to solve a specific problem. In this project, I have decided to use a Bidirectional Long Short Memory (BiLSTM) algorithm with a Convolutional Neural network algorithm (CNN). Bidirectional Long Short-Term Memory networks (Bidirectional LSTM or BiLSTM) and Convolutional Neural Networks (CNNs) are typically used for different types of data—LSTMs for sequential data like text or time series, and CNNs for grid-like data such as images. However, it is possible to combine these architectures for image classification tasks by using a combination of feature extraction with CNNs and sequence modeling with Bidirectional LSTMs. This approach is often referred to as a hybrid CNN-LSTM model or CNN-LSTM ensemble. The rationale behind using both algorithms is because there are multiple images for one patient in the dataset. BiLSTM will keep track of the sequence of images per patient, while the CNN does the classification.

BiLSTM

Bidirectional Long Short-Term Memory networks (Bidirectional LSTM or BiLSTM) are a variant of the Long Short-Term Memory (LSTM) neural network architecture. LSTMs are a type of recurrent neural network (RNN) designed to capture long-range dependencies and sequential patterns in data, making them particularly well-suited for tasks involving sequences, such as time series prediction, natural language processing, and speech recognition. Bidirectional LSTMs enhance the standard LSTM by processing the input sequence in both forward and backward directions, which allows them to capture information from past and future time steps simultaneously. The have the following characteristics:

1. **Forward and Backward Passes**:
   * In a Bidirectional LSTM, the input sequence is processed in two directions: forward and backward.
   * The forward pass starts from the beginning of the sequence, while the backward pass starts from the end.
   * At each time step, the forward LSTM cell computes hidden states by considering past information, while the backward LSTM cell computes hidden states by considering future information.
2. **Combining Hidden States**:
   * The outputs (hidden states) of both the forward and backward LSTMs are combined at each time step. Typically, this is done using concatenation.
   * The combined hidden states capture information from both the past and future, providing a more comprehensive understanding of the input sequence at each time step.
3. **Improved Sequence Understanding**:
   * Bidirectional LSTMs are effective at capturing bidirectional dependencies in sequential data.
   * They can capture context from earlier and later parts of the sequence, which can be especially valuable in tasks where understanding the entire sequence is important.

In terms of mathematical representation, a Bidirectional LSTM can be defined as a composition of two standard LSTMs: one that processes the input sequence from left to right (forward LSTM) and another that processes it from right to left (backward LSTM). The hidden states of both LSTMs are then combined at each time step. Here's a mathematical representation of a Bidirectional LSTM:

forward LSTM:

* At each time step t*t*, the forward LSTM computes the following:
  + **it(f)=σ(Wix(f)xt+Wih(f)ht−1(f)+bi(f))*it*(*f*)​=*σ*(*Wix*(*f*)​*xt*​+*Wih*(*f*)​*ht*−1(*f*)​+*bi*(*f*)​) (Input Gate)**
  + **ft(f)=σ(Wfx(f)xt+Wfh(f)ht−1(f)+bf(f))*ft*(*f*)​=*σ*(*Wfx*(*f*)​*xt*​+*Wfh*(*f*)​*ht*−1(*f*)​+*bf*(*f*)​) (Forget Gate)**
  + **ot(f)=σ(Wox(f)xt+Woh(f)ht−1(f)+bo(f))*ot*(*f*)​=*σ*(*Wox*(*f*)​*xt*​+*Woh*(*f*)​*ht*−1(*f*)​+*bo*(*f*)​) (Output Gate)**
  + **gt(f)=tanh⁡(Wgx(f)xt+Wgh(f)ht−1(f)+bg(f))*gt*(*f*)​=tanh(*Wgx*(*f*)​*xt*​+*Wgh*(*f*)​*ht*−1(*f*)​+*bg*(*f*)​) (Candidate Cell State)**
  + **ct(f)=ft(f)⊙ct−1(f)+it(f)⊙gt(f)*ct*(*f*)​=*ft*(*f*)​⊙*ct*−1(*f*)​+*it*(*f*)​⊙*gt*(*f*)​ (Cell State)**
  + **ht(f)=ot(f)⊙tanh⁡(ct(f))*ht*(*f*)​=*ot*(*f*)​⊙tanh(*ct*(*f*)​) (Hidden State)**

Backward LSTM:

* At each time step t*t*, the backward LSTM computes the following:
  + **it(b)=σ(Wix(b)xt+Wih(b)ht+1(b)+bi(b))*it*(*b*)​=*σ*(*Wix*(*b*)​*xt*​+*Wih*(*b*)​*ht*+1(*b*)​+*bi*(*b*)​) (Input Gate)**
  + **ft(b)=σ(Wfx(b)xt+Wfh(b)ht+1(b)+bf(b))*ft*(*b*)​=*σ*(*Wfx*(*b*)​*xt*​+*Wfh*(*b*)​*ht*+1(*b*)​+*bf*(*b*)​) (Forget Gate)**
  + **ot(b)=σ(Wox(b)xt+Woh(b)ht+1(b)+bo(b))*ot*(*b*)​=*σ*(*Wox*(*b*)​*xt*​+*Woh*(*b*)​*ht*+1(*b*)​+*bo*(*b*)​) (Output Gate)**
  + **gt(b)=tanh⁡(Wgx(b)xt+Wgh(b)ht+1(b)+bg(b))*gt*(*b*)​=tanh(*Wgx*(*b*)​*xt*​+*Wgh*(*b*)​*ht*+1(*b*)​+*bg*(*b*)​) (Candidate Cell State)**
  + **ct(b)=ft(b)⊙ct+1(b)+it(b)⊙gt(b)*ct*(*b*)​=*ft*(*b*)​⊙*ct*+1(*b*)​+*it*(*b*)​⊙*gt*(*b*)​ (Cell State)**
  + **ht(b)=ot(b)⊙tanh⁡(ct(b))*ht*(*b*)​=*ot*(*b*)​⊙tanh(*ct*(*b*)​) (Hidden State)**

Combining Hidden States:

* At each time step t*t*, the hidden states from the forward (ht(f)*ht*(*f*)​) and backward (ht(b)*ht*(*b*)​) LSTMs are concatenated:
  + **ht=[ht(f),ht(b)]*ht*​=[*ht*(*f*)​,*ht*(*b*)​]**

The final hidden states ht*ht*​ represent the bidirectional context information at each time step t*t*. These states can be used for various sequence modeling tasks, such as sequence classification, named entity recognition, machine translation, and more.

CNN

A Convolutional Neural Network (CNN) is a deep learning architecture designed specifically for processing structured grid-like data, such as images and videos. CNNs have become the cornerstone of computer vision and image recognition tasks due to their ability to automatically learn and extract hierarchical features from input data.

The mathematical representation of a Convolutional Neural Network (CNN) can be described in terms of its key components, including convolutional layers, pooling layers, activation functions, and fully connected layers. Below, I'll provide a simplified mathematical representation of these components:

1. **Convolution Operation**:
   * Let X*X* be the input data or feature map, and W*W* be a convolutional filter (kernel).
   * The convolution operation at a specific location (i,j)(*i*,*j*) is represented as:

(X∗W)ij=∑m=0M−1∑n=0N−1X(i+m)(j+n)⋅Wmn(*X*∗*W*)*ij*​=*m*=0∑*M*−1​*n*=0∑*N*−1​*X*(*i*+*m*)(*j*+*n*)​⋅*Wmn*​

* + Here, M*M* and N*N* are the dimensions of the filter W*W*, and (i+m)(*i*+*m*) and (j+n)(*j*+*n*) iterate over the input data X*X* to compute the dot product between the filter and the input data at each location.

1. **Pooling Operation**:
   * Pooling is often used to downsample the feature maps. Max pooling is a common operation, where the maximum value in a local region is selected.
   * If X*X* is the input data, the max pooling operation can be represented as:

MaxPooling(X)ij=max⁡(X(i⋅s)(j⋅s),X(i⋅s)(j⋅s+1),…,X(i⋅s+r)(j⋅s+s−1))MaxPooling(*X*)*ij*​=max(*X*(*i*⋅*s*)(*j*⋅*s*)​,*X*(*i*⋅*s*)(*j*⋅*s*+1)​,…,*X*(*i*⋅*s*+*r*)(*j*⋅*s*+*s*−1)​)

* + Here, s*s* is the stride (the step size for moving the pooling window), and r*r* and s*s* represent the dimensions of the pooling window.

1. **Activation Function**:
   * The Rectified Linear Unit (ReLU) activation function is commonly used in CNNs. It can be mathematically represented as:

ReLU(x)=max⁡(0,x)ReLU(*x*)=max(0,*x*)

1. **Fully Connected Layer**:
   * In fully connected layers, a set of weights W*W* and biases b*b* are learned. The output Y*Y* of a fully connected layer is computed as:

Y=ReLU(XW+b)*Y*=ReLU(*XW*+*b*)

* + Here, X*X* represents the input to the fully connected layer, W*W* is the weight matrix, and b*b* is the bias vector.

1. **Output Layer**:
   * The output layer is typically used for classification or regression tasks. The output O*O* can be computed based on the specific task. For classification with C*C* classes, it might involve using a softmax activation:

O=Softmax(XW+b)*O*=Softmax(*XW*+*b*)

where SoftmaxSoftmax is the softmax function that converts the logits into class probabilities.

The implementation of this combined model is discussed in the implementation section.

# **Implementation or Results**

Predicting breast cancer using machine learning is a critical application in healthcare. It typically involves using a dataset of patient information and medical features to train a machine learning model that can classify breast cancer cases as benign (non-cancerous) or malignant (cancerous). As mentioned in the previous chapter of this project work, we employed a BiLSTM and two dimensional convolutional neural network architecture to predict the presence of breast cancer.

To achieve the aim of this research work, some requirements need to be met for the implementation of the proposed solution. These requirements are discussed in the sub sections below:

**4.1 Software Requirements**

We used Anaconda as the development environment which is a lunch tool for scientific analysis using Python in the Anaconda environment. The interface of Anaconda is shown in figure 4.1 below:

A screenshot of a computer

Description automatically generated

**Figure 4.1 the anaconda interface**

**4.2 Hardware Requirement**

- Pentium IV processor (minimum)

– 256MB-4GB RAM/ memory space

- 10GB hard disk space (minimum)

- CD ROM driver or DVD driver

**4.3 Split Dataset to Train and Test Sets**

The dataset for this project is split into a training and testing set. The training set was used to train the model, while the testing set was used to validate the model. Figure 4.2 shows the function that performs the splitting.

A screenshot of a computer code

Description automatically generated

Figure 4.2 dataset splitting

It is seen that there are four folders after the splitting was performed where the “1” represents the presence of cancer (malignant) and the “0” represents the absence.

**4.4 Model Architecture**

After performing the necessary pre-processing task and the splitting of the dataset, the next thing is to define the architecture of the network models. As stated before, the model is a combination of the CNN model and the BiLSTM model. While the LSTM model performs feature extraction, the extracted feature is parsed to the CNN to make a classification.

The CNN model has 50 layers and has 32 x 4 dimensions. It is particularly optimized to perform this task. The fine-tuning task was to choose a good learning rate to ensure that the gradient descent algorithm of the CNN can converge. The feature vector of the last max pooling of the CNN is used as the input for the LSTM.

The LSTM has a dimension of 2048 and 2048 hidden layers with 0.4 dropout rate. The LSTM processes the breast cancer images considered by their arrangement. It ensures that images from the same patient are grouped and processed with temporal analysis. Leaky Relu and Softmax are the two-activation function used in this model. The output for training and evaluating the model is shown in figure 4.3.

A table of numbers and a line

Description automatically generated with medium confidence

Figure 4.3 model training

After training the model, the accuracy and loss function as shown in figure 4.3 is show in a graph. This is discussed in the next section.

**4.5 Result**

## In machine learning, classification is a common task where the goal is to categorize or classify data points into predefined classes or categories. Once the classification model is trained and evaluated, its performance is modelled to obtain classification results that helps us understand how well the model has performed. The accuracy and loss function was used for this purpose. Figure 4.4 shows the accuracy and figure 4.5 shows the loss function of the model.

A graph of a line

Description automatically generated with medium confidence

Figure 4.4 accuracy of the model

A graph of loss and loss

Description automatically generated

Figure 4.5 loss function of the model

## **Related Work**

A confusion matrix is a performance evaluation tool used in classification tasks to understand the performance of a machine learning model by showing the counts of true positive, true negative, false positive, and false negative predictions. It's a way to visualize the performance of a classification algorithm and gain insights into its accuracy and error types.

In a confusion matrix, each row represents the actual class labels, while each column represents the predicted class labels. The four main components of a confusion matrix are:

1. **True Positives (TP):** The cases where the model correctly predicted the positive class.
2. **True Negatives (TN):** The cases where the model correctly predicted the negative class.
3. **False Positives (FP):** The cases where the model predicted the positive class, but the actual class was negative (a Type I error).
4. **False Negatives (FN):** The cases where the model predicted the negative class, but the actual class was positive (a Type II error).

Using the values in the confusion matrix, we can calculate various performance metrics, such as:

* **Accuracy:** The proportion of correct predictions (TP and TN) out of all predictions. Accuracy = (TP + TN) / (TP + TN + FP + FN)
* **Precision (Positive Predictive Value):** The proportion of true positive predictions out of all positive predictions made by the model. Precision = TP / (TP + FP)
* **Recall (Sensitivity, True Positive Rate):** The proportion of true positive predictions out of all actual positive instances. Recall = TP / (TP + FN)
* **F1-Score:** The harmonic mean of precision and recall. It balances precision and recall when one is more important than the other. F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)
* **Specificity (True Negative Rate):** The proportion of true negative predictions out of all actual negative instances. Specificity = TN / (TN + FP)
* **False Positive Rate:** The proportion of false positive predictions out of all actual negative instances. False Positive Rate = FP / (FP + TN)

The result of the Precision of the model is shown in figure 4.6

**A screenshot of a computer program

Description automatically generated**

**Figure 4.6 model accuracy**

**Evaluation of the proposed solution**

Before we begin the evaluation of the proposed result, we consider the research work already done on this problem domain with respect to the proposed solution. We discuss this below:

**Bogdan et al**

Bogdan, Michał, Andrzej, Zbigniew, Bogusław, Paweł, Jakub, Bogusław (2019) conducted an investigation on Breast Cancer Classification on Histopathological Images Affected by Data Imbalance Using Active Learning and Convolutional Neural Network. They reported an output of the neural network on unlabeled samples and used it to calculate weighted information entropy. The output is utilized as uncertainty score for automatic selecting both samples with high and low conﬁdence. on the initial training pool consisting of only labeled data, the algorithm based on entropy achieved 86% accuracy, whereas algorithm based on weighted entropy achieved 83.9% accuracy. In seventh iteration the entropy-based algorithm learned on 3035 samples from PL pool and 3146 training samples from PH pool, and achieved on such training data 93.3% classiﬁcation accuracy, whereas weighted entropy-based algorithm learned on 3035 samples from PL pool and 3335 samples from PH pool and achieved 94.62% classiﬁcation accuracy. The result is presented in table 5 below:

|  |  |  |
| --- | --- | --- |
| **Run Time** | **Pool** | **Accuracy** |
| **Initial** | **PL** | **86%** |
| **Final** | **PH** | **83%** |
| **Initial** | **PL** | **93.3%** |
| **Final** | **PH** | **94.6%** |

**Table 5 accuracy results of Bogdan et al method**

**Result comparison**

We now compare the proposed solution with the work of Bogdan et al method. The result of the comparison is presented below:

**Table 1 Accuracy result comparison**

|  |  |  |
| --- | --- | --- |
| **MATRIC** | **PROPOSED** | **BOGDAN ET AL** |
| **ACCURACY** | **99.07%** | **93.97%** |

**Figure 4.7 the graph of the accuracy result comparison**

It can be appreciated from figure 4.7 above that the designed model achieved a better accuracy result on average when compared with that of Bogdan et al. the result may be influenced by a number of reasons which may include the dataset, the preprocessing method, and the model used for analysis. But I would also like to present the fact that the use of bidirectional long short-term memory to actively learn the dataset played an important role in the outcome of the results.

# **Conclusion**

Using a Bidirectional Long Short-Term Memory (LSTM) network for breast cancer prediction is an interesting application of deep learning. LSTMs are a type of recurrent neural network (RNN) that is particularly effective at capturing sequential patterns and dependencies in time series or sequential data. Bidirectional LSTMs take into account both past and future context when making predictions, making them well-suited for tasks involving temporal relationships.

Before starting the project, the breast cancer data was collected and preprocessed. The processing involved encoding categorical variables, normalize numerical variables, and handle missing values. The data was organized into sequences, where each sequence represents a patient's image data over a certain time period. The sequence length was decided, which determines the number of time steps to consider for each patient.

The next step was to create a Bidirectional LSTM model using a deep learning framework. The input to the model is the sequences of image data, and the output is a prediction of the breast cancer. Multiple LSTM was stacked for improved representation learning. Split the data into training, validation, and test sets. The Bidirectional LSTM model was trained using the training data and its performance using the validation data.

Different hyperparameters were experimented with such as the number of LSTM units, learning rate, batch size, etc., to optimize the model's performance. Dropout and batch normalization was used to prevent overfitting. The trained model was evaluated on the test dataset to assess its generalization performance. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were calculated to understand how well the model predicts breast cancer.

## **Reflection**

## Implementing a Bidirectional LSTM for breast cancer prediction requires both domain expertise and deep learning knowledge. Collaborating with healthcare professionals and experienced data scientists can help ensure the success of such projects.

Predicting breast cancer using Bidirectional LSTM models presents several challenges, some of which are specific to the medical domain and the characteristics of the data. Here are some key challenges I encountered:

1. **Data Availability and Quality:**
   * Medical datasets is limited in size due to privacy concerns and data scarcity.
   * Imbalanced datasets, where the number of positive cases (cancer) is much smaller than negative cases (non-cancer), can lead to biased models.
   * Noisy or incomplete data, missing values, and inconsistent annotations can affect model training and performance.
2. **Feature Engineering:**
   * Selecting relevant features from medical data is challenging due to the complex nature of medical images.
   * Deciding on the right temporal granularity and how to represent the breast cancer image as a sequence can impact model effectiveness.
3. **Temporal Relationships:**
   * Medical data often has complex temporal relationships. Capturing long-term dependencies requires proper sequence modeling techniques like LSTMs.
   * Determining the appropriate sequence length and time window for predictions can be difficult.
4. **Overfitting and Generalization:**
   * Overfitting can be a concern, especially with limited data. Applying techniques like regularization and cross-validation is crucial.
   * Generalizing the model to new breast cancer image populations or different hospital settings can be challenging due to variations in mammogram data collection practices.
5. **Interpretability and Explainability:**
   * Deep learning models like LSTMs are often considered black boxes. Explaining model decisions and predictions to medical professionals is important for trust and adoption.
   * Interpreting the features and patterns that the model learns from medical data can be complex.
6. **Ethical and Regulatory Considerations:**
   * Medical data is sensitive and subject to strict regulations. Ensuring compliance with data privacy and security laws is essential.
   * Addressing issues of bias and fairness in model predictions is critical to avoid perpetuating health disparities.
7. **Computational Resources:**
   * Bidirectional LSTMs can be computationally intensive and might require powerful hardware for training and inference, especially with large medical datasets.

Addressing these challenges requires a combination of technical expertise, domain knowledge, collaboration with healthcare professionals, and adherence to ethical and regulatory guidelines. Close collaboration between data scientists, medical experts, and stakeholders is key to developing accurate and reliable predictive models for breast cancer.

## Future Work

Predicting breast cancer using machine learning is an ongoing and dynamic field. As technology and methodologies evolve, there are several areas of future work that hold promise for improving the accuracy, interpretability, and clinical relevance of breast cancer prediction models:

1. **Advanced Deep Learning Architectures:**
   * Beyond traditional LSTMs, exploring other advanced architectures like Transformers or attention mechanisms could lead to more effective sequence modeling and improved performance.
2. **Multi-Modal Data Fusion:**
   * Integrating various types of data, such as imaging data (mammograms, MRIs), genomic data, and patient records, can provide a holistic view of a patient's health and improve predictive accuracy.
3. **Transfer Learning and Pretrained Models:**
   * Pretraining models on large medical datasets or related tasks can help with transfer learning, where the pretrained knowledge is fine-tuned for specific breast cancer prediction tasks.
4. **Uncertainty Estimation:**
   * Incorporating uncertainty estimation techniques can provide confidence intervals around model predictions, which is important for clinical decision-making.
5. **Interpretable AI:**
   * Developing techniques to make deep learning models more interpretable and explainable is crucial for building trust and adoption in clinical settings.
6. **Personalized Risk Assessment:**
   * Moving beyond binary predictions, models that can provide personalized risk scores for patients can guide more tailored interventions and screening schedules.

The future of breast cancer prediction using machine learning lies in the convergence of deep learning advancements, medical expertise, ethical considerations, and clinical validation. As these areas continue to develop, the potential for improved patient care and outcomes becomes more substantial.

# **References**

1. Adetokunbo A. A Adenowo, Basirat A. Adenowo Software Engineering Methodologies: A review of the Waterfall Model and object Oriented Approach International Journal of Scientific & Engineering Research, Volume 4 Issue 7, July 2013.
2. Ali Al Bataineh; A comparative Analysis of Nonlinear Machine Learning Algorithms for Breast cancer Detection International Journal of Machine Learning and Computing vol 9, No 3, June 2019.
3. Bennett. S, McRobbin S. & farmers Object Oriented Systems Analysis and Design Using UML Berchire Mc Graw Hill Education 2002
4. Etingo & Voropai, Most breast cancers are detected as a lump/mass on the breast, or through self examination or mammography 2006.
5. Hickey, A. M. & Davis, A. M (2003)Requirements Elicitation Technique Selection: Model of two Knowledge Intensive software Development Processes. In systems Sciences, proceeding of the 36th Annual Hawali International Conference, 6-9 Jan 2003
6. Judith Hurwitz and Daniel Kirsch. Machine learning for Dummies. John Wiley and Sons, Inc. 2018
7. Kononenko, I. “Machine Learning for Medical Diagnosis: History, state of the art and perspective”, Artificial Intelligence in Medicine, vol 23 no 1,pp 89-109, 2001.
8. Lee C. H., (2010), “Breast cancer screening with imaging: recommendations from the society of breast imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer,” J. Am. Coll. Radiol. 7(1), 18–27.10.1016/j.jacr.2009.09.022
9. M.karabatak, “A new classifier for breast cancer detection based on Naire Bayesian”, measurement, vol 72 , pp.32-36 2015.
10. Machine Learning and Applications in Breast Cancer Diagnosis and Prognosis Design 2018, 2, 13 www.mdpi.com/journal/designs
11. Machine learning with Applications in Breast Cancer Diagnosis and Prognosis 2013.
12. Mohenssen Mohammed, Muhammad badruddin Khan and Elihab Bashier Mohammed Bashier. Machine learning Algorithms and applications CRC Press Taylor & Francis Group Bocca Raton London New York. 2017.
13. Oliver Theobald. Machine Learning for Absolute Beginners Second Edition. 2017
14. Patrizia Ferroni, Fabio M. Zanzotto, Silva Riondino, Noemi, Scarpato, Fiorella Guadagni & Mario Roselli: Breast Cancer Prognosis Using a Machine Learning Approach Cancer 2019, 11, 328
15. Pfleeger, S. L & Atlee. J.M. (2006). Software Engineering Theory and Practice, 3rd Edition US: Prentice Hall.
16. Pressman, R.S (2005). Software Engineering: A Practitioner Approach, 6th Edition. Singapore: Mc Graw- Hill .
17. Saria Ettachi & Huda Kutrani Breast cancer diagnosis and prediction using Machine learning and Data Mining Techniques: Journal of Dental and Medical Science Volume 18, Issues 4 ser 20 (April 2019), pp 85-94wwwjosrjournals.org. A review 2019.
18. Siegel, Miller, and Jemal, (2016) “Cancer statistics, 2016,” CA Cancer J. Clin. 66(1), 7–30.10.3322/caac.21332
19. Wembin Yue, Zidong Wang, Hongwei Chen, Annette Payne and Xiaohui Liu
20. Weston AD & Hood. Systems biology proteomics and the future of health care: toward predictive preventive and personalised medicine. J proteome Res, 3;179-96 (2004)
21. World health Organisation, “WHO position paper on mammography Screening”, WHO Health Organisation, Geneva 2014.
22. Al-Masni, M. A., & Kim, D. H. (2021). CMM-Net: Contextual multi-scale multi-level network for efficient biomedical image segmentation. *Scientific reports*, *11*(1), 1-18.
23. El-Dahshan, E. S. A., Mohsen, H. M., Revett, K., & Salem, A. B. M. (2014). Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm. *Expert systems with Applications*, *41*(11), 5526-5545.
24. Koshki, A. S., Zekri, M., Ahmadzadeh, M. R., Sadri, S., & Mahmoudzadeh, E. (2020). Extending contour level set model for multi-class image segmentation with Application to Breast Thermography Images. *Infrared Physics & Technology*, *105*, 103174.
25. Heidari, M., Mirniaharikandehei, S., Danala, G., Qiu, Y., & Zheng, B. (2020, March). A new case-based CAD scheme using a hierarchical SSIM feature extraction method to classify between malignant and benign cases. In *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications* (Vol. 11318, pp. 309-315). SPIE.
26. Porwal, P., Pachade, S., Kadethankar, A., Joshi, A., Patwardhan, V., Kamble, R., ... & Meriaudeau, F. (2018, August). Automatic segmentation of optic disc by gradient minimization based approach. In*2018 International conference on intelligent and advanced system (ICIAS)* (pp. 1-5). IEEE.

# **Appendices**

A collage of cells

Description automatically generated

**Figure 7.1 sample image dataset**

A screenshot of a computer screen

Description automatically generated

Figure 7.2 sample folders in the dataset

A screenshot of a computer program

Description automatically generated

Figure 7.3 image folder merging function

A screenshot of a computer

Description automatically generated

Figure 7.4 new dataset.

A close-up of a microscope

Description automatically generated

Figure 7.5 resized image data.

A blurry image of a pink surface

Description automatically generatedA pink and purple pixelated square

Description automatically generated

Figure 7.6 gaussian and random noise filter application.

A graph with pink lines

Description automatically generated with medium confidence

Figure 7.7 distribution of IDC in the dataset.

A graph of patches

Description automatically generated

Figure 7.8 distribution of patches per patient.

A graph with blue and orange squares

Description automatically generated

Figure 7.9 count of IDC in dataset.

A group of maps of different countries/regions

Description automatically generated

Figure 7.10 patches distribution using binary objective visualization.

A screenshot of a computer

Description automatically generated

**Figure 7.11 the anaconda interface**

A screenshot of a computer code

Description automatically generated

Figure 7.12 dataset splitting

A table of numbers and a line

Description automatically generated with medium confidence

Figure 7.13 model training

A graph of a line

Description automatically generated with medium confidence

Figure 7.14 accuracy of the model

A graph of loss and loss

Description automatically generated

Figure 7.15 loss function of the model

**A screenshot of a computer program

Description automatically generated**

**Figure 7.16 model accuracy**

**Figure 7.17 the graph of the accuracy result comparison**

A screenshot of a phone

Description automatically generated

**Figure 7.18 teamwork project management tool.**