BREAST CANCER DETECTION USING DEEP LEARNING

A THESIS SUBMITTED TO
THE FACULTY OF ARCHITECTURE AND ENGINEERINGOF
EPOKA UNIVERSITY

BY

LAURA MUÇARAKU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTSFOR
THE DEGREE OF MASTER OF SCIENCEIN
COMPUTER ENGINEERING

JUNE, 2024

Approval sheet of Thesis

This is to certify that we have read this thesis entitled **“Breast Cancer Detection using DeepLearning”** and that in our opinion it is fully adequate, in scope and quality, as a thesis for thedegree of Master of Science.

____

Assoc. Prof. Dr. Arban Uka
Head of Department
June, 27, 2024

Examining Committee Members:

Assoc. Prof. Dr. Dimitrios Karras	(Computer Engineering)	__________________
Prof. Dr. Gëzim Karapici	(Computer Engineering)	__________________
Dr. Florenc Skuka	(Computer Engineering)	__________________

I hereby declare that all information in this document has been obtained and presentedin accordance with academic rules and ethical conduct. I also declare that, as requiredby these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Surname: Laura Muçaraku

Signature: _______________

ABSTRACT

BREAST CANCER DETECTION USING DEEP LEARNING

Muçaraku, Laura
M.Sc., Department of Computer Engineering
Supervisor: Dr. Florenc Skuka

In nearly 95% of the countries worldwide, breast cancer is the main reason of female deaths. The impact that this disease has on human body, depends on the stage in whichit is diagnosed, being a life-taking disease if not diagnosed in time. This Thesis makes an analysis on both traditional and revolutionary methods used for Breast Cancer Detection andClassification, and proposes the best model for different scenarios, based on the availabilityof data, human expertise, and time limitations. Available datasets that contain samples of Breast Cancer cells are also analyzed, and all the sources are collected and provided. The methods analyzed are classified into three main categories: Supervised, Unsupervised, and CNN methods. Four methods are analyzed and tested with Breast Cancer Wisconsin Diagnostic (WDBC) dataset from the first category: Random Forest, K-Nearest Neighbor, NaiveBayes, and Support Vector Machine. From the Unsupervised Learning Methods, are analyzed and tested with the same dataset: Auto-encoders, and Self-Organizing Maps. Two CNN models, UNet and ResNet are also built and tested with Breast Ultrasound Images Dataset. Each method is tested several times with different parameter values, with the aim of finding the combination of parameters that generates the best results for the available datasets. From the Supervised Methods Support Vector Machine achieved the highest ac- curacy of 99%. Auto Encoder won against the SOM as a Unsupervised Method with an accuracy of 98%, and within the CNN methods, UNet performed better with an accuracy of 97.44%.

Keywords: *Breast Cancer, Deep Learning, Model Comparison, Evaluation Metrics*

ABSTRAKT

DETEKTIMI I KANCERIT TE GJIRIT DUKE P ËRDORUR DEEP LEARNING

Muçaraku, Laura
Master Shkencor, Departamenti i Inxhinierisë Kompjuterike
Udhëheqësi: Dr. Florenc Skuka

Kanceri i gjirit është arsyeja kryesore e humbjes së jetës për gratë në rreth 95% të vendeve në mbarë botën. Impakti që kjo sëmundje ka në trupin e njeriut varet ngusht-ësisht nga shkalla e sëmundjes në momentin e diagnostifikimit. Si rrjedhojë, gjetja e metodave për identifikimin e shpejtë dhe të saktë të kësaj sëmundjeje është esenciale. Kjo Tezë do të bëjë një analizë të thellë të metodave që lidhen me klasifikimin e qelizave kanceroze si malinje apo beninje. Metodat e analizuara në këtë tezë do të ndahen në tre kategori: Metodat e Mësimit të Mbikëqyrur, Metodat e Mësimit të Pambikëqyrur dhe Modelet e Rrjetave Neurale Konvolucionale. Modelet Random Forest, K-Nearest Neighbor, Na¨ıve Bayes dhe Support Vector Machine do të testohen nga kategoria e parë duke përdorur datasetin Breast Cancer Wisconsin Diagnostic (WDBC). Me të njëjtin dataset do të testohen nga kategoria e dytë Auto-Encoders dhe Self Organizing Maps. Dy modelet CNN: UNet dhe ResNet do të testohen duke përdorur datasetin Breast Ultrasound Images Dataset. Cdo metodë do të testohet disa herë me kombinime të ndryshme të parametrave që pranon, për të gjetur parametrat me të cilët metoda performon më mirë për klasifikimin e saktë të qelizave kanceroze. Nga Metodat e Mësimit të Mbikëqyrur, metoda SVM arriti saktësinë më të lartë prej 99%. Metoda Auto Encoder fitoi përballë metodës SOM me një saktësi prej 98%. Midis dy modeleve CNN të testuar, saktësia më e lartë u arrit prej modelit UNet, me vlere 97.44%.

Fjalë kyçe: *Kanceri i Gjirit, Modele të Mbikëqyrura, Modele të Pambikëqyrura, Modelet e Rrjetave Neurale Konvolucionale, Metrika Vlerësimi*

I dedicate this Thesis to my family, as a special thanks for their endless support and encouragement

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and appreciation to my advisor, Dr.Florenc Skuka, for his constant support, direction, and inspiration during my research. ThisThesis was completed due to his deep expertise and insights.

I would also like to thank all the professors of Department of Computer Engineering at EPOKA University, who have shared their knowledge and expertise with us during these twoyears, and have contributed to building a robust academic background.

Lastly, I would like to thank my family and friends for always believing and supportingme.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction to Breast Cancer

Breast cancer is a disease that occurs to people when the growth of breast cells is abnormaland they form tumors [2]. This type of cancer is responsible for the death of 670’000 womenworldwide in 2022. Based on the World Health Organization, in average there are 2.3 million cases of breast cancer diagnosed yearly, making this type of cancer the most common canceramong adults. In majority of countries worldwide, nearly 95% of the countries, breast canceris the main reason of female deaths.

Breast cancer, same as all other deadly diseases, has huge impact on the next generationstoo. Based on a study conducted on 2020 by the International Agency for Research on Cancer, nearly 1 million children were orphaned because of the death of their mothers by this disease. These children, who have experienced the lost of their parent because of breast cancer, are more likely to experience health, educational and financial disadvantages throughouttheir lives.

1.2 Problem Statement

In 2023 the National Breast Cancer Inc has released some interesting, yet terrifying facts about breast cancer. Based on it, every 2 minutes a woman in United States is diagnosed with breast cancer. Upon diagnosis, the stage of an individual’s breast cancer is determinedto assess its extent and whether it has metastasized beyond the breast. In this point it is important to point out the significance of early detection, as it is easier to cure the diseaseif it is diagnosed in early stage and if it is localized only in the breast part of the body.

Otherwise, if the cancer has been spread in other parts of the body as well, the survival chances reduce significantly. This is where Deep Learning algorithms and their efficiency come to help. Table 1.1 [3] presents the likelihood of the patients being alive

5 years after cancer detection, grouped on the type and stage of the cancer at diagnosis. This estimation is known as 5-year relative survival rate.

As it can be seen from Table 1.1, the highest probability of surviving from breast cancer, 99%, is if it is diagnosed in an early stage. That is why it is important for researchers to find new and more sophisticated methods for cancer detection and classification, which result toreliable and fast diagnosis.

Table 1. 1 5-year Relative Survival Rate for Breast Cancer Patients

Breast Cancer (SEER) Stage	**5-Year Relative Survival Rate**
Localized (invasive cancer has not spread outside of the breast)	99%
Regional (cancer has spread outside of the breast to nearby structuresor lymph nodes)	86%
Distant (cancer has spread to other parts of the body, such as lungs,liver, or bones)	30%
All SEER stages combined	91%

1.3 Objectives of the Thesis

This Thesis aims to analyze traditional and recent Deep Learning models that are used to diagnose Breast Cancer cells of female patients and propose the best model based on the specific situation: available dataset, and available human expertise. The thesis proposes thebest Deep Learning model to accommodate these scenarios:

- If there is enough human expertise as to label the available dataset, and represent itinto meaningful numerical values.
- If the available dataset is numeric, and there is no possible human intervention to label it into the desired class labels.
- If the available dataset consists of complex images, where feature extraction is complex and needs to be automated

The results generated by these Deep Learning models can be used to assist doctors ontheir daily jobs to save the lives of humans, so the accuracy of these algorithms is critical.

CHAPTER 2

WHAT IS BREAST CANCER

Breast cancer is a disease that occurs to people when the growth of breast cells is abnormaland they form tumors [2].

Breast cancer cells have their origin inside the milk ducts or inside of the milk-producinglobules of the breast. In the earliest stage of the Breast Cancer, which is known as situ, this disease is not considered to be life-threatening. But, as cancer cells infiltrate surrounding breast tissue, they create tumors, resulting in the formation of lumps or thickening. This is the reason why it is extremely important to diagnose this disease in its earliest stage, as the probabilities of curing it are much higher than diagnosing it in the latest stages.

Breast cancers that are invasive are able to spread to adjacent lymph nodes or other organs. This process is known as metastasis, which when caught in a late stage can be life- threatening.

Treatment strategies vary from one person to another, depending on the type of cancer, the diagnosis stage, and the extend of its spread. A combination of surgery, radiation therapy,and medications forms the basis of treatment.

2.1 Breast Cancer Stages

The stage in which a patient finds himself diagnosed with Breast Cancer, can be expressed as a number in the scale of 0 to 4. Stage 0 refers to non-invasive cancer that is not spread outside of its original location. Stage 4 refers to invasive cancers that have spread outside oftheir original location, breast, and are present in other parts of the body too [4].

Stage numbers are calculated based on three characteristics of cancerous cells:

1. **T:** The cancer tumor’s size and whether or not it has spread to surrounding tissues

2. **N:** Whether there is cancer in the lymph nodes or not
3. **M:** Whether the disease has progressed to other organs outside of the breast

2.2 The highest risk of being diagnosed with Breast Cancer

Breast cancer is a disease which typically occurs in females, rather than males. Based on theWorld Health Organization, only 0.5% up to 1% of breast cancer cases are intended to occurin men. Nevertheless, if a man is diagnosed to have breast cancer, he should follow the sameprocess as a woman having the same disease for curing it.

Other risk factors more than gender that contribute to Breast Cancer disease are age, obesity, family history of breast cancer, alcohol, reproductive history, and postmenopausal hormonal therapy. The risk of being diagnosed with Breast Cancer increases by the increaseof age or by the increase of body weight more than normal. Also being abusive with alcoholhas the same effect. But even though all these factors have an impact in the development ofthis disease, in half of the cases worldwide, the patients do not have any risk factor present,other than the gender (female) and the age (over 40 years old). This is why it is very important for all women worldwide to be informed about the presence of this disease and to do periodic controls of their body. BRCA1, BRCA2, and PALB-2 gene mutations are the mostprevalent inherited high penetrance gene variants that substantially increase the risk of breastcancer. Should a woman have mutations found in these primary genes, she might wish to consider risk reduction strategies such as having both breasts surgically removed [2].

2.3 Signs and Symptoms

The symptoms of the Breast Cancer become more understandable when the cancer stage increases. In the beginning of the disease, so in the early stage of the cancer, most people donot experience any symptom. Some of the most known symptoms of Breast Cancer include:

- Breast enlargement or lump, often painless.
- Changes in the dimensions, shape, or appearance of the breast.
- Skin changes such as dimpling or redness.

- Alterations in the nipple’s or the surrounding skin’s (areola) look.
- Strange or bloody nipple outflow.

By the time passing, cancerous cells may spread all over the body and reach other organs,which may include brain, bones, lungs, liver, etc. If this is the case, new symptoms arise, other than those mentioned above. Some of the new symptoms may include bone pain or headaches.

2.4 Treatment

Doctors dealing with patients with Breast Cancer implement a combination of treatments to cure the patients and to reduce the risk of cancer recurrence. These treatments include:

- Surgical procedures to eliminate the breast tumor.
- Radiation therapy aimed at reducing the risk of recurrence in the breast and adjacenttissues.
- Medications designed to kill cancerous cells and prevent their spread. These medica- tions include hormonal therapies, targeted biological therapies or chemotherapy.

The earlier a patient begins the treatment, the higher are the chances of the treatment being more effective.

Patients receiving medicinal treatment for breast cancer may receive it either before surgery (referred to as ”neoadjuvant”) or after surgery (referred to as ”adjuvant”), dependingon the biological subtyping of the tumors. For malignancies that express the estrogen receptor (ER) and/or progesterone receptor (PR), endocrine (hormone) therapy, such as tamoxifen or aromatase inhibitors, is probably beneficial. These oral drugs virtually totally remove thechance of these ”hormone-positive” cancers coming back in the future; they should be takenfor five to ten years. Endocrine therapies can cause menopause symptoms, albeit they are typically well tolerated [2].

Chemotherapy is also necessary for ”hormone receptor negative” tumors, which, unlessthey are very tiny, do not express the ER or the PR. In most cases, hospitalization is not necessary for breast cancer chemotherapy patients unless there are complications [2].

CHAPTER 3

LITERATURE REVIEW

This chapter reviews some of the most recently published studies related to breast cancer detection. The methods used in each study will be analyzed, and the results of each methodwill be compared to each other to find out the advantages of each method used.

3.1 Methodology Overview

This section makes a brief analyzes about common methodologies and techniques used in Literature for Breast Cancer Detection using Deep Learning. Methods used for this purpose can be categorized based on different characteristics and features. Sharma, Shubham and Aggarwal, Archit [5] have made a separation of Deep Learning algorithms into three categories, and toward this Thesis we will proceed with that classification:

- **Supervised Learning**: where the data is labeled and it is known beforehand. Supervised Learning techniques generate a function predicting outputs based on input observations.
- **Unsupervised Learning**: where the data is unlabeled and it is differentiated based onsome characteristics or common features. The algorithm then should act based on thisinformation, without external guidance.
- **Convolutional Neural Networks**: These biologically inspired computer models can achieve much higher performance than previous AI iterations on popular machine learning tasks. Because of their precise yet simple architecture, CNNs are mostly used to deal with difficult image-driven pattern recognition problems and provide a more straightforward way to start working with ANNs [6].

Figure 3.1 shows graphically the categorization of methodologies in this Thesis.

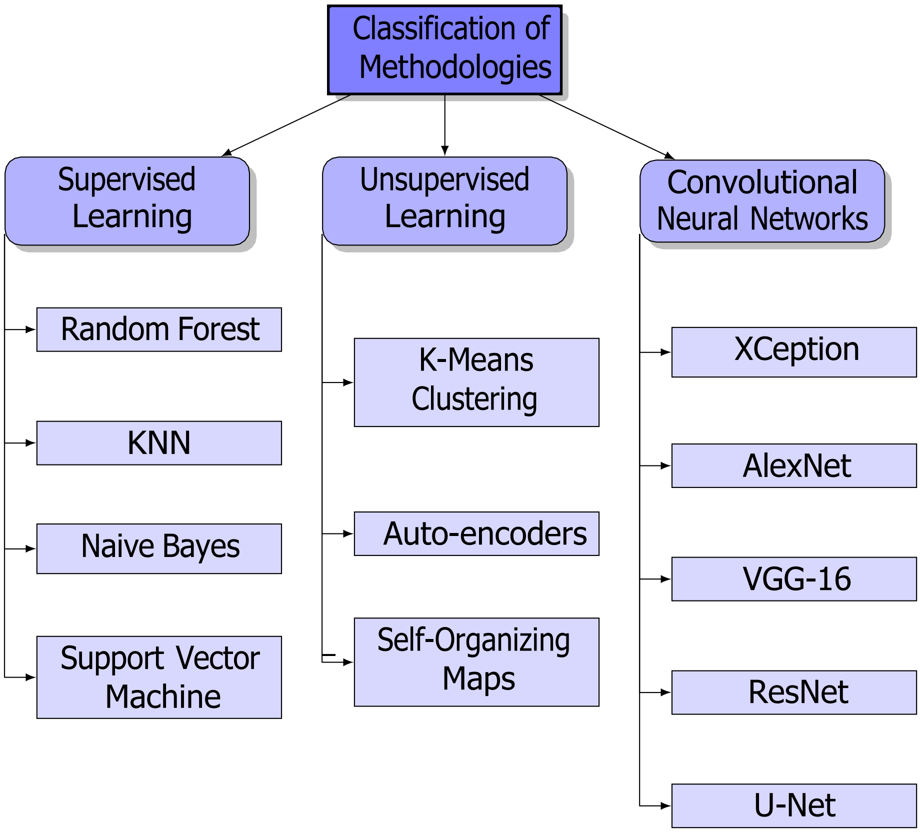


Figure 3. 1 Classification of Methodologies for Breast Cancer Detection into three categories: Supervised, Unsupervised, and Convolutional Neural Networks

3.2 Supervised Learning Techniques used for Breast Cancer Detection

Among supervised algorithms used for Breast Cancer Detection, this Thesis analyzes Random Forest, K-Nearest-Neighbor (KNN), Na¨ıve Bayes, and Support Vector Machines (SVM).

3.2.1 Random Forest

Random Forest method is based on the ground technique of recursion. The training of this model is done by building a a large number of decision trees, and then the model combinesthe predictions of these decision trees for more reliable and accurate results. In each instanceof the iterations, a random sample size N is selected from the data-set.

The creator of Random Forest model, Leo Breiman, created it to solve over-fitting and reduce the variance in Decision Trees. This approach was novel because it merged the outputof training several models into one, more potent learning model and applied the statistical technique of bootstrapping for the first time [7].

The random forest algorithm has been extremely successful with impressive results in both classification and regression tasks [8]. The algorithm of the Random Forest is given inAlgorithm 1, and its general architecture is shown graphically in Figure 3.2.

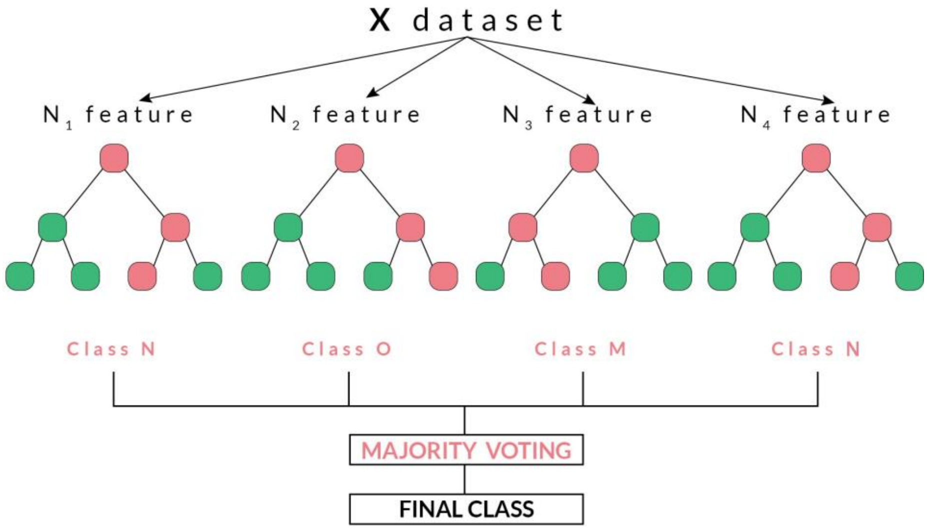


Figure 3. 2 Random Forest Architecture

Random Forest is a useful supervised learning technique which is proven to perform well, both in efficiency and effectiveness in the task of Breast Cancer Detection. Sharma, Shubham and Aggarwal, Archit and Choudhury, Tanupriya have tested this model with Wisconsin dataset, which contains in total 569 instances and 32 features for each instance. They have used 70% of the dataset for training, which equals to 389 instances, and 30% of the dataset for testing, which equals 171 instances. They have published the results in [5] and the model hasachieved the accuracy of 94.74%. Out of 171 predictions, the total number of correct Benign predictions was 103, the number of correct Malignant predictions was 59, the number of benign instances that were misclassified as Malignant was 5, and the number of Malignant instances that were misclassified as Benign was 4.

Algorithm 1 Random Forest Algorithm

Input: Training set D_n, number of trees $M > 0$, $a_n \in \{1, \dots, n\}$, $m_{\text{try}} \in \{1, \dots, p\}$, nodesize $\in \{1, \dots, a_n\}$, and $x \in X$.
Output: Prediction of the random forest at x.
for $j = 1, \dots, M$ **do**
  Select a_n points, with (or without) replacement, uniformly in D_n. In the followingsteps, only these a_n observations are used.
  Set $P = (X)$, the list containing the cell associated with the root of the tree.
  Set $P_{\text{final}} = \emptyset$, an empty list.
  while $P = \emptyset$ **do**
    Let A be the first element of P.
    if A contains less than nodesize points or if all $\mathbf{X_i} \in A$ are equal **then**
      Remove the cell A from the list P.
      $P_{\text{final}} \leftarrow$ Concatenate(P_{final}, A).
    else
      Select uniformly, without replacement, a subset $M_{\text{try}} \in \{1, \dots, p\}$ of cardi-nality m_{try}.
      Select the best split in A by optimizing the CART-split criterion along thecoordinates in M_{try} (see text for details).
      Cut the cell A according to the best split. Call A_L and A_R the two resultingcells.
      Remove the cell A from the list P.
      $P \leftarrow$ Concatenate(P, A_L, A_R).
    end if
  end while
  Compute the predicted value $m_n(\mathbf{x}; \Theta_j, D_n)$ at $\mathbf{x}$ equal to the average of the Y_i fallingin the cell of $\mathbf{x}$ in partition P_{final}.
end for
Compute the random forest estimate $m_{M,n}(x; \Theta_1, \dots, \Theta_M, D_n)$ at the query point x.

3.2.2 K-nearest-neighbor

K-nearest-neighbor is the second supervised learning method that has been applied to the detection of breast cancer. It is a technique that has proven to be successful in both classification and regression tasks. KNN was first invented by Evelyn Fix and Joseph Hodges in 1951, and was then improved later by Thomas Cover [9].

KNN is one of the data mining strategies that is ranked in the top 10 for data mining[10].One way to define KNN algorithm is as an algorithm that uses the data sets nearby to decidewhere a given data set belongs [5]. Each data point in the training set has both featuresand a labeled output, and the algorithm uses this information to make predictions for new instances. This technique is mostly used for regression and classification. Its architecture isshown in Figure 3.3.

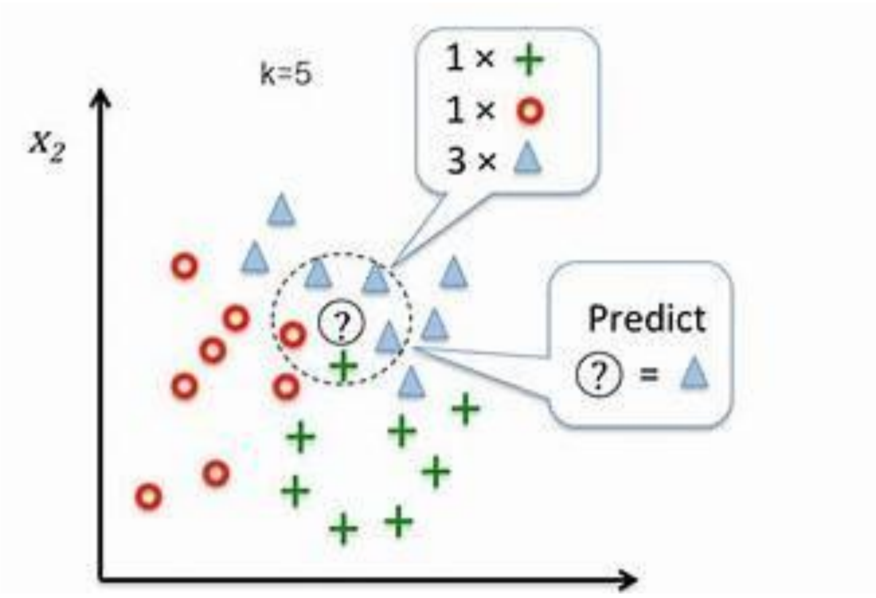


Figure 3. 3 K-Nearest Neighbor Architecture

Algorithm 2 presents the general algorithm of the KNN:

Algorithm 2 K-Nearest Neighbors Algorithm

for all the unknown samples *UnSample*(i) **do**
  for all the known samples *Sample*(j) **do**
    compute the distance between *UnSamples*(i) and *Sample*(j)
  end for
  find the k smallest distances
  locate the corresponding samples *Sample*(j_1), . . . , *Sample*(j_k)
  assign *UnSample*(i) to the class which appears more frequently
end for

The accuracy of the KNN algorithm in Breast Cancer Detection is calculated to be 95.9%[5]. The test of this algorithm was done with Wisconsin dataset, by using 70% of the datasetfor training, which equals to 389 instances, and 30% of the dataset for testing, which equals171 instances. Out of 171 predictions, the total number of correct Benign predictions was 107, the number of correct Malignant predictions was 57, the number of benign instances that were misclassified as Malignant was 1, and the number of Malignant instances that weremisclassified as Benign was 6.

3.2.3 Na¨ıve Bayes

The third supervised learning technique that can be used for Breast Cancer Detection is Na¨ıve Bayes. This technique is based on Bayes' theorem, and it is a probabilistic machine learningalgorithm, mainly used for classification tasks. This model tries to find the probability that an event will occur, given that another event has already occurred. This can be expressed mathematically with Equation 3.1:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{3.1}$$

Naıve Bayes classifier makes simplifying assumptions, as indicated by the name "Naive." Considering the class label, the classifier makes the assumption that the features used to de-scribe an observation are conditionally independent. The name "Bayes" honors the ReverendThomas Bayes, a theologian and statistician from the 18th century who developed the Bayestheorem [11]. The advantages of this algorithm include that it is both effective and efficientin practice. This mainly because it is a very easy algorithm to implement, and it can also bescaled with every dataset available. The Naive Bayes Architecture is shown graphically in Figure 3.4.

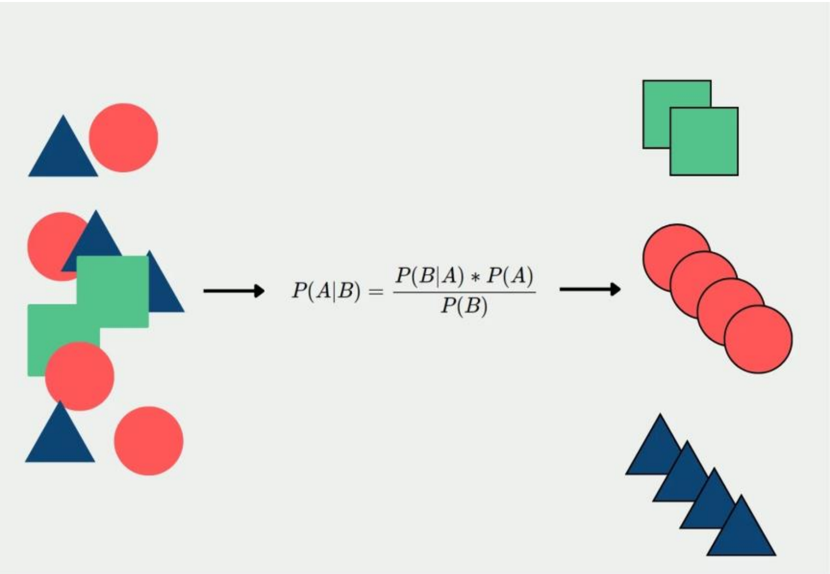


Figure 3. 4 Naive Bayes Architecture

Naive Bayes has been compared with KNN and Random Forest for Breast Cancer detection [5], and in terms of accuracy, Naive Bayes is the supervised learning model with thelowest result, being 94.47%. The authors have worked with Wisconsin dataset, by using 70% of the dataset for training, which equals to 389 instances, and 30% of the dataset for testing,which equals 171 instances. Out of 171 predictions, the total number of correct Benign pre-dictions was 101, the number of correct Malignant predictions was 54, the number of benigninstances that were misclassified as Malignant was 7, and the number of Malignant instancesthat were misclassified as Benign was 9.

3.2.4 Support Vector Machine

Support Vector Machine is a Deep Learning model that is mostly used for classification andregression tasks. The main reason why Support Vector Machine is widely used is because it is easy to use, has high accuracy in both classification and regression tasks, and requires less computational power. Since SVM is a Supervised Learning model, it works by taking labeled data as input and then classifies each instance of the input into one of the target classes (output). The way how it does this, is by finding a hyperplane that separates an N- dimensional space into enough sub-spaces such that each instance of the labeled input is setto one sub-space, and all the instances of the same class belong to the same sub-space. If the plane is two-dimensional, the hyperplane is a simple line which splits the plane into twoparts. Each class of the dataset lies on either side of the line [12]. Finding the ideal boundaryto divide the data into distinct classes is the SVM's goal while handling classification challenges. When choosing the boundary, it takes into consideration the distance between this boundary and the data points from each class that are closer to it. This distance is known asmargin, and the closest data points are called support vectors. This architecture is also shownin Figure 3.5, which is retrieved from [13].

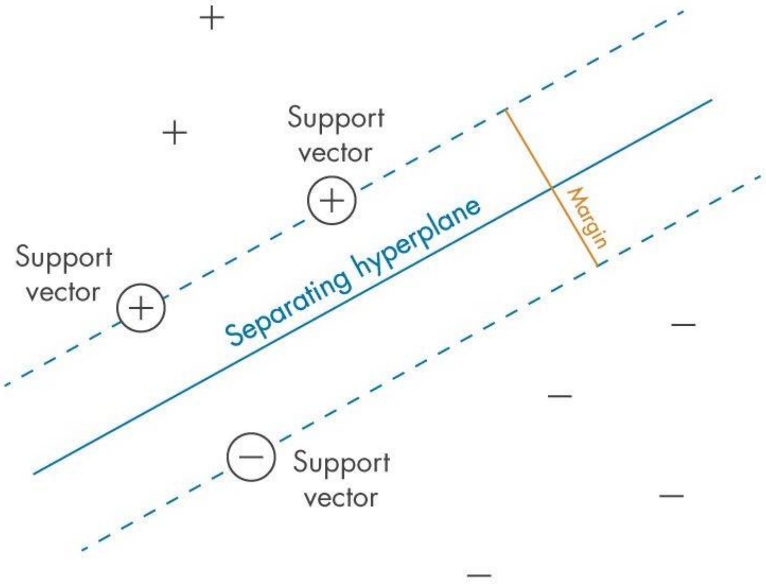


Figure 3. 5 Support Vector Machine Architecture

The Pseudo code of the Support Vector Machine is shown in Algorithm 3, where the input of the model is a training dataset, and the generated output are two parameters of the hyperplane: weights and bias.

Algorithm 3 Support Vector Machine Algorithm

```
Initialize w = 0 (or a random vector) and b = 0
Choose a learning rate α and regularization parameter λ
Repeat until convergence:
 for each training example (x_i, y_i) in D do
     Compute the margin: margin = y_i · (w · x_i + b)
     if margin < 1 then
         Update w: w = w − α · (w · λ − y_i x_i)
         Update b: b = b + α · λ · y_i
     else
         Update w: w = w − α · w
     end if
end for
Check convergence criteria (e.g., change in w or b is small)
Output: Hyperplane parameters w and b
```

The most important parameter that is required by the SVM algorithm is the kernel, whichmay take four different values: linear, polynomial, Gaussian RBF, and sigmoid. The choiceof the kernel type depends on data distribution. Linear and polynomial kernels perform wellon datasets in which the data is linearly separable. Sigmoid kernel does not perform as well as linear and polynomial kernel in linearly separable data, but it gives better results in nonlinear data. Yet, even in nonlinear data Gaussian RBF kernel generally produces better results than sigmoid kernel. The Gaussian RBF kernel is actually a universal kernel functionwhich has a very good performance in all datasets, despite their distributions [14].

Chen, Mingqi and Jia, Yinshan have tested the model for Breast Cancer Detection, and have compared its performance with four different kernel functions [14]. They have worked with Wisconsin dataset with a distribution of 70% of the available data for training, and 30% of the available data for testing. Their results show that the kernel function with the lowest accuracy for this dataset is sigmoid kernel with an accuracy of 95.32%, followedby polynomial kernel with an accuracy of 96.95%, and linear kernel with an accuracy of 97.66%. The winning kernel function is RBF which scored an accuracy of 98.25%.

3.2.5 A comparison of Supervised Learning Techniques for Breast Cancer Detection

Table 3.1 shows a comparison between the four above-mentioned Supervised Learning models. In terms of time complexity, KNN is the fastest algorithm. The time complexity at test time for kNN is O(1) without data preprocessing. In case of Na¨ıve Bayes and SVM, the time complexity depends on the number of training sets and the dimensions of the data, so the number of attributes that each instance of the data has. In Table 3.1, N stands for the numberof training examples, and d stands for the number of the features. Variable K in the RandomForest algorithm represents the number of variables randomly drawn at each node. In termsof the type of the problems where each algorithm performs best, Na¨ıve Bayes is the only Supervised Learning model that exclusively addresses classification problems. kNN, RandomForest, and SVM on the other hand can handle both classification and regression problems.The models can also be compared based on the type of predictions they make. Algorithms that simplify the function to a known form, make strong hypotheses about distribution of the data, and have a fixed number of parameters are referred to as Parametric Deep Learning algorithms. Other algorithms that do not make hypotheses about the distribution of the data,and do not have a fixed number of parameters are known as Non-Parametric Deep Learningalgorithms. Na¨ıve Bayes and Support Vector Machine algorithms can be expressed as both a parametric and non-parametric model. KNN and Random Forest algorithms on the otherhand are Non-Parametric models.

Table 3. 1 Comparison of KNN, Random Forest, Naive Bayes, and SVM

	KNN	Na¨ıve Bayes	Random For-est	SVM
Time Complexity (Training Phase)	O(I)	O(Nd)	O(MKNlog2N)	O(Nd)
Problem Type	Classification and Regression	Classification	Classification and Regression	Classification and Regression
Model Parame-ter	Non Paramet-ric	Parametric/Non Parametric	Non Paramet-ric	Parametric/Non Parametric

3.3 Unsupervised Learning Techniques used for Breast Cancer Detection

Unsupervised Learning Techniques are considered those techniques in which the model should try to find patterns in an unlabeled dataset and with little human oversight [15]. Thesetype of machine learning algorithms are useful in cases where labeled data is impossible to be found, or when the human expertise is unable to label the data at hand. There are three tasks where the Unsupervised Learning Models find usage: **Clustering**, **Association**, and **Dimensionality Reduction**. Clustering is a method of unsupervised learning that explores the features of the dataset with the aim of finding similarities and differences between the features. Then, the instances that have the highest similarity between features are grouped together and are said to belong to the same class. Figure 3.6 is a visual representation of howClustering in Unsupervised Learning Algorithms work.

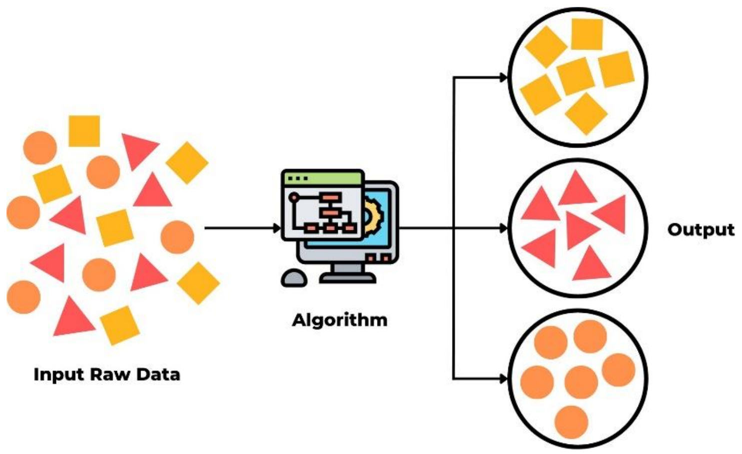


Figure 3. 6 Clustering in Unsupervised Learning

Another task that is performed by Unsupervised Learning models is Association. It triesto find associations or relationships between a group of items in the dataset. The purpose ofthese models is to find a set of combinations which occur together more often than would be expected by chance. The third task of Unsupervised Learning models is Dimensionality Reduction, which is responsible for reducing the dimensions/features in the dataset, withoutlosing information. This technique is very useful when the available datasets are large anddifficult to interpret. By reducing the less important dimensions of such dataset, it becomes more easy to visualise it, analyse and interpret.

Throughout the years, unsupervised learning techniques have been investigated and tested with the aim of detecting Breast Cancer cells as benign and malignant instances. Actually, having labeled data especially in medicine fields is quite difficult and also expensive, and in most cases researches need to work with unlabeled data. This type of data is easier tobe found, and does not require human intervention to label it, which is also error-prone. Nevertheless, in order to make use of these unlabeled datasets, unsupervised dimension re- duction algorithms need to be used. Three unsupervised learning models that are analyzed and compared in this section are: K-Means Clustering, Auto-encoders, and Self-OrganizingMaps.

3.3.1 K-Means Clustering

K-Means Clustering is one of the classical Unsupervised Deep Learning models. As thename suggests, it divides the available data into K clusters, by grouping together instances of the data with similar features, and putting in different clusters instances of the data with different features. It works with unlabeled data, and tries to find meaning within the input data,by examining the input data characteristics' and by trying to find meaning and relationshipsbetween these characteristics.

To understand better how K-means work, we have used an image retrieved from [16] and shown in Figure 3.7 .

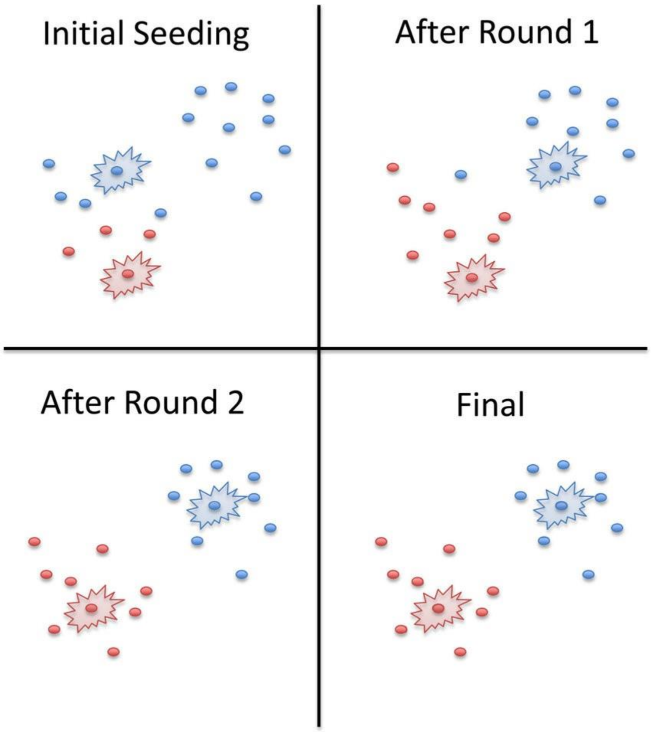


Figure 3. 7 K Means Cluster Algorithm

K-means is a partitioning clustering algorithm, which works by dividing the dataset intomutually exclusive groups [17]. Mutually exclusive means non overlapping groups. It first receives the unlabeled input data, and then requires the user to initialize K clusters. The fact that the method depends on human intervention for the initialization of the number or clusters is considered as one of the main disadvantages of the method by Radha, R., and Rajendiran, P. [18]. After the number of clusters has been initialized, the method assigns one centroid to each cluster. The data point in each cluster's center is known as the centroid.After initializing the number of clusters and the clusters' centroids, it then repeatedly assignseach data point to the nearest available centroid by finding the minimum distance. This is done by utilizing a distance function, where the most popular one is the Euclidean distancefunction [17]. The Euclidean distance equation is shown in Equation 3.2 .

$$d(x_i - y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (3.2)$$

Another Distance Function that can be used to measure the distance between each pointin the dataset from the centroids of the clusters is the Manhattan's distance. It's equation is shown in Equation 3.3.

$$d(x_i - y_i) = \sum_{i=1}^{n} |x_i - y_i| \tag{3.3}$$

The parameters of the K-Means Clustering algorithm are as explained below:

- **Nr. of clusters:** This parameter specifies the number of clusters in which the model will separate the data
- **Nr. of initial attempts:** Specifies the number of times the model will initialize its centroids.
- **Maximum nr. of iterations:** Specifies the maximum possible number of iterations before the model reaches convergence
- **Verbose:** This parameter indicates whether the model will communicate with the user during training for providing information or not

Bichen Zheng, Sang Won Yoon and Sarah S. Lam [19] have implemented the K-Meansmodel with the Wisconsin dataset and they discovered the number of optimal clusters to bethree. After reducing the dimensions of the initial dataset of 30 features into only 6 features,they implemented the SVM classifier with sigmoid kernel to test the accuracy of the the model with the new dataset. The accuracy of the model was calculated to be 97.38%. This high accuracy indicates that the K-means algorithm has a high performance, even though itreduced the number of initial features from 30 to 6.

3.3.2 Auto-encoders

An Auto-encoder neural network is an unsupervised learning model that tries to convert inputs into outputs with the least amount of distortion as one of the many deep learning techniques [20]. It comprises two fundamental components:

1. an encoder that transforms the n-dimensional input space into an m-dimensional hid-den space (lower dimensional representation)

2. a decoder that endeavors to reconstruct the initial input space from this hidden space.

Since Auto Encoders are unsupervised learning algorithms that are used for feature ex- traction and dimensionality reduction, the number of dimensions in the hidden space is lowerthan that of the input space, signifying that the hidden state encapsulates a low-dimensional,yet meaningful representation of the input data. The Architecture of a simple Auto Encodercan be seen in Figure 3.8.

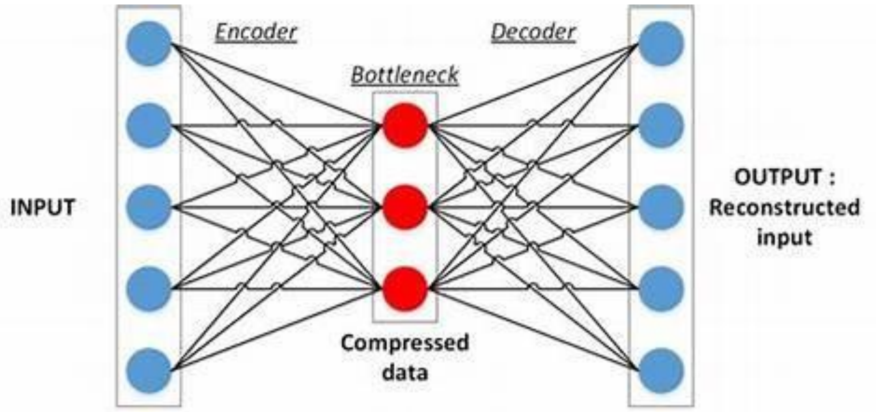


Figure 3. 8 Auto Encoders Architecture

Auto Encoders are used in the problem of Breast Cancer Detection as algorithms that make feature extraction from the available unlabeled datasets.

An auto encoder algorithm can be used for several goals, such as: dimensionality reduction and feature extraction, anomaly detection, image denoising, generative models, etc [21].

Yawen Xiao [20] developed a deep stacked auto-encoder (SAE) model, which is a com-bination of multiple layers of auto-encoders. In this model, each layer receives as input the output generated by its proceeding layer. Yawen Xiao utilized this improved model of Auto-Encoders for extracting the most significant data from Wisconsin dataset, and then used thisnew dataset with reduced dimensions as input to a SVM model to classify the samples as either benign or malignant. The reconstruction error of the features was calculated and based on it, the reconstruction error was minimized when using only 15 features from the initial dataset of 30 features. This new dataset with only 15 features was given to the SVM classifier with linear kernel as input, and the accuracy of the hybrid model was 98.25%.

3.3.3 Self-Organizing Maps

Self-organizing maps (SOMs) represent a category of unsupervised learning models extensively employed in machine learning literature for tasks such as data visualization, nonlineardimensionality reduction, and clustering. SOM is an unsupervised learning model that usesan artificial neural network to map a high-dimensional space into a lower-dimensional one [22]. Words are the weights of the neural network that discover the mappings between the high- and low-dimensional regions. The competitive learning paradigm and the SOM modelare related. Usually, the resulting map is a 2D lattice of neurons, which is obtained by non-linearly mapping high-dimensional input instances into a 2D surface. Among other things, SOM mapping is notable for its ability to maintain the topological properties of the input space, which guarantees that neurons in close proximity are allocated to instances that sharesimilarity in the high-dimensional input space. This unique property of SOMs leads to the formation of clusters representing similar input instances after the model has been trained. Figure 3.9 presents graphically the Architecture of Self Organizing Maps.

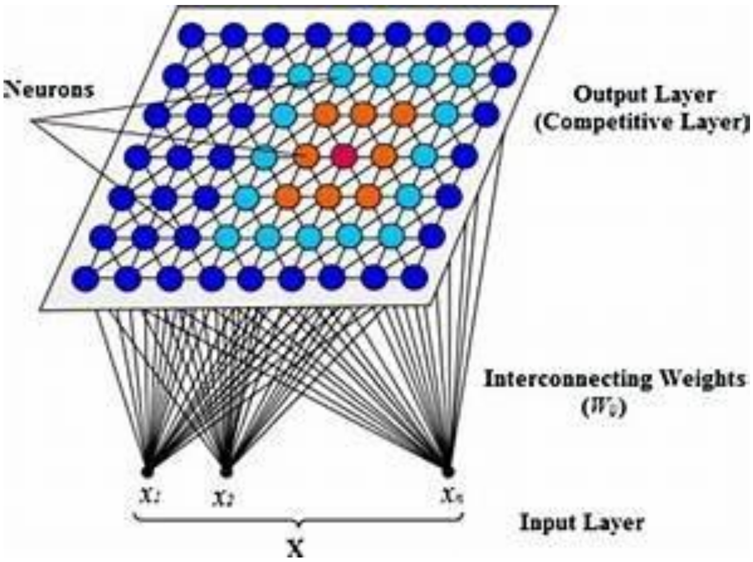


Figure 3. 9 Self Organizing Maps Architecture

Each Self Organizing Map model expects some parameters in order to be initialized.

These parameters include:

- **Grid size of the SOM**: Define how many rows and how many columns will the SOMhave.

- **Nr of epochs**: Number of iterations through the dataset for training the model. In eachiteration, the model updates its weights.
- **Sigma (σ)**: Determines the dimensions of the area surrounding the winning neuron during training.
- **Learning rate (α)**: The initial learning rate of the model, which determines the up-dates of the weight during training.

Oprea, Alina E. and Strungaru, Rodica and Ungureanu, G. Mihaela [23] used the SOM model to extract the most significant features from the Mini–MIAS (Mammographic ImageAnalysis Society) database, so to extract the tumor areas, if any. First they preprocessed theimages by removing the noise and improved the highlighting of possible areas of interest. Then they applied the SOM model on the processed data and applied model evaluation by calculating the detection rate and false positive rate. The evaluation was done by comparing the results of the SOM model with MIAS (dataset) annotation. The detection rate was calculated to be 81% and the false positive rate was compared to be 39%.

3.4 Convolutional Neural Network Techniques used for Breast Cancer Detection

In addition to supervised and unsupervised learning techniques mentioned above, there are other type of Deep Learning algorithms which have proven to be effective in detecting and classifying Breast Cancer cells. CNNs, a subset of deep learning algorithms, have demon- strated exceptional capabilities in discerning intricate patterns within medical images, mak-ing them well-suited for the complex task of identifying cancerous abnormalities. Some CNN methods used for Breast Cancer Detection are: XCeption algorithm, AlexNet, VGG- 16, ResNet50, LeNet, and U-Net.

3.4.1 VGG-16 and ResNet50 Models

Both VGG16 and ResNet50 are models of Convolutional Neural Network. A convolutionalneural network's architecture is made up of an input layer, several hidden layers, and one output layer. Numerous filters, each smaller in size than the input, are included in each convolution layer and conduct the convolutions on the image individually. These filters pickup patterns throughout the whole image [1].

Both VGG-16 and resNet 50 are CNN models that are trained on the ImageNet database,which has over a million sample images. With 16 layers, the VGG-16 network can classify images into 1000 different object categories. The network requires images of input size 224x224 pixels [1]. The main disadvantage of this model is the degradation problem, whichstates that the accuracy of the model reduces rapidly if the network depth increases. The Architecture of the VGG-16 model is shown in Figure 3.10.

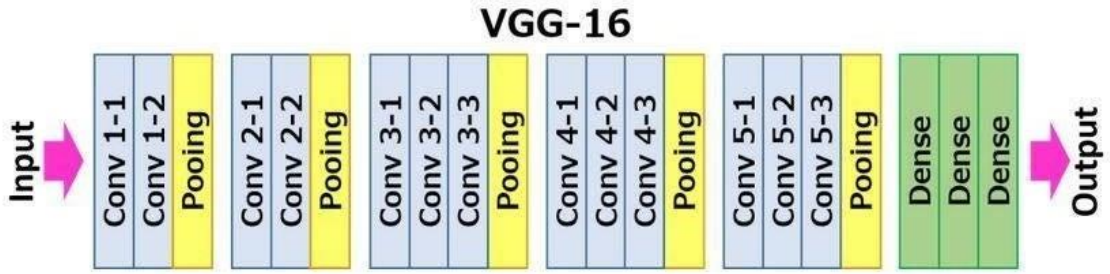


Figure 3. 10 Architecture of VGG-16

ResNet model on the other side was first introduced in [24] to address the degradation problem infused in VGG-16. ResNet has a max depth of 152 layers and it introduces the concept of residual blocks and Shortcut Connections. Shortcut Connections indicate that oneor more layers can be skipped within the model. The layers of the model are connected withone another and they can transfer their input to the next layer. The Architecture of ResNet-50is presented graphically in Figure 3.11. In this Figure, the formula F(x) +x is actually a feedforward neural network with shortcut connections. The aim of shortcut connections within the ResNet model is to add their outputs to the stacked layer outputs in order to execute identity mapping. This optimized architecture of the model is more efficient and requires less computational power in comparison with VGG-16.

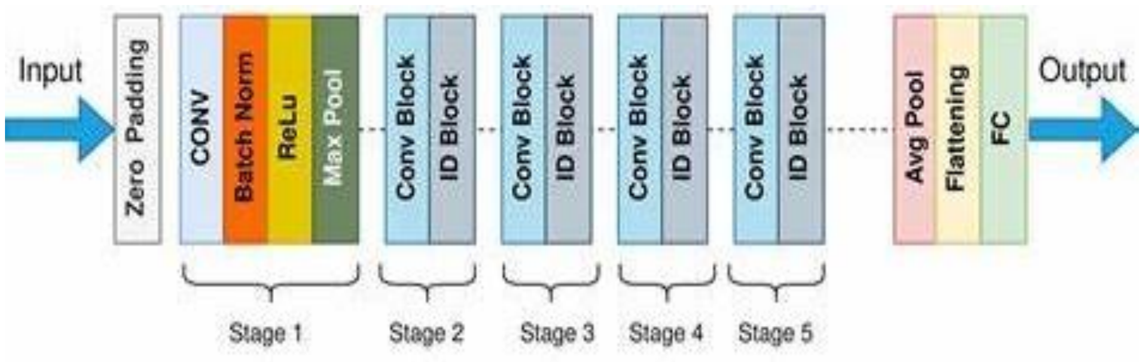


Figure 3. 11 Architecture of ResNet50

Ismail, Nur Syahmi and Sovuthy, Cheab [1] have tested and compared both VGG-16 and resNet models for Breast cancer detection using mammography images, retrieved fromImage Retrieval in Medical Application (IRMA) dataset. This dataset contains 931 images that are diagnosed as normal, and 584 abnormal images, that can be either benign or malignant, all of size 128x128 pixels. Before implementing the models, Ismail, Nur Syahmi and Sovuthy, Cheab have preprocessed the data by resizing the images to 224 x 224 pixels,and have transformed all gray-scaled images into 3-channel RGB. After data preprocessing, the models have been tested by using a 30-70 distribution of data for testing and training, respectively. Then model evaluation is performed, using three performance evaluation matrices: precision, recall and accuracy. The results of the models are shown in table 3.2.

Table 3. 2 Performance Evaluation of VGG-16 and ResNet-50 in [1]

Measure	**VGG-15**	**ResNet-50**
Precision	89%	88%
Recall	99%	94%
Accuracy	94%	91.7%

3.4.2 U-Net Model

U-Net is a Convolutional Neural Network model, widely used for image segmentation tasks. It was first introduced in 2015 by Ronneberger et al. [25]. Its architecture is shown in Figure3.12, retrieved from [26]. UNet is an extended version of the fully convolutional network, which achieves very high accuracy and precise localisation due to its architecture. As it can be seen from Figure 3.12 this model has a U-shaped architecture with two parts: one contracting path known as encoder (left), and another expanding path known as decoder (right). The contracting path of the model is responsible for capturing context from the input image, and the expanding path then enables precise localization of this image.

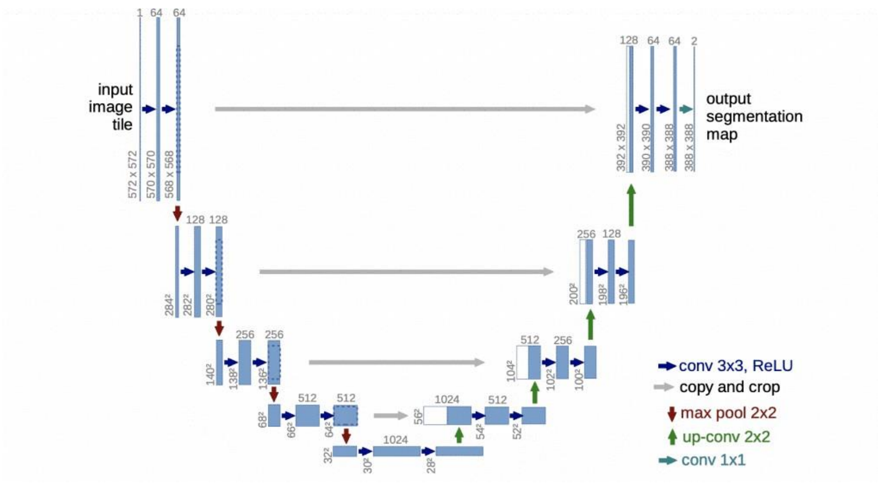


Figure 3. 12 Basic UNet Architecture

The desired features of the original image are retained from the UNet model even after shape recognition, through the ability of the model to transfer feature information from the encoder to the decoder [26]. The architecture of the model applies skip connections, andit also employs an overlap-tile strategy. Through this strategy, the model segments images into overlapping tiles instead of working with the whole original images. These tiles are overlapped in order to avoid false predictions and to maintain continuity.

The UNet model consists of two main components: The Contracting path responsible for feature extraction and context capturing, and the Expanding path responsible for preciselocalization

The Contracting Path of the model (Encoder) has these components:

- *Convolutional Layers:* Each block has two consecutive 3 x 3 Convolutional Layers
- *Activation Function:* A ReLu activation function proceeds each Convolutional Layer,which allows the system to deal with more complax patterns in data by introducing non linearities.
- *Max pooling:* A 2 x 2 max pooling operation is used by the model after each convolutional layer, with stride 2. In the max pooling layer the model reduces the spatial dimensions of the data by half. This operation is very useful within the model, becauseit makes the latter invariant to small shifts or even distorions in the data.
- *Feature Doubling:* The last layer of the Contracting path doubles the number of feature channels, with the aim of increasing the models' capability to learn from the data

despite the spatial reductions performed in the max pooling layer.

The Expanding Path of the model (Decoder) has these components:

- Up-sampling of the feature map
- A 2x2 convolution which reduces by half the feature channels
- Two 3x3 convolutions, which are followed by a ReLu activation

The last layer of the UNet model is the final 1x1 convolution which maps the feature vector at the last stage of the Decoder, to the target number of classes for segmentation.

In the task of detecting breast cancer cells as either being malignant or benign, this model has been tested by Mirya Robin, Jisha John, and Aswathy Ravikumar in [27] with the BreakHis dataset available in Kaggle repository. They have used 80% of the dataset fortraining the model and 20% for testing. The training accuracy of the trained model was 94.35% and the validation accuracy was 93.9%.

3.4.3 XCeption Model

The xCeption model is a deep convolutional neural network (CNN) architecture primar-ily designed for image classification tasks. It is an extension of the Inception architecture, which was originally introduced by Google. The xCeption model aims to enhance the repre-sentational power of the network by introducing a new concept called ”depthwise separableconvolutions.”

XCeption Algorithm has been tested by Abunasser, B.S. et al. [28] using the BreakHist dataset from Kaggle depository. They have separated it into three categories: training, val- idating, and testing. 60% of the data is used for training, 20% for validating, and 20% for testing. Xception model achieved Training Accuracy of 99.78%, Validating Accuracy of 98.59% and Testing Accuracy of 97.60%. In the customized model Training Loss was 0.00315, Validating Loss was 0.07326, and Testing Loss was 0.09518. The model required2944 seconds for training and 5.32 seconds for testing.

3.4.4 AlexNet Model

Another Convolutional Neural Network Technique that can be used for Breast Cancer Detec-tion is AlexNet. This method is named AlexNet after one of its inventors, Alex Krizhevsky[29]. The Work flow of this technique using data augmentation and transfer learning is shown in Figure 3.13. The steps of this algorithm include:

1. Acquiring and preprocessing images
2. Transfer learning with finetuning pre-trained models
3. Classification of the data into target class labels

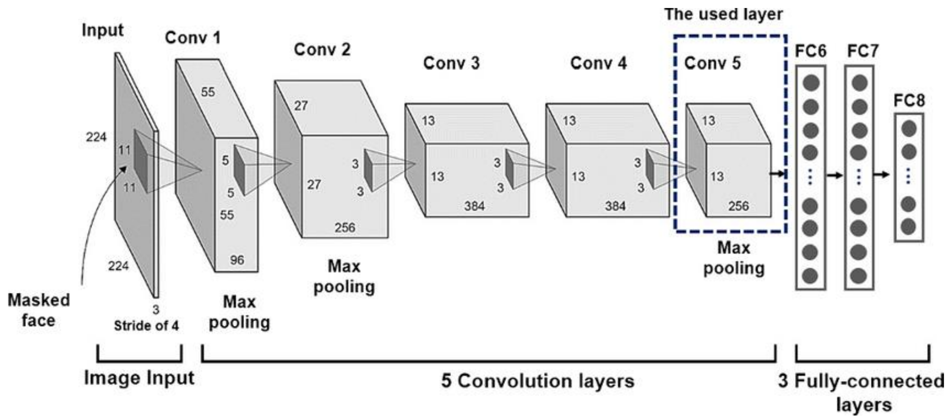


Figure 3. 13 AlexNet Architecture

A. Titoriya and S. Sachdeva [30] have tested the AlexNet model using the BreakHis dataset, which contains images in different magnifying factors: 40x, 100x, 200x, and 400x.The images required by AlexNet method must be in size 227x227x3. In order to fit the dataset into the required format by the method, the authors have made several transforma- tions and processings in the input data.

The highest accuracy of 96.8% was achieved at 40x magnification factor, followed by97.9% at 100x, 96.7% at 200x, and 95.4% at 400x.

3.5 Evaluation Metrics

Evaluation Metrics play a crucial role in the area of Breast Cancer Detection, as the perfor-mance and reliability of these models is of a high importance, and the lack of accuracy can be life threatening. Evaluation Metrics are used as quantitative measures to understand the performance of each model, and to compare each model with one another. In this section wediscuss some of these evaluation metrics. TP in the formulas stands for True Positive, TN stands for True Negative, FP stands for False Positive, FN stands for False Negative, TPF stands for True Positive Fraction, and FPs/image stands for False Positive per Image:

1. **Accuracy** measures the frequency at which the model correctly predicts the result. Its formula is given in Equation 3.4

$$\boldsymbol{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \tag{3.4}$$

2. **Precision** measures the frequency at which the model correctly predicts the results ofthe positive class. Its formula is given in Equation 3.5

$$\boldsymbol{Precision} = (TP/(TP + FP)) \tag{3.5}$$

3. **Recall** quantifies the capacity of the model to accurately identify every positive sam-ple. Its formula is given in Equation 3.6

$$\boldsymbol{Recall} = (TP/(TP + FN)) \tag{3.6}$$

4. **F-1 Score** is a measure of the harmonic mean of precision and recall. Its formula is given in Equation 3.7

$$\boldsymbol{F1Score} = 2x\,(PrecisionxRecall)/(PrecisionRecall) \tag{3.7}$$

5. **Adjusted Rand Index (ARI)** measures the similarity or dissimilarity between be-tween two clusters in unsupervised learning [31]. The cluster is compared with the ground truth.
6. **True Positive Fraction** compares the total number of detected cells to the total numberof actual cells. Its formula is given in Equation 3.8

$$\boldsymbol{TPF} = Precision = Numbero\,f\,TPs/numbero\,f\,totalsamples \tag{3.8}$$

7. **False Positive per Image** It's formula is given in 3.9

$$\textbf{\textit{FPs/image}} = Number of FPs/number of images \tag{3.9}$$

8. **Minimum Inter-neuron Distance (MID)** is a common statistic for assessing a trainedSelf Organizing Map's performance. It indicates the lowest distance between two neurons on the grid. Lower MID values show that the data is more arranged and that the model has been successful in preserving the correlation between the data points and capturing the underlying structure of the data.

3.6 Available Datasets for Breast Cancer Detection

This section analyzes and describes some free, available datasets that can be used for BreastCancer Detection.

3.6.1 Breast Cancer Wisconsin Diagnostic (WDBC)

Wisconsin Diagnosis Breast Cancer (WDBC) is an available dataset used for Breast CancerDetection, retrieved from UCI Machine Learning Repository [32]. It is available in .csv format and it contains 569 instances with 32 attributes. The first attribute of the dataset contains a unique number which identifies the instance and does not have any other medicalmeaning related to the instance. The second attribute of this dataset indicates the target value: M for malignant cases, and B for benign cases. Out of all the instances of the dataset,212 instances belong to malignant images, and the rest of 357 belong to benign images. It then has 30 real-valued input features, each of which represents a specific characteristic of the single instance. The input features are categorized into 10 attributes, and each attribute is represented by three indicators: mean value, standard error, and maximum value. The attributes of this dataset are:

- Radius
- Texture (standard deviation of gray-scale values)
- Area

- Perimeter
- Symmetry
- Compactness
- Smoothness (local variations in radius length)
- Concavity
- Concave points
- Fractal dimension

The dataset is linearly separable, and it can be downloaded online from this link: WisconsinDiagnosis Breast Cancer (WDBC). Figure 3.14 shows a visual representation of some of thecharacteristics of the dataset. Target indicates the class label, where 0 stands for malignant and 1 stands for benign.

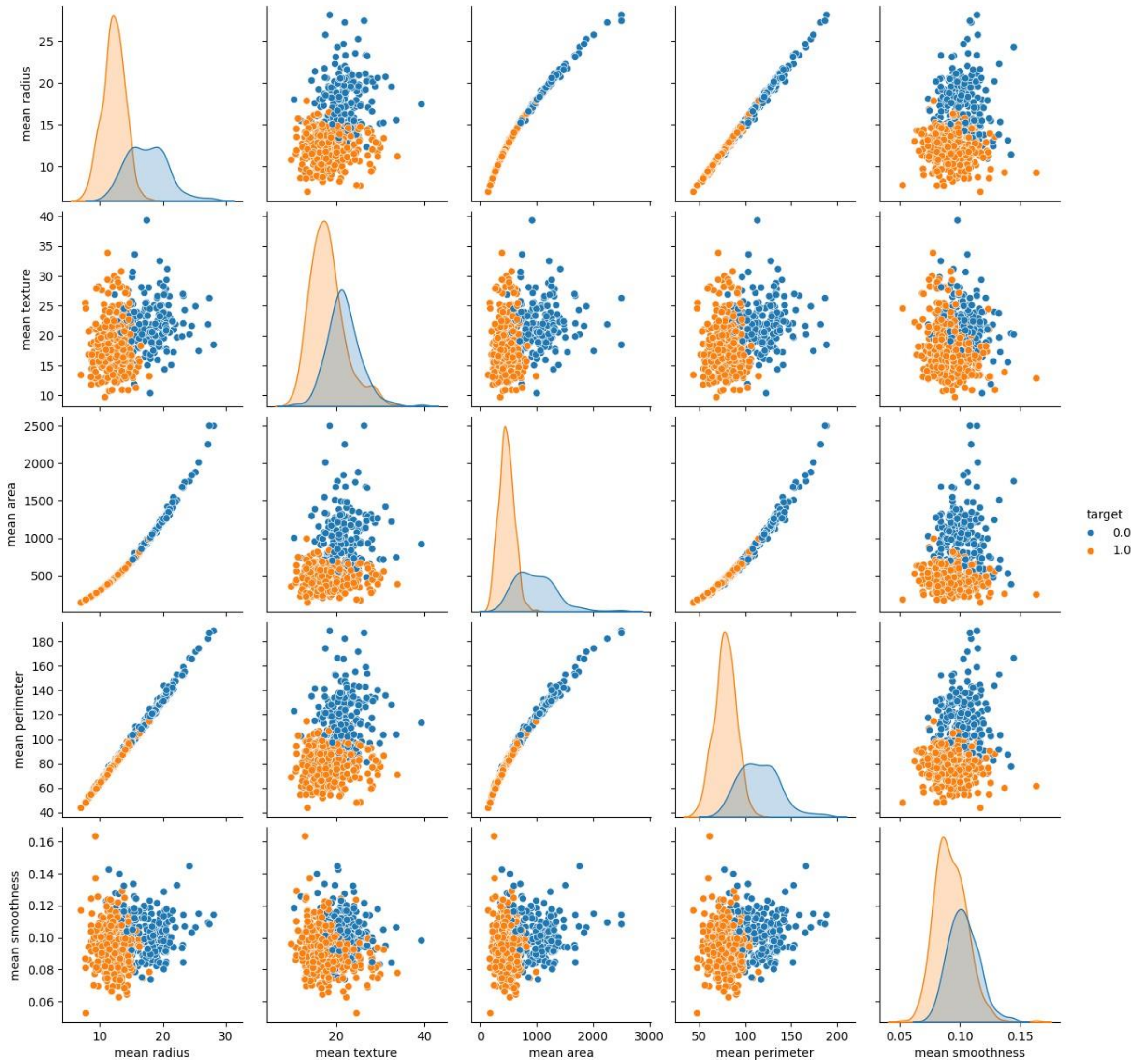


Figure 3. 14 WDBC Characteristics

3.6.2 Breast Cancer Wisconsin Original

Breast Cancer Wisconsin (Original) is an available dataset used for Breast Cancer Detection,retrieved from UCI Machine Learning Repository [33]. It has been obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Dr. Wolberg has reported periodically all the clinical cases he has had. He has first reported 367 instances ofthe dataset in January 1989, and he has continued to do so up until November 1991, where the instances reached the number 699. The dataset contains 699 instances, one identification number for the record, the class label (2 for benign, 4 for malignant), and 9 real-valuedattributes. 9 real-valued input features of the dataset include:

- Clump Thickness (Att 1)
- Uniformity of Cell Size (Att 2)
- Uniformity of Cell Shape (Att 3)
- Marginal Adhesion (Att 4)
- Single Epithelial Cell Size (Att 5)
- Bare Nuclei (Att 6)
- Bland Chromatin (Att 7)
- Normal Nucleoli (Att 8)
- Mitoses (Att 9)

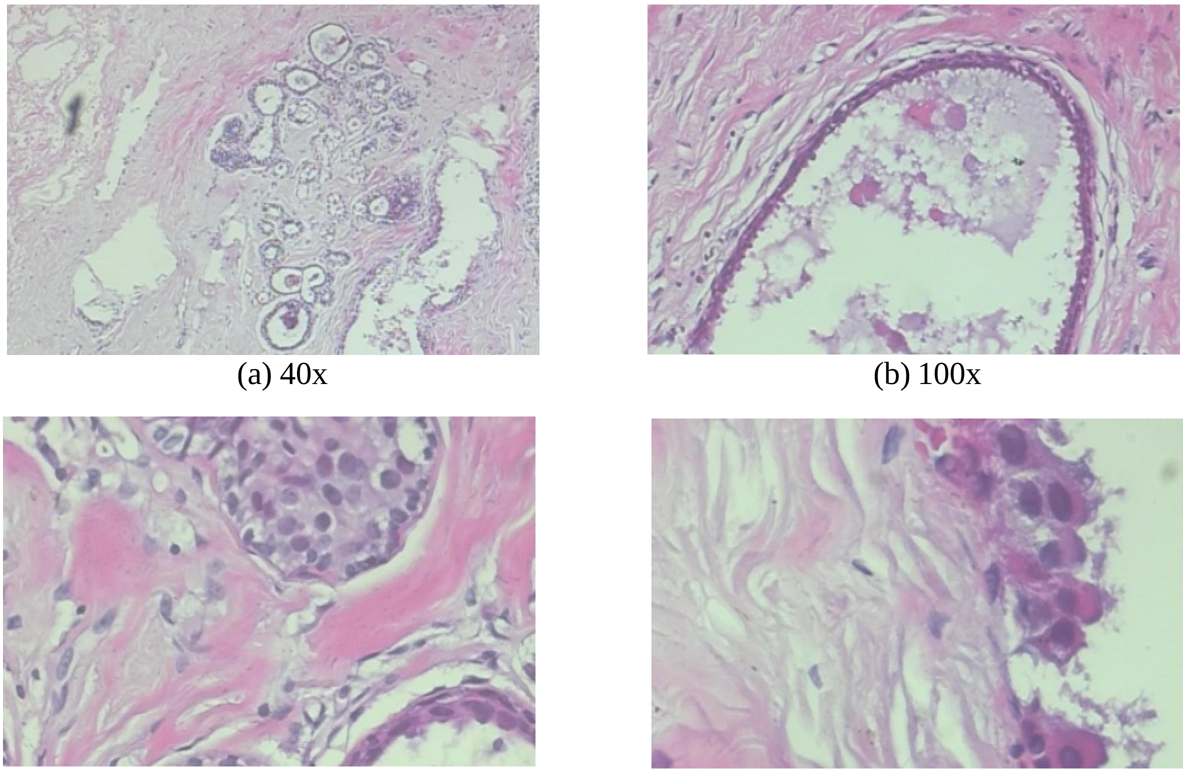
This dataset has a distribution of 65.5% benign and 34.5% malignant, which correspondsto 458 benign instances and 241 malignant instances. In addition, there are 16 instances in the dataset, each of which contains a single missing value. This missing attribute value can be distinguished by the value '?'. The dataset can be downloaded here. Some samples of theBreast Cancer Wisconsin Original Dataset are shown in Table 3.3.

Table 3. 3 Three random samples from the Breast Cancer Wisconsin Original Dataset

ID	Att 1	Att 2	Att 3	Att 4	Att 5	Att 6	Att 7	Att 8	Att 9	Label
1091262	2	5	3	3	6	7	7	5	1	4
1096800	6	6	6	9	6	?	7	8	1	2
1099510	10	4	3	1	3	3	6	5	2	4

3.6.3 Breast Cancer Histopathological Database (BreakHis)

The Breast Cancer Histopathological Database (BreakHis) includes 7909 unique microscopic images of breast cancer tissue, collected from 82 individuals at different magnifyingfactors [34]. The dataset contains 2480 images that correspond to benign cases, and 5429 images that correspond to malignant cases. All the images are three channel RGB with 8- bit depth in each channel, 700X460 pixels, and in PNG format. This dataset was built in collaboration with the P/D Laboratory – Pathological Anatomy and Cytopathology, Parana,Brazil". The BreakHis dataset can be downloaded here. Figure 3.15 represents an image of a benign breast cancer tissue, and Figure 3.16 represents a malignant breast cancer tissue atdifferent magnifying factors.



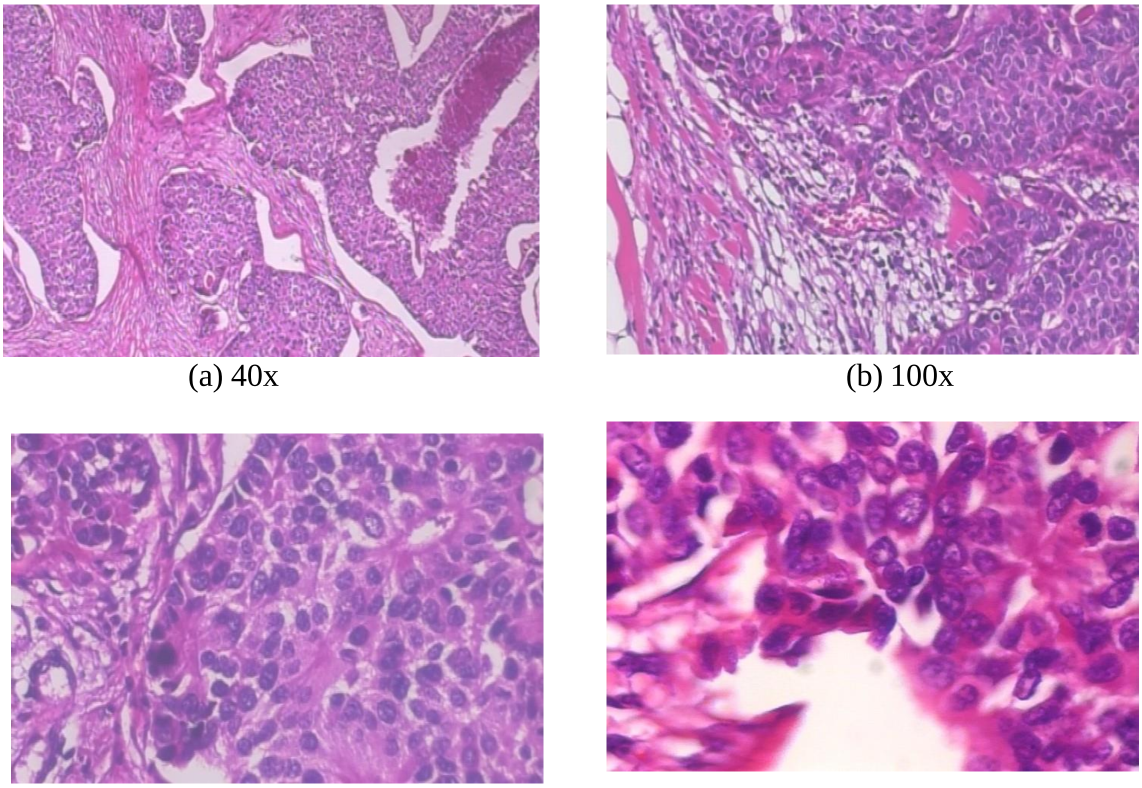
(a) 40x

(b) 100x

(c) 200x

(d) 400x

Figure 3. 15 Benign Breast Cancer Tissue



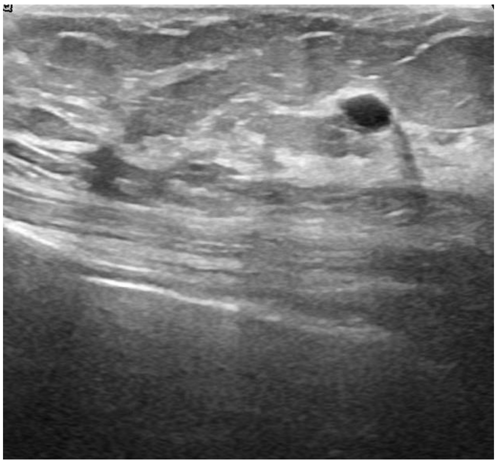
(a) 40x (b) 100x

(c) 200x (d) 400x

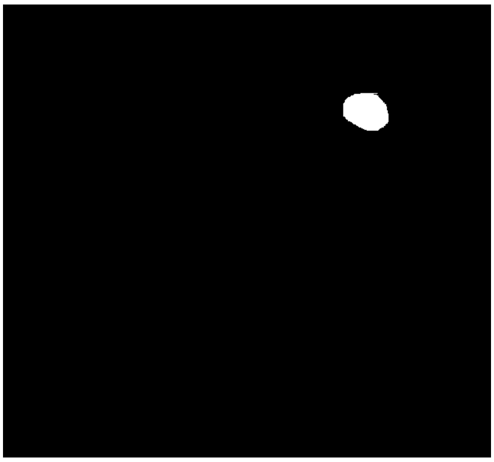
Figure 3. 16 Malignant Breast Cancer Tissue

3.6.4 Breast Ultrasound Images Dataset

Breast Ultrasound Images [35] has been collected in 2018 and it contains images that rep- resent breast ultrasounds of 600 different women in ages between 25 and 75 years old. In total there are 780 available images with an average image size of 500x500pixels in PNG format. Each image in this dataset is labeled and belongs to one of the three classes: normal,benign, and malignant. This dataset can be downloaded here. Original images in the datasetare also associated with ground truth images. Figure 3.17 shows an example of an image labeled as benign (a), and its respective Ground Truth (b). Figure 3.18 shows an example ofa Breast Cancer image labeled as malignant(a), and its respective Ground Truth image(b). Figure 3.19 shows an image labeled as normal (a) and its ground truth (b).

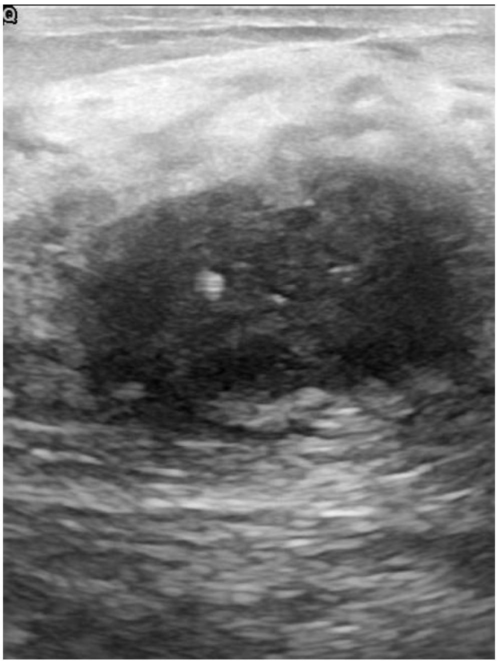


(a) Benign Breast Cancer ultrasound im-age

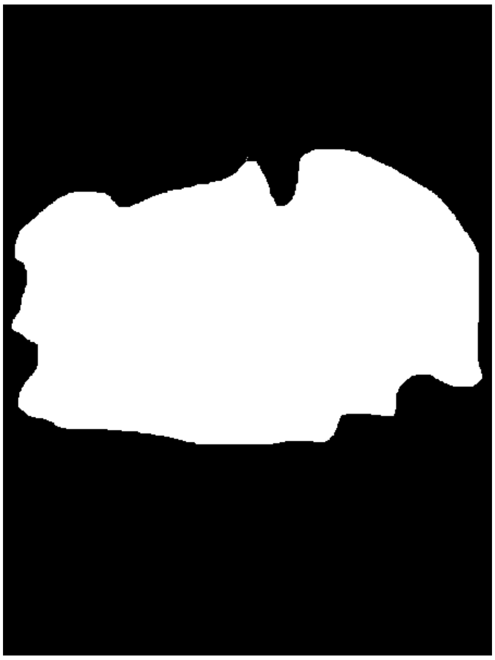


(b) Benign Ground Truth

Figure 3. 17 Benign Breast Cancer ultrasound image and Ground Truth

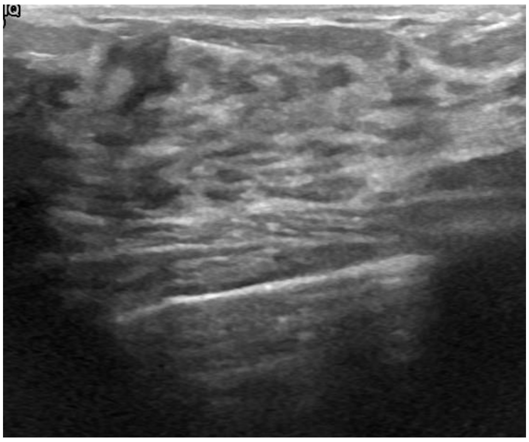


(a) Malignant Breast Cancer ultrasound image



(b) Malignant Ground Truth

Figure 3. 18 Malignant Breast Cancer ultrasound image and Ground Truth



(a) Normal Breast Cancer ultrasound image

(b) Normal Ground Truth

Figure 3. 19 Normal Breast Cancer ultrasound image and Ground Truth

CHAPTER 4

METHODOLOGY

4.1 Proposed Methodology

The methodology proposed in this Thesis is shown in Figure 4.1. After retrieving the datasets from the online repositories, and pre-processing them, we have tested each model on specificdatasets: Supervised Learning Models (SLM) have been tested with labeled data (Wisconsindataset); Unsupervised Learning Models (ULM) have been tested with unlabeled data (Wis-consin dataset after dropping the target column); CNN models have been tested with imagedata, using the Ultrasound Images Dataset. Each model has been trained with different parameters to see with which combination of parameters it performs best.

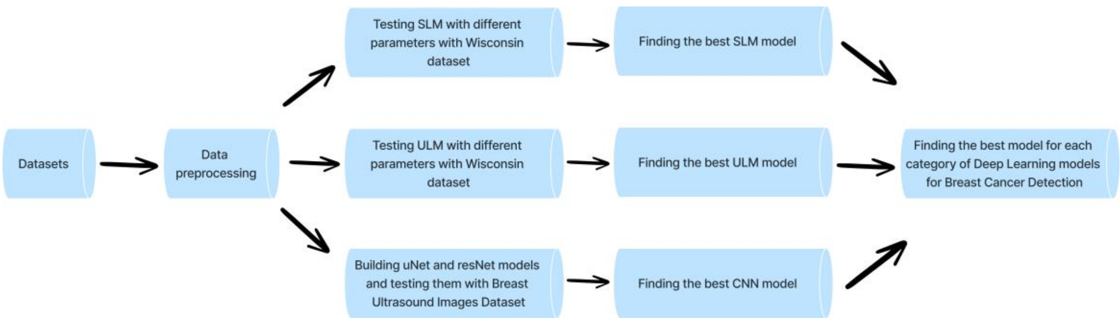


Figure 4. 1 Proposed Methodology

The approach we have followed in this Thesis is this: we have tested 4 Supervised Learning models Random Forest, Naive Bayes, SVM, and KNN with different parameter values for each, and we have compared them in terms of the time needed for training and validating, as well as in terms of accuracy, precision and recall. The model that outperforms the othershas been chosen as the best Supervised Model for Breast Cancer Detection. We have followed the same approach for Unsupervised Learning methods, where we have tested Auto Encoders, Self-Organizing Maps, and K-Means clustering. Then we have tested two CNN models: resNet and uNet, and we have compared their performance based on the training time and validation time, as well as based on the accuracy and loss of each model. In the endof the Thesis all these models are tested and compared with one another, and there will be one winning model for each category in the task of Breast Cancer detection.

The purpose of the proposed methodology is to answer the following questions:

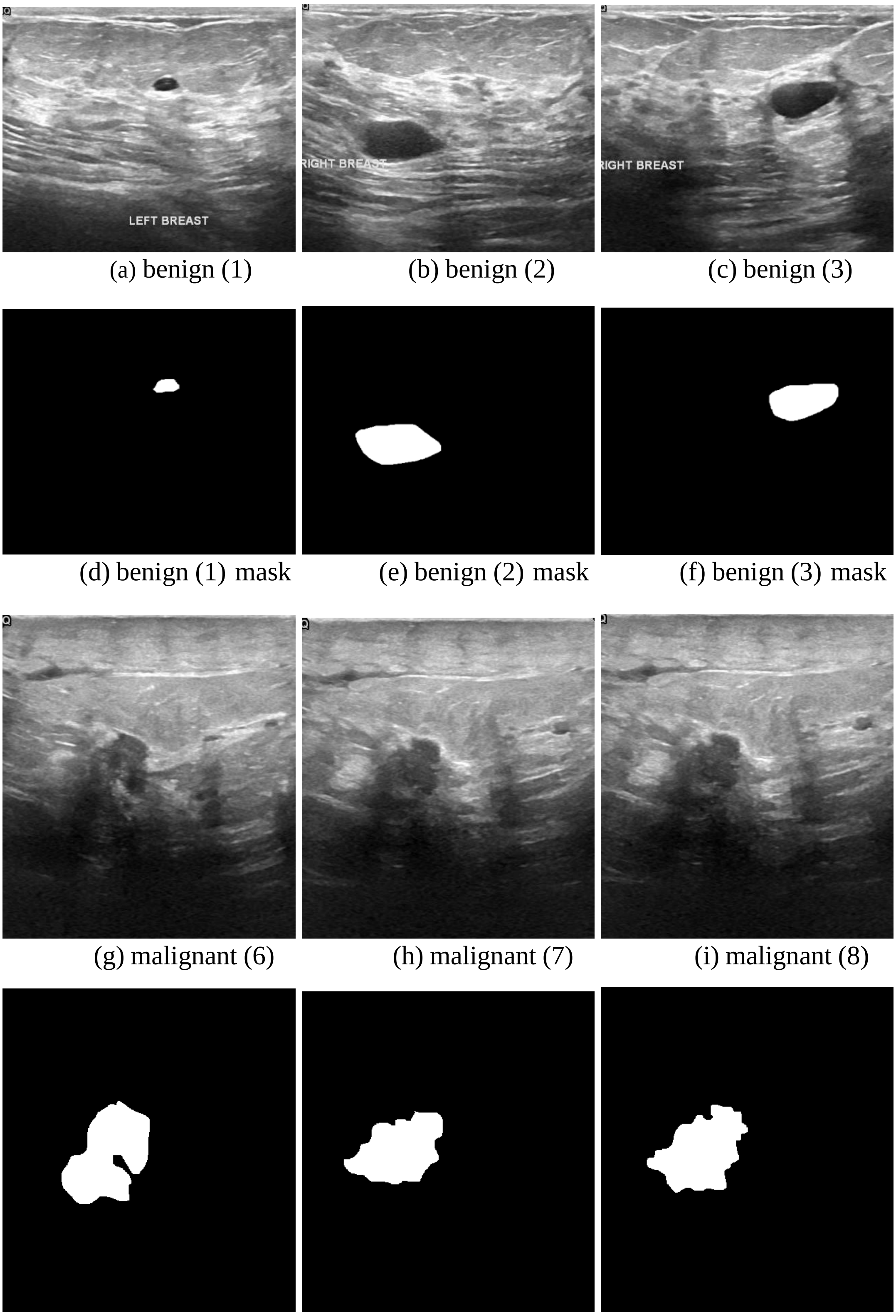
1. What is the best model that should be used for Breast Cancer Detection if labeled datais available, and human expertise to structure the data and represent it into meaningfulnumerical values is possible?

2. What is the best model that should be used for Breast Cancer Detection if human intervention for labeling data is not possible and the available dataset is unlabeled, yetnumeric?

3. What is the best model that should be used for Breast Cancer Detection if the availabledataset consists of complex images, where feature extraction is complex and needs tobe automated?

4.2 Datasets used

In this Thesis we work with two of the datasets mentioned in section 3.6: Breast Cancer Wis- consin Diagnostic (WDBC), retrieved from UCI Machine Learning Repository, and Breast Ultrasound Images Dataset, retrieved from Kaggle. We use the first dataset for Supervised and Unsupervised Learning methods, which require numerical data. The second dataset is used with two CNN models: UNet and ResNet, which require image data. Figure 4.2 showssamples of Breast Ultrasound Images dataset. Wisconsin Original Dataset has instances with 32 numerical attributes each, and this makes it impossible to provide here some samples. Butthe dataset can be downloaded from UCI Machine Learning Repository in the link providedin Section 3.6.1.



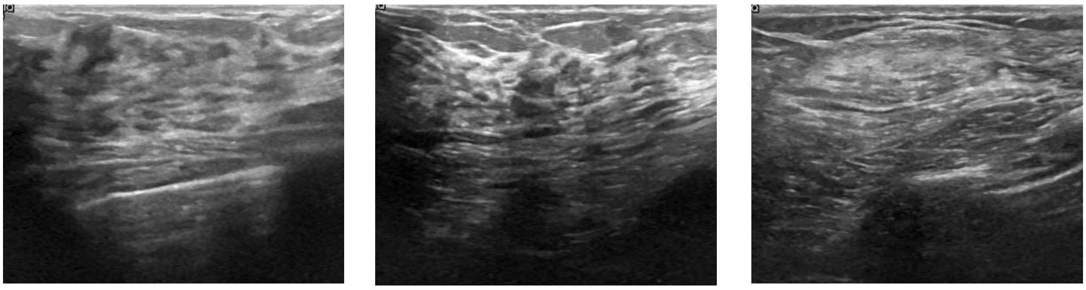
(a) benign (1) (b) benign (2) (c) benign (3)

(d) benign (1) mask (e) benign (2) mask (f) benign (3) mask

(g) malignant (6) (h) malignant (7) (i) malignant (8)

(j) malignant (6) mask (k) malignant (7) mask (l) malignant (8) mask

Figure 4. 2 Samples from Breast Ultrasound Images Dataset: Real Images and respectiveMasks



(a) normal (1) (b) normal (2) (c) normal (3)

(d) normal (1) mask (e) normal (2) mask (f) normal (3) mask

***Figure 4.* 3** Samples from Breast Ultrasound Images Dataset: Real Images and respective Masks (Normal class)

4.3 Data preprocessing

For each one of the datasets used, Wisconsin and Breast Ultrasound Images Dataset, we haveused several pre-proccesing techniques in order to make the datasets more useful and benefitthe most from them. For the first dataset used by Supervised and Unsupervised models, Wisconsin dataset, we have used dimensionality reduction by dropping the first column. This column is dropped because it is an identification code that simply identifies the recordwithin the dataset, but does not represent any valuable information related to the disease. Then we have used Label Encoding to transform the categorical target values ’M’ and ’B’ into numerical values 0 and 1. Table 4.1 shows the target values before and after applying this normalization technique. Row (a) shows the target values before applying Label Encoding,and row (b) shows target values after applying Label Encoding.

Table 4. 1 Target values of the Wisconsin Dataset before (a) and after (b) applying Label Encoding

(a)	’M’	’M’	’B’	’M’	’B’	’B’
(b)	’1’	’1’	’0’	’1’	’0’	’0’

We have also normalized the dataset using min max scaling technique, because it contains features whose range of values vary widely. Data normalization is achieved by using the **MinMaxScaler** class of the **sklearn.preprocessing** library in Python. Table 4.2 shows some attribute values of this dataset before and after using MinMaxScaler for scaling the values.

Table 4. 2 Attribute values of the Wisconsin Dataset before (a) and after (b) applying Min- MaxScaler

	radius	texture	perimeter	area	smoothness	compactness	concavity
(a)	441	17.27	25.42	112.4	928.8	0.08331	441
(b)	0.3185	0.4614	0.3207	0.1843	0.7198	0.5429	0.2194

For the Breast Ultrasound Images Dataset we have used two pre-processing techniques:Image Resizing and Image Conversion. Image Resizing is used to resize the images to the required size for each model, and Image Conversion is used to convert the images into RGBfor the resNet model, and grayscale for uNet model.

4.4 Architectures of the models

This section provides the source codes for the models that are tested in this Thesis, and provides the architecture of two CNN models that are customized and tested with Breast Ultrasound Images dataset. The source codes can be found in Table 6.1. We have used these source codes as the ground base for our testing, and have made several changes whennecessary, explained in the Experimental Results chapter.

Table 4. 3 Open source codes for Supervised and Unsupervised Learning methods for BreastCancer Detection

Method	**GitHub Link**
KNN	Bhttps://github.com/Manishnir/Breast-Cancer-Prediction-using-KNN
Naive Bayes	https://github.com/shaadclt/Breast-Cancer-Detection-NaiveBayesClassifer
Random Forest	https://github.com/jimschacko/Breast-Cancer-Detection-using-Random-Forest
SVM	https://github.com/mayorofdata/Breast-Cancer-Classification-using-Support-Vector-Machine
Auto Encoder	https://github.com/mainak-ghosh/AutoEncoder
SOM	https://github.com/sethns/Self-Organizing-Maps

4.4.1 Architecture of UNet

In this Thesis we have worked with a UNet model that utilized the Breast Ultrasound ImagesDataset described in 3.6.4. Each image in this dataset is resized into a size of 128x128 pixels, is labeled into one of the three classes: benign, malignant, or normal, and it is also associatedwith its mask image. We have used 80% of this dataset to train the model, and 20% to test the model's performance. The architecture of this model is shown in Figure 4.4.

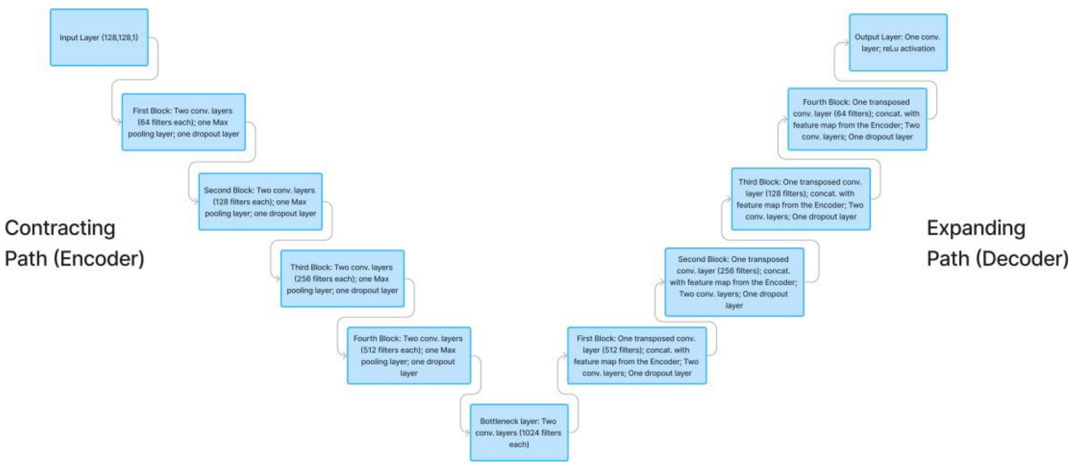


Figure 4. 4 Architecture of the UNet model used in this Thesis

4.4.2 Architecture of ResNet

The second CNN model tested With Breast Ultrasound Images dataset is ResNet. This dataset contains original images of size 500x500pixels. Since ResNet model expects 3 channel input images of size 224 x 224, we modified the dataset by preproccesing it. We applied two preprocessing techniques: **Image Resizing** to resize the images from 500 x 500pixels into 224 x 224 pixels, and **Image Conversion** to convert any gray-scale images into 3-channel RGB images. We have then used OneHotEncoder to convert categorical labels intoa one-hot encoded format. Now that the dataset is ready to be used by the ResNet model, we have loaded the available pre-trained model using **tf.keras.applications.ResNet50**. The advantage of using this pre-trained model as a starting point for our new model, is that this model is trained with a very large dataset (ImageNet), and owns all of the feature extractioncapabilities gained from it. To make use of the weights learned from this dataset, we have used ***weights="imagenet"***. To froze the base model so that it only learns the weights once in order to save time and space, we have used ***trainable = False***. In addition, in order to be able to customize the base model for our dataset, we have excluded its top layers by using ***include top=False***. The architecture of the ResNet model we have used in this Thesis is shown in Figure 4.5.

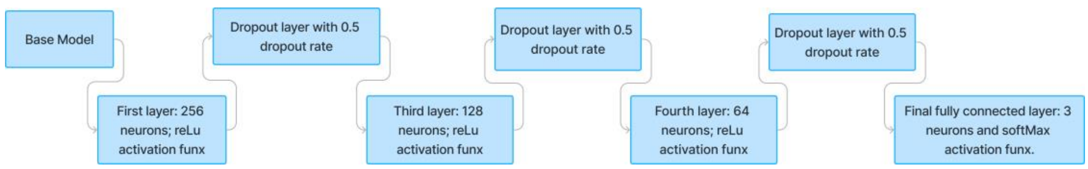


Figure 4. 5 Architecture of the ResNet model used in this Thesis

4.5 Evaluation Metrics

The evaluation metrics that are be used for Supervised and Unsupervised models are: accu-racy, precision, recall and F-1 score. For CNN models we have used the history of models' accuracy and the history of models' loss.

4.6 Implementation Details

All the methods tested in this Thesis are implemented in Python, and run in Google Colabenvironment using a T4 GPU runtime.

CHAPTER 5

EXPERIMENTS AND RESULTS

In this chapter we discuss and explain all the experiments that we have done with Supervised,Unsupervised, and CNN models. The conditions in which these experiments are done are given in details, as well as the results of each experiment.

5.1 Results of Supervised Methods for Breast Cancer Detection

In this section we provide the results of each of the three supervised methods mentioned in Section 3.2.

5.1.1 K-Nearest Neighbor

The supervised method K-Nearest Neighbor is tested using both the Breast Cancer Wisconsin Diagnostic (WDBC) Dataset, and the Breast Ultrasound Images Dataset. A very important parameter of the K-Nearest Neighbor Algorithm is the K-Value. This value indicates the number of neighbors that the model considers before making the decision. Since the initial value of this parameter directly impacts the results and therefore the effectiveness of the algorithm, we have used different values to compare the results.

We have firstly tested the algorithm by using a K-Value=5 with Wisconsin dataset. Theaccuracy of the method under these conditions is 0.96, the time it takes for training the modelis 0.0039 sec and the time for predicting the results is 0.0096 sec. The Confusion Matrix forthis method is shown graphically in Table 5.1.

Table 5. 1 Confusion Matrix for K-Nearest Neighbor Classifier with K-Value=5: WisconsinDataset

	Predicted Negative	**Predicted Positive**
Actual Negative	71	0
Actual Positive	5	38

We tested again KNN model with a K-Value=5 with Breast Ultrasound Images dataset.The accuracy of the method is calculated to be 0.89, the time it takes for training the modelis 0.2091 seconds, and the time for predicting the results is 24.4128 seconds. The ConfusionMatrix for this method is shown in Figure 5.1.

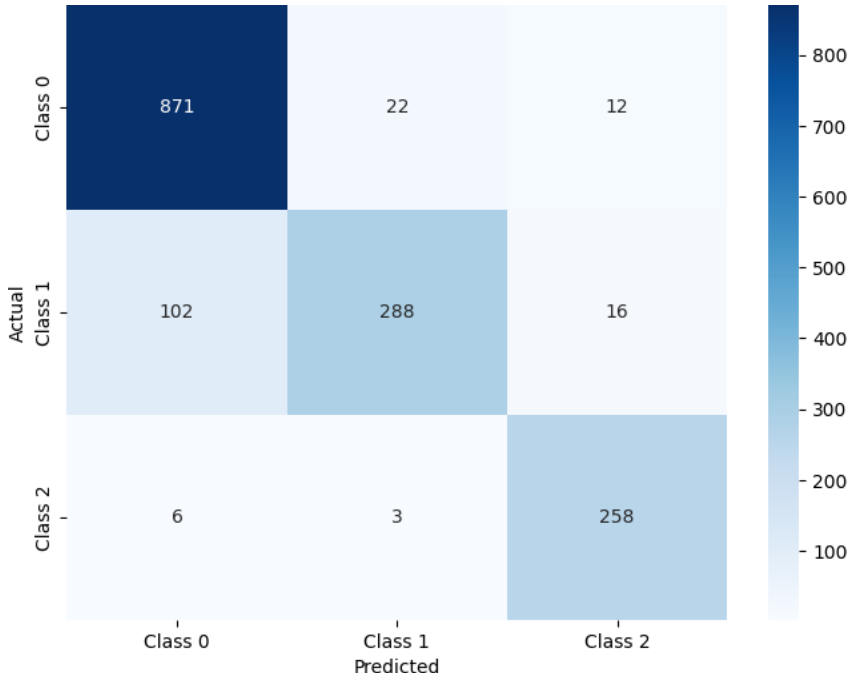


Figure 5. 1 Confusion Matrix for K-Nearest Neighbor Classifier with K-Value=5: BreastUltrasound Images Dataset

Research shows that a high value of K typically reduces the effect of noise in classification, whereas a small value of K increases the sensitiveness of the model to local variationsin the data [36] . Therefore, to see the actual impact that the K-Value has in both Wisconsinand Breast Ultrasound Images datasets, we have tested again the algorithm using two different values of K: 3 and 9. The accuracy of the model with a K-Value=3 for the Wisconsin dataset is calculated to be 0.93, and for Breast Ultrasound Images dataset is calculated to be 0.95. For Breast Ultrasound Images dataset, the training time with this value of K is 0.2193seconds, and the prediction time is 25.7287 seconds. The accuracy of the model with a K- Value=7 for the Wisconsin dataset is calculated to be 0.96, whereas for Breast Ultrasound Images is calculated to be 0.83. Its training time when tested with Breast Ultrasound Im- ages dataset is 0.2065 seconds, and its prediction time is 26.4246 seconds. Tables 5.2, and 5.3 present graphically the Confusion Matrices of these testings for Wisconsin dataset, and Figures 5.2 and 5.3 present the Cofusion Matrices for Breast Ultrasound Images dataset.

Table 5. 2 Confusion Matrix for K-Nearest Neighbor Classifier with K-Value=3

	Predicted Negative	**Predicted Positive**
Actual Negative	68	3
Actual Positive	5	38

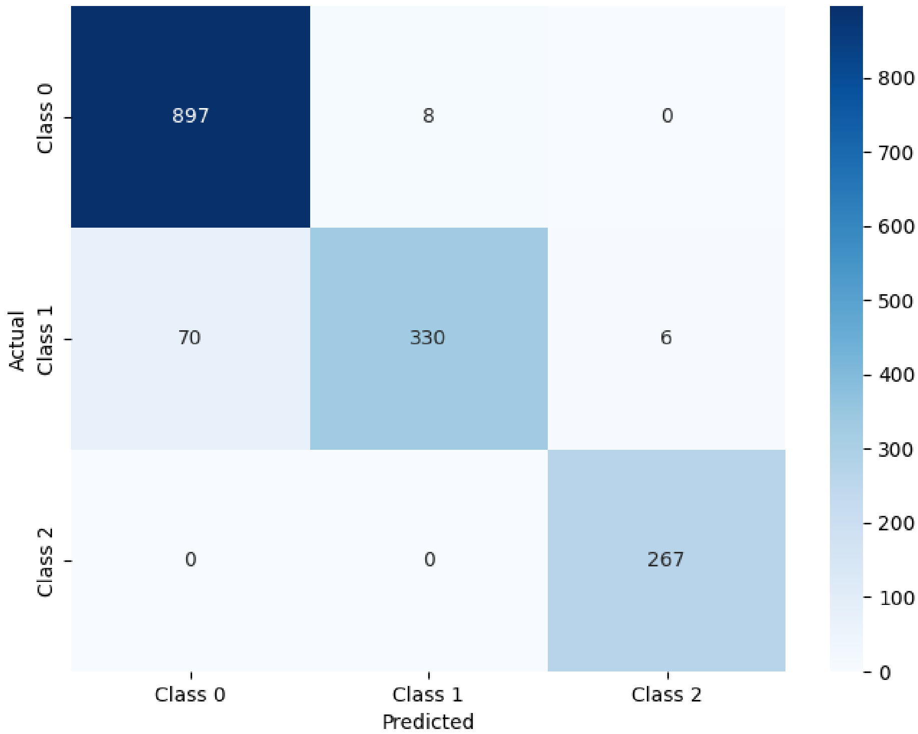


Figure 5. 2 Confusion Matrix for K-Nearest Neighbor Classifier with K-Value=3: BreastUltrasound Images Dataset

Table 5. 3 Confusion Matrix for K-Nearest Neighbor Classifier with K-Value=7

	Predicted Negative	**Predicted Positive**
Actual Negative	70	1
Actual Positive	4	39

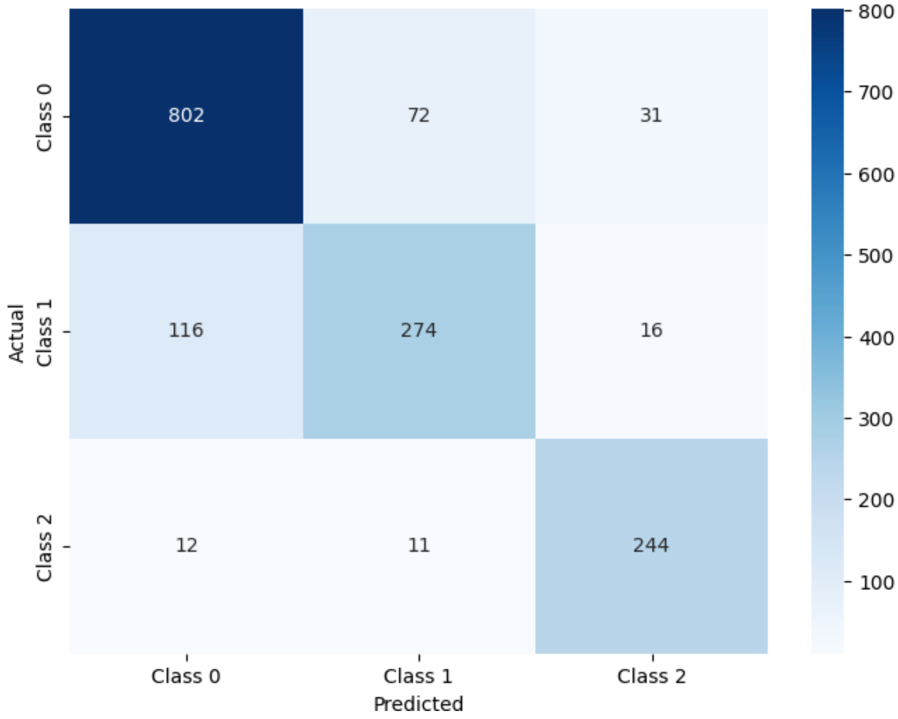


Figure 5. 3 Confusion Matrix for K-Nearest Neighbor Classifier with K-Value=7: BreastUltrasound Images Dataset

All the results of the KNN algorithm with different values of K parameter for Wisconsindataset are shown in Table 5.4. For each evaluation metric included, the highest value is highlighted. In terms of accuracy, the KNN algorithm performed better with both K-values 5 and 7, for which it reached the maximum accuracy of 0.96. In terms of Precision, the bestperformance was achieved using a K-value=7 for class 0 and using a K-value=5 for class 1.The precision for these two cases was respectively 0.95 and 1.00. The highest value of recall(1.00) for class 0 was achieved by using K-value of 5, and the highest value of recall (0.91)for class 1 was achieved by using a K-value of 7. For F-1 score the results were the same for both Class 0 and Class 1 when using K-value 5 and K-value 7. For both of these values,the highest value of F-1 score for Class 0 was 0.97, and the highest value for class 1 was 0.94. In conclusion, when tested with the Wisconsin dataset, KNN model performs best withK-value 5 and 7.

Table 5. 4 Performance of KNN with different K-values: Wisconsin Dataset

		K-value=3	K-value=5	K-value=7
Accuracy		0.93	**0.96**	**0.96**
Precision	Class 0	0.93	0.93	**0.95**
	Class 1	0.93	**1.00**	0.97
Recall	Class 0	0.96	**1.00**	0.99
	Class 1	0.88	0.88	**0.91**
F-1 Score	Class 0	0.94	**0.97**	**0.97**
	Class 1	0.90	**0.94**	**0.94**

Results of the K-NN algorithm with Breast Ultrasound Images Dataset are shown inTable 5.5. For this dataset, KNN model performs best with K-value=3.

Table 5. 5 Performance of KNN with different K-values: Breast Ultrasound Images Dataset

		K-value=3	K-value=5	K-value=7
Accuracy		**0.95**	0.89	0.83
Precision	Class 0	**0.93**	0.89	0.86
	Class 1	**0.98**	0.92	0.77
	Class 2	**0.98**	0.90	0.84
Recall	Class 0	**0.99**	0.96	0.89
	Class 1	**0.81**	0.71	0.67
	Class 2	**1.00**	0.97	0.91
F-1 Score	Class 0	**0.96**	0.92	0.87
	Class 1	**0.89**	0.80	0.72
	Class 2	**0.99**	0.93	0.87

5.1.2 Naive Bayes

The second method that is tested using both Breast Cancer Wisconsin Diagnostic (WDBC) Dataset, and Breast Ultrasound Images dataset is Naive Bayes. Depending on the type of the dataset' features, and the probability distribution, we can use different variants of Naive Bayes classifiers. First, we have tested the method using Gaussian Classifier with Wisconsindataset. The accuracy of the method when implemented using the Gaussian Classifier on this dataset is 0.97. The time needed for training the model is 0.0026 seconds, and the time it takes to predict the results 0.0026 seconds. The results of this type of method for the Wisconsin dataset are shown graphically in Table 5.6.

Table 5. 6 Confusion Matrix for Gaussian Naive Bayes Classifier: Wisconsin Dataset

	Predicted Negative	**Predicted Positive**
Actual Negative	71	0
Actual Positive	3	40

We tested again the method using Gaussian Classifier with Breast Ultrasound Images dataset. The accuracy of the method when implemented using the this dataset is 0.38. The time needed for training the model is 5.1149 seconds, and the time it takes to predict the results 1.5542 seconds. The results of this type of method for Breast Ultrasound Images dataset are shown graphically in Figure 5.4.

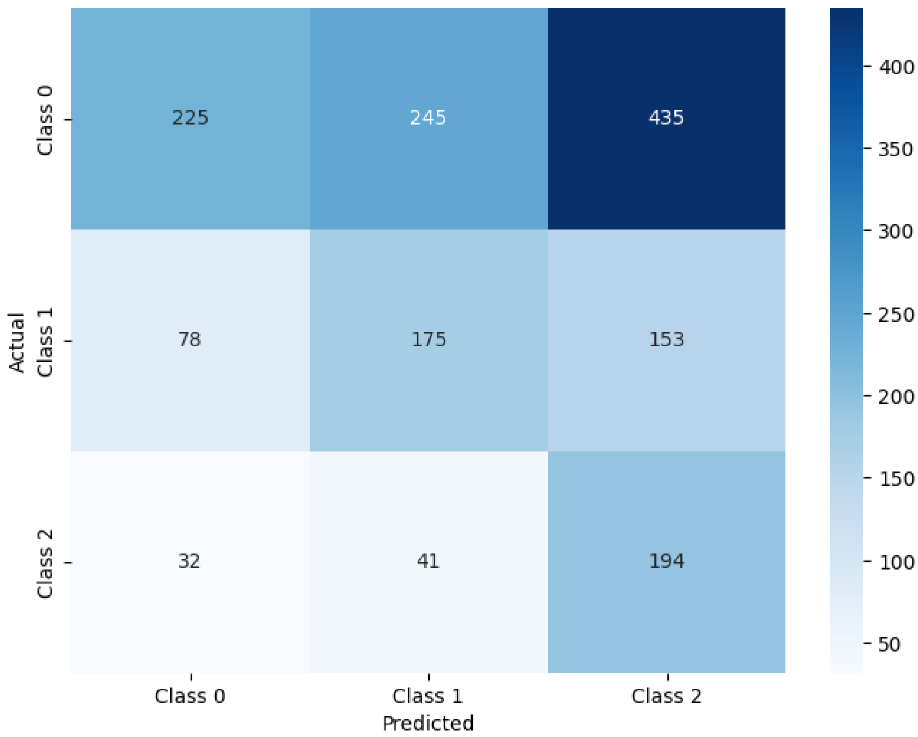


Figure 5. 4 Confusion Matrix for Gaussian Naive Bayes Classifier: Breast Ultrasound Im-ages Dataset

With the aim of finding the best parameters which maximise the accuracy of the method, we have also tested Naive Bayes using Multinomial and Bernoulli Classifiers on both datasets. The first one is more suitable for datasets where the features represent counts or frequencies, whereas the last one is more suitable for binary features. The accuracy of the model using the Multinomial Classifier on Wisconsin dataset is 0.94, the time needed for training the datais 0.1246 seconds, and the time needed for prediction is 0.0035 seconds. When using the Bernoulli Classifier on Wisconsin dataset, the accuracy of the model is calculated to be 0.62,the time needed to train the data is 0.0053 seconds, and the time needed to predict the new data is 0.0027 seconds. The results of such testings can be seen in Table 5.7, and Table 5.8.

Table 5. 7 Confusion Matrix for Multinomial Naive Bayes Classifier: Wisconsin dataset

	Predicted Negative	Predicted Positive
Actual Negative	71	0
Actual Positive	7	36

Table 5. 8 Confusion Matrix for Bernoulli Naive Bayes Classifier: Wisconsin dataset

	Predicted Negative	Predicted Positive
Actual Negative	71	0
Actual Positive	43	0

The same testings are performed on Breast Ultrasound Images dataset also. The accuracyof the model using the Multinomial Classifier on this dataset is 0.29, the time needed for training the data is 18.4128 seconds, and the time needed for prediction is 0.2639 seconds. When using the Bernoulli Classifier on Breast Ultrasound Images dataset, the accuracy of the model is calculated to be 0.23, the time needed to train the data is 17.8780 seconds, andthe time needed to predict the new data is 0.4367 seconds. The results of such testings can be seen in Figures 5.5 and 5.6.

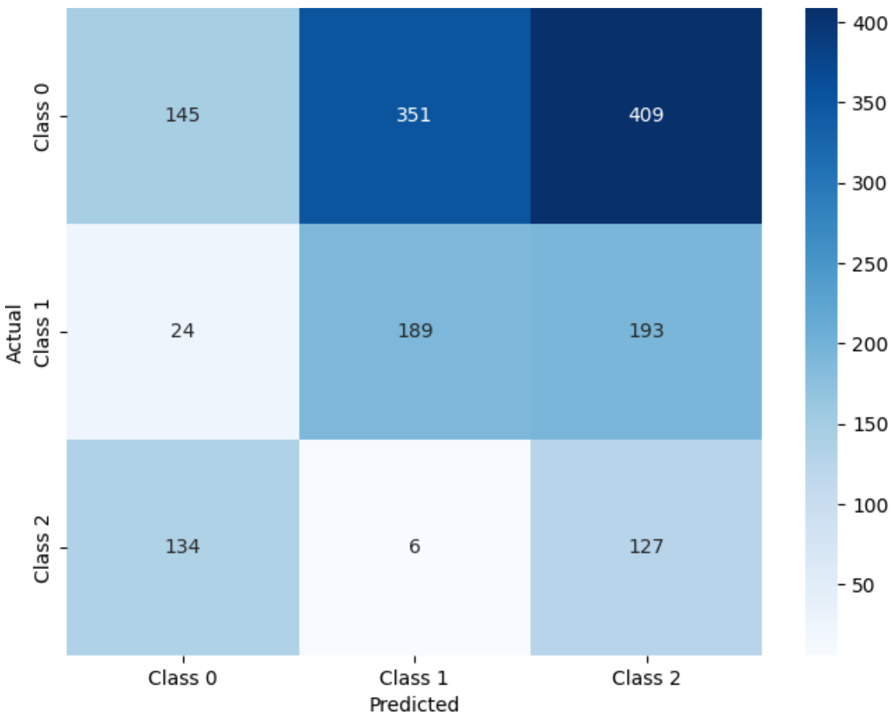


Figure 5. 5 Confusion Matrix for Multinomial Naive Bayes Classifier: Breast UltrasoundImages Dataset

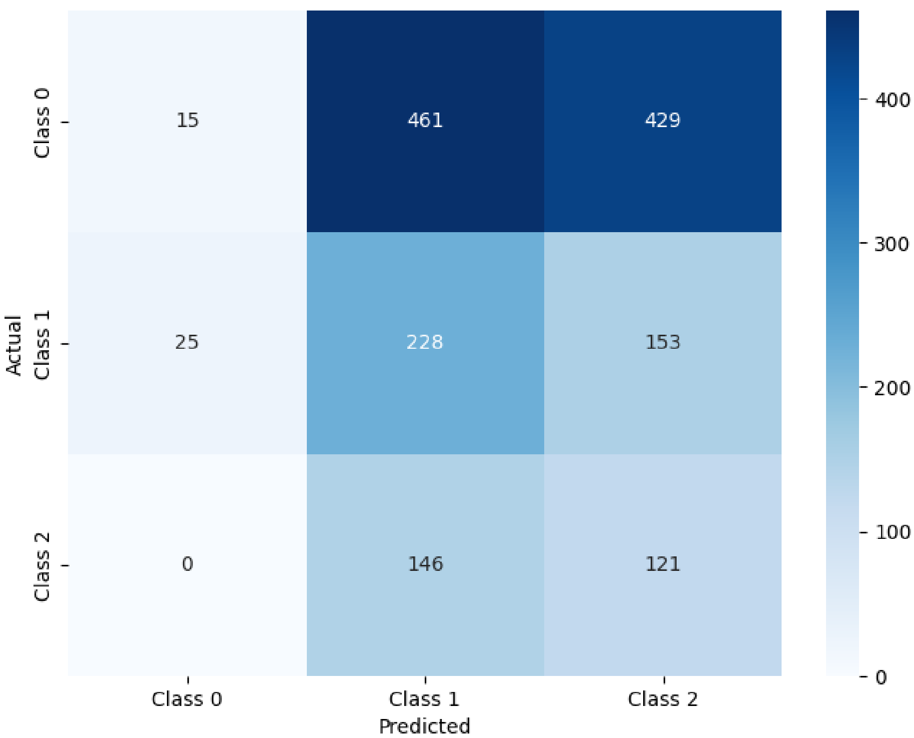


Figure 5. 6 Confusion Matrix for Bernoulli Naive Bayes Classifier: Breast Ultrasound Im-ages Dataset

We have compared all the results together and have shown them in Table 5.9 for Wis-consin dataset and n Table 5.10 for Breast Ultrasound Images dataset. For the Wisconsin dataset, the Naive Bayes algorithm achieved its highest accuracy of 0.97 when implementedwith the Gaussian classifier. The highest Precision for class 0 was also achieved when usingthe Gaussian Classifier, with its maximum value of 0.96. Whereas for Class 1, Precision reached the maximum value of 1.00 for both the Gaussian and the Multinomial Classifier. The highest value of Recall for class 1 was also reached when using the Gaussian classifier,with the value of 0.93. For class 0 on the other hand, Recall scored 1.00 with all of the classifiers used. F-1 Score achieved its highest value of 0.98 for class 0 and 0.96 for class 1 whenimplemented with the Gaussian classifier as well. So, it can be said that for the Wisconsin Dataset, the Naive Bayes model performs best with Gaussian Classifier.

Table 5. 9 Performance of Naive Bayes with different Classifiers: Wisconsin Dataset

		Gaussian	Multinomial	Bernoulli
Accuracy		**0.97**	0.94	0.62
Precision	Class 0	**0.96**	0.91	0.62
	Class 1	**1.00**	**1.00**	0.00
Recall	Class 0	**1.00**	**1.00**	**1.00**
	Class 1	**0.93**	0.84	0.00
F-1 Score	Class 0	**0.98**	0.95	**0.77**
	Class 1	**0.96**	0.91	0.00

For the Breast Ultrasound Images dataset, the Naive Bayes algorithm achieved its highestaccuracy of 0.38 when implemented with the Gaussian classifier. The highest Precision for class 0 was also achieved when using the Gaussian Classifier, with its maximum value of 0.67. For Class 1, Precision also reached the maximum value of 0.38 when implemented with the Gaussian classifier. For Class 2, the highest value of Precision was 0.25, again when implemented with Gaussian classifier. The highest value of Recall for Class 0 was also reached when using the Gaussian classifier, with the value of 0.25. For class 1 on the other hand, Recall scored 0.56 with Bernoulli classifier. For Class 2, the highest value of Recall was 0.73, achieved with Gaussian classifier. F-1 Score achieved its highest value of 0.36 for Class 0, 0.40 for Class 1, and 0.37 for Class 2 when implemented with the Gaussianclassifier as well. So, it can be said that even for the Breast Ultrasound Images Dataset, the Naive Bayes model performs best with Gaussian Classifier.

Table 5. 10 Performance of Naive Bayes with different Classifiers: Breast Ultrasound Im-ages Dataset

		Gaussian	Multinomial	Bernoulli
Accuracy		**0.38**	0.29	0.23
Precision	Class 0	**0.67**	0.48	0.38
	Class 1	**0.38**	0.35	0.27
	Class 2	**0.25**	0.17	0.17
Recall	Class 0	**0.25**	0.16	0.02
	Class 1	0.43	0.47	**0.56**
	Class 2	**0.73**	0.48	0.45
F-1 Score	Class 0	**0.36**	0.24	0.03
	Class 1	**0.40**	**0.40**	0.37
	Class 2	**0.37**	0.26	0.25

5.1.3 Random Forest

Random Forest is the third Supervised method that is tested using the Breast Cancer Wis- consin Diagnostic (WDBC) Dataset, and Breast Ultrasound Images Dataset. It is a method that expects three different hyper-parameters:

1. N: Number of decision trees in the forest.
2. M: Maximum depth of trees.
3. min: Minimum number of samples required to split a node.

These hyper-parameters can be set either implicitly or explicitly. If we do not specify explicitly the values of these parameters, they take default values. We have tested the method with different values for these parameters, and we refer to each test case as: Default, Scenario1, and Scenario 2. Table 5.11 shows the parameter values we have used for the Random Forest model in each test case we have simulated.

Table 5. 11 Parameter values for each test case with Random Forest Model

Parameter	Default	Scenario 1	Scenario 2
N	100	1000	5
M	None	2	120
min	2	5	10

First, we have tested the Random Forest model without explicitly specifying the hyper- parameters. The accuracy of the method with default hyper-parameter' values when tested with Wisconsin dataset is 0.96. The needed time to train the model on this dataset is 0.2358seconds, and the time needed for prediction is 0.0092 seconds. The Confusion Matrixfor Random Forest Classifier with Default Hyper-parameter values, tested with Wisconsin dataset is shown in Table 5.12.

Table 5. 12 Confusion Matrix for Random Forest Classifier with Default Hyper-parameter values: Wisconsin Dataset

	Predicted Negative	**Predicted Positive**
Actual Negative	70	1
Actual Positive	3	40

We have tested again the Random Forest model with default hyper-parameter values with Breast Ultrasound Images Dataset. The accuracy of the method with default hyper- parameter' values when tested with this dataset is 0.96958. The needed time to train the model on this dataset is 67.1449 seconds, and the time needed for prediction is 0.1950 seconds. The Confusion Matrix for Random Forest Classifier with Default Hyper-parameter values, tested with Breast Ultrasound Images Dataset is shown in Figure 5.7.

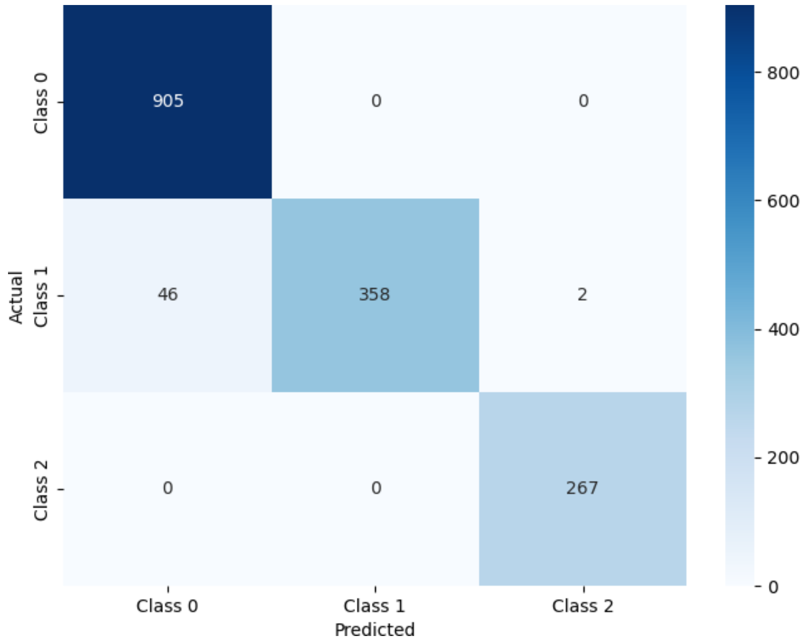


Figure 5. 7 Confusion Matrix for Random Forest Classifier with Default Hyper-parameter values: Breast Ultrasound Images Dataset

Then we have tested again the Random Forest Classifier, this time by explicitly specifying the values of the hyper-parameters in Scenario 1. The training time of the Method in Scenario 1 when tested with Wisconsin dataset is increased considerably. From an initial time of 0.2358 seconds with default parameters, with explicitly set parameters it reached 3.0435 seconds. The time needed for prediction is 0.0916 seconds. All the other evaluation metrics, including the accuracy do not seem to change. The accuracy of the model is again 0.96, and the Confusion Matrix of the model in this scenario can be seen in Table 5.13.

Table 5. 13 Confusion Matrix for Random Forest Classifier in Scenario 1: Wisconsin Dataset

	Predicted Negative	Predicted Positive
Actual Negative	70	1
Actual Positive	3	40

With the hyper-parameters in Scenario 1, we tested again the Random Forest Classifier,now with Breast Ultrasound Images dataset. The training time of the Method in Scenario 1 when tested with this dataset is 96.1781 seconds. The time needed for prediction is 0.3137 seconds. The accuracy of the model is reduced considerably, reaching the value of 0.6102, and the Confusion Matrix of the model in this scenario can be seen in Figure 5.8.

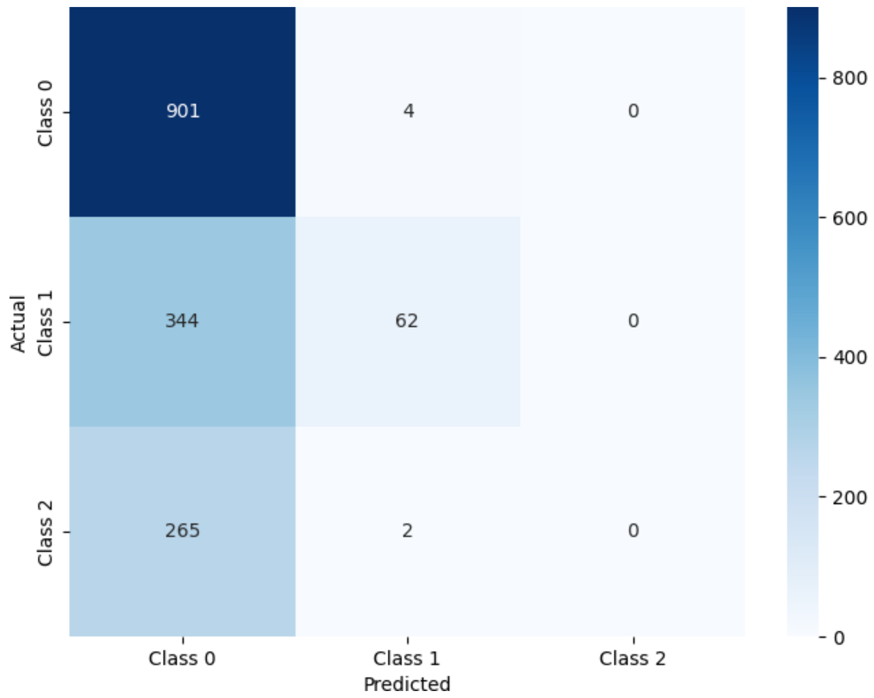


Figure 5. 8 Confusion Matrix for Random Forest Classifier in Scenario 1: Breast UltrasoundImages Dataset

We tested again the model with different parameter values, now with those in Scenario 2. The training time of the Method in Scenario 2 with Wisconsin dataset is reduced consid-erably. From an initial time of 0.2358 seconds with default parameters, to 3.0435 seconds in Scenario 1, now it reached 0.0440 seconds. The time needed for prediction is 0.0034 seconds. The accuracy in this scenario is increased to 0.97, and its Confusion Matrix can beseen in Table 5.14.

Table 5. 14 Confusion Matrix for Random Forest Classifier in Scenario 2: Wisconsin Dataset

	Predicted Negative	**Predicted Positive**
Actual Negative	70	1
Actual Positive	2	41

Random Forest Classifier is tested again with parameter values in Scenario 2, by using the Breast Ultrasound Images Dataset. The training time of the Method in Scenario 2 with this dataset is 6.9669 seconds. The time needed for prediction is 0.4293 seconds. The accuracy in this scenario is 0.9378, and its Confusion Matrix can be seen in Figure 5.9.

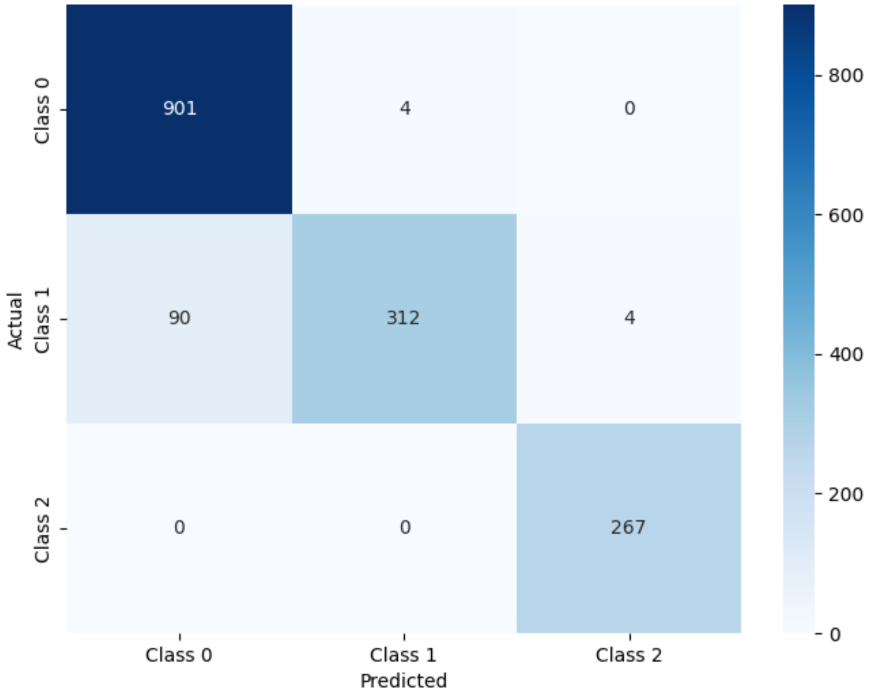


Figure 5. 9 Confusion Matrix for Random Forest Classifier in Scenario 2: Breast UltrasoundImages Dataset

We have compared all the results of the Random Forest testings for both Wisconsin andBreast Ultrasound Images dataset, and have shown them in Table 5.15 and 5.16. Random Forest has performed better in Scenario 2 for the Wisconsin Dataset, in term of all the evaluation metrics used.

Table 5. 15 Performance of Random Forest with different Parameter Values: Wisconsin Dataset

		Default	Scenario 1	Scenario 2
Accuracy		0.96	0.96	**0.97**
Precision	Class 0	0.96	0.96	**0.97**
	Class 1	**0.98**	**0.98**	**0.98**
Recall	Class 0	**0.99**	**0.99**	**0.99**
	Class 1	0.93	0.93	**0.95**
F-1 Score	Class 0	0.97	0.97	**0.98**
	Class 1	0.95	0.95	**0.96**

Table 5. 16 Performance of Random Forest with different Parameter Values: Breast Ultra-sound Images Dataset

		Default	Scenario 1	Scenario 2
Accuracy		**0.96**	0.61	0.94
Precision	Class 0	**0.95**	0.6	0.91
	Class 1	**1.00**	0.91	0.99
	Class 2	**0.99**	0.0	**0.99**
Recall	Class 0	**1.00**	**1.00**	**1.00**
	Class 1	**0.88**	0.15	0.77
	Class 2	**1.00**	0.00	**1.00**
F-1 Score	Class 0	**0.98**	0.75	0.95
	Class 1	**0.94**	0.26	0.86
	Class 2	**1.00**	0.00	0.99

5.1.4 Support Vector Machine

Support Vector Machine is the last method that falls under the Supervised Learning modelsin this Thesis. We have tested it using both Wisconsin and Breast Ultrasound Images dataset, with 20% of the data used for testing, and 80% used for training. We have simulated again three scenarios for the SVM model: one using its default values, and two other scenarios byusing a combination of parameter values. Table 5.17 shows the parameter values for each test we have made.

Table 5. 17 Parameter values for each test case with SVM Model

Parameter	Default	Scenario 1	Scenario 2
C (Regularization Parameter)	1.0	100	50
Kernel	rbf	linear	poly
Gamma rbf	scale	0.0	0.0
Polynomial kernel coeff.	0.0	0.0	3
Class Weight	none	None	balanced

The accuracy of the model with default parameter values, tested with Wisconsin datasetis calculated to be 0.98. The time needed to train 80% of the WDBC dataset is 0.0021 seconds, and the time needed to test the rest 20% of the dataset is 0.0109. Table 5.18 shows the Confusion Matrix for Support Vector Machine tested under these conditions.

Table 5. 18 Confusion Matrix for SVM with Default Parameter Values: Wisconsin Dataset

	Predicted Negative	**Predicted Positive**
Actual Negative	46	2
Actual Positive	0	66

We tested again the Support Vector Machine model with default parameter values, withBreast Ultrasound Images dataset. Its accuracy with this dataset is calculated to be 0.91. The time needed to train 80% of the WDBC dataset is increased drastically to 1027.0339 seconds, and the time needed to test the rest 20% of the dataset is 617.6630. Figure 5.10 shows the Confusion Matrix for Support Vector Machine tested under these conditions.

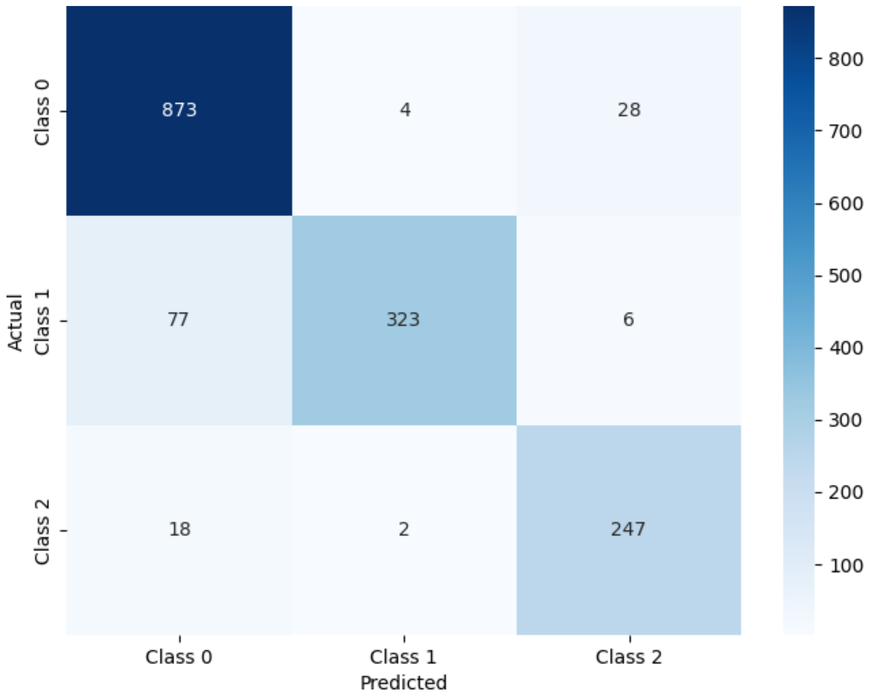


Figure 5. 10 Confusion Matrix for SVM with Default Parameter Values: Breast UltrasoundImages Dataset

After calculating the accuracy of the Support Vector Machine model with default parameter values, we have then changed these values into Scenario 1.

The accuracy of the model with these explicitly set parameter values for Wisconsin dataset is 0.97. The training time is 0.0086 seconds, and the prediction time is 0.0123 seconds. Table 5.19 shows the Confusion Matrix for Support Vector Machine tested under the conditions in Scenario 1.

Table 5. 19 Confusion Matrix for SVM in Scenario 1: Wisconsin Dataset

	Predicted Negative	**Predicted Positive**
Actual Negative	46	2
Actual Positive	1	65

The accuracy of the model with the explicitly set parameter values for Breast UltrasoundImages dataset in Scenario 1 is 1.00. The training time is 979.8132 seconds, and the prediction time is 375.9158 seconds. Figure 5.11 shows the Confusion Matrix for Support Vector Machine tested with Breast Ultrasound Images dataset under the conditions in Scenario 1.

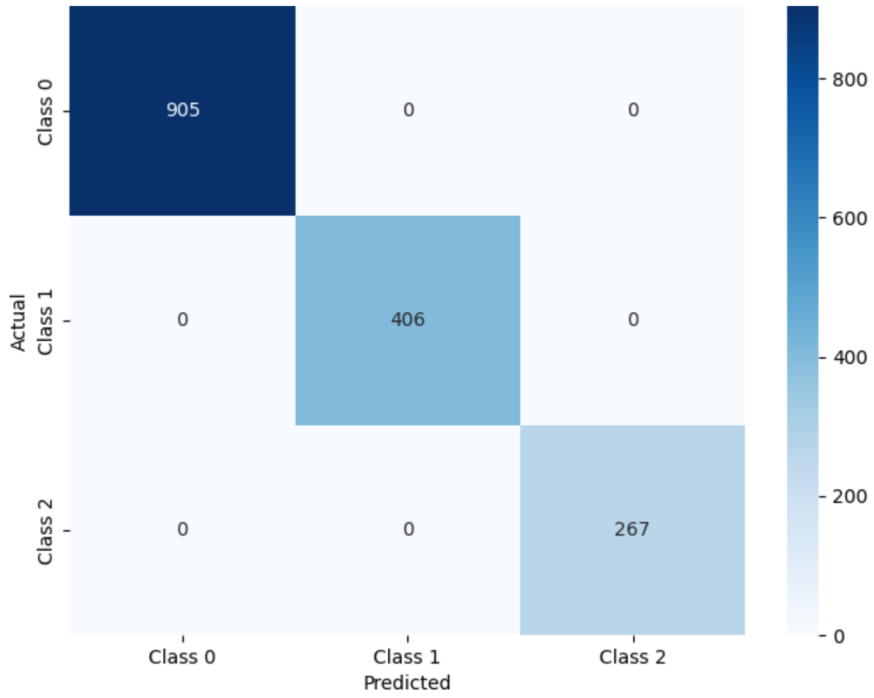


Figure 5. 11 Confusion Matrix for SVM win Scenario 1: Breast Ultrasound Images Dataset

We simulated Scenario 2 for the SVM model, and its accuracy with these explicitly set parameter values for the Wisconsin dataset is 0.96. The training time is 0.0026 seconds, andthe prediction time is 0.0103 seconds. Table 5.20 shows the Confusion Matrix for Support Vector Machine tested under the conditions in Scenario 2.

Table 5. 20 Confusion Matrix for SVM in Scenario 2

	Predicted Negative	**Predicted Positive**
Actual Negative	46	2
Actual Positive	2	64

We have compared all the results of the Support Vector Machine model for Wisconsin dataset, and have shown them in Table 5.21. The results show that for this dataset, Support Vector Machine model performs best with default parameter values.

Table 5. 21 Performance of SVM with different Parameter Values: Wisconsin Dataset

		Default	Scenario 1	Scenario 2
Accuracy		**0.98**	0.97	0.96
Precision	Class 0	**1.00**	0.98	0.96
	Class 1	**0.97**	**0.97**	**0.97**
Recall	Class 0	**0.96**	**0.96**	**0.96**
	Class 1	**1.00**	0.98	0.97
F-1 Score	Class 0	**0.98**	0.97	0.96
	Class 1	**0.99**	0.98	0.97

5.1.5 Performance comparison for supervised learning methods

From all the supervised methods tested with Wisconsin dataset with 80% of the data used fortraining, and 20% used for testing, Support Vector Machine outperformed the other modelswith an accuracy of 0.98. When changing the dataset separation to 60% used for training, and 40% used for testing, SVM still outperformed the other models, and its accuracy increased to 0.99. Table 5.33 compares the accuracy of all the supervised methods tested in all scenarios.

Table 5. 22 Accuracy comparison of Supervised Learning models for Wisconsin dataset

	Scenario 1		Scenario 2		Scenario 3	
	20%-80%	40%-60%	20%-80%	40%-60%	20%-80%	40%-60%
KNN	0.93	0.93	0.96	0.96	0.96	0.97
Naive Bayes	0.97	0.95	0.94	0.93	0.62	0.65
Random Forest	0.96	0.97	0.96	0.96	0.97	0.95
SVM	0.98	**0.99**	0.97	0.98	0.96	0.34

5.2 Results of Unsupervised Methods for Breast Cancer Detection

This section provides the results of two unsupervised methods analyzed in the thesis with different parameter values. The methods that are tested are: Auto Encoder, and Self OrganizingMaps.

5.2.1 Auto Encoder

The Auto Encoder algorithm is tested using the Breast Cancer Wisconsin Diagnosis (WDBC) dataset, with pre-processing techniques explained in Section 5.1.

First it is created one input layer in order to retrieve the data. Since this dataset has30 features that come as input to the algorithm, it is created an input layer with 30 input nodes, where each node represents one feature of the dataset. Then the input data is encodedusing a dense layer with 3 nodes and the ReLu activation function. The encoding of the data transforms it into a lower-dimensional representation, with only 3 dimensions. This is known as the hidden layer. In order to reconstruct again the original input after it has been encoded, the algorithm uses after the encoding layer another dense layer with 30 nodes anda sigmoid activation function.

The optimization algorithm that is used is Adam optimizer with a 0.01 learning rate. Theloss function is set to MSE (Mean Squared Error). The way how Auto Encoders are trained isby iterating and iterating multiple times through the entire dataset. In every single iteration,the methods tries to learn the features and the characteristics of the dataset, and then uses this information during the testing phase. One complete pass by the model through the entiredataset is known as an epoch. We have trained the Auto Encoder model with 500 epochs, so the model makes 500 iterations through the entire dataset. We are referring to the above scenario as **Scenario 1** when interpreting the results of the Auto Encoder. We have simulated two other scenarios for the Auto Encoder model, and the parameter values for each scenarioare given in Table 5.23.

Table 5. 23 Parameter values for each test case with Auto Encoder Model

Parameter	Scenario 1	Scenario 2	Scenario 3
Input layer nodes	30	30	30
Hidden layer nodes	3	10	15
Output layer nodes	30	30	30
Input activation function	reLu	Sigmoid	Sigmoid
Output activation function	Sigmoid	Sigmoid	Tanh
Optimization algorithm	Adam	Adam	Adam
Learning rate	0.01	0.02	0.02
Loss function	MSE	MSE	MSE
Epochs	500	250	500

In order to calculate the accuracy and other evaluation metrics of the model by using theencoded representation, instead of the real input data, we have used the KNN model. The accuracy of the model under Scenario 1 is calculated to be 0.97. This algorithm takes more time to be trained, in comparison with Supervised Algorithms that are tested and explained above. The time it needs to be trained is 37.0435 seconds, and the time it takes to predict theresults is 0.0083 seconds. The Confusion Matrix for Auto Encoder in Scenario 1is shown graphically in Table 5.24.

Table 5. 24 Confusion Matrix for Auto Encoder in Scenario 1

	Predicted Negative	**Predicted Positive**
Actual Negative	54	1
Actual Positive	2	34

The model loss of the Auto Encoder in Scenario 1 is shown in Figure 5.12

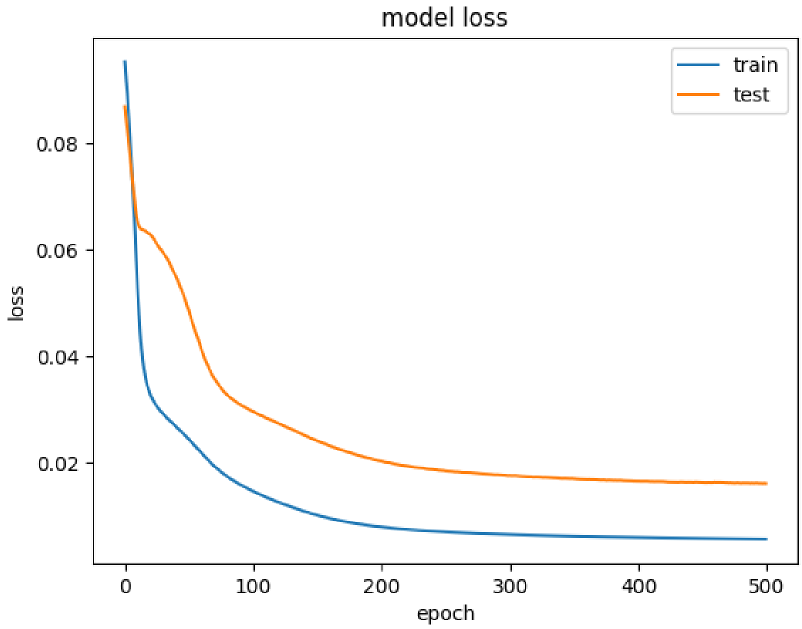


Figure 5. 12 Model Loss for Auto Encoder in Scenario 1

We have tried to change the number of layers and other parameters of the model, now with the values in Scenario 2. The model loss of the Auto Encoder in Scenario 2 is shown inFigure 5.13

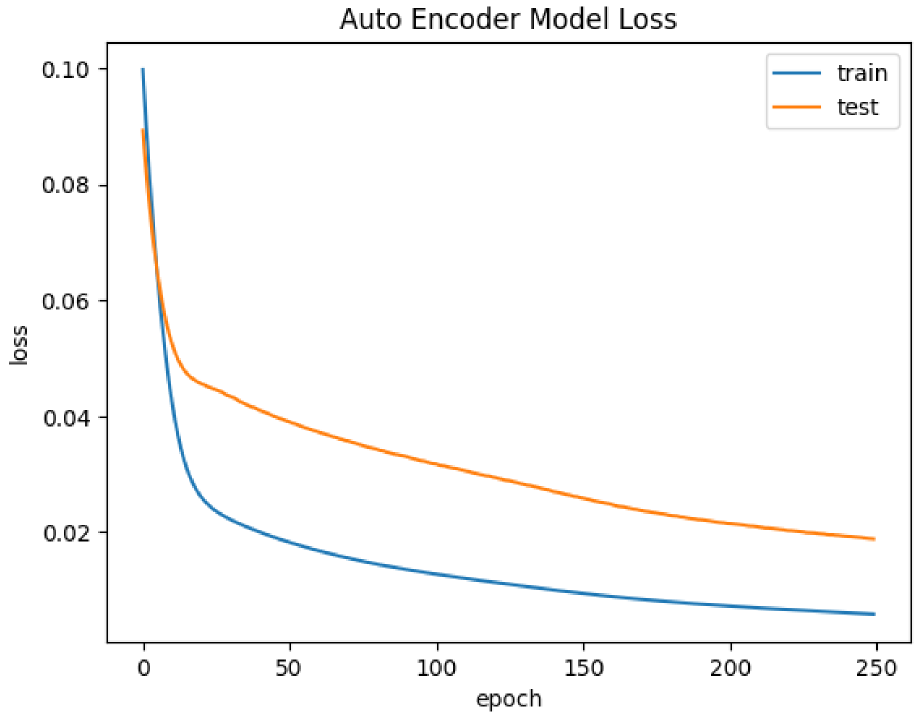


Figure 5. 13 Model Loss for Auto Encoder in Scenario 2

The accuracy of the model under the conditions in Scenario 2 is increased by 1% in comparison with Scenario 1, with the value 0.98. The Confusion Matrix of this method in Scenario 2 is shown graphically in Table 5.25.

Table 5. 25 Confusion Matrix for Auto Encoder in Scenario 2

	Predicted Negative	**Predicted Positive**
Actual Negative	54	1
Actual Positive	1	35

Auto Encoder model is tested again in Scenario 3. The model loss of the Auto Encoderin Scenario 3 is shown in Figure 5.14

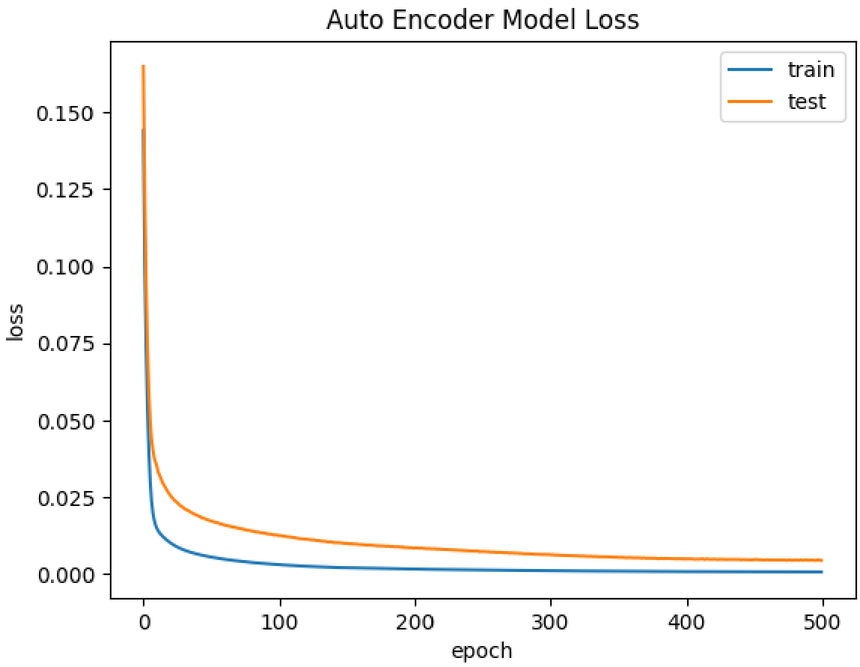


Figure 5. 14 Model Loss for Auto Encoder in Scenario 3

The accuracy of the model under the conditions in Scenario 3 equals the accuracy of themodel in Scenario 1. The Confusion Matrix of this method in Scenario 3 is shown graphically in Table 5.26.

Table 5. 26 Confusion Matrix for Auto Encoder in Scenario 3

	Predicted Negative	**Predicted Positive**
Actual Negative	55	0
Actual Positive	3	33

We have compared all the results of the Auto Encoder model testings and have shown them in Table 5.27. For the Wisnonsin Dataset, the Auto Encoder model performs best withparameters in Scenario 2.

Table 5. 27 Comparison of the Performance of Auto Encoder model

		Scenario 1	Scenario 2	Scenario 3
Accuracy		0.97	**0.98**	0.97
Precision	Class 0	0.96	**0.98**	0.95
	Class 1	0.97	0.97	**1.00**
Recall	Class 0	0.98	0.98	**1.00**
	Class 1	0.94	**0.97**	0.92
F-1 Score	Class 0	0.97	**0.98**	0.97
	Class 1	0.96	**0.97**	0.96

5.2.2 Self Organizing Maps

Self-Organizing Maps (SOM) is the second Unsupervised Learning Algorithm that is testedusing WDBC dataset. To test the Self Organizing Map model in Python, it is necessary to install the minisom library. SOMs work as grids and expect the values for width and height.We have set both the dimensions of the Self Organizing Maps to be 10 units/neurons. The other parameter that needs to be defined for the SOM to work properly and to generate efficient results, is σ , which determines the influence that the neighboring neurons have during weight updates. We have first set the σ to be 1. In addition, we have set the learning rate (α) of the algorithm to 0.5, which means that the weights of the model are adjusted by 50% during training based on the input data. SOM algorithm works with multiple iterations/epochs through the entire dataset, and in the first Scenario **(Scenario 1)** we have decided to work with 500 iterations. During each iteration through the dataset, it is computed the distance between the input space X and all the code words. The code word with the smallest distance is then selected, and it is known as the winner unit/neuron or best matching unit (BMU).

We refer to the conditions mentioned above for the Self Organizing Map model as **Scenario 1**. Three scenarios are simulated in total for SOM model, and the parameter values foreach scenario can be seen in Table 5.28.

Table 5. 28 Parameter values for each test case with SOM Model

Parameter	Scenario 1	Scenario 2	Scenario 3
Grid Size	10x10	15x15	20x20
Sigma (σ)	0.5	1.5	1
Learning rate (α)	1	0.8	0.6
Epochs	500	250	350

We have tested the SOM model in Scenario 1. The time it takes the model to be trainedis 0.0461 seconds, and the time it takes to predict the results is 0.0180 seconds.

Figure 5.15 shows the MID of the SOM model tested in Scenario 1.

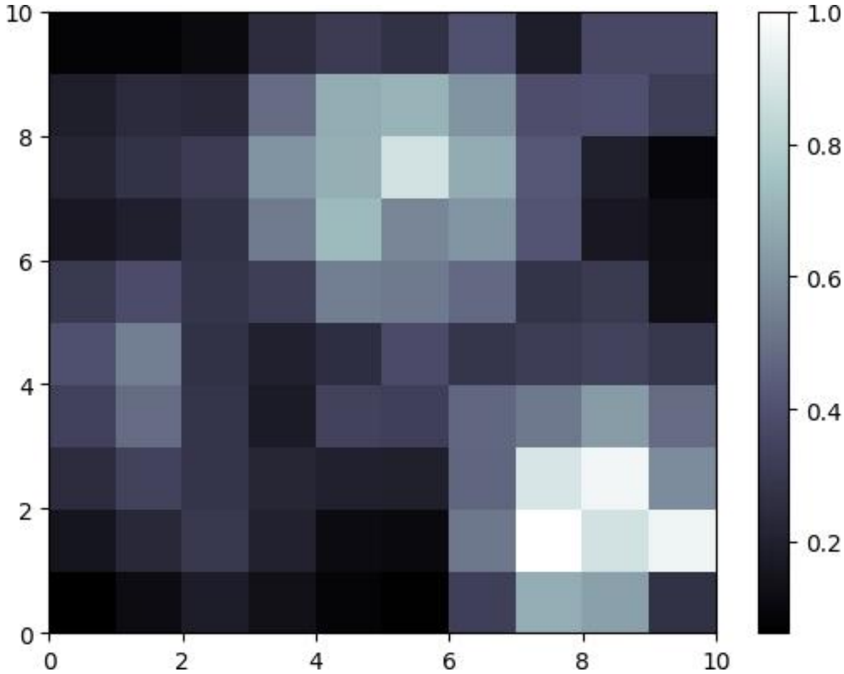


Figure 5. 15 MID of the SOM model in Scenario 1

After the SOM model is trained in Scenario 1, it has generated the labels in Figure 5.16. In this figure, red circles represent Class 0 and green squares represent Class 1.

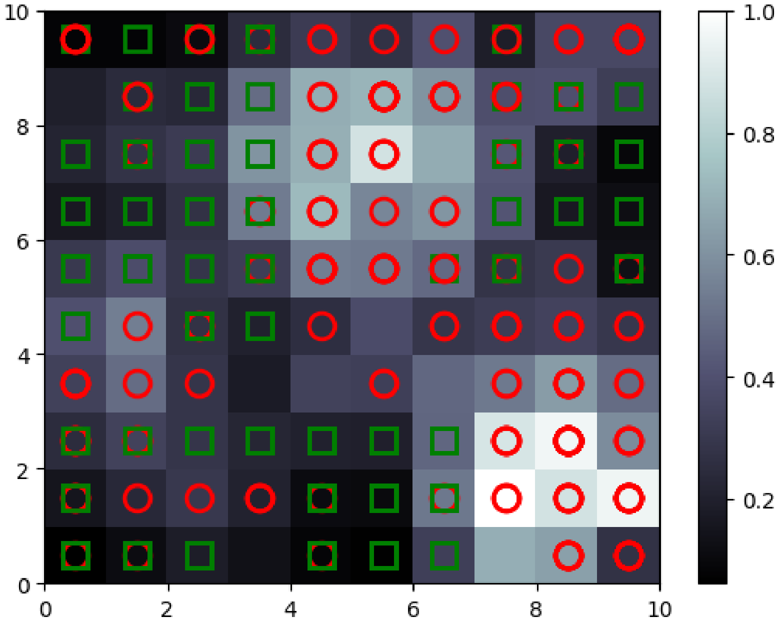


Figure 5. 16 U-matrix visualization of the SOM model in Scenario 1

Since SOM is an unsupervised machine learning model, whose task is to find the most meaningful features of the data, we have incorporated it with KNN classifier with K-value=5to calculate the accuracy and other evaluation metrics of the model. So the representation ofthe input data that is generated by the SOM model is compared with the real input data fromthe Wisconsin dataset, and from this comparison are calculated the Evaluation Metrics. Theaccuracy of the model in Scenario 1 is calculated to be 0.91. Table 5.29 shows the ConfusionMatrix of the SOM in Scenario 1.

Table 5. 29 Confusion Matrix for SOM in Scenario 1

	Predicted Negative	**Predicted Positive**
Actual Negative	41	6
Actual Positive	4	63

The results of the SOM model in Scenario 1 are not very satisfying, and we have tried tochange the parameters of the model in order to improve its performance with those values inScenario 2.

The accuracy of the SOM in Scenario 2 increased by 1% in comparison with the accuracyin Scenario 1. In Scenario 2 the accuracy is 0.92. The Confusion Matrix for SOM in Scenario2 is shown in Table 5.30.

Table 5. 30 Confusion Matrix for SOM in Scenario 2

	Predicted Negative	**Predicted Positive**
Actual Negative	42	5
Actual Positive	4	63

Figure 5.17 shows the MID of the SOM model tested in Scenario 2, and Figure 5.18 shows the U-matrix visualization of the SOM model, so the labels it has generated for the unlabeled dataset.

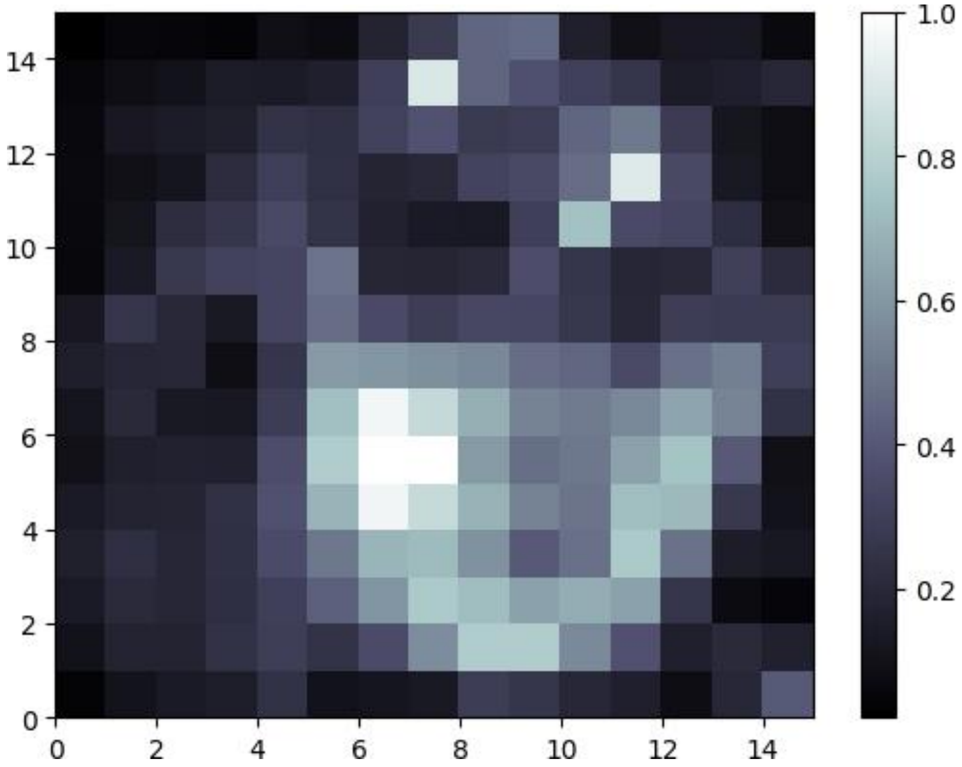


Figure 5. 17 MID of the SOM model in Scenario 2

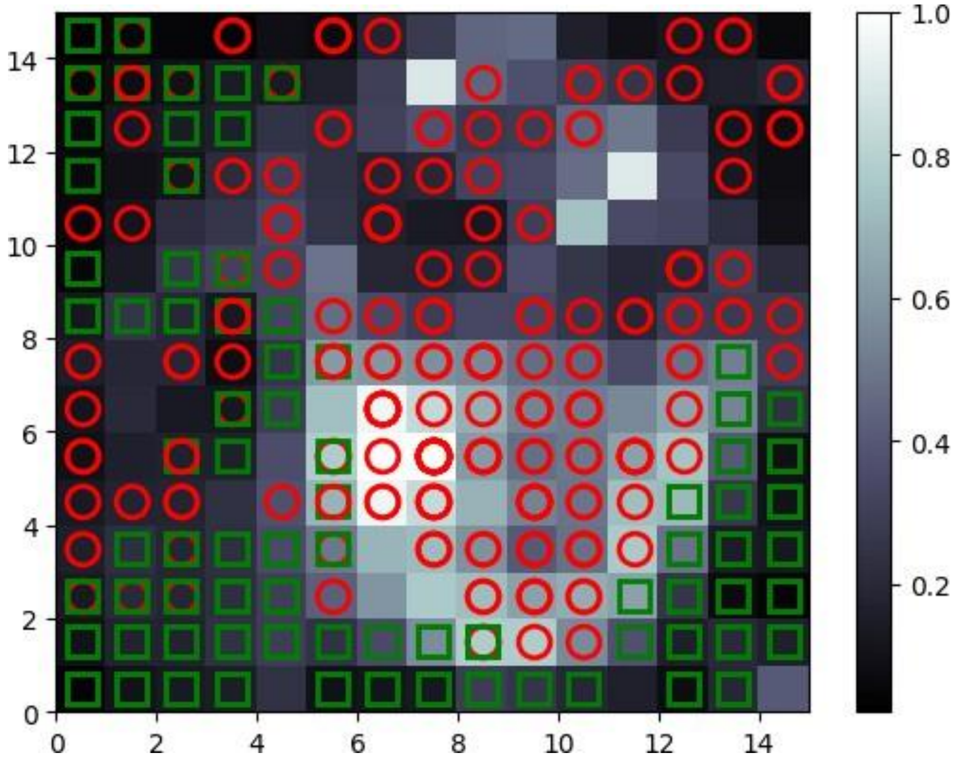


Figure 5. 18 U-matrix visualization of the SOM model in Scenario 2

We have tested once again the model in Scenario 3. The accuracy of the SOM model inScenario 3 is calculated to be 0.76, so lower than the accuracy in Scenario 1 and Scenario 2.Table 5.31 shows the Confusion Matrix for this model in Scenario 3.

Table 5. 31 Confusion Matrix for SOM in Scenario 3

	Predicted Negative	**Predicted Positive**
Actual Negative	27	20
Actual Positive	7	60

Figure 5.19 shows the MID of the SOM model tested in Scenario 3, and Figure 5.20 shows the U-matrix visualization of the SOM model, so the labels it has generated for the unlabeled dataset.

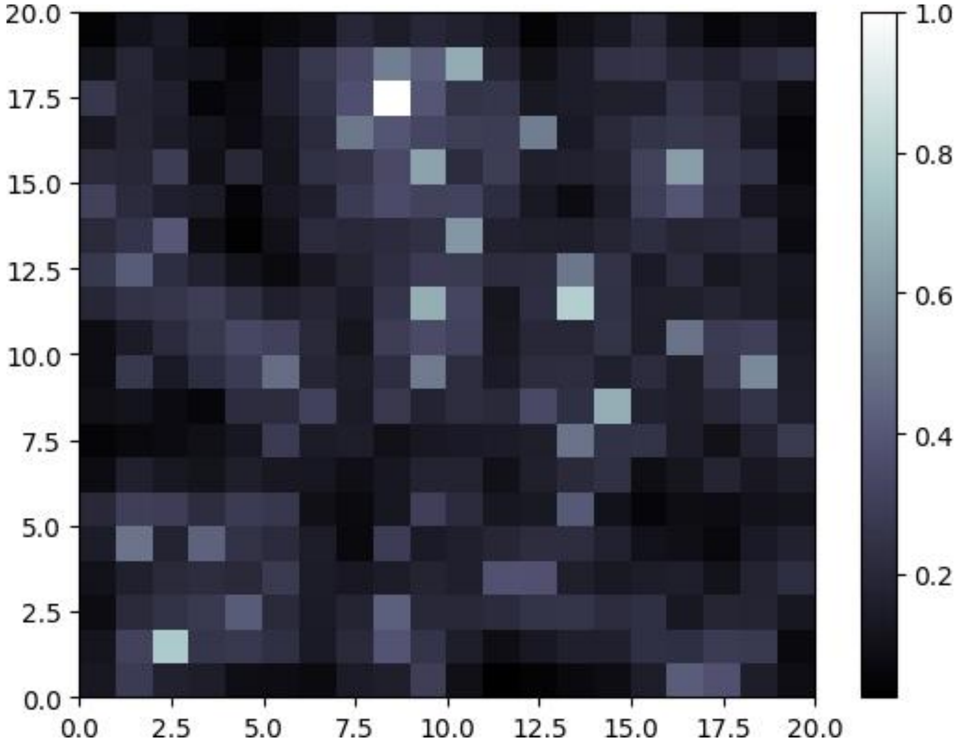


Figure 5. 19 MID of the SOM model in Scenario 3

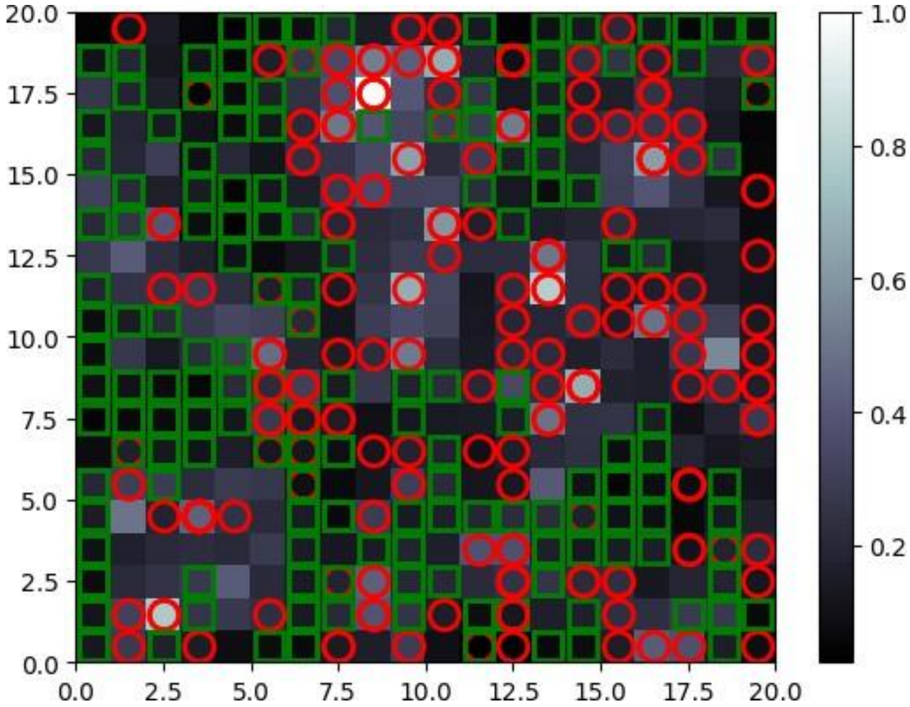


Figure 5. 20 U-matrix visualization of the SOM model in Scenario 3

After we have tested the SOM model in the three Scenarios explained above, we havemade a comparison in order to understand under what conditions/parameter values the SOM model performs best with Wisconsin Dataset. As it can be seen in Table 5.32 Scenario 2 has maximised the performance of the model in terms all evaluation metrics used: Accuracy,Precision, Recall, and F-1 Score.

Table 5. 32 Comparison of the Performance of SOM model

		Scenario 1	Scenario 2	Scenario 3
Accuracy		0.91	**0.92**	0.76
Precision	Class 0	**0.91**	**0.91**	0.79
	Class 1	0.91	**0.93**	0.75
Recall	Class 0	0.87	**0.89**	0.57
	Class 1	**0.94**	**0.94**	0.90
F-1 Score	Class 0	0.89	**0.90**	0.67
	Class 1	**0.93**	**0.93**	0.82

5.2.3 Performance comparison for unsupervised learning methods

Within the two unsupervised methods tested with Wisconsin dataset with 80% of the data used for training, and 20% used for testing, Auto Encoder model outperformed SOM with an accuracy of 0.98. We performed again all experiments for unsupervised learning modelsusing a different split of the Wisconsin dataset, this time 60% for training, and 40% for testing. Table 5.35 compares the accuracy of Auto Encoder and SOM tested in all scenarios,with each dataset split. The highest value of accuracy is achieved in Scenario 2 by the Autoencoder model, when tested with a dataset split of 80% and 20%. With this separation of thedata points within the dataset, Auto Encoder model maximised its accuracy to 0.98.

Table 5. 33 Accuracy comparison of Unsupervised Learning models for Wisconsin dataset

	Scenario 1		**Scenario 2**		**Scenario 3**	
	20%-80%	40%-60%	20%-80%	40%-60%	20%-80%	40%-60%
Auto Encoder	0.97	0.91	**0.98**	0.94	0.97	0.93
SOM	0.91	0.91	0.92	0.85	0.76	0.65

5.3 Results of CNN models for Breast Cancer Detection

This section provides the results of two CNN models for Breast Cancer Detection: UNet andResNet.

5.3.1 ResNet

We have tested the ResNet model that we built by using 20% of the data for testing, and 80%for training, and the used training parameters are:

- Batch size: 16
- Epochs: 30
- Patience: 4
- Optimizer: Adam
- Loss function: categorical crossentropy
- Evaluation metric: Accuracy

Under these conditions, the proposed model achieved a training accuracy of 93.18%, anda validation accuracy of 80.80%. The required time to train the model was 28.07 seconds, and the required time to test the model was 30.44 seconds. Figure 5.21 shows graphically the history of model's accuracy and loss.

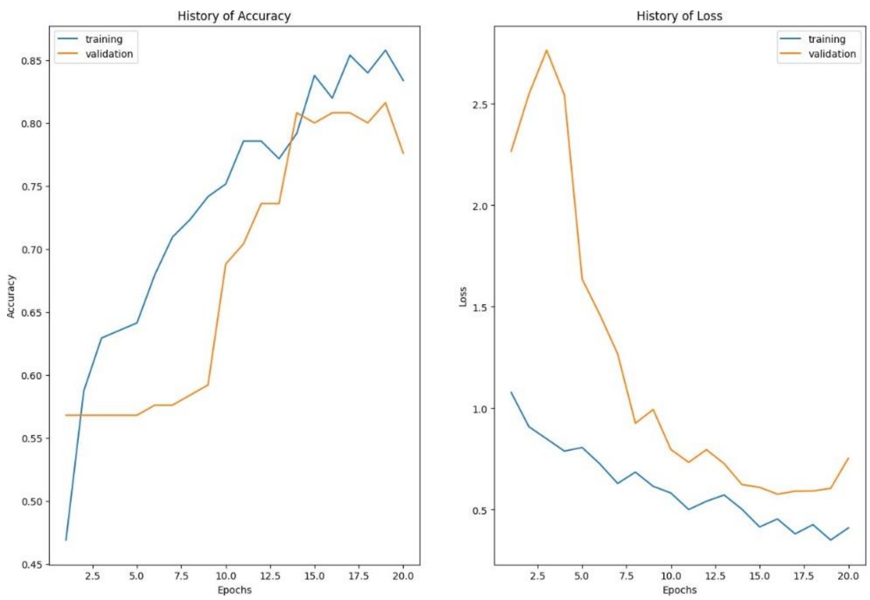


Figure 5. 21 ResNet Accuracy and Loss

5.3.2 U-Net Model

We have compiled the UNet model with different number of epochs, and Adam optimizer with 0.00005 learning rate. We have used MSE for the model' loss, and accuracy to evaluate the performance of the model. The training accuracy of the model under these conditions achieved its maximum value of 0.9867 with 80 epochs, whereas the maximum value for thevalidation accuracy was 0.9744 with 60 epochs. The model needed 800.2316 seconds to betrained, and 1.7739 seconds to predict 156 test images with 60 epochs. Figure 5.22 shows themodel accuracy, and Figure 5.23 shows the model loss for 60 epochs, chosed as the numberof epochs that maximised the validation accuracy of the model.

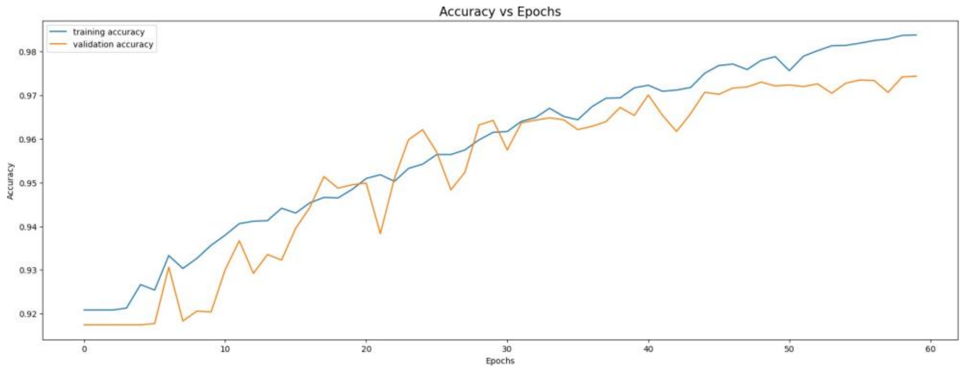


Figure 5. 22 UNet Model Accuracy with 60 epochs

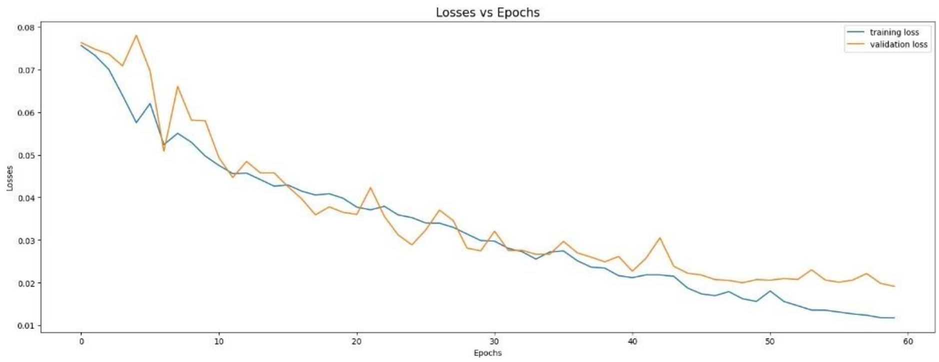


Figure 5. 23 UNet Model Loss with 60 epochs

All the results of the UNet model, tested under different number of epochs, using theBreast Ultrasound Images Dataset are shown in Table 5.34.

Table 5. 34 Performance Evaluation of UNet Model with different number of epochs

	40 epochs	**50 epochs**	**60 epochs**	**80 epochs**
Validation Acc.	0.9661	0.9681	**0.9744**	0.9719
Validation Loss	0.0271	0.0243	**0.0192**	0.0223

To visualize the results of the UNet model with the Breast Ultrasound Images dataset, we are providing some samples of original images, their respective masks, and the predictions that UNet model has made for each image. These results can be seen in Figure 5.24, 5.25, 5.26, 5.27, 5.28, 5.29.

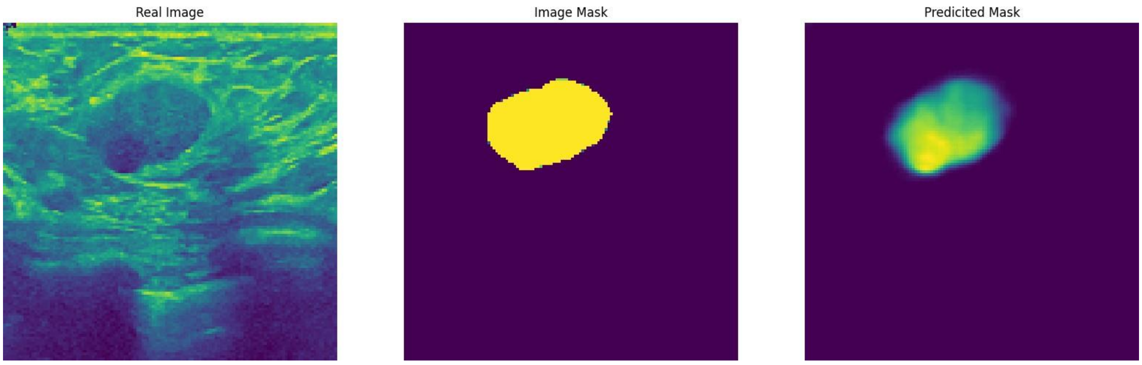


Figure 5. 24 UNet Results: Single benign image, its mask, and UNet's prediction

Figure 5. 25 UNet Results: Single malignant image, its mask, and UNet's prediction

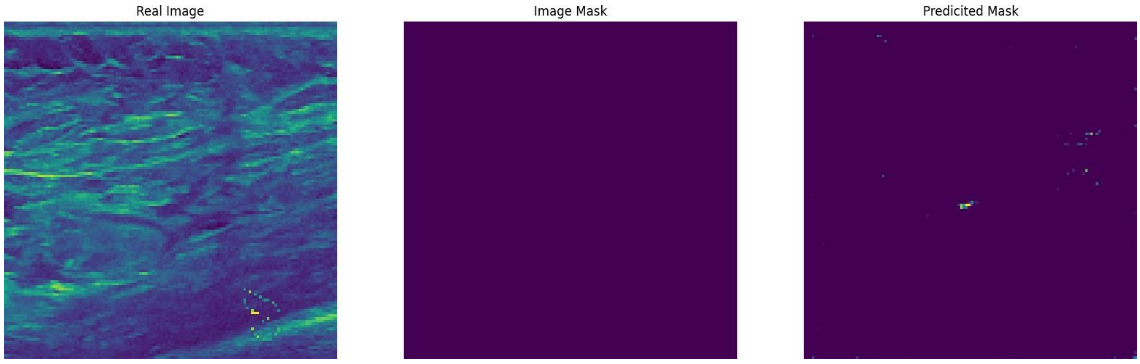


Figure 5. 26 UNet Results: Single normal image, its mask, and UNet's prediction

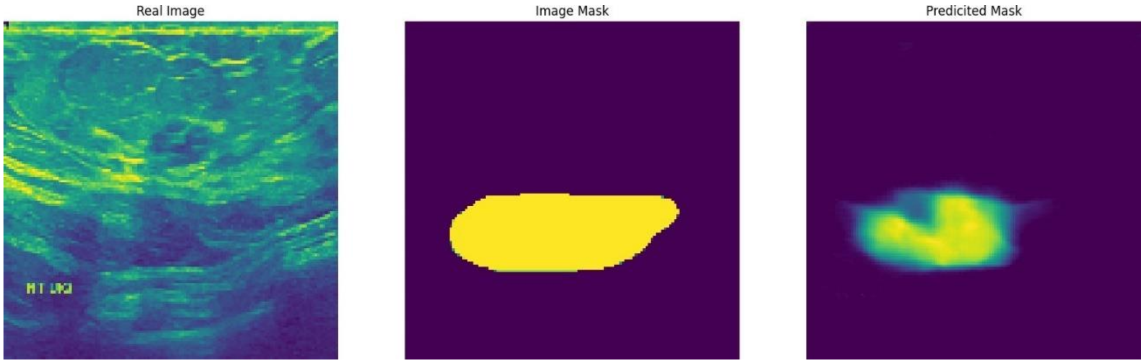


Figure 5. 27 UNet Results: Single benign image, its mask, and UNet's prediction

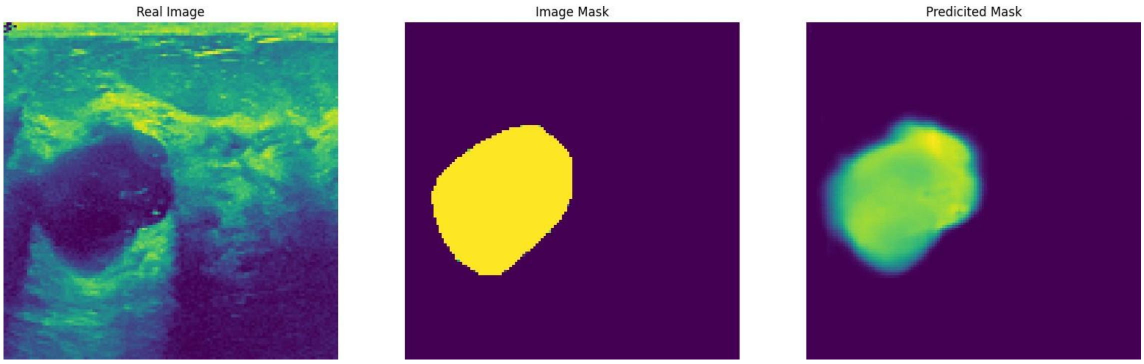


Figure 5. 28 UNet Results: Single benign image, its mask, and UNet's prediction

5.3.3 Performance comparison for CNN models

Both UNet and ResNet models are tested with Breast Ultrasound Images dataset. The re- sults for each model, implemented with the architecture and parameters explained above, are shown in Table 5.35. UNet model achieved higher accuracy in comparison with ResNetmodel.

Table 5. 35 Accuracy comparison for CNN models with Breast Ultrasound Images Dataset

	UNet	**ResNet**
Validation Accuracy	0.9661	0.9681
Validation Loss	0.0271	0.0243

CHAPTER 6

DISCUSSION

In this chapter we discuss and analyze the results of the Thesis, as well as the limitations wehave faced.

6.1 Best Models for each Category

The best results within each category of deep learning methods in terms of accuracy, can beseen graphically in Table 6.1. Within four of the Supervised Learning models tested with Wisconsin numerical dataset for Breast Cancer Detection, Support Vector Machine achievedthe highest accuracy with the value 98% with 20%-80% dataset split, and the accuracy of 99% with 40%-60% dataset split with these combination of parameter values:

- **C (Regularization parameter)**: 1.0
- **Kernel**: rbf
- **Gamma (for RBF kernel)**: scale
- **Kernel Coefficient (for polynomial kernel)**: 0.0
- **Class Weight**: None

The required time for training the SVM model was 0.0021 seconds, and the required timefor testing was 0.0109 seconds.

From the Unsupervised Learning models tested again with Wisconsin numerical dataset,the model that achieved the highest accuracy was Auto Encoder, with the value 98%. This accuracy value was achieved using these combination of parameters:

- Nr. of input layer nodes: 30

- Nr. of hidden layer nodes: 10
- Nr. of output layer nodes: 30
- Input activation function: Sigmoid
- Output activation function: Sigmoid
- Optimization algorithm: Adam optimizer with a 0.02 learning rate
- Loss function: MSE (Mean Squared Error)
- Nr. of epochs: 250

From the CNN models, UNet outperformed ResNet with a validation accuracy of 97.44%.

Table 6. 1 The best model within each category of deep learning methods for Breast CancerDetection: The accuracy, training time, and testing time for each

Category	Model	Accuracy	Training Time (s)	Validating Time (s)
Supervised	SVM	98%	0.0021	0.0109
Unsupervised	Auto Encoders	98%	19.8015	0.1855
CNN	UNet	97.44%	800.2316	1.7739

6.2 Limitations of the study

Even though the results achieved in this Thesis seem to be promising, there is still placefor improvement. The biggest challenge and limitation we faced within this Thesis was the inability to find a more updated dataset, with a larger number of images and more diverse ones. The availability of such a dataset would make the models more general and ableto consider datasets of different size and characteristics. Nevertheless, the models would need to consider more features of the data, and would need to carefully distinguish the most significant features in order to prevent over-fitting. Thus, the performance of these models on larger datasets needs to be investigated.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Summary of findings and contributions

This Thesis analyzed the importance of Breast Cancer Detection for helping doctors in dis-ease diagnosis, without significant reliance in human interpretation. We considered several Deep Learning Techniques, divided them into three different categories, and proposed one best model for each category, suitable for different possible scenarios. If labeled data is avail-able, and human expertise to structure the data and represent it into meaningful numerical values is possible, we proposed Support Vector Machine as the best Supervised model for Breast Cancer Detection, which in this Thesis achieved an impressive classification accuracyof 99%. If human intervention for labeling data is not possible and the available dataset is unlabeled, yet numeric, we proposed Auto Encoders as the best Unsupervised model, whoseaccuracy also achieved the impressive result of 98%. If the available dataset consists of com- plex images, where feature extraction is complex and needs to be automated, we proposed UNet, which in this Thesis achieved the accuracy of 97.44%. The contribution of this Thesisto recent research in the field of Breast Cancer Detection lies in the practical insight that it provides for model selection, based on available data and human expertise. This is importantnot only to researchers, but also to clinicians, as a reference point in their ongoing battle against breast cancer.

7.2 Future Work

Despite promising results, the proposed models need to be investigated further with larger and more diverse datasets, with the aim of generalizing them to perform well even if the available data is complex and not structured. Future work could focus on enhancing the available datasets, as well as on deeper investigation for alternative evaluation metrics formore comprehensive model comparison.

REFERENCES

[1] N. S. Ismail and C. Sovuthy, "Breast cancer detection based on deep learning technique," in *2019 International UNIMAS STEM 12th Engineering Conference (EnCon)*, 2019, pp. 89–92, accessed: 29 November 2023.

[2] WorldHealth Organization, "Breast cancer," https://www.who.int/news-room/fact-sheets/detail/breast-cancer#:~:text=Breast%20cancer%20is%20a%20disease, the%20body%20and%20become%20fatal., accessed: 08 December 2023.

[3] American Cancer Society, "Survival rates for breast cancer (2024) american cancer society," accessed: 22 May 2024.

[4] BreastCancer.org, "Breast cancer stages," https://www.breastcancer.org/ pathology-report/breast-cancer-stages, accessed: 08 December 2023.

[5] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 114–118, accessed: 20 January 2024.

[6] M. Li, "Research on the detection method of breast cancer deep convolutional neural network based on computer aid," in *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 2021, pp. 536–540, accessed: 05 March 2024.

[7] C. Bento, "Random forests algorithm explained with a real-life example and some python code," Jan 2022, "Accessed: 08 March 2024".

[8] G. Biau and E. Scornet, "(2016) a random forest guided tour," in *(2016) A Random Forest Guided Tour*, 2016, pp. 000 099–000 104, accessed: 28 January 2024.

[9] R. K. Barwal and N. Raheja, "A classification system for breast cancer prediction using svof-knn method," in *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2022, pp. 765–770, accessed: 12 January 2024.P. P. Mucherino, A. and P. Pardalos, "K-nearest neighbor classification," in *K-NearestNeighbor Classification*, 2021, pp. 000 099–000 104, accessed: 16 December 2023.

[10] S. G and R. G, "A novel and robust breast cancer classification based on histopatholog- ical images using naive bayes classifier," in *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, 2023,pp.

1–8, accessed: 12 January 2024.

[11] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, “Prediction of breast cancer using support vector machine and k-nearest neighbors,” in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 226–229, accessed: 09 May 2024.

[12] S. Tiwari, “Support vector machine machine learning al- gorithm with example and code,” 2023, accessed: 09 May 2024. [Online]. Available: https://www.codershood.info/2019/01/10/ support-vector-machine-machine-learning-algorithm-with-example-and-code/

[13] M. Chen and Y. Jia, “Support vector machine based diagnosis of breast cancer,” in *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2020, pp. 321–325, accessed: 12 May 2024.

[14] S. Naveen, N. V. Kashyap, V. P. Kulkarni, S. A, and M. S. Chakradhar, “Breast can- cer prediction using unsupervised learning technique k-means clustering algorithm,” in*2023 2nd International Conference on Vision Towards Emerging Trends in Communi-cation and Networking Technologies (ViTECoN)*, 2023, pp. 1–6, accessed: 13 January2024.

[15] “Flowchart of self evolving k means cluster algorithm,” accessed: 11 May 2024. [Online]. Available: https://mavink.com/post/89E95618B800C55242CCEB64964CFA8330AM4F8EAF/k-means-clustering-algorithm-in-matlab

[16] P. Praveen and B. Rama, “An empirical comparison of clustering using hierarchical methods and k-means,” in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2016, pp. 445–449, accessed: 11 May 2024.

[17] R. Radha and P. Rajendiran, “Using k-means clustering technique to study of breast cancer,” in *2014 World Congress on Computing and Communication Technologies*, 2014, pp. 211–214, accessed: 11 May 2024.

[18] S. W. Y. Bichen Zheng and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms,” *Expert Systems with Applications*, vol. 41, no. 4, Part 1, pp. 1476–1482, 2020, accessed: 12 May 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0957417413006659

[19] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, “Breast cancer diagnosis using an unsupervised feature extraction algorithm based on deep learning,” in *2018 37th Chinese Control*

Conference (CCC), 2018, pp. 9428–9433, accessed: 09 May 2024.

[20] A. B. O. V. Silva and E. J. Spinosa, “Graph convolutional auto-encoders for predicting novel lncrna-disease associations,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 4, pp. 2264–2271, 2022, accessed: 09 May 2024.

[21] M. P. K, M. Selvan, D. R, D. Suganthi, B. U. Maheswari, and P. Madan, “Use of self-organizing kohonen maps for quantization of tomography images,” in *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, 2023,pp. 1452–1455, accessed: 20 April 2024.

[22] A. E. Oprea, R. Strungaru, and G. M. Ungureanu, “A self organizing map approach to breast cancer detection,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 3032–3035, accessed: 14 May 2024.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016,pp. 770–778, accessed: 17 May 2024.

[24] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *ArXiv*, vol. abs/1505.04597, 2015, accessed: 18 May2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:3719281

[25] J. S. Suri, M. Bhagawati, S. Agarwal, S. Paul, A. Pandey, S. K. Gupta, L. Saba, K. I. Paraskevas, N. N. Khanna, J. R. Laird, A. M. Johri, M. K. Kalra, M. M. Fouda, M. Fatemi, and S. Naidu, “Unet deep learning architecture for segmentation of vascularand non-vascular images: A microscopic look at unet components buffered with prun-ing, explainable artificial intelligence, and bias,” *IEEE Access*, vol. 11, pp. 595–645, 2023, accessed: 18 May 2024.

[26] M. Robin, J. John, and A. Ravikumar, “Breast tumor segmentation using u-net,” in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1164–1167, accessed: 18 May 2024.

[27] B. S. Abunasser, M. R. J. AL-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, “Breast cancer detection and classification using deep learning xception algorithm,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022, accessed: 18 December 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.0130729

[28] S. S. Boudouh and M. Bouakkaz, “Breast cancer: Using deep transfer learning tech-niques alexnet convolutional neural network for breast tumor detection in mammog-

raphy images," in *2022 7th International Conference on Image and Signal Processing and their Applications (ISPA)*, 2022, pp. 1–7, accessed: 14 January 2024.

[29] A. Titoriya and S. Sachdeva, "Breast cancer histopathology image classification using alexnet," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, 2019, pp. 708–712, accessed: 19 December 2023.

[30] C.-J. . R. G. Sundqvist, M., "Adjusting the adjusted rand index. a multinomial story," 2023, p. 327–347, accessed: 11 May 2024. [Online]. Available: https://doi.org/10.1007/s00180-022-01230-7

[31] M.-O. S. N. Wolberg, William and W. Street, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1995, accessed: 07 Feb 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[32] W. Wolberg, "Breast Cancer Wisconsin (Original)," UCI Machine Learning Repository, 1992, accessed: 07 May 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original

[33] Bukun, "Breast cancer histopathological database (breakhis)," Mar 2020, accessed: 11Jan 2024. [Online]. Available: https://www.kaggle.com/datasets/ambarish/breakhis

[34] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," Kaggle, Feb 2020, accessed: 16 May 2024. [Online]. Available: https://www.kaggle.com/example/dataset

[35] Z. Lubis, P. Sihombing, and H. Mawengkang, "Optimization of k value atthe k-nn algorithm in clustering using the expectation maximization algorithm," *IOP Conference Series: Materials Science and Engineering*, vol. 725, no. 1, p. 012133, jan 2020, accessed: 27 April 2024. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/725/1/012133