COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR BREAST CANCER DETECTION

M.SC. PROJECT

BY

AFOLAYAN JESUTOFUNMI ONAOPE

13CD002816

SUPERVISOR

PROFESSOR ADEBIYI AYODELE

CO-SUPERVISOR

DR. (MRS) ADEBIYI MARION

DEPARTMENT OF COMPUTER SCIENCE,

LANDMARK UNIVERSITY, OMU-ARAN.

MAY, 2021.

COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR BREAST CANCER DETECTION

AFOLAYAN JESUTOFUNMI ONAOPE

(13CD002816)

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE, COLLEGE OF PURE AND APPLIED SCIENCES, LANDMARK UNIVERSITY, OMU-ARAN, NIGERIA.

IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF MASTERS OF SCIENCE (M.Sc.) IN COMPUTER SCIENCE.

MAY, 2021

DECLARATION

I, **JESUTOFUNMI ONAOPE AFOLAYAN**, a M.Sc. student in the Department of Computer Science, Landmark University, Omu-Aran, hereby declare that this thesis entitled **"COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR BREAST CANCER DETECTION",** submitted by me, is based on my original work. Any material(s) obtained from other sources or work by any other persons or institutions have been duly acknowledged.

…….

Student's Full Name and Matriculation Number

…….

Signature and Date

CERTIFICATION

This is to certify that this dissertation has been read and approved as meeting the requirements of the Department of Computer Science, Landmark University, Omu-Aran, Kwara State, Nigeria, for Award of M.Sc. Degree.

Professor A.A. Adebiyi — Date
(Supervisor)

Dr.(Mrs.) M.O. Adebiyi — Date
(CO-Supervisor)

Dr. (Mrs.) M.O. Adebiyi — Date
(Head of Department)

Professor P.A. Idowu — Date
(External Examiner)

DEDICATION

This project is dedicated to my family for their support, love, prayers and encouragement all through the course of this research work.

ACKNOWLEDGEMENT

I give glory to God Almighty who grants me with immeasurable bounties, blessings and adequate wisdom to put in for my M.Sc Degree program in Computer Science and gave me the power to achieve all what I have accomplished so far, and seek in the future.

My deepest and indebited gratitude goes to my supervisors Prof. Ayodele Adebiyi and Dr. (Mrs) Marion Adebiyi, who have been of inestimable assistance during this research and for their noble character, cooperation, devoting guidance, career development, recommendation, and correction at various phases of the research work.I am thankful for all your keen effort and long hours spent reading and correcting this manuscript as well as my papers.They have imparted me with wealth of knowledge in machine learning and computational data analysis, and also taught me the essential skills of a MSc student and researcher. I have been fortune to study under their tutelage

Special thanks to Dr. Arowolo, Dr. Asani, and Mrs. Ogundokun for their support in the course of this project.This thesis would not have been achievable without you.

I am also grateful to my parents, Engr & Mrs. J.A. Afolayan, for their love, prayers, financial support, discipline, and encouragement. My appreciation also goes to my older sister, Oluwatomisin Afolayan for her care, support, and love.

Finally, I am grateful to all other well-wishers whose names are not mentioned. The love they have shown is deeply appreciated.

ABSTRACT

Death from cancer is one of humanity's biggest problem, though there are many ways of stopping it before it occurs, there are still no cure forms of cancer. Due to recent population growth in clinical research, effective diagnosis of cancer is significant. The rate of death from breast cancer is increasing significantly with the rapid growth of the population. Cancer of the breast is one of the major cancer-related deaths amongst women globally. Survival rates differ across the numerous health treatments provided, comprising of surgery, chemotherapy, surgical procedures, and radiation treatment. To facilitate quick treatment and also to achieve more reliable outcomes, data analysis approaches employed for the detection and treatment of cancer of the breast have to be improved.The aim of this study is to carry out a comparative analysis of machine learning techniques for breast cancer detection.

This study was analyzed using The Wisconsin breast cancer datasets from an online UCI machine-learning repository.Feature selection was carried out through Particle Swarm Optimization algorithm (PSO), this algorithm helped pick relevant features from the raw dataset to eliminate and reduce noises for a better outcome and then a reduced dataset was achieved.Three(3) machine learning algorithms for classification was used namely: The support vector machine (SVM), artificial neural networks (ANNs), and decision tree (DT), for classification purpose, and these classifiers were used for further analysis on the reduced dataset to simulate the model.

The performance metrics used for evaluating the model includes: precision, sensitivity, specificity, accuracy, F-score, false acceptance rate, error rate, and false-rejection rate. The model was simulated using Matlab 2015 version.

The result from the evaluation phase in terms of performance metrics reveals that ANNs achieved the highest accuracy, sensitivity, precision, and F-score, and recall of 97.13%, 99.10%, 96.49%, 97.77%, and 99.09% respectively, and ANN also produced the lowest false acceptance rate, error rate, and false rejection rate of 0.0450, 0.0666 and 0.0090 respectively.This study will be beneficial to medical practitioners and clinicians in decision-making.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

PSO	Particle swarm optimization
ANNs	Artificial Neural networks
SVM	Support vector machine
DT	Decision Tree
UCI	University of California Irvine Repository
FNA	Fine Needle Aspirate
MATLAB	Matrix Laboratory
AUC	Area under curve
ROC	Receivers operating characteristics curve

CHAPTER ONE

1. 0 INTRODUCTION

1.1 Background to the Problem

The body's cells maintain a cycle of recovery form. The body's normal operating process is generally maintained by the controlled growth and death rate of cells, although this is not always the case.(Miller & Zachary, 2017). Occasionally, an uncommon occurrence occurs in which a few cells begin to develop abnormally. Cancer is caused by the abnormal growth of cells, which may begin in any area of the body and spread to other parts. (R.M. Weinberg, 2013)

In the human body, several forms of cancer may be framed (Das *et al.*, 2000)Females are more susceptible to the cancer of the breast than men based on the structure of the human body. The Age, family history, breast mass, weight, alcohol consumption, and gender are among the multiple reasons for the causes of breast cancer.(Momenimovahed & Salehiniya, 2019) Lobules, ducts, nipples, and fatty tissues make up a woman's breasts; milk is produced in the lobules and transported to the nipple through ducts. Epithelial tumors usually grow within the lobules and ducts, and then cancer develops in the breast (Feng *et al.*, 2018)As soon as cancer starts in the breast, it disseminates across the human body.

Breast cancer is a heterogeneous tumor with a variety of biochemical behaviors, clinicopathological features, and molecular characteristics. In recent years, researchers have gained a better understanding of multistep carcinogenesis and the critical function of genetic engineering in the detection, treatment, and prevention of breast cancer. This

necessitates a broadening of breast cancer prevention, screening, and management methods (Arif Harahap *et al.,* 2017). However, the rapid increase in the rate of cancer of the breast progression, recovery order has skyrocketed as a result of advances in care thanks to advanced technology.

(Siegel *et al.*, 2018). Despite this, breast cancer remains one of the most common causes of cancer-related death in women all around the world. The latest treatments used, such as chemotherapy, surgery, radiotherapy, and hormone therapy all have different survival rates (Akram *et al.*, 2017). Nonetheless, a patient's reaction to a procedure varies depending on several reasons that are being investigated (Tabl *et al.*, 2019)

Previously, doctors investigated the variables affecting breast cancer survival rates using critical programming projects such as Microsoft Excel, Statistical package for social sciences (SPSS), and Statistics and evidence(Bhoo-Pathy *et al*., 2015). These traditional predictive methods are not as adaptable when it comes to detecting new variables and creating innovative and integrative representations. Due to the shortcomings of traditional statistical analyses, various machine learning (ML) techniques, such as classification and regression trees (CART), decision trees (DT), support vector machines (SVM), and artificial neural networks, and have been widely used in this sector. (Melillo *et al*., 2018).

Machine learning is a form of data processing that allows for the computerization of model structure and learning from given data. It is a branch of artificial intelligence that is based on the ability of computers to learn through understanding. Without some formal code, machines may use pattern recognition to find hidden bits of information through machine learning. Machine learning has been used in research on binary and multiple cancer

classification with genomic knowledge, as opposed to traditional biological and computational approaches.(Angermueller *et al.*, 2016)

The use of bioinformatics methods to classify genes useful for cancer diagnosis and prognosis prediction can help patients get faster care. Because of the vast number of gene expression data available, cancer data analysis is valuable but difficult. As a result, only specific elements that can express a patient's state of well-being must be extracted (Shon *et al*., 2020) Additionally, the development of effective characterization models based on the derived genes aids in cancer patient early detection and prognosis prediction. Gene modifications facilitate a cell to replicate at an exponential rate, permeate normal surrounding cells, and spread throughout the body, resulting in cancer. Studies also identified genes involved in spinal muscular atrophy, inherited nonpolyposis colon cancer, and autism by using machine learning algorithms to correctly predict patients' disease state by studying mutations only in the gene pattern. (Zhou *et al*., 2017)

(Sharma & Rani, 2021) found several promising machine learning applications in bioinformatics and genomics science. PathAI, for example, was developed for digital pathology after artificial intelligence was used to analyze image data from breast cancer patients, lowering the error rate of diagnosing metastasized cancer through deep learning. Also, (Huml *et al.*, 2013)combined gene data with pathology imaging data to examine the survival rate of patients with brain tumors. The accuracy of survival rate estimation was found to be very good in this research. Deep learning convolutional neural networks were also shown to estimate survival rates with greater accuracy than a pathologist-based diagnosis.

Another study used machine learning to interpret gene-related big data to predict the degree of risk of nearly 20 cancers (Ferroni *et al*., 2019). Various data processing tools have been used over the years. (Nguyen *et al*., 2013)for example, used a rough set (RS) based support vector machine classifier (RS_ SVM) to diagnose breast cancer, and their results found that the (RS SVM) classifier had a 96.87 percent accuracy rate.

Furthermore, (Stoean & Stoean, 2013) indicated that hybridized support vector machines and evolutionary algorithms obtained the correct classification of 97% for diagnostic and 79% for predictive. Subsequently, classification and regression tree, decision tree, artificial neural networks, and support vector machine are some algorithms that were applied to diagnose breast cancers by(Yue *et al*., 2018). The experiment results revealed that the support vector machine obtained high classification accuracy compared to other classifiers. Based on artificial neural networks, (Alexis Marcano *et al.*, 2011)deduced a method named artificial metaplasticity on multilayer perceptron AMMLP based on the biological meta-plasticity property of neurons and Shannon's information theory. As reported by(Alexis Marcano *et al.*, 2011), a total of 99.26% accuracy was obtained using the AMMLP. More so, the result of (Vijayalakshmi & Priyadarshini, 2017) on the utilization of artificial neural networks on breast cancer image classification depicts the accuracy of Back Propagation Neural Network (BPPN) and radial basis neural networks (RBFN) to be 59.0% and 70.4% respectively.

This research work focuses majorly on a comparative study of machine learning techniques for breast cancer detection, and this model was simulated using the Matlab platform.

1.2 Statement of the Problem

Gene expression data processing can lead to important biological breakthroughs. Most of the research into detecting differentially expressed genes has concentrated on the more dramatic variations, but the more subtle differences in the results may have gone unnoticed (Zhao *et al*., 2016). Computational approaches for analyzing these results have enormous potential for discovering gene regulatory targets, cancer prediction, and drug production (Jackson *et al.*, 2020). However, the high dimensionality and noise associated with these data make these activities difficult. Furthermore, a dimensionality curse" arises from the discrepancy between a large number of genes and a typically limited number of samples. Using gene expression, a variety of algorithms have been used to differentiate normal cells from irregular cells (Reddy, 2015) even though much research has been done on cancer diagnosis using gene expression evidence, there is still a vital need to: Improve accuracy

There has been some progress with machine learning approaches for dimensionality reduction and classification of gene expression results. There are also some obstacles to decoding the most important cues for classification purposes (Kourou *et al*., 2015). Recently, there have been attempts to characterize samples based on gene expression results using single-layer nonlinear dimensionality reduction techniques (Onderwater, 2015).

To date, only a few studies have used machine learning approaches to forecast personalized breast cancer risk or compared the predictive precision and efficiency of these methods to models widely used in clinical practice. In this light, this study prefers to test genetic data (breast cancer) using three machine-learning algorithms (Support Vector Machine, artificial neural network, Decision Tree) and choose the best one.

1.3 Aim and Objectives of the Study

This aim of this study is to undertake a comparative analysis of machine learning techniques for breast cancer detection. The Specific objectives of the study are:

i. Identify relevant features for the model.
ii. Implement a data-preprocessing analysis using PSO on Matlab to select significant features from the Breast cancer dataset to obtain a subset of data.
iii. Classify the developed model using the subset data obtained from 2
iv. Simulate the developed model using the MATLAB platform.
v. Evaluate and Compare the performance of the machine learning algorithms using identified performance metrics.

1.4 Justification for the Study

This study will help clinicians, health workers, genome, computers scientist, and future researchers in decision making on breast cancer related works for detection purposes. It also pose the best machine learning technique that have showed to be more accurate in breast cancer detection. This will help facilitate machine learning technique that can be used when biomedical technicians are designing machine for breast cancer detections.

1.5 Scope of the Study

This study is limited to three (3) machine-learning algorithms only, the breast cancer dataset used is restricted to the UCI machine-learning repository (Wisconsin breast cancer dataset) whose features are gotten from a fine needle aspirate (FNA) of a breast mass. These algorithms were analyzed with a MATLAB software tool, on a Windows operating

system. Feature Selection was achieved using: Particle Swarm Optimization Algorithm (PSO).

1.6 Significance of the Study

The findings from this study could be significant to health workers and medical researchers, and computers scientist.

To the health workers, findings from this study may help proper decision-making by health practitioners. Findings from this study could be beneficial to computer scientists such that they could get into a self-learning mode on the machine learning that performs best without explicit programming.

1.7 Definition of Major Terms

The following terms and variables used in the Study have been operationally defined as follows:

Breast cancer: refers to a group of breast tumor subtypes that have different genetic and cellular backgrounds and clinical characteristics**.** An invasive tumor develops in the mammary gland.

Genetic Data: Refers to personal data relating to inherited or acquired genetic characteristics of a natural person acquired through DNA or RNA analysis.

Machine learning: is a branch of artificial intelligence (AI) that allows systems to learn and develop without being directly programmed**.**

Artificial Neural Network: ANN is a subset of machine learning that has become increasingly important in current research and growth. Machine learning is the ability of a computer to understand the nature of data through the use of a mathematical or statistical model.

Support Vector Machine: is a supervised learning model with associated learning algorithms for classification and regression analysis that analyzes results.

Classification and Regression Tree: Refer to a predictive modeling problem where the class label is predicted for a given example of input data

Decision Tree: An important tool for classification and estimation, as well as for promoting decision making in sequential decision problems, is the decision tree.

Particle Swarm Optimization: This is a computational method that optimizes a problem by iteratively trying to improve a candidate solution about a given quality measure

1.8 Thesis Layout

The remaining part of this thesis is organized as follows:

Chapter 2: Examine the conceptual issues related to the Study, and the chapter further reviews the theoretical review and lastly, it identifies the gaps in the reviewed literature.

Chapter 3: It was done under the following headings: Research Methodology, Research Approach, Research Framework, Data collection, Experimental Dataset, Feature Selection, Classification, Research tool, Matlab, System configuration, and Performance Metrics.

Chapter 4: It consists of Implementation, results, and discussion

Chapter 5: It entails summary, conclusion and proffer appropriate recommendations based on the findings of the Study.

CHAPTER TWO

2.0 LITERATURE REVIEW

This chapter focuses on analyzing the relevant kinds of literature that capture the variable on the use of machine learning techniques to analyze genetic data (breast cancer). It entails the conceptual review, empirical review, and theoretical review. The conceptual review provides clarifications and discussions of concepts related to the subject matter such as breast cancer; machine learning techniques which include: artificial neural networks, support vector machine, decision tree and; feature selection techniques using particle swarm optimization algorithm; and machine learning cancer prognosis. The theoretical review provides insights on relevant theories that are related to the subject matter, while the review of related works provided detailed information on previous researches on the subject matter.

2.1 Conceptual Issues

2.1.1 Breast Cancer

Breast cancer is the leading cancer in females all over the world. Breast cancer is caused due to the abnormal growth of some cells in the breast. Several techniques have been introduced for the correct diagnosis of breast cancer. Breast screening or mammography(Mori *et al.*, 2017) is a technique to diagnose breast cancer. It is used to check the nipple status of women through X-rays. Generally, it is almost impossible to detect breast cancer at the initial stage due to the small size of the cancer cell seen from

outside. It is possible to diagnose cancer at the early stage through mammography, and this test takes just a few minutes.

Conferring to the American Cancer Society(*Breast Cancer:Statistics, Approved by the Cancer.Net Editorial Board.*, 2017), the ladies are affected by breast cancer in comparison to all other cancers already introduced. Estimation shows that the ladies will be affected with intrusive breast cancer approximately 252,710 and around 63,410 females will be detected within situ breast cancer in the United States in 2017. Men also have a greater chance of breast cancer. An estimation for men is that they will be affected by this cancer approximately 2470 in the United States in 2017. Another estimation shows that about 41,070 persons will die from this cancer in 2017. Recent statistics in the UK reports that 41,000 women are affected by breast cancer every year whereas only 300 men are affected by this disease.

Hazard factors for developing breast cancer include being female,heftiness, an absence of physical exercise, liquor addiction, hormone substitution treatment during menopause, ionizing radiation, early age from the start of a monthly cycle, having kids in late life or more seasoned age, having an earlier history of bosom disease, and a family history of breast cancer.(Mensah, 2014).

It is more common in developing nations, and it affects women 100 times more than men.(Momenimovahed & Salehiniya, 2019). The most prominent symptom of breast cancer is a lump that looks different from the rest of the breast tissue. When a human feels a lump with their fingers, this can be discovered. Mammograms, on the other hand, detect the early stages of breast cancer.(Mason, 2017). Breast cancer can be detected by lumps of

the lymph nodes in the armpits, a rash on or around a nipple, leakage from nipple/s, thickening different from the rest of the breast tissue, one Breast becoming greater or smaller, a nipple shifting location or form or becoming twisted, skin puckering or dimpling. Paget's disease of the breast is another sign of breast cancer.(Mason, 2017). This condition manifests as eczema-like surface changes on the nipple tissue, such as redness, discoloration, or slight flaking. Tingling, swelling, heightened sensitivity, burning, and discomfort are all signs of Paget's disease of the breast as it progresses. A discharge from the nipple is also possible. Around half of the women with Paget's disease of the breast still have lump(s) in their breast.(Basu *et al.*, 2008).

Fiery Breast Cancer is uncommon (just observed in under 5% of bosom malignant growth finding), yet the forceful type of bosom disease is described by the swollen, red territories framed on the head of the Breast.(Okocha *et al.*, 2018). The enhanced visualizations of Inflammatory Breast Cancer are an after-effect of a blockage of lymph vessels by malignant growth cells. This kind of bosom malignancy is found normally in younger ages, stout ladies and African American ladies. As provocative bosom disease isn't present as an irregularity and there may be a deferral in conclusion. (Okocha *et al.*, 2018).

Threatening tumors may lead to metastatic tumors, which are optional tumors that grow after the primary tumor has spread beyond its original location.(Franchi, 2020). The symptoms of metastatic bosom malignant development can vary depending on the region of metastasis. The bone, liver, lung, and mind are the most common sites for metastasis. When malignant development has progressed to this point, it is classified as a phase 4 disease; tumors in this stage are often fatal. Unexpected weight loss, bone and joint pain, jaundice, and neurological side effects are all common symptoms of stage 4 cancers. Since

these side effects may be symptoms of a variety of illnesses, they are referred to as ambiguous side effects.(Franchi, 2020). The majority of bosom issues' side effects, such as many abnormalities, do not reflect secret bosom malignant development. Fewer than 20% of lumps are malignant, and benign bosom ailments including mastitis and fibro adenoma of the bosom are becoming more common causes of bosom problem side effects.(Franchi, 2020).

2.1.2 Machine Learning Technique

Knowledge tests are linked to the general concept of derivation by machine learning (ML), which is a branch of Artificial Intelligence. Each learning procedure has two stages: i) estimating obscure conditions in a system from a given dataset, and (ii) using the evaluated conditions to predict new framework yields. ML has also been seen to be an intriguing area of biomedical research with several applications, where valuable speculation is obtained by searching an n-dimensional space for a given set of natural examples using different methods and calculations. (Vamathevan *et al.*, 2019). There are two primary sorts of ML strategies known as (I) regulated learning and (ii) solo learning. In managed learning, a marked arrangement of preparing information is utilized to assess or map the information to the ideal yield.(Jason Brownlee, 2019).

Conversely, under the unaided learning strategies, no marked models are given, and there is no thought of the yield during the learning procedure. Thus, it is up to the learning plan/model to discover designs or find the gatherings of the information. (İlhan, A., Gülersoy, 2019).

In administered learning, this method can be thought of as a grouping issue. The undertaking of characterization alludes to a learning procedure that sorts the information

into a lot of limited classes. Two other regular ML assignments are relapse and bunching.(Vamathevan *et al*., 2019). On account of relapse issues, a learning capacity maps the information into a genuine worth variable. Accordingly, for each new example, the estimation of a prescient variable can be assessed because of this procedure.(Vamathevan *et al*., 2019). Grouping is a typical solo assignment where one attempts to discover the classifications or bunches to portray the information. Because of this procedure, each new example can be appointed to one of the distinguished bunches concerning the comparative attributes that they share. (Kashyap, 2019)

Assume we've collected clinical reports relevant to bosom disease and are attempting to predict whether a tumor is harmful or benign based on its size. The ML query would have implied a prediction of whether the tumor is dangerous or not (1 = Yes, 0=No). Figure 2.1 specifies the protocol for determining whether a tumor is harmful or not. The documents in the vicinity show some misclassification of the kind of tumor caused by the technique.

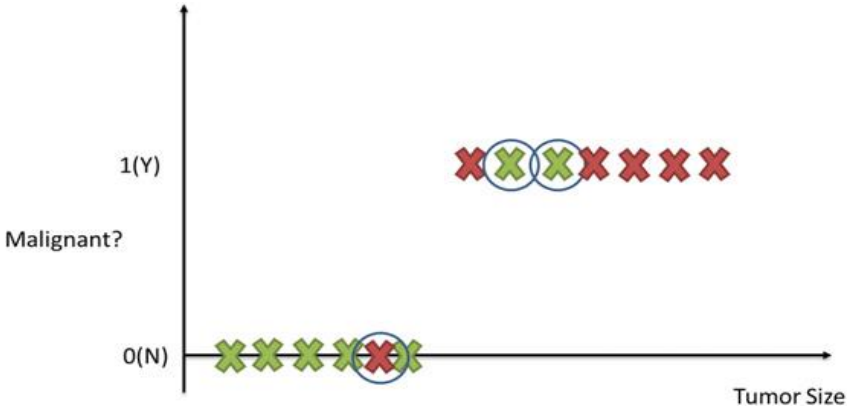


Figure 2. 1: Supervised learning classification

Tumors are symbolized by the letter X and are graded as benign or malignant. The tumors that have been misclassified are shown in the circled illustrations.(Kourou *et al.,* 2015).

Semi-administered realizing, a combination of supervised and solo instruction, is another widely used ML technique. It creates an exact learning model by combining named and unlabeled data. When there are more unlabeled datasets than labeled, this type of learning is usually used. Knowledge experiments are used to determine the essential pieces of a machine learning technique. Each example is illustrated with a few highlights, and each feature has a variety of attributes. Furthermore, understanding ahead of time what kind of data is being used allows for the proper selection of apparatuses and methods to be used for their analysis. A few information-related problems allude to the essence of the data and the preprocessing measures needed to make it more ML-friendly. The presence of clamor, glitches, missing or duplicate details, and one-sided unrepresentative information are also examples of information consistency problems. (Kourou *et al.,* 2015).

When the accuracy of the evidence is increased, the nature of the resulting inquiry is usually improved as well. Furthermore, preprocessing procedures that rely on changing the facts should be used to make the raw data more appropriate for further investigation. There are a variety of protocols and methodologies for knowledge preprocessing that rely on manipulating data to make it more suitable for a specific machine learning technique. The most important methodologies within these methods are (i) dimensionality reduction, (ii) include determination, and (iii) highlight extraction. When there are a large number of highlights in a dataset, dimensionality has numerous advantages. ML calculations works better when the dimensionality is lower (Pang-Ning,Tan *et al*., 2005).

Furthermore, the decrease of dimensionality can wipe out unimportant highlights, diminish clamor, and can deliver progressively vigorous learning models because of the inclusion

of fewer highlights. The dimensionality decreases by choosing new highlights, which are a subset of the old ones is known as highlight choice.

Notwithstanding, the utilization of highlight determination procedures may bring about explicit vacillations concerning the making of prescient component records. ML procedures' principal goal is to create a model that can be utilized to perform characterization, forecast, estimation, or some other comparable assignment. The most widely recognized undertaking in the learning process is the arrangement. As referenced already, this learning capacity characterizes the information thing into one of a few predefined classes. (Jenny *et al.*, 2020)

When an order model is created, by methods for ML procedures, preparing and speculation mistakes can be delivered. The previous alludes to misclassification mistakes on the preparation information while the last on the normal blunders on testing information.(Kashyap, 2019). A decent arrangement model should fit the preparation set well and precisely order all the occurrences. If the test mistake paces of a model start to increment, although the preparation blunder rates decline, then the wonder of model overfitting happens. This circumstance is identified with model intricacy, implying that the preparation blunders can be decreased if the model unpredictability increments. The perfect multifaceted nature of a model not helpless to overfitting is the one that delivers the most reduced speculation mistake. A conventional technique for breaking down a learning calculation's normal speculation mistake is the predisposition fluctuation deterioration. The predisposition segment of a specific learning calculation gauges the mistake pace of that calculation. (Kashyap, 2019).

When a grouping model has utilized at least one ML method, it is essential to assess the classifier's presentation. The presentation investigation of each proposed model is estimated as far as affectability, particularity, exactness, and territory under the bend (AUC).(Kourou *et al.*, 2015). Affectability is characterized as the extent of genuine positives that the classifier effectively sees, while particularity is given by the extent of genuine negatives that are accurately distinguished. The quantitative measurements of exactness and AUC are utilized for surveying the general execution of a classifier. In particular, precision is a measure identified with the complete number of right expectations. Despite what might be expected, AUC is a proportion of the model's presentation, which depends on the ROC bend that plots the tradeoffs among affectability and 1-particularity. (Kourou *et al.*, 2015)

2.1.3 Feature Selection

Classification is a crucial role in machine learning, and it entails categorizing individual occurrence in a data set into separate classes according to the knowledge represented by its attributes.(Kadhim *et al* 2018).It's impossible to tell which attributes are useful without previous experience. As a consequence, several functions, including important, irrelevant, and redundant features, are typically added to the data collection. Irrelevant and obsolete elements, on the other hand, are useless for classification. Due to the huge search area, known as "the curse of dimensionality," they can also degrade classification results (Gheyas & Smith, 2010). This problem can be solved using feature filtering, which selects only the most important features for classification. Variable selection, also known as function selection, will minimize the number of features, reduce training time, simplify

learned classifiers, and increase classification accuracy by removing and minimizing irrelevant and redundant features.(Unler & Murat, 2010).

The process of selecting a subset of specific features (variables, predictors) for use in model construction is referred to as feature selection, attribute selection, or variable subset selection in machine learning and statistics. Feature selection methods are used for a variety of purposes, which include:

1. Model simplification for easy interpretation (James *et al*., 2013)
2. Lesser training time.
3. Avoiding the curse of dimensionality.
4. Enhanced generalization by reducing overfitting(Bermingham *et al*., 2015)
5. Reduction of varieties (James *et al.*, 2013)

The basic premise of a function selection method is that the data contains certain items that are outdated or obsolete, and hence can be discarded without risking any information loss.(Bermingham *et al*., 2015) Since one important feature can be unnecessary in the presence of another suitable feature for which it is closely associated, the terms redundant and irrelevant are used interchangeably(Isabelle & Andre, 2003). It's important to differentiate feature selection techniques from feature extraction techniques. (Sarangi *et al.,* 2020) Feature extraction generates new features from the functions of the original features, while feature selection only returns a subset of them. In domains with a large number of features and a small number of samples or data points, feature selection methods are often used.

Wrappers, filters, and embedded approaches are the three major types of function discovery algorithms, and these evaluation criteria differentiate between them. (Isabelle & Andre, 2003).

Wrapper method: To score function subsets (Phuong *et al*., 2006) used a predictive model. It tests subsets of variables, allowing it to identify potential interactions between variables, unlike filter approaches.

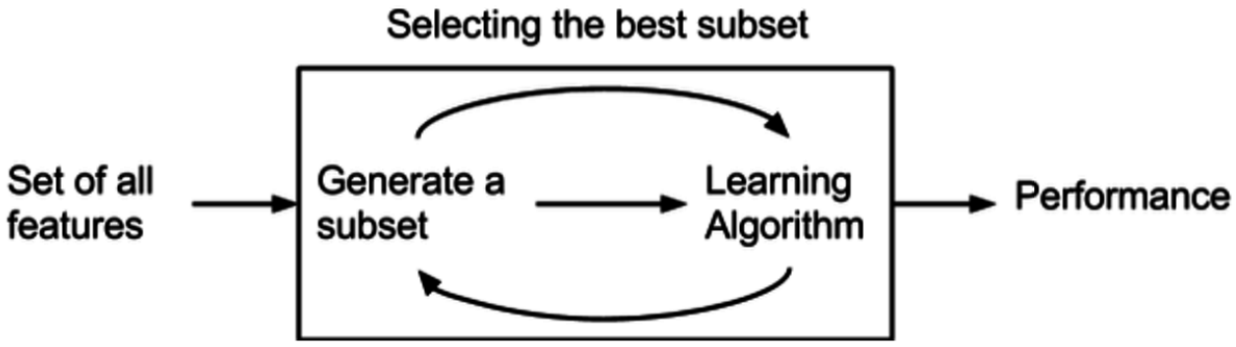


Figure 2. 2: Wrapper method for feature selection(Phuong *et al.*, 2006)

Filter Methods: (Zhang *et al*., 2013) scored a function subset using a proxy metric rather than the error rate, but they are typically less computationally expensive than wrappers. Nonetheless, they generate a feature set that isn't tailored to a particular kind of predictive model.

Methods of the filter sort pick variables independent of the model. They are solely dependent on general characteristics such as the association with the predictor to be predicted. The least interesting variables are suppressed by filtering techniques. The other variables will be used to characterize or forecast data using a classification or regression

model. These methods are particularly efficient in terms of calculation time and are resistant to overfitting. (Julie, 2013).

When filter methods are used without taking into account the relationships between variables, they are prone to selecting redundant variables. More complex features, such as the FCBF algorithm, attempt to mitigate this problem by eliminating strongly correlated variables.(Yu & Liu, 2013).

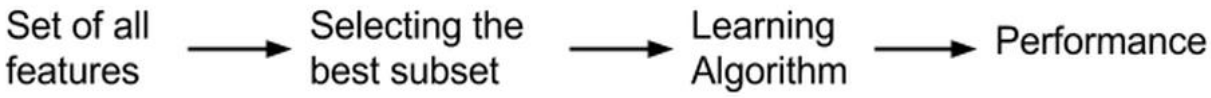


Figure 2. 3: Filter method (Yu & Liu, 2013)

Embedded Method: The embedded approach employs a set of techniques for feature selection during the model creation process. Recently, embedded approaches have been proposed that attempt to incorporate the benefits of the previous methods. A learning algorithm, such as the FRMT algorithm, uses the variable selection mechanism to simultaneously perform feature selection and classification..(Saghapour *et al.*, 2017)

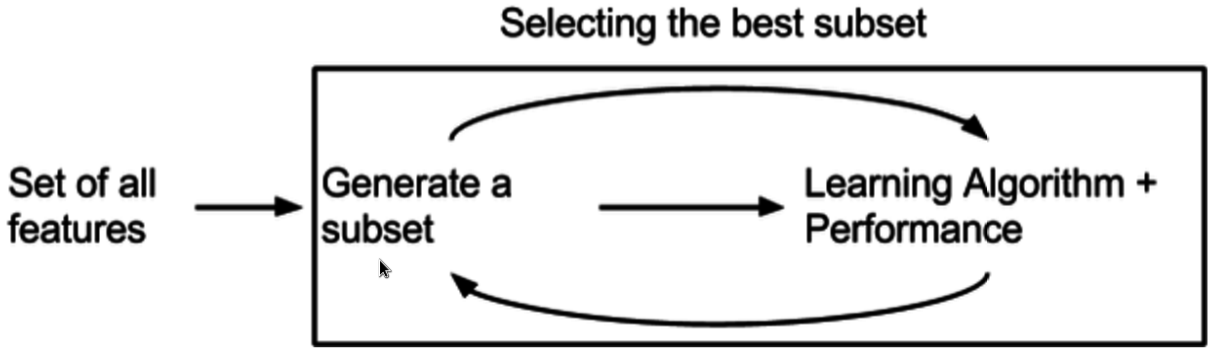


Figure 2. 4: An embedded method for feature selection (Saghapour *et al*., 2017)

Since features can work in complex ways, feature selection can be difficult. When used in conjunction with other features, an individually important function can become redundant (irrelevant). As a consequence, an optimal function subset should consist of a series of complementary features that cover the various properties of the classes to differentiate them correctly. Because of the large search room, selecting features is often challenging. If the number of available data sets capabilities increases, the scale of the search space expands exponentially. (Isabelle & Andre, 2003)

A strong global search technique is needed to assist in the resolution of feature selection issues. The global searchability of evolutionary computation (EC) techniques is well-known. Optimization of particle swarms (Kennedy & Eberhart.); is a modern EC approach that focuses on swarm intelligence. (R.C *et al*., 2011) Some EC algorithms, such as genetic algorithms (GAs) and genetic programming, are computationally more expensive and take longer to converge than PSO (GP). As a result, PSO has proved to be useful in a wide range of applications, including feature selection (Mohemmed & Zhang, & Johnston, 2009).

2.1.4 Particle Swarm Optimization

PSO is a problem-solving algorithm that iteratively attempts to refine a solution for a given quality metric(Coppo Leite, 2019). For optimum function collection, PSO is proposed and introduced. PSO is a global search technique that is both accurate and reliable. It is an appropriate algorithm for solving feature selection problems because it has better representation, the ability to scan wide spaces, is computationally less costly, simpler to implement, and has fewer parameters (R.C *et al.*, 2011).

Eberhart and Kennedy invented the PSO, an evolutionary computational method in 1995. Social habits such as bird flocking and fish learning inspire PSO. PSO is based on the idea

that information is enhanced by social contact in a population where the thought is both personal and social(Chen *et al*., 2014). Due to (i) simple feature encoding, (ii) global search capability, (iii) rational computationally, (iv) fewer parameters, and more straightforward implementation, it is a suitable algorithm for feature selection problems. For the reasons mentioned above, the PSO is used to choose features. The principal space is the search space in which PSO was used to discover and pick a subset of principal components or core features. Particles in PSO reflect candidate solutions in the quest space and form a swarm, which is also known as a population. The particle swarm is generated by randomly scattering 1s and 0s. If the principal component is 1, every particle is chosen, and the central variable of 0 is ignored. As a result, each particle represents a particular subset of the principal components. The particle swarm is generated at random. By updating its location and velocity, it is then pushed in the quest space or principal space to find the best subset of functions. The current status of particle *i* and its velocity are expressed in (1) and (2):
(1) and (2) express the current state of particle I and its velocity:

xi = xi1; xi2;..... xiD equation (1)

(1) Where D denotes the dimension of the primary search space.

vi = vi1, xi2,..... viD equation (2)

(2) Equation is used to measure it's velocity and speed. (3) Invasive detection feature collection using particle swarm optimization

vidi+1 = w*v1id +c1*r1d*(pid– x2id) +c2* *(pgd– x1id), xidi+1 = x1id + vidi+1, xidi+1 = x1id + vidi+1, xidi+1 = x1id + vidi+1, xidi+1 = x1id

T denotes the process's tth iteration, and d denotes the search space's dth dimension. c1 and c2 are acceleration constants, and W is inertial weight. r1 and r2 are randomly distributed random values in the range [0,1]. The elements of pbest and gbest in the dth dimension are represented by pid and pgd.

Before a stopping criterion is met, such as a maximum number of iterations or a good fitness value, the position and velocity values of each particle are continuously updated to find the best combination of features, or the algorithm stops when a predefined criterion is met, such as a good fitness value or a predefined maximum number of iterations.

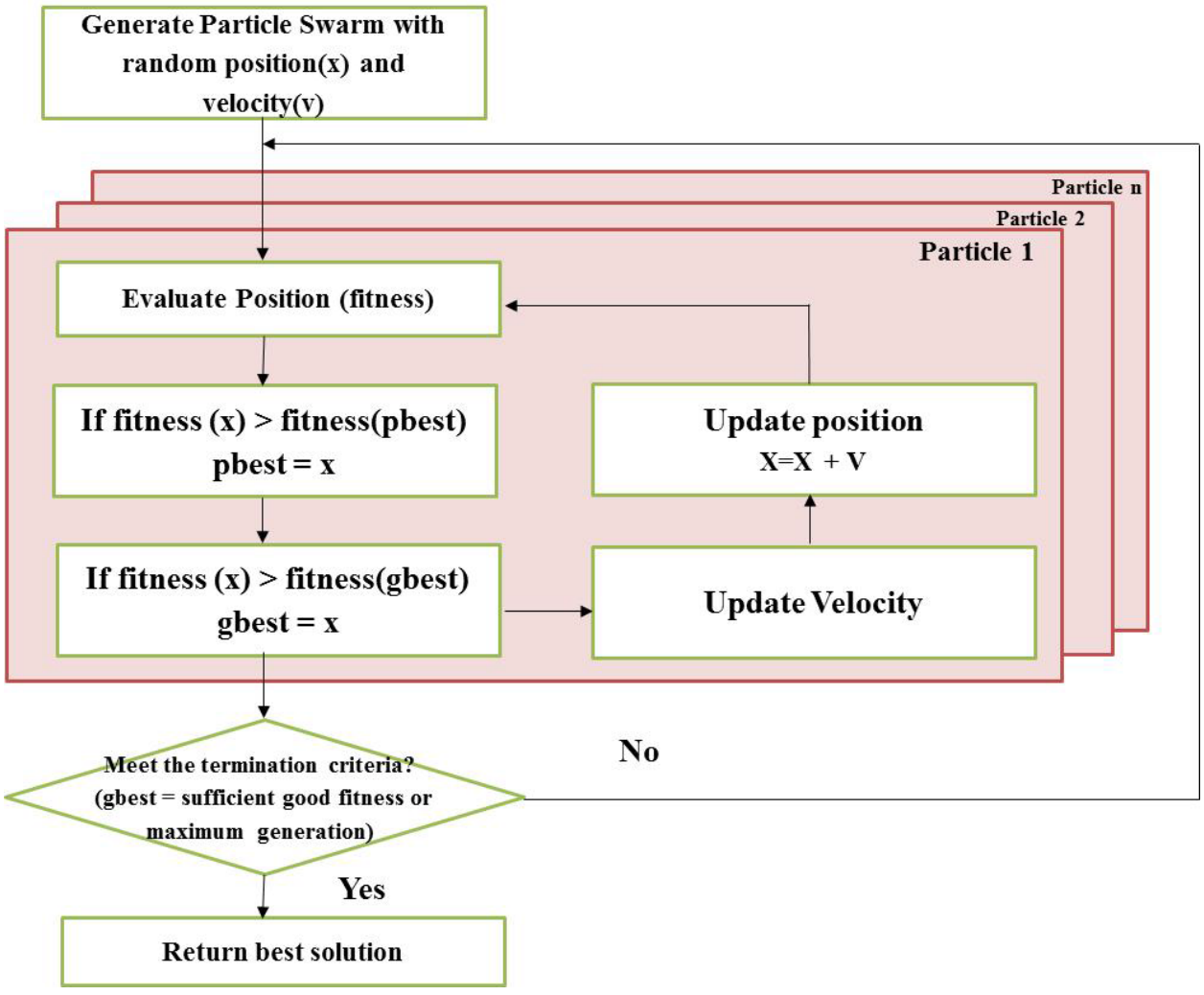


Figure 2. 5: Particle Swarm Optimization process (Sakri *et al.*, 2018)

2.1.5 Support Vector Machine

A well-known AI method that can solve both order and relapse problem is the Support Vector Machine (SVM) calculation. A non-probabilistic twofold straight classifier, an SVM preparing calculation constructs a model that appoints new guides to one of two classifications given a large number of preparing models, each set apart as having a position with one of two classifications (even though strategies, for example, Platt scaling exist to utilize SVM in a probabilistic order setting)(S, 2019).

The models are represented as points in space in an SVM model, to isolate instances of various groups by a reasonable hole that is as wide as possible under the circumstances. New models are then planned into the equal vacuum, with the assumption that they will be classified according to which side of the hole they fall. Aside from straightforward characterization, SVMs can easily execute a non-straight grouping using the part stunt, ensuring that their contributions to high-dimensional component spaces are well designed(Ben-Hur & Guyon).

SVMs are useful in text and hypertext order because they can eliminate the need for named planning events in both the traditional inductive and Trans-inductive approaches. Configurations Aid vector machines are used in a few techniques for superficial semantic parsing(Lee *et al*., 2010)

SVMs may also be used to categorize images. After just three or four pertinence input stages, test results reveal that SVMs achieve fundamentally better pursuit exactness than traditional query refinement plans. This is true for picture division systems as well, even those that use a modified adaptation SVM that employs Vapnik's advantageous method. (Prashar & Harish, 2015). The classification of satellite data, such as SAR data, using

administered SVM and hand-composed characters can be seen using SVM. (Zhou *et al*., 2017)

SVM is a fantastic method for creating a classifier. It involves deciding between two groups and allowing for the prediction of marks from at least one element vector(Asri *et al*., 2016). The hyperplane, or preference limit, is oriented in such a way that it is above what anyone would think conceivable from the closest knowledge focuses from any of the groups. The closest focuses are referred to as bolster vectors.

Given a named training dataset, xi Rd and yi (1, +1), where xi is a feature vector representation and yi is the classmark (negative or positive) of a training compound i.

wxT+ b=0 can then be defined as the ideal hyperplane.

The weight vector is w, the input feature vector is x, and the bias is b.

The w and b would satisfy the following inequalities for all elements of the training set:

If yi=1, wxiT+ b +1$wxiT+ b \leq -1$ if $yi=-1$

The objective of training an SVM model is to find the w and b so that the hyperplane separates the data and maximizes the margin 1 / || w ||2.

Vectors xi for which $|yi|$ ($wxiT+ b$) = 1 will be termed support-vector.

Another application of SVM is the kernel method, which can be used to model higher-dimensional, non-linear models. In a non-linear problem, a kernel function could add more dimensions to the raw data, turning it into a linear problem in the higher-dimensional space. (Huang et al., n.d.). Briefly, a kernel function could help do certain calculations faster, which would otherwise need computations in high dimensional space.

It is defined as:

$K (x, y) = <f(x), f(y)>$

The kernel function is K, and the n-dimensional inputs are x and y. The function f is used to convert an n-dimensional input into an m-dimensional space. The dot product is denoted by x, y>. We could measure the scalar product between two data points in a higher-dimensional space using kernel functions without having to calculate the mapping from the input space to the higher-dimensional space directly. In certain cases, computing the kernel is easy, but computing the inner product of two feature vectors in a high-dimensional space is difficult. Also, simple kernels' feature vectors will balloon in size, and kernels like the Radial Basis Function (RBF) kernel (KRBF(x, y) = exp (-||x - y||2) have infinite-dimensional feature vectors. The kernel, on the other hand, is almost trivial to compute. The kernel function is chosen, among other things, may have a significant impact on the efficiency of an SVM model (Shujun *et al.*, 2018). However, there is no way to know which kernel is better for a particular pattern recognition problem. Trial and error is the only way to find the perfect kernel. Starting with a basic SVM, and then playing with various standard kernel features, this can be accomplished. One kernel may be stronger than the others depending on the nature of the problem. Cross-validation can be used to choose an ideal kernel function from a fixed number of kernels in a statistically robust manner.

2.1.6 Artificial Neural Network

ANN is a form of machine learning that has grown in importance in recent research and development. Machine learning refers to a computer's ability to comprehend the existence of data using a quantitative or computational model.(Ahuja, 2019). An artificial neural network, on the other hand, is not a simple process; it necessitates a different approach and deals with complex trends in vast amounts of data. This necessitates a thorough

understanding of methods as well as a well-structured algorithm.(Abiodun *et al*., 2018) ANN uses supervised learning, unsupervised learning, and reinforcement learning methods in the implementation process. Until now, the neural network learning algorithm has been the subject of study and accepted by communities.

2.1.6.1 A brief history of artificial neural network

The human biological brain, which consists of up to 60 trillion interconnected sets of neurons to conduct network patterns of decision making, is the inspiration for ANN. The artificial neural network method starts with very basic entangled neurons that function as a single processor, based on this fundamental principle. The neuron model of Mc Culloch and Pits was used to implement the idea.(Negnevitsky, 2005). A single layer of input, operation, and output components form the basis of an ANN. As a result, ANN works as a dynamic mathematical algorithm to achieve an optimal result for certain datasets or problem parts, based on a basic principle of the information processing period.

ANN is one of the most effective artificial intelligence tools for common data mining problems like classification and regression. Many studies have shown that ANN is effective in detecting breast cancer. This form, however, has several drawbacks:

ANN has multiple hidden levels, hidden nodes, learning thresholds, and activation features that must be tuned at the start of the training process.

2. Due to the dynamic design and parameter update mechanism in each iteration, the training process takes a long time and has a high computational cost.

3. It may get stuck in local minima, making optimum efficiency impossible to guarantee.

2.1.6.2 Learning paradigms

Supervised learning, unsupervised learning, and reinforcement learning are the three main learning paradigms. Each one is linked to a specific learning objective.

2.1.6.2.1 Supervised learning

In supervised learning, a mixture of paired inputs and optimal outputs is used. The learning goal of each input is to produce the desired output. In this case, the cost feature is concerned with excluding erroneous deductions. The mean-squared error is a typical investment that aims to reduce the average squared error between the network's output and the expected output (Bertsekas & Tsitsiklis, 1996). Pattern detection (also known as classification) and regression are two features that are well suited for supervised learning (also known as function approximation). Sequential data may also benefit from supervised learning (e.g., handwriting, speech, and gesture recognition) (Deng *et al.*, 2020).This can be compared to learning with a "teacher" in the form of a function that gives constant input on the consistency of the solutions so far

2.1.6.2.2 Unsupervised learning

Unsupervised learning necessitates the use of input data, as well as the cost feature, such data functions, and the performance of the network. The mission (model domain) and any previous expectations decide the cost function (the implicit properties of the model, its parameters, and the observed variables)(Morocho-Cayamcela *et al.*, 2019). As an example, consider the constant model and the cost. You get a return that is equal to the data average when you reduce this expense implementing the expense function can be a lot more difficult. Its form varies depending on the application: in compression, it may be related to

mutual information between and, while in mathematical modeling, it may be related to the model's corresponding probability, given the specifics are provided.(Morocho-Cayamcela *et al*., 2019)

2.1.6.2.3 Reinforcement learning

The aim of reinforcement learning is to balance the network (create a policy) in order to take actions with the lowest long-term (expected cumulative) cost. The world reacts to the agent's behavior at any given moment with an observation and an immediate cost. Dependent on certain (usually unknown) rules. The rules and the long-term cost can only be measured in certain situation.

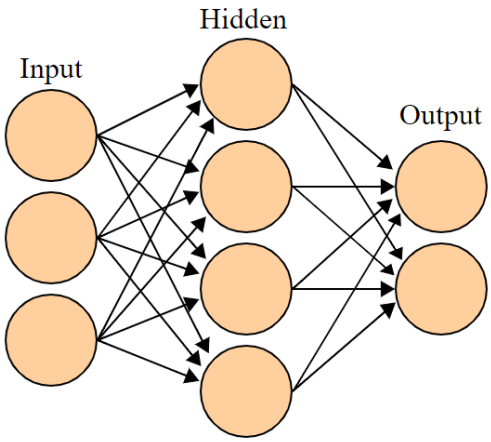


Figure 2.6: An artificial neural network architecture

Neural networks can be utilized in various fields. The undertakings to which artificial neural networks are applied will in general fall inside the accompanying general classifications:

- ·Function estimate, or relapse investigation, including time arrangement forecast and demonstrating.

- Classification, including example and grouping acknowledgment, curiosity location, and consecutive decision making.
- Data preparing, including sifting, bunching, dazzle signal division, and pressure.

Nonlinear framework distinguishing proof and control (vehicle control, process control), game-playing and decision-making (backgammon, chess, dashing), design recognition (radar frameworks, face ID, object recognition), grouping recognition (motion, discourse, written by hand text acknowledgment), clinical determination, money-related applications, and dashing are several of the application zones of ANNs (Billings, 2013). For example, a semantic profile of a customer's preferences may be created using images prepared for object recognition*(Szymon et al.,* n.d.)

The demonstrating of computerized pictures for bosom malignancy characterization is accomplished utilizing artificial neural networks. Artificial Neural Network is prepared on the pictures(Mehdy *et al.,* 2017). The engineering of the ANN depends on a feed-forward network. This engineering is prepared by Back-proliferation calculation. The acquired outcomes are additionally improved by utilizing the Radial Basis work network; the spiral premise work network's performance outperformed the ANN.

2.1.7 Classification Trees

A classification challenge with a sample of n observations distributed on a parameter Y with values of 1, 2,...k and a p predictor parameter, X1,...Xp, as well as a classification challenge with a sample of n observations distributed on a parameter Y with values of 1, 2,...k and a classification challenge with a sample of n observations distributed on a parameter Y with values of 1, 2,...k and a classification challenge with a sample of n observations distributed on The aim is to develop a model that can forecast Y values from

new x values. In theory, the solution is to divide the X space into k disjoint sets, A1, A2,..., Ak, so that the predicted value of Y is j if X belongs to Aj, where j = 1, 2,..., k. (Cover & Hart, 1967). If the p values are high, these techniques produce sets Aj with stepwise linear and nonlinear constraints, which are easier to decode.

Aj are rectangular sets obtained by the periodic distribution of a data set of x component at a time using classification tree techniques. The sets became easier to decipher as a result of this. THAID splits a node by looking exhaustively over both X and S for the split X S that minimizes the cumulative impurity of its two child nodes. The set S is an interval of the form (, c] if X takes ordered values. S is a subset of the values taken by X if not otherwise specified.

Each child node's data is subjected to the process in a recursive manner. If the proportional decrease in impurity is below a pre-determined threshold, splitting comes to a halt. The simple steps are pseudocoded in Algorithm 1.

1st algorithm Pseudocode for exhaustive search tree creation

1. Begin at the top of the tree, with the root node.

2. Find the set S that minimizes the number of the node impurities in the two child nodes for each X, and select the split X S that gives the smallest total X and S for each X.

3. Exit if a stopping condition is met. Otherwise, repeat steps 2 and 3 for every child node.

These methods are strictly followed by the later developed algorithms such as C4.5 and CART(Breiman *et al.*, 1984). For its impurity feature, C4.5 uses the disorderliness of its sets, while CART uses the Gini index, which is a generalization of the binomial variance. Unlike THAID, they grow an unnecessarily large tree first, then prune it to a smaller size to reduce the misclassification error estimation. CART uses 10-fold (default) cross

validation, while C4.5 estimates error rates using a heuristic algorithm. RPART is the R implementation of CART. (Therneau and Atkinson, 2008), which we use in the examples below. The exhaustive quest strategy, despite its ease of use and style, has one drawback. Notice that an ordered variable with m distinct values has (m 1) splits of the form X c, while an unordered variable with m distinct unordered values has (2m1 1). If all other factors were similar, parameters with distinct values would have a better chance of being chosen. This bias impedes the integrity of the final conclusions drawn from the structure. CRUISE, GUIDE, and QUEST (Kim & Loh, 2001), capitalized on the methods of FACT algorithm (Loh and Vanichsetakul, 1988) through the use of two-steps methods based on the tests to divide each node. For instance, each x is tested for connection with *y* and the most relevant parameter is picked and secondly, an exhaustive search is performed for the set *S*. This particular selection is absolutely free of bias since the X parameters have the same likelihood of been selected if each is not affected by Y. Furthermore, large computations is saved since the search for S is conducted only on the X parameters selected. GUIDE, CRUISE and QUEST applies the chi squared tests for unordered parameters and ANOVA for ordered parameters which is an unbiased technique, employs the permutation tests.

Pseudocode for the GUIDE algorithm is given in Algorithm 2

2. The CRUISE, GUIDE, and QUEST trees are pruned the same way as CART.

Algorithm 2 Pseudocode for GUIDE classification tree construction

1. Begin at the top of the tree, with the root node.

2. Or each ordered variable X, group its values in the node into a limited number of intervals to transform it to an unordered variable X_.

Set X_ = X if X is not ordered.

3. On the data in the node, perform a chi squared test of independence of each X_ vector versus Y and compute its significance likelihood..

4. Choose the X_ vector with the lowest significance likelihood..

5. To break the node into two child nodes, find the split set {X* S*} that minimizes the number of Gini indexes.

6. If a stopping criterion is reached, exit. Otherwise, apply steps 2–5 to each child node.

7. Prune the tree with the CART method.

CHAID (Kass, 1980), on the other hand, takes a different view. If X is an ordered parameter, the data values in the node are separated into ten intervals, each with its own child node. If X is unordered, each value of X is allocated to a child node. Then, to bind pairs of child nodes, CHAID uses more precise checks and Bonferroni corrections. There are two results from this approach. For starters, certain nodes may have more than two child nodes. Second, the procedure is skewed against choosing variables with few distinct values due to the linear nature of the experiments and the inexactness of the corrections. In the nodes, CRUISE and GUIDE can also fit bivariate linear discriminant models, and CRUISE can also fit bivariate kernel density and nearest neighbor models. GUIDE can also create ensemble models using techniques like bagging (Breiman, 1996) and random tree.

2.1.8 Regression trees

Regression and classification tree have similar characteristics, except that the *Y* parameter are ordered values and a regression model best describe each of the nodes to result to

predicted values of y. AID (Morgan & Sonquist, 1963) is the first regression tree algorithm having its impurity node to be the sum of squared deviations from the mean and the node determining the mean of y. This results to stepwise constant models. However, its interpretation is relatively simple, the accuracy of these models often lags behind that of models with more smoothness. This is impracticable computationally, although, this can be corrected by the generation of stepwise linear equations that fitted into every candidate split. M5'(Ian & Eibe, 2005), an adaptation of a regression tree algorithm by Quinlan (Ian & Eibe, 2005)employs a more robust method to construct stepwise linear equations. It first constructs a stepwise tree and then fits a linear regression model to the data in each leaf node. Since the tree is similar to the stepwise constant model, the yielding trees tend to be greater than those derived from other stepwise linear tree techniques. GUIDE (Loh, 2002) employs classification tree approach to solve the regression problem. At each node, it fits a regression model to the data and computes the residuals.

2.1.9 Decision Trees

In sequential decision problems, decision trees are primarily used for classifying, forecasting, and encouraging decision making. This method examines three different kinds of decision trees in depth. (Hastie *et al.,* 2001). The first is a recommendation algorithm built on a collection of knowledge nodes; the second is classification and regression trees; and the third is survival trees. (Dey, 2019)

When a decision-making process requires a series of decisions, the problem becomes more difficult to imagine and execute(Song & Lu, 2015). In these situations, decision trees are important graphical methods. They make for intuitive problem understanding and can help with decision-making.

A decision tree is a graphical model that depicts choices and their consequences. There are three kinds of nodes in a decision tree (Berry & Linoff, 2008)

1. Decision node: Sometimes shown as squares, these nodes indicate where choices can be taken. Both distinct choices available at a node are represented by lines originating from a rectangle.

2. Chance node: Often depicted as a circle with random outcomes. Chance consequences are situations that could happen but are outside the decision maker's grasp.

3. Terminal node: A triangle or a line with no further decision nodes or chance nodes is also used to describe a terminal node. The results of the decision-making process are represented by terminal nodes.

For example, a hospital that performs esophagostomies (surgical removal of all or part of the esophagus) for patients with esophageal cancer would like to have a guideline for what constitutes an appropriate lymphadenectomy in terms of the total number of regional lymph nodes extracted throughout surgery(Hastie *et al*., 2001). Pathology, according to the hospital, should direct such a procedure (available to the surgeon prior to surgery). Histopathologic cell type (squamous cell carcinoma or adenocarcinoma); histopathologic grade (a crude measure of tumor biology); and depth of tumor invasion should all be included in this details (PT classification). When the histopathologic grade is poorly differentiated and the number of nodes to be removed varies by cell type, it is thought that the number of nodes to be removed should increase for more highly invasive tumors.

In this case, the decision tree is mostly made up of chance effects, which are the products of pathology (cell type, grade, and tumor depth).

The only decision the surgeon has to make is whether or not to conduct the esophagostomy. If the surgeon decides to operate, he or she moves from left to right down the decision line on the graph, using pathology details to assess the terminal node(Bhukya & Ramachandram, 2010) The number of lymph nodes to be destroyed is the terminal node, or final result.

In certain cases, decision trees can be used to make the best choices. To do so, the decision tree's terminal nodes must be given terminal values (sometimes called payoff values or endpoint values). One method is to assign values to each judgment and chance branch, and then define a terminal value as the number of the branch values that contribute to it. Tree values are determined by following terminal values from right to left after terminal values have been allocated. Multiply the importance of chance outcomes by their likelihood to get their value. The sum of these values is the total for a chance node (Kele? & Segal, 2002). The cost of each choice in each decision line is subtracted from the cost already determined to determine the worth of a decision node. This importance reflects the decision's advantage

2.1.9.1 Machine Learning and Cancer Prognosis

In the past two decades, a wide range of ML methods and highlight choice algorithms have been broadly applied to infection forecast and expectation (Saeid & Eslaminejad, 2016). A large portion of these works utilizes ML techniques for displaying the movement of cancer and recognize enlightening elements that are used subsequently in a characterization plot. Moreover, in practically all the investigations quality articulation profiles, clinical factors just as histological boundaries are included in a correlative way to be taken care of as a contribution to the prognostic strategy.

According to the current PubMed findings for the topic of ML and cancer, 7510 papers have been published as of today. For the identification of tumors as well as the forecast/guess of a cancer type, the vast majority of these distributions employ at least one machine learning algorithm and integrate data from heterogeneous hotspots. In the last decade, a growing trend has emerged in the application of other-directed learning procedures, specifically SVMs and BNs, to cancer prediction and forecasting (Saeid & Eslaminejad, 2016)These characterization algorithms have been widely used in cancer science to address a wide range of problems. Previously, the most popular evidence used by doctors to make an informed judgment about cancer prediction were histological, clinical, and population-based data (Bi *et al.*, 2019)

The joining of highlights, for example, family ancestry, age, diet, weight, high-chance propensities, and introduction to ecological cancer-causing agents assume a basic job in foreseeing the improvement of cancer. Even though this sort of large-scale data alluded to a few factors so standard measurable strategies could be utilized for expectation purposes, anyway these kinds of boundaries don't give adequate data to settling on powerful decisions. With the quick approach of genomic, proteomic, and imaging innovations, another sort of sub-atomic data can be gotten. Atomic biomarkers, cell boundaries just as the statement of specific qualities have been demonstrated as extremely educational markers for cancer forecast. The nearness of such High Throughput Technologies (HTTs) these days has created tremendous measures of cancer data that are gathered and are accessible to the clinical examination network.

Notwithstanding, the exact expectation of a sickness result is one of the most fascinating and testing assignments for doctors. Thus, ML techniques have become a mainstream

apparatus for clinical analysts. These strategies can find and distinguish examples and connections between them, from complex datasets, while they can viably anticipate future results of a cancer type. Also, highlight determination strategies have been distributed in writing with their application in cancer (Baylin & Jones, 2016).

2.2 Theoretical Review

The theoretical framework that is guiding this study is based on Incorporating Unlabeled Data and Interaction in the Learning Process, Similarity-based Learning, Clustering via Similarity Functions, and General Technical Theme.

2.2.1 Incorporating Unlabeled Data and Interaction in the Learning Process

In addition to a list of labeled examples from the underlying data distribution, Passive Semi-Supervised Learning (van Engelen & Hoos, 2020) is a typical setting for incorporating unlabeled data into the learning process, where the learning algorithm would use a (usually much larger) number of unlabeled examples from the same distribution. Several semi-supervised learning algorithms have been developed, and several promising experimental results have been obtained. However, the underlying assumptions of these methods are quite different.(C A Padmanabha Reddy *et al.*, 2018) For example, others believe that data is organized in groups, while others assume that the grouping rule is self-consistent. A significant roadblock to change has been the lack of clarity on whether certain common rules underpin all of these methods. Standard learning models, in particular, are unable to justify their usefulness (the PAC model or the Statistical Learning Theory framework). (Balcan *et al*., 2008) develops a detailed analytical paradigm for reasoning about semi-supervised learning that can be used to reason for many of the diverse methods taken in the machine learning field over the last decade.

Balcan's model is useful for answering questions like "Under what circumstances and to what extent will unlabeled data help?" “How much data does one expect to use to perform well?” and “How much data does one hope to use to perform well?” as well as to create algorithms with demonstrably better guarantees than those previously created.

Active Learning is a second method for integrating unlabeled data into the learning process that has been widely common in recent years(Dasgupta *et al.,* 2005). In this case, the learning algorithm is much more efficient because it can interactively request labels for unlabeled instances of its choice. The expectation is that by deliberately leading inquiries to insightful samples, a successful classifier can be taught with even fewer marks. (Balcan *et al.,* 2008) provided some new experimental findings for this paradigm.

2.2.2 Similarity-based Learning

Kernel methods have risen in popularity and been a burgeoning field of study in recent years, owing to their practical utility in working with a wide range of data types as well as their theoretical basis. These approaches make use of named examples and interface with the data using a pairwise function known as a kernel that also meets some mathematical requirements. In terms of viewing kernels as tacit mappings, a proven principle exists for these approaches, but it does not fit the functional intuition. A good kernel for a given problem is one that forms a normal notion of similarity in that domain. The inability to use theory as a tool for constructing useful kernels for new application domains has been hampered by this discrepancy between theory and intuition. Balcan (2008) offered theoretical evidence for the general intuition that a good kernel function is one that serves as a good measure of similarity by developing more intuitive and operational reasons for desired properties of good kernel functions.

2.2.3 Clustering via Similarity Functions

(Balcan & Blum, 2006) also offered a fresh take on the well-known Clustering dilemma. Machine learning's recent developments in fields of computational biology and gene discovery have also taken more traditional learning approaches like Clustering to the fore. The learning algorithm in this case does not use labelled data at all, but rather a similarity measure between pairs of items, with the intention of revealing some unknown hidden structure in the data. Since such issues are common in science, clustering has gotten a lot of coverage in a lot of different fields for a long time. Despite the development of a plethora of clustering algorithms, the issue of which approach is ideally suited to a particular form of data or what requirements are required to achieve highly accurate solutions remains unanswered. Existing theory has been flaky, relying on firm hypotheses regarding cluster uniformity or maximizing distance-based target functions that are only tangentially connected to the true objectives.

2.2.4 General Technical Theme

Balcan's theory work, in addition to helping physicians, advances the state of the art of machine learning theory. (Balcan & Blum, 2006).On a computational level, many of the models used to analyze these learning paradigms (e.g., semi-supervised learning or learning and clustering using similarity functions) use data-dependent concept spaces, which is expected to be a significant line of research in machine learning in the coming years. To deliver a broad range of results, these models focus on a wide range of insights and techniques from Algorithms and Complexity, Empirical Processes and Statistics, Optimization, and Geometry and Embeddings..(Lavrač *et al*., 2020)

2.3 Review of Related Works

Works relating to the use of machine-learning methods to solve medical breast cancer diagnosis were reviewed in this section.

(S. Lee *et al.*, 2020) By anticipating conditions among BCAC iCOGS SNPs, proposed a powerful machine learning method to discern gathering of interacting single nucleotide polymorphisms (SNPs) that contribute most to the breast cancer (BC) danger. An inclination tree boosting technique was used, supplemented by a flexible iterative SNP scan to find dynamic non-direct SNP-SNP partnerships and, finally, a set of associating SNPs with strong BC threat prescient capacity. The study also suggested using a support vector machine framed by the well-known SNPs to differentiate between BC cases and controls. In distinguishing BC cases and controls in KBCP, OBCS, and blended KBCP-OBCS test sets, our method achieves mean normal exactness (mAP) of 72.66, 67.24, and 69.25, respectively.

These results outperform the mAPs of 70.08, 63.61, and 66.41 obtained using a polygenic hazard score model based on 51 identified BC-related SNPs in the KBCP, OBCS, and consolidated KBCP-OBCS test sets, respectively. The 200 recognized KBCP SNPs from the proposed technique perform well in characterizing estrogen receptor-positive (ER+) and negative (ER) BC cases in both KBCP and OBCS results, according to the BC subtype investigation. Further, a biological investigation of the distinguished SNPs uncovers qualities identified with significant BC-related systems, estrogen digestion, and apoptosis.

(Tabl *et al.*, 2019) led an examination on "a machine learning approach for distinguishing quality biomarkers controlling the treatment of breast cancer, they present a progressive machine learning framework that predicts the 5-year survivability of the patients who

experienced, however, explicit treatment; the classes are based on the blend of two sections that are the survivability data and the given treatment. For the survivability data part, it characterizes whether the patient endures the 5 years or perished. While the treatment part signifies the treatment has been taken during that span, which incorporates hormone treatment, radiotherapy, or medical procedure, that thoroughly shapes six classes.

(Prasetyo *et al.*, 2014) worked on breast cancer diagnosis using artificial neural networks with extreme learning techniques. They used breast cancer Wisconsin dataset and made used of K-fold cross validation to experiment their results. Their result showed that extreme learning machine neural networks (ELM ANN) has better generalized classifier model than Gradient-based back propagation artificial neural networks (BP ANN). Although the specificity rate was slightly lower than BP ANN while ELM ANN had a higher sensitivity and accuracy rates

(Bhardwaj & Tiwari, 2015) developed a hereditarily enhanced neural network (GONN). By presenting new hybrid and transition administrators, they advanced neural network computing. They used WBCD to evaluate their work, looking at classification exactness, affectability, particularity, disarray structure, ROC bends, and AUC under ROC bends of GONN with old-style model and conventional Back spread model with old-style model and traditional Back spread model. This method provides a good level of precision classification. However, it continues to be better by using a larger dataset than WBCD, as well as highlight extraction, to make GONN more efficient for continuous Breast Cancer detection.

(Ashraf & Siti, 2018) developed a PC-based breast cancer treatment plan. To increase both the precision and network composition, the technique used a multilayer perceptron (MLP)

neural network based on improved non-ruled arranging hereditary calculation (NSGA-II). In comparison to other methods, this study increases classification accuracy. MLP will stall out in neighborhood minima in any situation.

(Ak, 2020) used computer visualization and machine learning applications to do a quantitative study of breast cancer identification and diagnosis. They used k-nearest neighbors, logistic regression, support vector machine, decision tree, nave bayes, rotation forest and random forest on their dataset to make predictions on breast tumor forms. They discovered that the logistic regression model had the best classification accuracy of 98.1 percent, and that the proposed solution improved accuracy results.

(Islam *et al.,* 2020) operated on a retrospective analysis using machine-learning methods to forecast breast cancer. Support vector machine (SVM), K-nearest neighbors, random forests, artificial neural networks (ANNs), and logistic regression were the five supervised machine-learning techniques they compared. The results of their study in terms of accuracy, sensitivity, specificity, precision, negative predicted value, false negative rate, false positive rate, F1 score, and Matthew's correlation coefficients show that ANN had the highest accuracy, precision, and F1 score of 98.57 percent, 97.82 percent, and 0.9890, respectively, while 97.14 percent, 95.65 percent, and 0.9890, respectively.

(Bataineh, 2019) A comparison of nonlinear machine learning algorithms for breast cancer diagnosis was carried out. On the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, a performance distinction is made between five non-linear machine learning algorithms: K-nearest neighbors (KNN), multilayer perceptron (MLP), Classification and regression trees (CART), support vector machine (SVM), and Gaussian nave bayes (NB), and In terms of classification test performance, specificity, and recall, they graded data based on the

reliability and efficacy of each algorithm. MLP had the highest accuracy of 96.70 percent, which was higher than the other four algorithms, and it also yielded the best results for K-fold cross validation in terms of accuracy, precision, and recall, with 99.12 percent, 99.00 percent, and 99.00 percent, respectively.

(Kanchanamani, 2016) completed a performance assessment and comparison of different machine learning methods for breast cancer diagnosis. They looked at Nave Bayes, SVM (Support Vector Machine), LDA, KNN, and MLP, which are all machine-learning algorithms. They used it to classify decomposed images, and they checked it against the MIAS database (Mammography Image Analysis Society). They also put the established framework into a tenfold cross validation process. In comparison to other techniques, the results suggest that SVM is an effective strategy. (Bazazeh & Shubair, 2016) Machine learning algorithms for breast cancer identification and diagnosis were compared in a report. Random Forest (RF), Support vector machine and Bayesian networks are three of the most common machine learning techniques (BN). They used the Wisconsin initial breast cancer dataset as a training set to test and compare the accuracy, memory, precision, and region of ROC of the three ML classifiers. Their findings revealed that SVMs perform best in terms of accuracy, specificity, and precision. RFs, on the other hand, have the best chance of accurately classifying tumors.

2.4 Gap Identified in the Literature

The following gaps were deduced from literature:

(Tabl *et al.*, 2019) showed that some of the potential biomarkers are strongly related to breast cancer survivability and cancer in general.

(Dhahri *et al*., 2019) proved that genetic programming can automatically find the best model by combining feature preprocessing methods and classifier algorithms.

(Arif Harahap *et al*., 2017) compared the neural network to other machine learning and concluded that the neural network improves classification accuracy than others.

(Houfani *et al.*, 2020) suggested a procedure on Wisconsin Original Breast Cancer (WBC) and WBCD in demonstrating its usefulness and reducing the estimation complexity.

Also, results from (Sakri *et al.,* 2018) showed that there is a need to improve the accuracy of their classifier because they currently have a very low accuracy of 76.3% without PSO normalization.

Thus, there is a need to enhance accuracy and even with PSO an accuracy of 81.3% is quite low.

The researchers above used machine-learning algorithms to conduct different analyses, but they did not compare the results of support vector machines, artificial neural networks, and Decision Trees. This thesis would test breast cancer genetic data using three machine learning algorithms (Artificial Neural Network, Support Vector Machine, and Decision Tree) and choose the best one based on performance metrics such as accuracy, sensitivity, specificity, precision, recall, F-score, error rate, false acceptance rate, and false rejection rate.

CHAPTER THREE

3.0 RESEARCH METHODOLOGY

3.1 Research Approach

The developed system consists of two stages: Feature Selection and Classification. The dataset is loaded, then subjected to the Particle Swarm Optimization (PSO) algorithm to eliminate noises and remove irrelevant features through the help of The MATLAB tool. The classifiers (Artificial Neural Networks, Support Vector Machines, and Decision Trees) receive the feedback from the feature selection phase, and the results are displayed on Matlab environment.

PSO for selecting features was combined with three classifiers (Artificial Neural networks, Support Vector Machine and Decision Tree) for classification purpose. These three techniques were added and used in model construction for performance of classification, results compared from each technique was be evaluated using nine performance metrics. The details of methodology deployed in this study are explained below:

1. PSO is used for selecting features and then applied to raw dataset to get a reduced dataset.
2. Classifiers will be affected on the reduced dataset
3. Evaluation of three classifiers using the following performance metrics: Accuracy, Sensitivity, specificity, Precision, F-score, Recall, False acceptance rate, Error rate and False rejection rate.
4. Classifiers were compared based on the results gotten from the evaluation of performance metrics. Figure 3.1 shows the overall system design of the study.

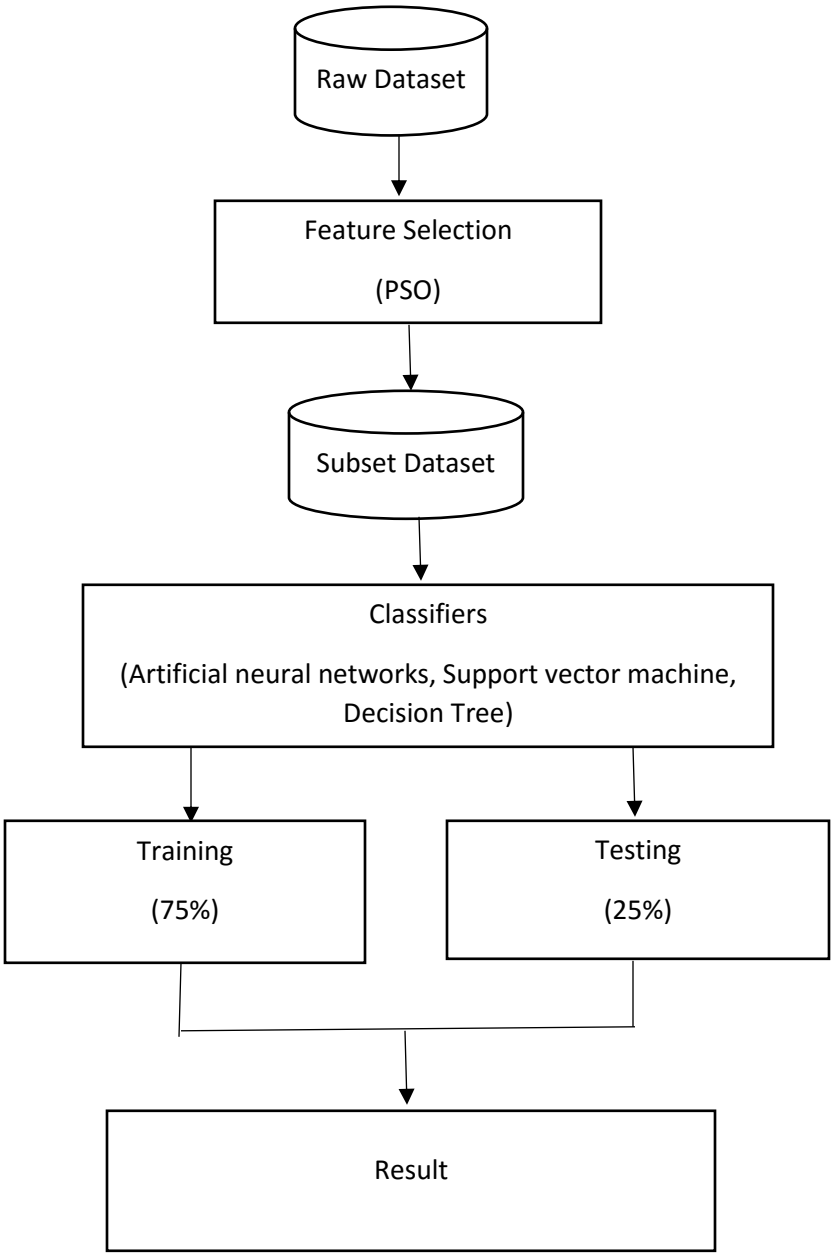


Figure 3. 1: Research Framework of the study

3.2 Experimental Dataset

The data source for this project was retrieved from the University of California Irvine Machine Learning Repository (Wisconsin Breast Cancer dataset). The features came from digitized image of a fine-needle aspirate of a breast mass that described the nucleus of the current image. A total of 699 cases in Wisconsin have been affected by the Wisconsin Diagnostic Breast Cancer (WDBC) site. FNA test measurements are represented by each observation. The research portion of the dataset was used 75% for training, while 25% was used for testing

3.3. Description of Datasets

The breast cancer dataset was obtained from the University of California, Irvine's machine-learning library. This dataset contains 699 cases, each of which is either not harmful or harmful. 458 (65.50 percent) of these cases are benign, while 241 (34.50 percent) are malignant. The dataset's class is divided into 0 and 1 cases, with 0 corresponding to the benign case and 1 corresponding to the malignant case. Attributes of the dataset are listed below.

Table 3. 1: Attributes of the Dataset

S/N	Attributes
1	Benign
2	Malignant
3	Clump thickness
4	Uniformity of cell size
5	Uniformity of cell shape
6	Marginal Adhesion
7	Single Epithecial cell size
8	Bare Nuclei
9	Bland chromatin
10	Normal Nuclei
11	Mitoses

3.4 Feature Selection

Feature selection is used to remove functions that are unnecessary or obsolete, which may decrease efficiency. Feature selection attempts to pick a limited number of appropriate features to achieve classification efficiency that is comparable to or superior than using all features. (Brank *et al.*, 2011)

The analysis of algorithms for minimizing data dimensionality in order to increase machine learning efficiency is known as feature selection. Feature selection tends to minimize M to M′ and M′ M for a dataset with N features and M dimensions (or features, attributes). It's

a common and effective method for reducing dimensionality (Brank *et al.*, 2011). It pinpoints the fields that are most crucial in forecasting a specific result.

Feature selection may resolve this issue by picking only suitable classification features. Variable selection, also known as feature selection, may decrease the number of features, simplify the learned classifiers, minimize training time and increase the efficiency of the classification by removing and lowering unnecessary and redundant features.(Unler & Murat, 2010)

In addition, the aim of selecting features is to find the best subset consisting of m characteristics selected from the total n characteristics topology. The dataset is converted to the classifier method after missing unimportant attributes, which is then divided into two parts: testing and training data.

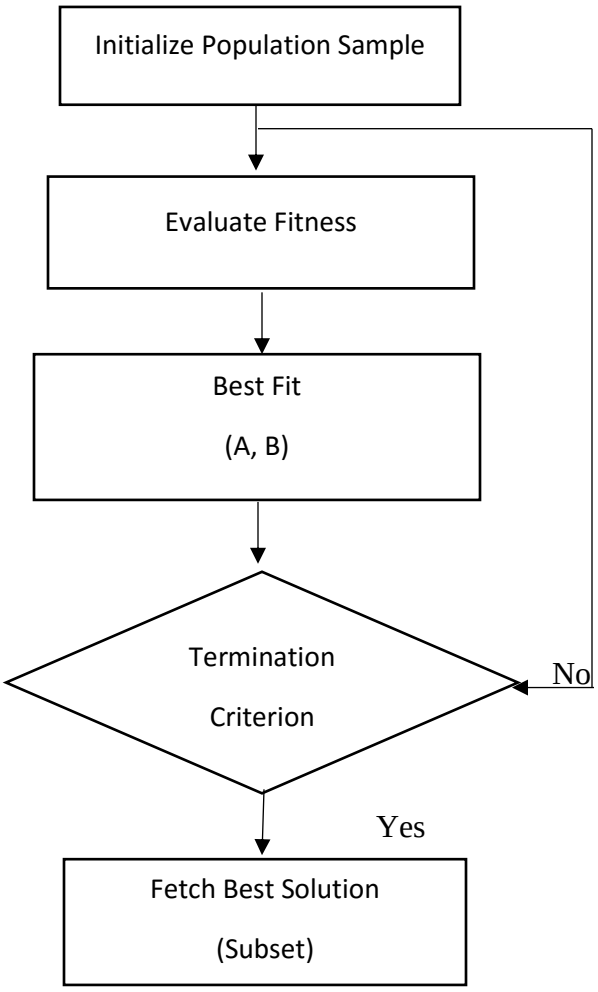


Figure 3. 2: Particle Swarm Optimization (Sakri *et al*., 2018)

Pseudo code for PSO

1) Initialize population sample

2) Assess each particle's fitness in (1)

3) Equate each particle's fitness assessment to that of the existing particles to obtain P-best

4) Calculate G-best by comparing fitness evaluation to the population's average previous best.

5) Update (4) and (2)

6) If the stopping condition is not met, proceed to phase 2.

7) End

3.5 Classification

The final performances are heavily influenced by classification models. On any given data set, different classifiers behave differently. The (Support Vector Machine:) definition is used in the classification module. It functions by identifying the best judgment boundary for separating data points into different classes, and then predicting the class of new findings based on that boundary. (Sacchet *et al*., 2015). SVM works with kernel.

Artificial neural network: ANN is one of the best artificial intelligence techniques for everyday machine learning tasks. An ANN's foundation is made up of a single layer of input, process, and output elements. ANNs works with nodes.

Decision Tree: The decision tree is a directed learning algorithm that represents the outcomes in an easily understandable tree structure, with visualization playing an important role in data analysis.("A Relative Analysis of Multi-Relational Decision Tree Learning Algorithm," 2017)Decision tree works with trees.)

They are machine-learning classifiers that can effectively deal with prediction and detection of breast cancer.

3.6. Performance Evaluation Metrics

The performance evaluation metrics of classifier is evaluated in terms of classification accuracy, sensitivity, Specificity, Precision, F-score, Recall, false acceptance rate, Error rate and false rejection rate. The terms are defined below and they are the general formulae universally accepted.

Accuracy= (TP + TN)/ (TP+TN+FP+FN) %

Sensitivity=TP/ (TP + FN) %

Specificity =TN/ (TN +FN) %

Precision = TP/ (TP + FP)

F-score= 2*TP/ (2*TP + FP +FN)

Recall: TP/TP + FN

Error rate: FN + FP /TP + FN

False acceptance rate: FP/ (FN + TN)

False rejection rate: FN/ (FN + TP)

Where:

TP is an acronym for (True Positives) which means correctly classified positives cases,

TN is an acronym for (True Negative) which means correctly classified negative cases,

FP is an acronym for (False Positives) which means incorrectly classified positive cases,

FN is an acronym for (False Negative) which means incorrectly classified negative cases.

Table 3.2: Research Objectives and Their Methodology

S/N	OBJECTIVES	METHODOLOGY
Objective 1	Identify relevant features for the model.	Through Extensive literatures.
Objective 2	Perform feature selection using PSO on breast cancer dataset as a data pre-processing step to obtain a subset data.	Implement particle swarm optimization algorithm.
Objective 3	Classify the developed model using the subset data obtained from 2	Artificial neural networks , Support Vector Machine (SVM), Decision Tree (DT)
Objective 4	Evaluate the developed model.	Confusion matrix

Objective 5	Compare the performance of the learning techniques using identified performance metrics.	The performance are compared using the following metrics: Accuracy; Sensitivity; Specificity; Precision; F-score; Recall; Error rate; False acceptance rate; False rejection rate

CHAPTER FOUR

4.0. RESULTS AND DISSCUSSIONS OF FINDINGS

4.1 Experimental Setup

This chapter focuses on analysis and interpretation of data using MATLAB tool and machine learning techniques. The data collected was compiled in Microsoft excel-2007 software. The MATLAB software was used for feature selection, after classification techniques were applied to the study. Machine-learning techniques for feature selection and classification were compared. Specifically, this section presents the results of the study. The developed model in this system was developed and implemented on MATLAB 2015a, a fourth generational programming language with object oriented based procedures.

4.2 Matlab Environment

This chapter discusses the system implementation and result of this chapter. The study uses PSO as a feature selection approach to select important features in the breast cancer dataset. The selected features were classified using three algorithms namely: Artificial neural network, Support Vector Machine and Decision tree. MATLAB environment was used to implement the model, figure 4.1 shows development environment and the procedure of executing the model.

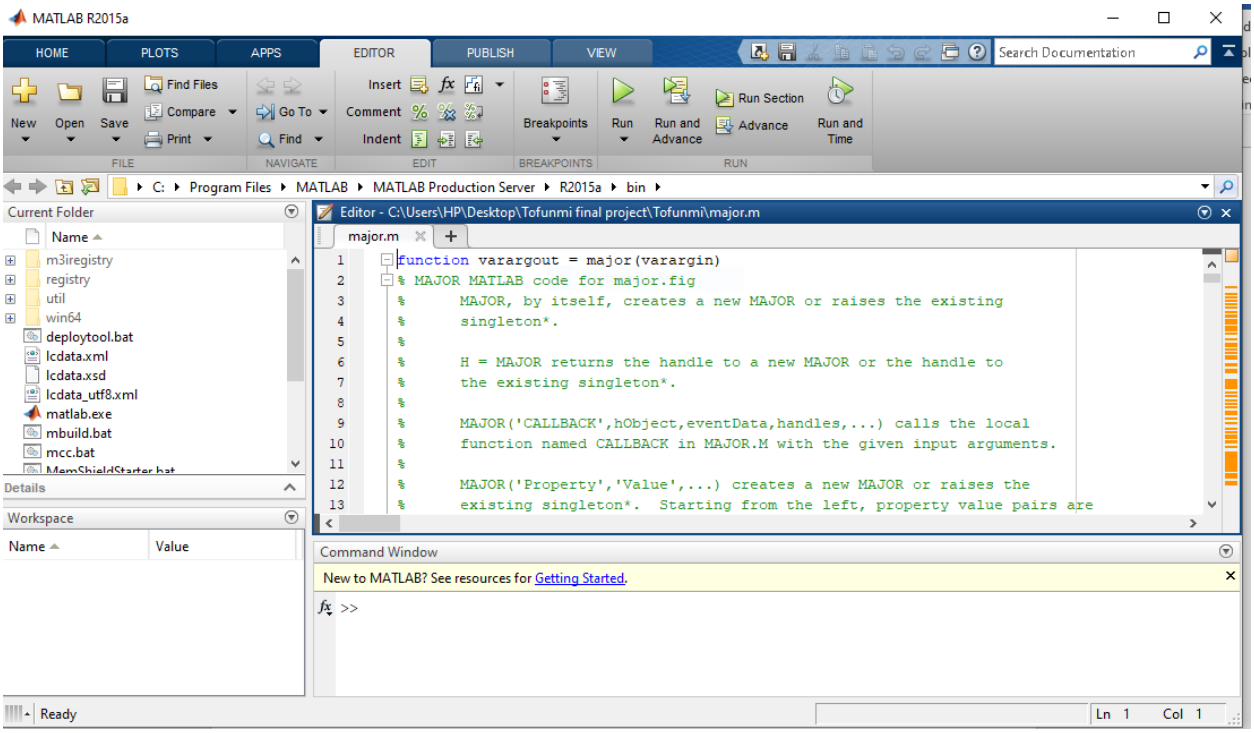


Figure 4 1: MATLAB IDE 2015a (Execution procedure of the study model)

4.3. User Interface

The user interface consists of three main functionalities and sub functionalities namely: (Sequence) for loading data functionality, Particle Swarm Optimization feature selection algorithm and selection time. The classification training method, run classification and the training time. Figure 4.2 shows the user interface for the implementation of the model.

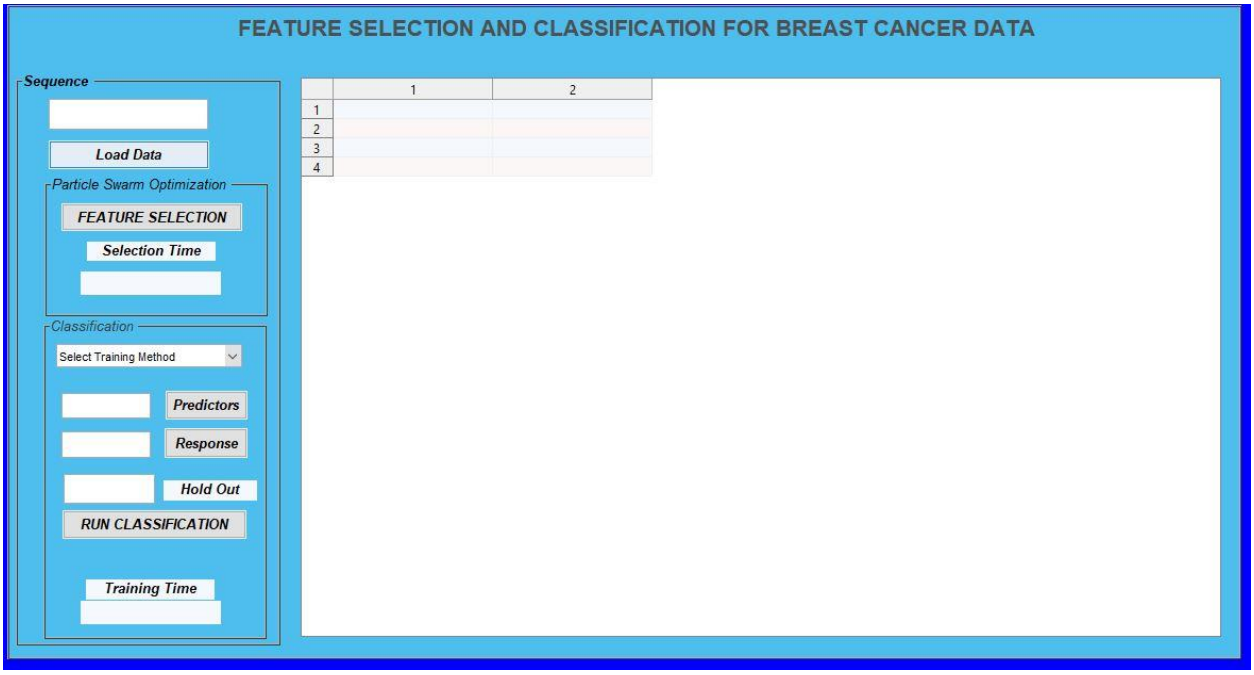


Figure 4 2: User Interface for loading dataset (Initial)

The load button is used for loading the data. The feature selection algorithm uses PSO to select relevant features from the loaded data. Figure 4.3 shows the loading of breast cancer datasets. It took 26.0153 Secs to fetch the relevant features from the given data.

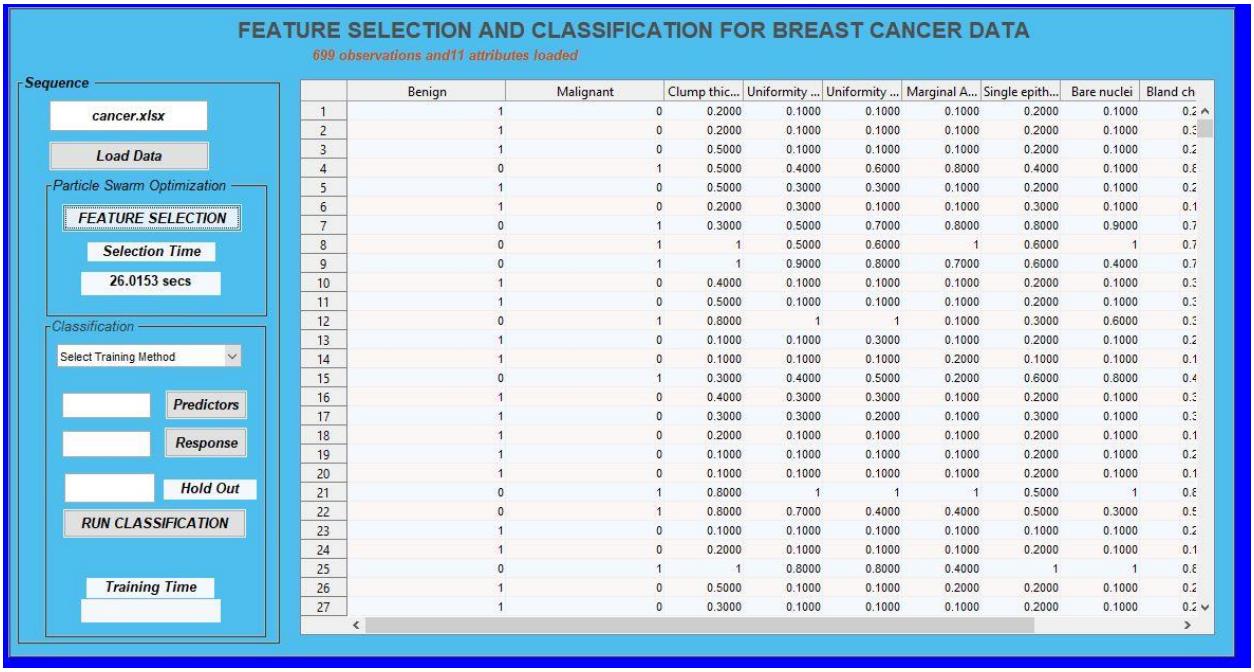


Figure 4 3: Breast cancer dataset during normalization

For better compilation and user friendliness, the MATLAB software developer tools (GUIDE) was used to create an interactive environment for easy readability, formatting and interactivity. The GUI development was sectioned into three major sections namely:

1. Load
2. Feature selection
3. Classifiers

Figure 4.3 shows when an initial data is loaded into the environment using the load data menu bar. The breast cancer dataset is loaded into the system with a total input data of 699 observations and 11 attributes.

4.4 Feature Selection

At the feature selection mode, features are selected using the Particle Swarm Optimization algorithm (PSO) the 699 observations were reduced to 284 entities after applying PSO, figure 4.4 shows the selected features. The obtained result output saved was passed into the classifiers in further analysis.

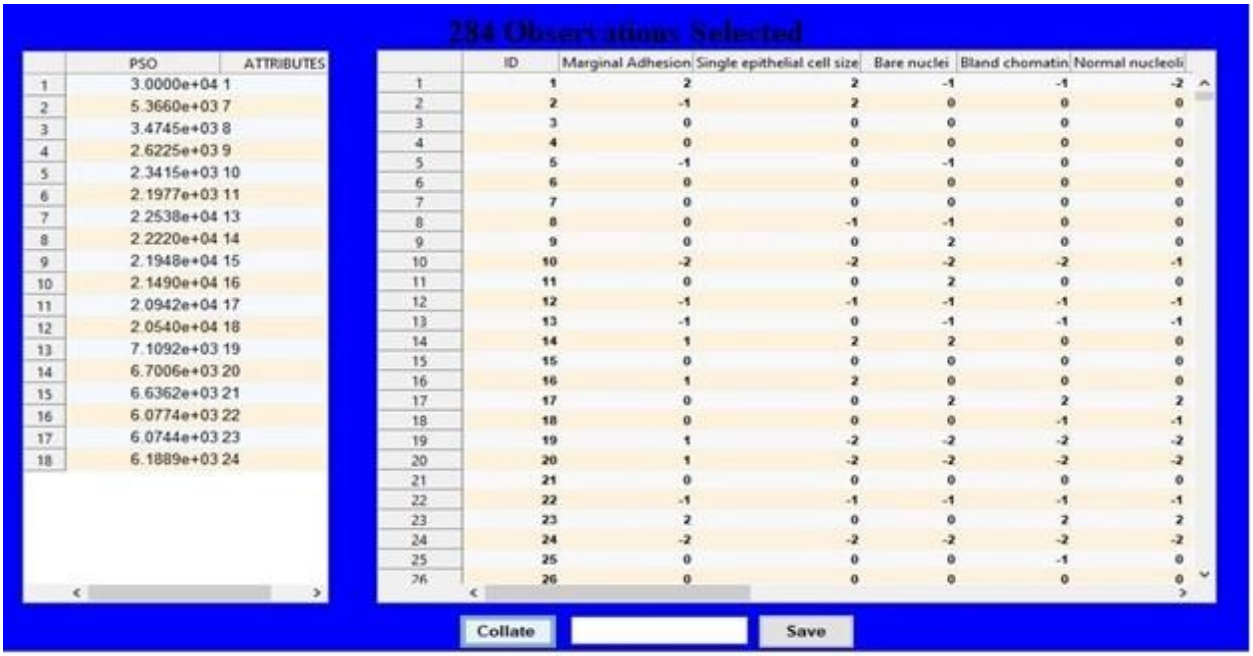


Figure 4 4: Feature selection interface using PSO

Figure 4.4 shows the processing of the loaded breast cancer data using feature selection algorithm known as Particle Swarm Optimization.

4.5 Classification

Three classifiers were used, (Artificial neural network, Support Vector Machine and Decision Tree) on the selected data. The procedures of analysis and the output of the result are discussed here.

The results of the selected features using PSO algorithm were first classified using Artificial Neural Network with a 0.25 hold out of which deduced that 75% of the selected features was used for training and 25% was used for testing.

4.6 Results and Discussion for ANN

Eight neurons were given to the ANN, it detected 10 hidden layers of nodes, with one output. Figure 4.5 shows the architecture of the input, and output of the experiment.

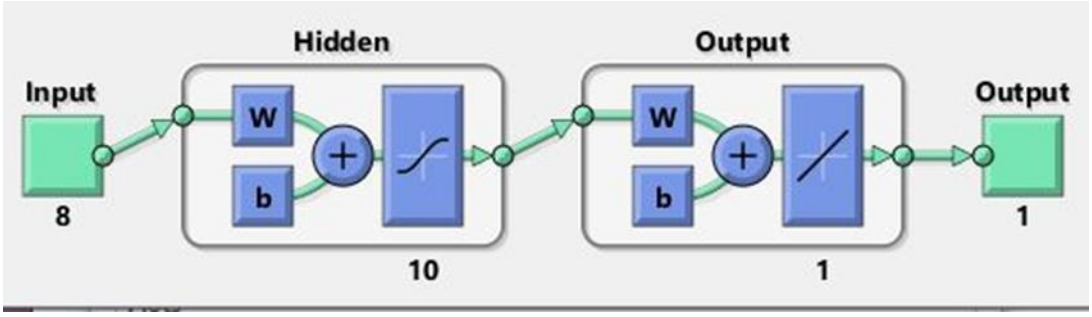


Figure 4 5: Artificial Neural Network architecture.

Configuration of neural network is as follows:

Type of network = Artificial Neural Network

No of inputs=1

Input layer=8

Hidden neurons layer=10

No of output =1

The procedure and result of ANN proves to be interesting, by giving the results of the gradients, Mean square error, the epochs iterations, time performance and the regressions. Figure4.6 shows the training procedures.

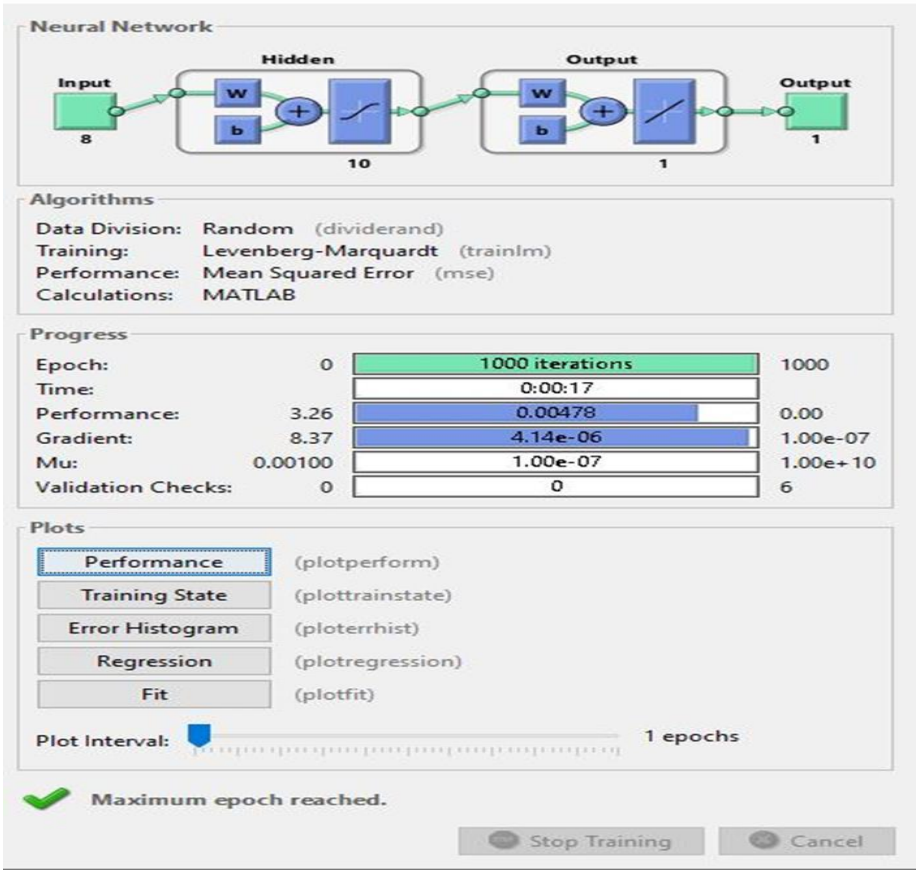


Figure 4 6: ANN Training process interface.

The gradients, mu control parameter for error convergence, and validation check results are reported in figure 4.7. The Gradient feature at 1000 Epoch helps in allowing the adjustment parameter of the network to minimize the output deviation, training was done for 4.137e-06.

The Mu feature at 1000 Epoch, training was achieved for 1e-07 and Val check feature at 1000Epoch, training was achieved with 0 validation checks, and the validation check provides an unbiased evaluation of our model to make it fit for training.

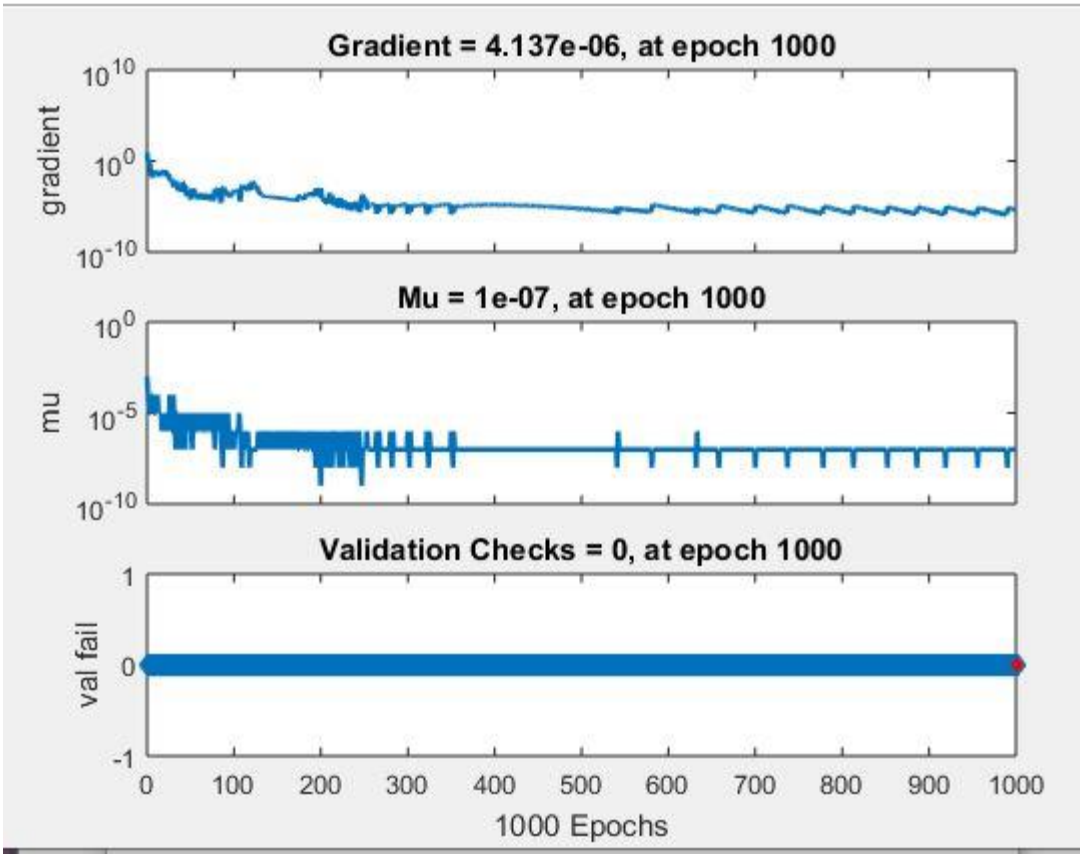


Figure 4 7: Illustration of the various ANN training state features on 1000Epochs

Figure 4.8 shows the regression plot of ANN after been trained. Output was plotted against target; the training produced a regression value of 0.99735.

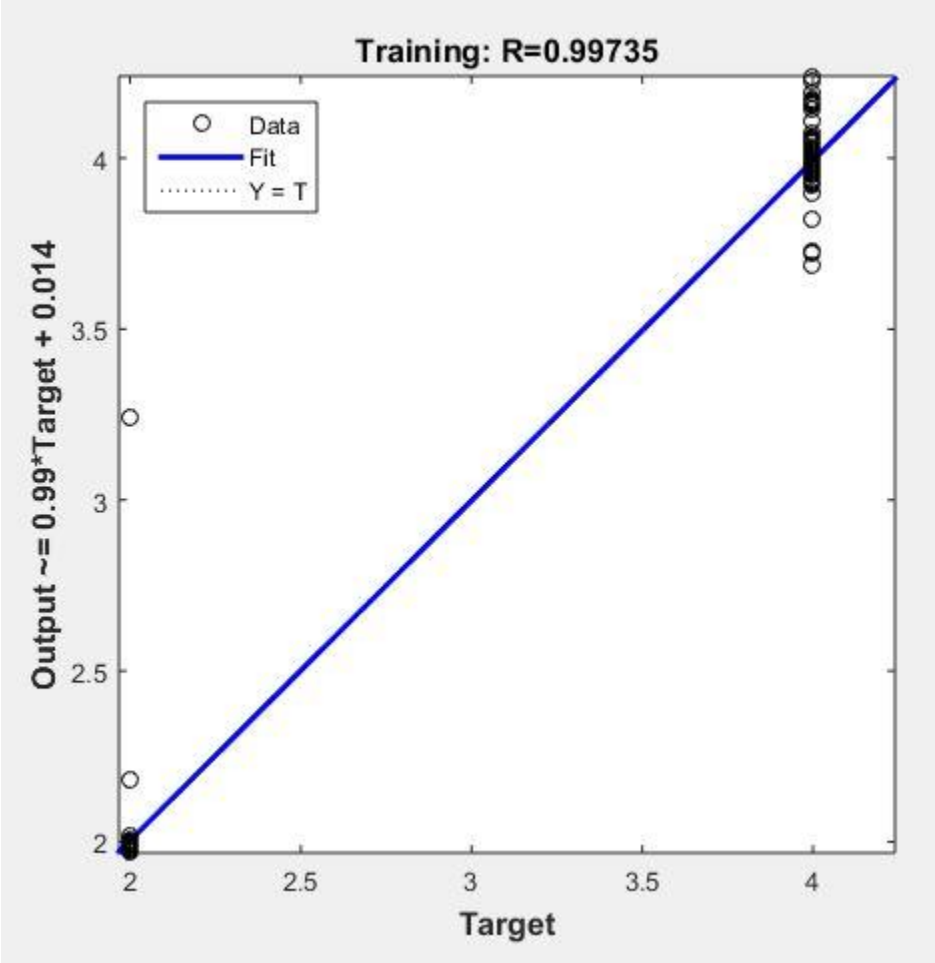


Figure 4 8: ANN Training regression chart.

Figure 4.9 is a screen shot of the Root mean square error (mse) at an epoch of 1000 and it gives its best training performance of 0.0047805.

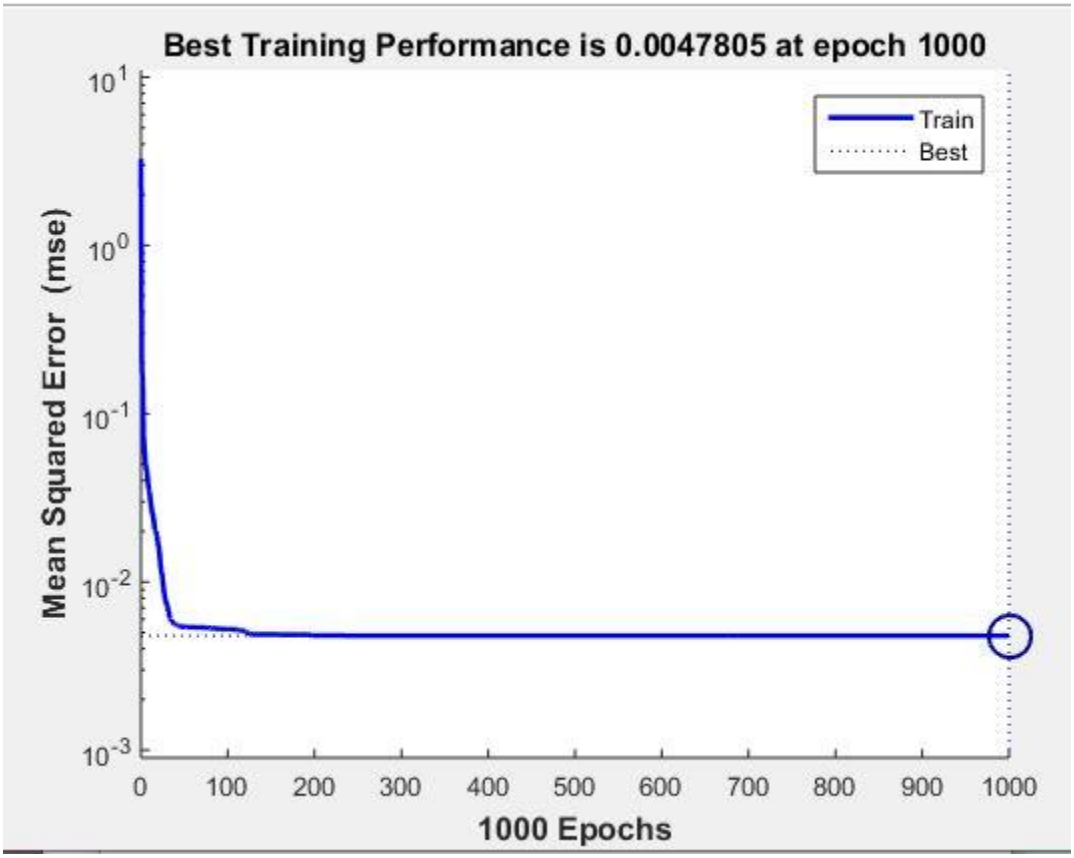


Figure 4 9: Mean Squared Error with best training performance of 0.0047805

The histogram did not capture any error between the point where the highest error was recorded at an iteration instance 350 and the next or nearest recorded error at 120 iteration instance.

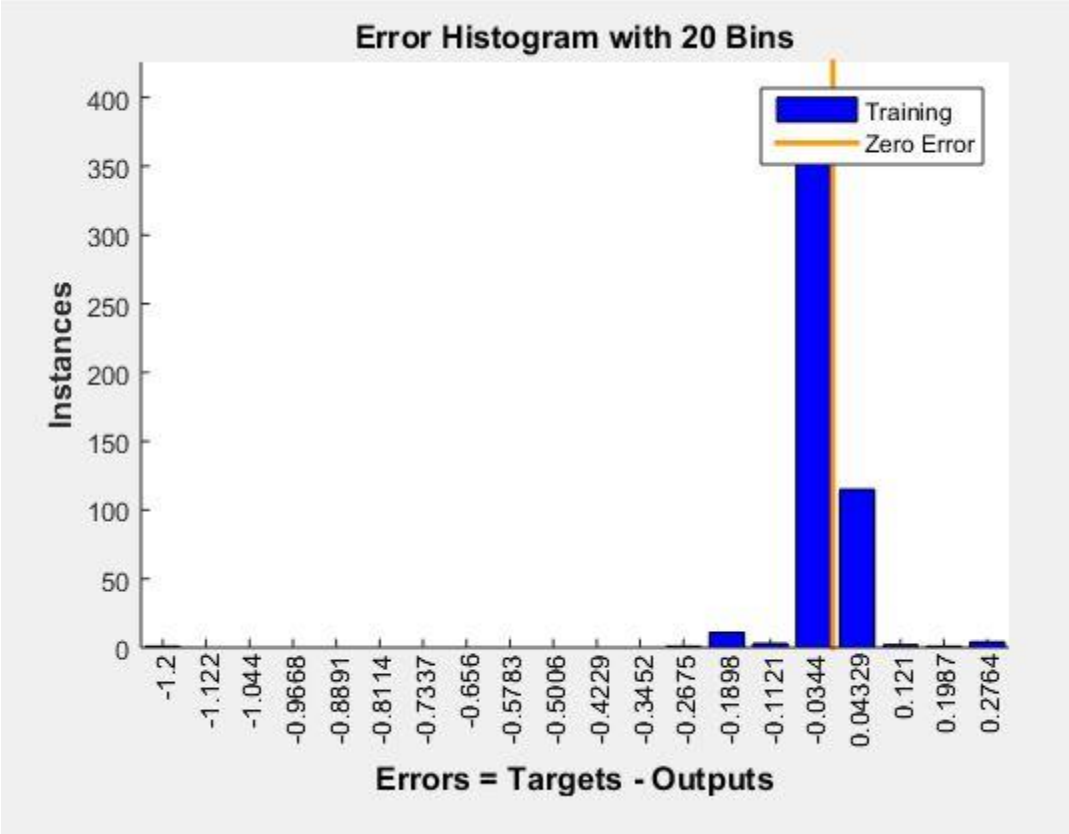
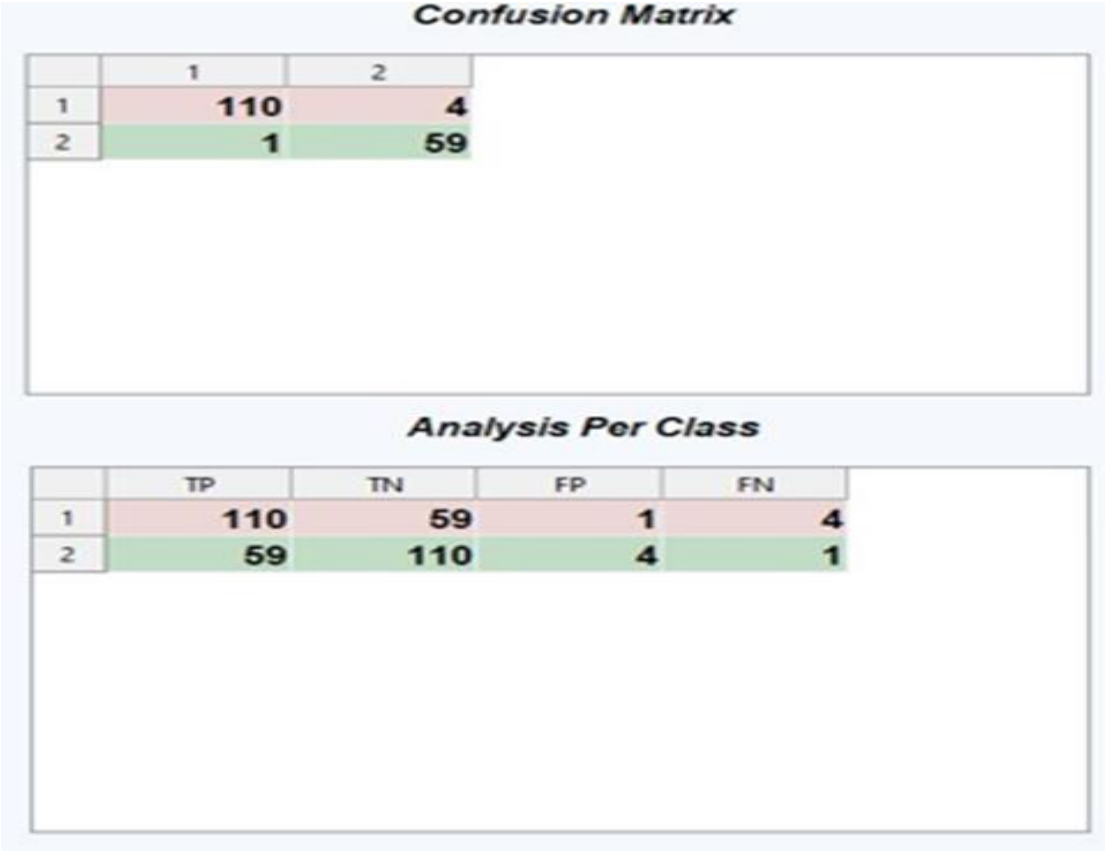


Figure 4 10: Error Histogram for ANN training process.

Figure 4.11 shows the confusion matrix result for the classification of PSO with ANN technique. The True positive yields 110, false negative yields 1, false positive yields 4, and true negative yields 59. TP=110 FP=4 TN=59 FN=1.



Confusion Matrix

	1	2
1	110	4
2	1	59

Analysis Per Class

	TP	TN	FP	FN
1	110	59	1	4
2	59	110	4	1

Figure 4 11: ANN confusion Matrix

The essence of generating a confusion matrix is to be able to compute for the assessment metrics.

ACCURACY: The most basic scoring factor is ACCURACY. It works out the percentage of cases that are properly categorized.

Accuracy= (TP + TN) (TP +TN +FP +FN)

(110 + 59) / (110 + 59 +4 + 1) = 169/174

Accuracy Rate= 0.97126

Percentage Accuracy rate = 97.13%

SENSITIVITY: The sensitivity score, also known as the Recall or True Positive, indicates how likely a sample with breast cancer characteristics would result in a positive test result.

Sensitivity = TP/ (TP + FN)

(110)/ (110 + 1)

110/111

Sensitivity rate =0.9909

Percentage Sensitivity Rate=99.10%

SPECIFICITY: The specificity, also known as the True negative, refers to a classifier's ability to recognize negative outcomes.

Specificity= TN/ (FP + TN)

59 / (4 + 59)

59/63

Specificity rate =0.9365

Percentage Specificity Rate= 93.65%

PRECISION: This is a relevant measure-retrieved example.

Precision =TP/ (TP + FP)

110/ (110 + 4)

110/114

Precision rate = 0.9649

Percentage Precision Rate= 96.49%

F-SCORE: It's a metric for assessing a model's precision and recall.

F-SCORE = 2*TP/ (2*TP + FP +FN)

2 *110/ (2*110 +4+1)

220/225

F-SCORE rate =0.9777

Percentage F-SCORE Rate =97.77%

RECALL: Is a metric on how many true positives are predicted from the total number of positives in the dataset. Sensitivity is another name for it.

RECALL= TP/TP+FN

110/ (110 +1)

110/111

Recall rate=0.9909

Percentage RECALL rate = 99.09%

ERROR RATE: Which is the percentage of times the forecast is incorrect.

ERROR RATE: FN+FP/TP +FN

1 +4+110 + 1

5/111

Error rate =0.0450

FALSE ACCEPTANCE RATE: This happens when we admit a user that we should have refused in the first place.

FAR= FP/ (FN + TN)

4 / (1 + 59)

4/60

=0.0666

FALSE REJECTION RATE: This happens when we refuse a user that should have been approved in the first place.

FRR= FN/ (FN +TP)

1 / (1 +110)

1/111

=0.0090

4.7 Result and Discussion for SVM

Figure 4.12 shows the scattered plot of the selected features for classification and how they were distributed across. Instances are seen to be more distributed across column 0 and 1 and sparingly distributed between column 8 and 9 also in between column 8 and 9, there is a single misfit instance represented with the (x)

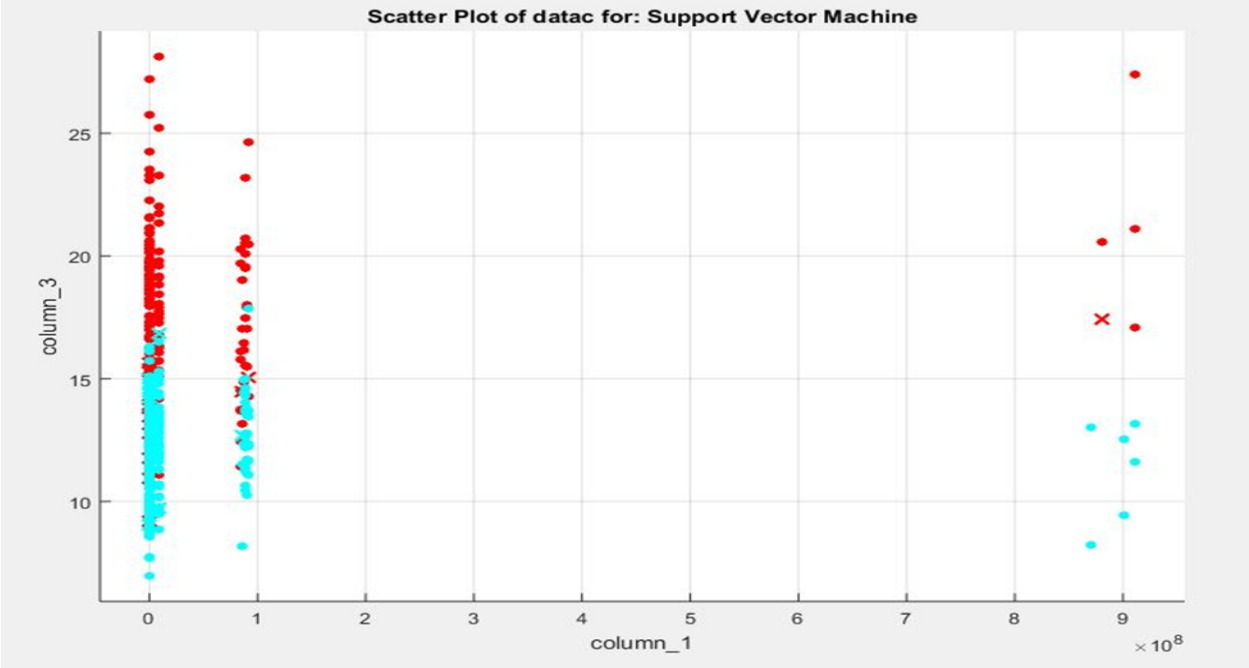


Figure 4 12: Scattered plot of PSO with SVM

Figure 4.13 shows the receiver operating characteristic curve for identifying the performance of SVM classifier. This curve plots two parameters: True positive rate axis over false positive rate axis. The curve starts from zero (0) at the false positive axis to one (1) at the same axis.

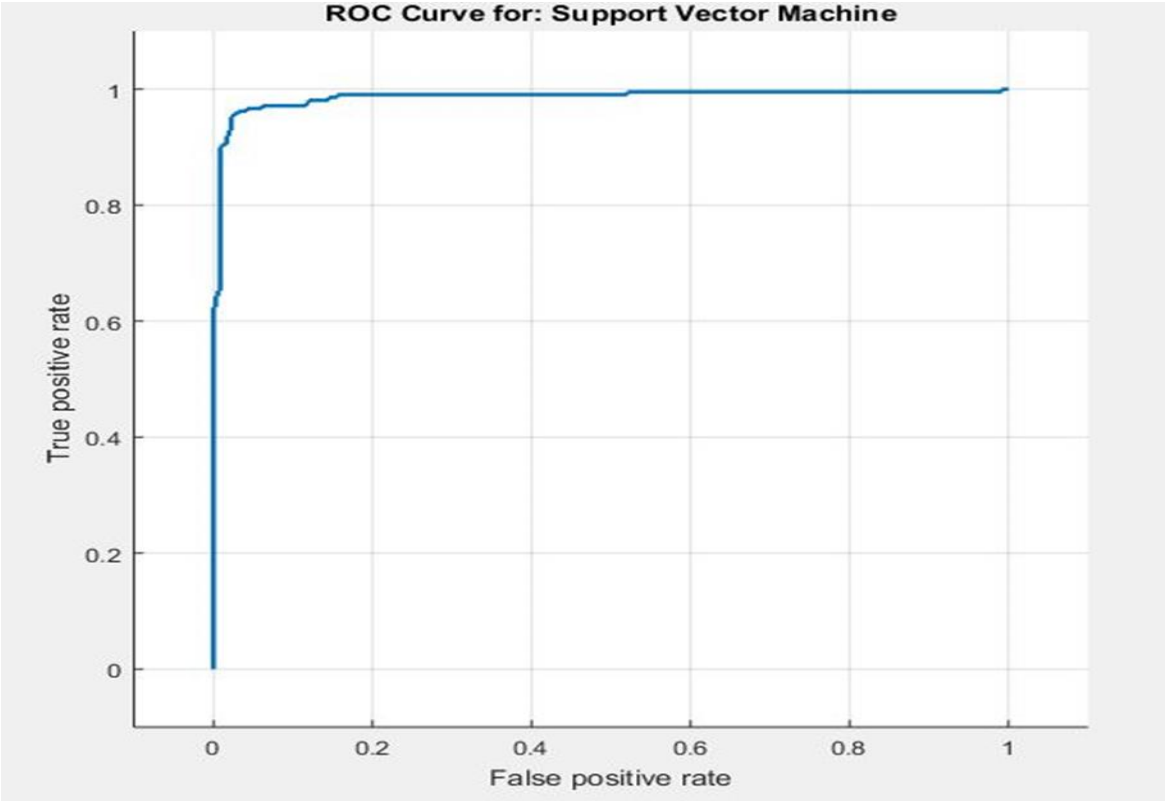


Figure 4 13: ROC curve for SVM

The Figure 4.14 shows the confusion matrix result for the classified components, which was extracted using PSO and SVM technique. The True positive yields 93.3% and false negative yields 2.2% likewise false positive yields 6.7% and true negative yields 97.8%. TP=196 FP=14 TN=351 FN=8

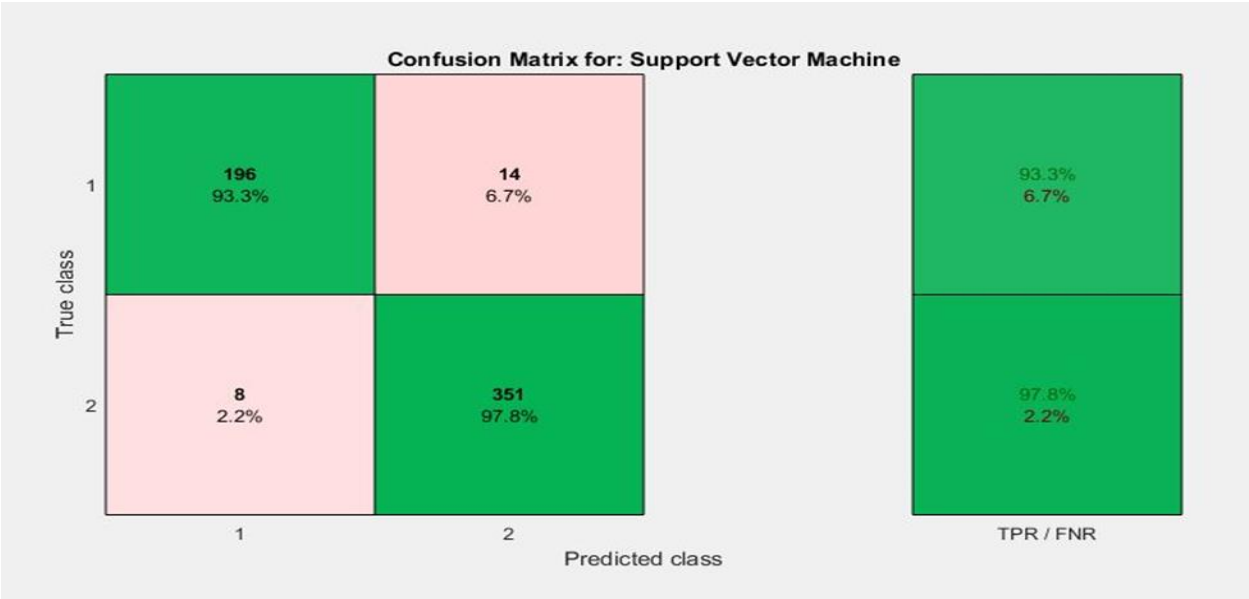


Figure 4 14: Confusion Matrix for SVM

ACCURACY

Accuracy= (TP + TN) (TP +TN +FP +FN)

(196 + 351) / (196 + 351 +14 + 8) = 547/569

Accuracy Rate= 0.9613

Percentage Accuracy rate = 96.13%

SENSITIVITY

Sensitivity = TP/ (TP + FN)

(196)/ (196 + 8)

196/204

Sensitivity rate =0.9607

Percentage Sensitivity Rate=96.07%

SPECIFICITY.

Specificity= TN/ (FP + TN)

351/ (14 + 351)

351/365

Specificity rate =0.9616

Percentage Specificity Rate= 96.16%

PRECISION

Precision =TP/ (TP + FP)

196/ (196 + 14)

196/210

Precision rate = 0.9333

Percentage Precision Rate= 93.33%

F-SCORE

F-SCORE = 2*TP/ (2*TP + FP +FN)

2 *196/ (2*196 +14+8)

392/414

F-SCORE rate =0.9468

Percentage F-SCORE Rate =94.68%

RECALL

RECALL= TP/TP+FN

196/ (196 +8)

196/204

Recall rate=0.9607

Percentage RECALL rate = 96.07%

ERROR RATE

ERROR RATE: FN+FP/TP +FN

8 +14/196 + 8

22/204

Error rate =0.107

FALSE ACCEPTANCE RATE

FAR= FP/ (FN + TN)

14 / (8 + 351)

14/359

=0.0389

FALSE REJECTION RATE

FRR= FN/ (FN +TP)

8(8 +196)

8/204

=0.03921

4.8 Results and Discussion for DT

Figure 4:15 depicts the selected feature and how they were distributed across plot. Instances are seen to be more distributed across column 0 and 1 and sparingly distributed between column 8 and 9 also in between column 8 and 9, there is no misfit instance.

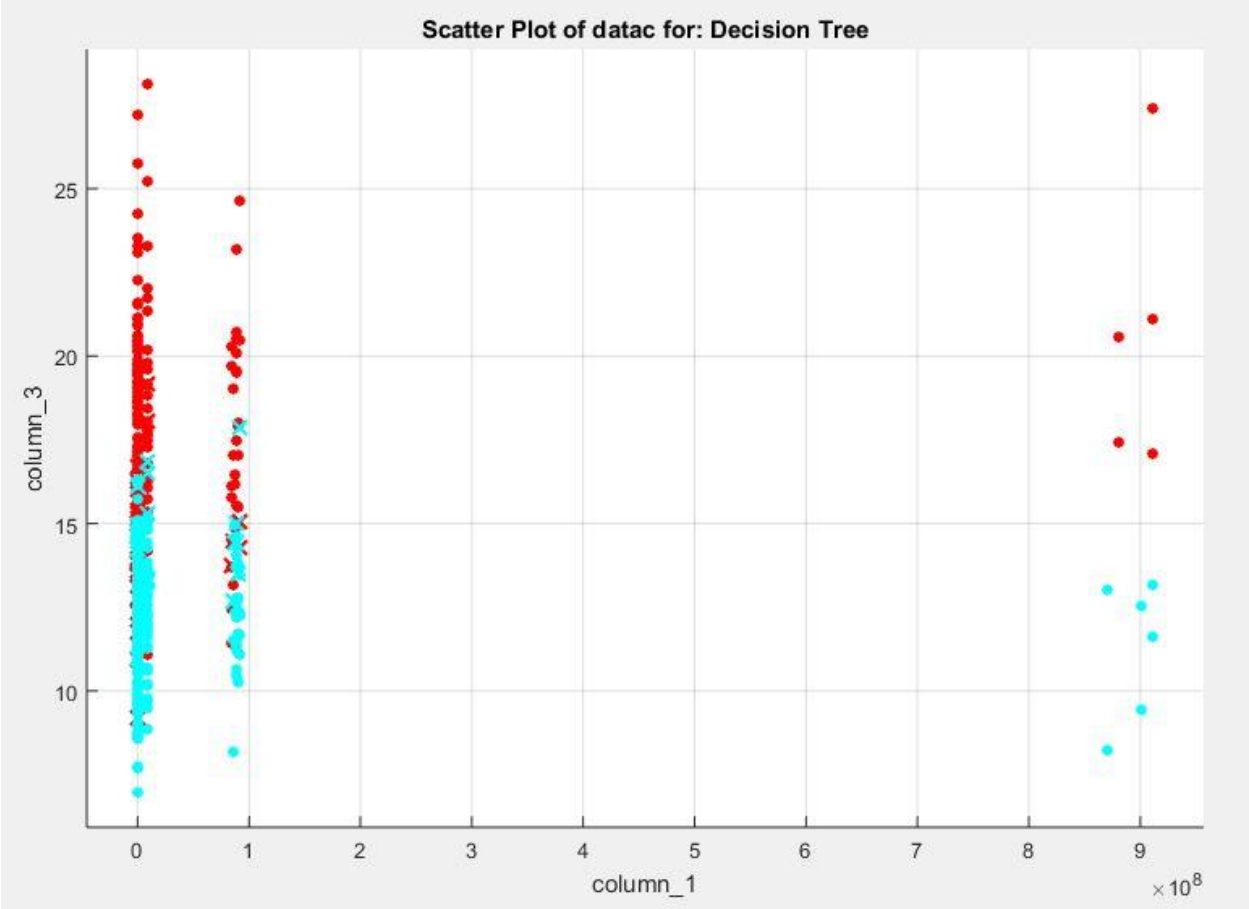


Figure 4 15: Scattered plot for PSO with DT

Figure 4.16: shows the receiver operating characteristic curve, this represents the performance of DT classifier at all classification threshold. This curve plots two parameters: True positive rate axis over false positive rate axis. The curve starts from zero (0) at the false positive axis to one (1) at the same axis.

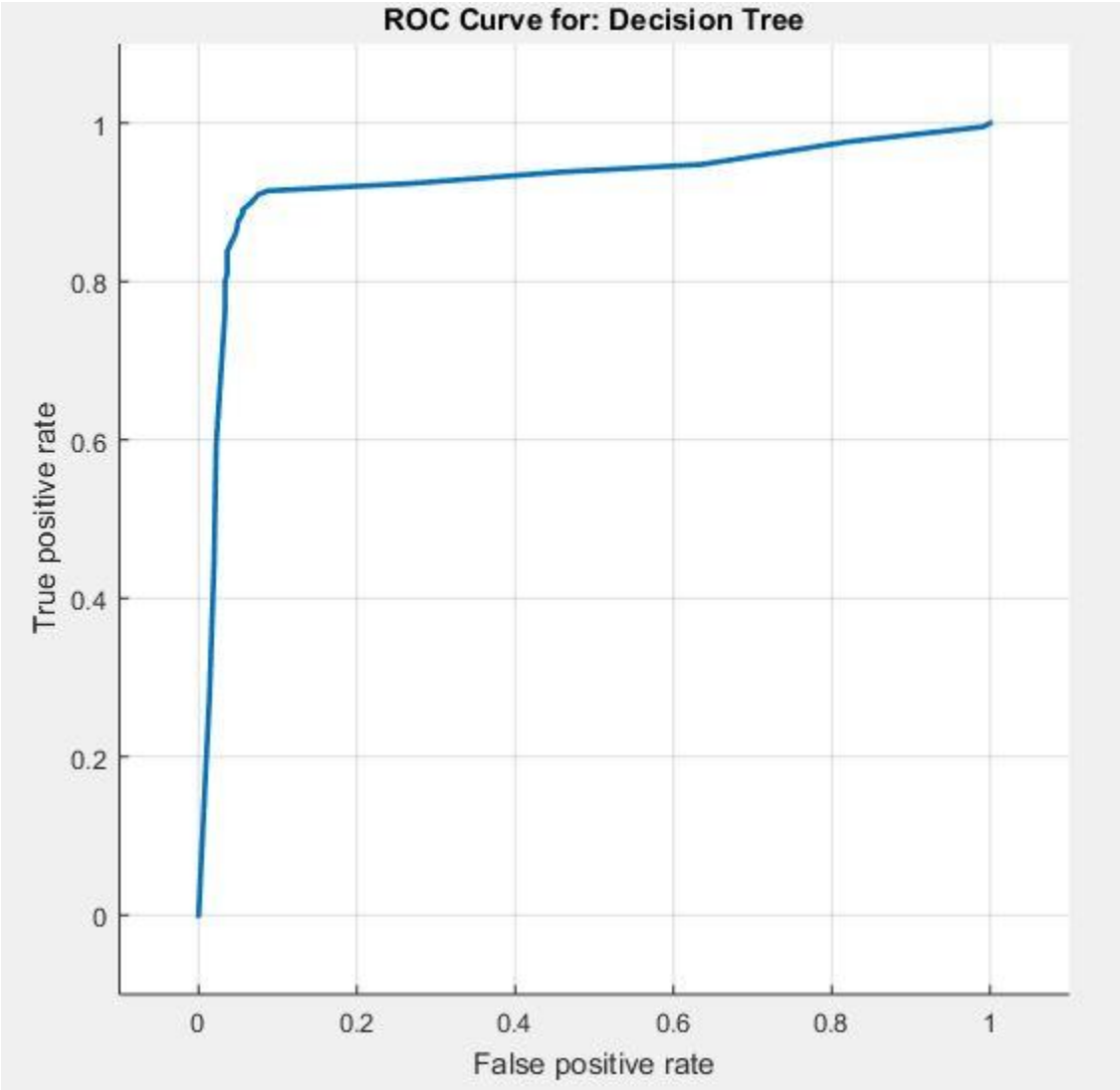


Figure 4 16: ROC curve for DT

Figure 4.17 shows the confusion matrix result for the classified components, which was extracted using PSO with DT technique. The True positive yields 87.6% and false negative yields 5% likewise false positive yields 12.4% and true negative yields 95%. TP=184 FP=26 TN=341 FN=18

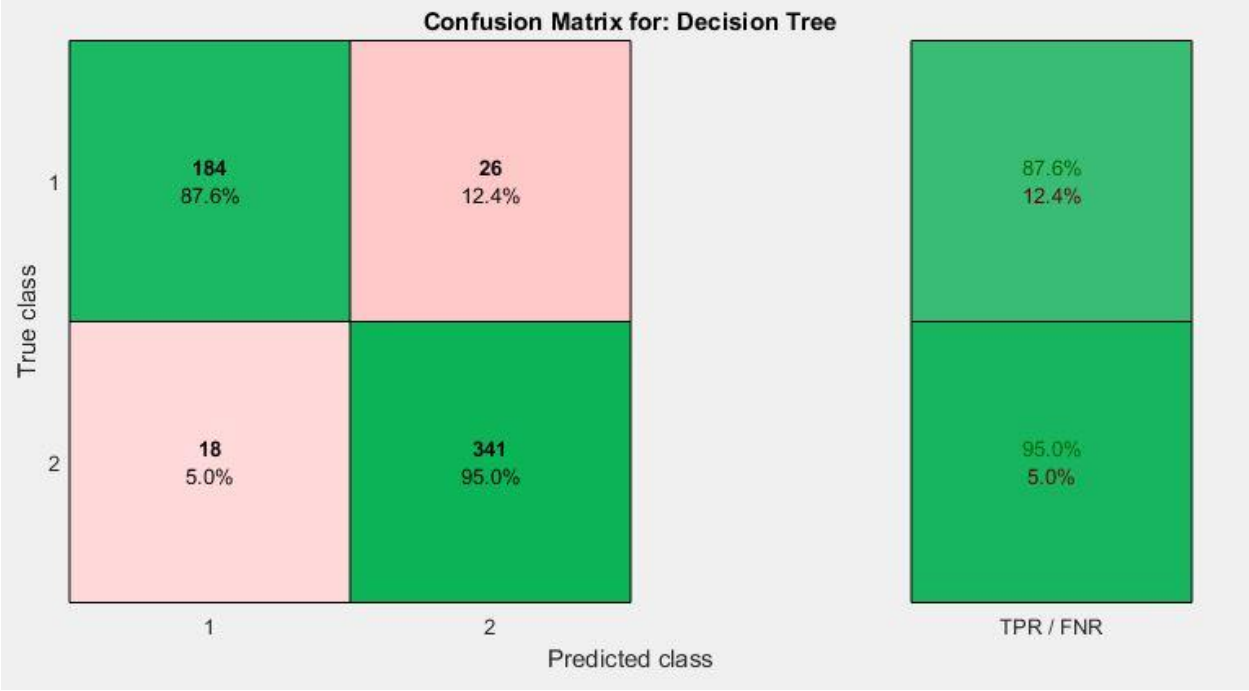


Figure 4 17: DT confusion Matrix

ACCURACY

Accuracy= (TP + TN) (TP +TN +FP +FN)

(184+ 341) / (184 + 341 +26 + 18) = 525/569

Accuracy Rate= 0.9226

Percentage Accuracy rate = 92.26%

SENSITIVITY

Sensitivity = TP/ (TP + FN)

(184)/ (184 + 18)

184/202

Sensitivity rate =0.9108

Percentage Sensitivity Rate=91.08%

SPECIFICITY

Specificity= TN/ (FP + TN)

341/ (26 + 341)

341/367

Specificity rate =0.9291

Percentage Specificity Rate= 92.91%

PRECISION

Precision =TP/ (TP + FP)

184/ (184 + 26)

184/210

Precision rate = 0.8761

Percentage Precision Rate= 87.61%

F-SCORE

F-SCORE = 2*TP/ (2*TP + FP +FN)

2 *184/ (2*184 +26+18)

368/412

F-SCORE rate =0.8932

Percentage F-SCORE Rate =89.32%

RECALL

RECALL= TP/TP+FN

184/ (184 +18)

184/202

Recall rate=0.9108

Percentage RECALL rate = 91.08%

ERROR RATE

ERROR RATE: FN+FP/TP +FN

18 +26/184 + 18

44/202

Error rate =0.217

FALSE ACCEPTANCE RATE

FAR= FP/ (FN + TN)

26 / (18 + 341)

26/359

=0.0724

FALSE REJECTION RATE

FRR= FN/ (FN +TP)

18(18 +184)

18/202

=0.08910

Table 4. 1: Comparative Evaluation of ANN, SVM and DT

SN	PERFORMANCE METRICS	ANN BASED METHOD	SVM BASED METHOD	DT BASED METHOD
1	Accuracy	97.13	96.13	92.26
2	Sensitivity	99.10	96.07	91.08
3	Specificity	93.65	96.16	92.91
4	Precision	96.49	93.33	87.61
5	F-SCORE	97.77	94.68	89.32
6	Recall	99.09	96.07	91.08
7	Error rate	0.0450	0.107	0.217
8	false Acceptance Rate	0.0666	0.0389	0.0724
9	false Rejection Rate	0.0090	0.03921	0.0891

The performance analysis of classification using PSO on breast cancer dataset shows that Artificial Neural Network achieves higher values in the datasets on maximization parameters in terms of its accuracy, sensitivity, precision, F-score, and Recall than other classifiers. Likewise, Artificial Neural Network produce a lower error rate, false acceptance rate and false rejection rate than other classifiers. In addition, Support Vector Machine performs better than DT in terms of its accuracy, sensitivity, specificity, precision, F-score, Recall than Decision Tree and it produces a lower error rate, false acceptance rate and false rejection rate then Decision Tree, while DT is the least classifier with the lowest result in term of all the above performance metrics and it produces the highest error throughout the training.

Table 4. 2: Comparison Of This Study With Other Techniques In Literature In Terms Of Accuracy

S/N	AUTHOR(S)	TECHNIQUES	ACCURACY RATE
1	Kanchanamani (2016)	They used Five machine learning techniques namely: Naïve bayes, KNN, MLP, LDA, and Support vector machine,	SVM gave the highest accuracy of 87.3% without tenfold cross validation and likewise with tenfold cross validation SVM gave an accuracy of 92.5%
2	Igbal *et al*,(2017)	They used random forest, Bayesian networks, and support vector machine	They had an accuracy of 97%.
3	Sakri *et al*.,(2018)	The techiques used includes; naïve bayes, Rep tree, K-nearest neighbors	Without PSO, Rep tree gave an accuracy of 76.3% with PSO, naïve bayes gave an accuracy of 81.3%
4	Vijayarajan *et al*, (2018)	They used KNN, J48, Naïve bayes, and SVM	Naïve bayes gave the highest accuracy of 95.99%
5	Bataineh,(2019)	The techniques used includes:K-nearest neighbors, Classification and regression tree, multilayer	MLP had the highest accuracy of 96.7%

		perceptron(MLP), support vector machine and Gaussian naïve bayes.	
6	Ganggayah *et al* (2019)	They used 6 different classifiers namely: Decision tree, Random forest, Neural networks, Extreme boost, Logistic regression, and SVM.	Random forest gave the highest accuracy of 82.70%
7	Afolayan .(2021)	Three (3) machine learning classifiers were used namely: Artificial neural networks, Support vector machine and Decision tree.	Artificial neural networks gave the highest accuracy of 97.13%

The table above comprises of six different experiments that were compared. The study developed a model with an accuracy of 97.13%. When compared with past literature, for instance, Kanchanamani (2016) had an accuracy of 92.5% using five classifiers, Igbal *et al* (2017) made use of three classifiers as well and had an accuracy of 97%, Sakri *et al.,(*2018) had an highest accuracy of 81.3% using three classifiers , Vijayarajan *et al,* (2018) used four classifiers and had an accuracy of 95.99% , Bataineh,(2019) had an accuracy of 96.7% when five classifiers were compared, lastly, Ganggayah *et al* (2019) had the highest accuracy of 82.70% when six classifiers were compared.As shown in table

4.2, the developed system used three algorithms for classification and one algorithm for features selection (PSO) and outperformed the listed previous research work.

CHAPTER FIVE

5.0 SUMMARY, CONCLUSION, AND RECOMMENDATION

5.1. Summary

This thesis worked on a comparative analysis of three (3) machine learning techniques for breast cancer detection, the Wisconsin breast cancer datasets was used to analyse the model. Three supervised algorithms for machine-learning approach was used for classification purpose, Particle Swarm Optimization algorithm (PSO) was used, to pick relevant features from the raw dataset to eliminate and reduce noises for a better outcome, the classification of the reduced data uses the support vector machine (SVM), artificial neural networks (ANNs), and decision tree (DT), on the obtained data from a UCI machine-learning repository. The model was simulated using The Matlab platform. The result reveals that ANNs gained the highest accuracy, sensitivity, precision, and F-score, and recall of 97.13%, 99.10%, 96.49%, 97.77%, and 99.09% respectively, and it also produced the lowest false acceptance rate, error rate, and false rejection rate of 0.0450, 0.0666 and 0.0090 respectively, this study will be more of more importance to medical practitioners and clinicians in decision-making.

5.2. Conclusion

The model was developed by combining feature selection with machine learning techniques (classifiers) for breast cancer detection. The system uses Particle Swarm Optimization (PSO), to reduce the dataset features and attributes, the reduced dataset was used on three classifiers namely: Artificial Neural Network (ANN), Support Vector

Machine (SVM) and Decision Tree (DT). Data processing techniques for the diagnosis and treatment of breast cancer need to be developed in order to enable faster treatment and also to produce more accurate results.

The PSO used in the system handles the correlation of the data more efficiently and makes our classifiers to obtain optimal result (in other words, it enhances our classifiers efficiency and reduces noise from our system.

The three classifiers all produced a higher accuracy value above 90% accuracy respectively and the results are displayed in a graphical format and tabular format. The system proposed a more accurate result when compared to other existing literatures. Out of all the three classifiers, ANN produced an accurate of 97.13% and it yielded more in terms of other performances metrics used when compared to other classifiers. Thus, the use of PSO feature selection increases the performance of the classifiers compared to the state of Art.

5.3. Major Contribution

This work contributes to knowledge by developing an efficient model for breast cancer detection. The model was evaluated and compared using three (3) machine learning techniques.

5.4. Future Works and Recommendation

The future work can be done by applying both feature selection and feature extraction to the raw dataset, to check if the model would produce a more accurate result. Further research can be done on why ANN produces a better result in terms of accuracy, specificity, sensitivity, recall, f-score, false acceptance rate and false rejection rate. Also deep learning machine learning could be used for classification to train and test the system with different

datasets. Future researchers can explore algorithms such Convolutional Neural Network among others.

REFERNCES

A Relative Analysis of Multi-Relational Decision Tree Learning Algorithm. (2017). In *International Journal of Science and Research (IJSR)* (Vol. 6, Issue 1). https://doi.org/10.21275/ART20164150

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, *7*, e7702. https://doi.org/10.7717/peerj.7702

Ak, M. F. (2020). A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. *Healthcare*, *8*(2), 111. https://doi.org/10.3390/healthcare8020111

Akram, M., Iqbal, M., Daniyal, M., & Khan, A. U. (2017). Awareness and current knowledge of breast cancer. Biological research. *Biological Research*, *50*(1), 33. https://doi.org/10.1186/s40659-017-0140-9

Alexis Marcano, C., Joel, Q., & Diego, A. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications*, *38*, 9573–9579.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. https://doi.org/10.15252/msb.20156651

Arif Harahap, W., Ramadhan, Khambri, D., Haryono, S., & Dana Nindrea, R. (2017). Outcomes of Trastuzumab Therapy for 6 and 12 Months in Indonesian National Health Insurance System Clients with Operable HER2-Positive Breast Cancer. *Asian Pacific Journal of Cancer Prevention : APJCP*, *18*(4), 1151–1156. https://doi.org/10.22034/APJCP.2017.18.4.1151

Ashraf, O. I., & Siti, M. S. (2018). Intelligent breast cancer diagnosis based on enhancedPareto optimal and multilayer perceptron neural network. *International Journalof Computer Aided Engineering and Technology, Inderscience*, *10*, 543–556.

Asri, H., Mousannif, H., Moatassime, H. Al, & Noel, T. (2016). Using Machine Learning

Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*, 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224

Balcan, M.-F., & Blum, A. (2006). On a theory of learning with similarity functions. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 73–80. https://doi.org/10.1145/1143844.1143854

Balcan, M.-F., Hanneke, S., & Wortman, J. (2008). *The True Sample Complexity of Active Learning.*

Basu, C. B., Wahba, M., Bullocks, J. M., & Elledge, R. (2008). Paget Disease of a Nipple Graft Following Completion of a Breast Reconstruction With a Nipple-Sharing Technique. *Annals of Plastic Surgery*, *60*(2), 144–145. https://doi.org/10.1097/SAP.0b013e31806a592b

Bataineh, A. Al. (2019). A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection. *International Journal of Machine Learning and Computing*, *9*.

Baylin, S. B., & Jones, P. A. (2016). Epigenetic Determinants of Cancer. *Cold Spring Harbor Perspectives in Biology*, *8*(9), a019505. https://doi.org/10.1101/cshperspect.a019505

Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 1–4. https://doi.org/10.1109/ICEDSA.2016.7818560

Ben-Hur, A., & Guyon, I. (n.d.). Detecting Stable Clusters Using Principal Component Analysis. In *Functional Genomics* (pp. 159–182). Humana Press. https://doi.org/10.1385/1-59259-364-X:159

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., & Haley, C. S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports*, *5*(1), 10312. https://doi.org/10.1038/srep10312

Berry, M. J., & Linoff, G. S. (2008). *Mastering data mining: The art and science of customer relationship management.*

Bertsekas, D. ., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*.

Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using Genetically Optimized Neural Network model. *Expert Systems with Applications*, *42*(10), 4611–4620. https://doi.org/10.1016/j.eswa.2015.01.065

Bhoo-Pathy, N., Peeters, P. H., Uiterwaal, C. S., Bueno-de-Mesquita, H. B., Bulgiba, A. M., Bech, B. H., Overvad, K., Tjønneland, A., Olsen, A., Clavel-Chapelon, F., Fagherazzi, G., Perquier, F., Teucher, B., Kaaks, R., Schütze, M., Boeing, H., Lagiou, P., Orfanos, P., Trichopoulou, A., … van Gils, C. H. (2015). Coffee and tea consumption and risk of pre- and postmenopausal breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort study. *Breast Cancer Research*, *17*(1), 15. https://doi.org/10.1186/s13058-015-0521-3

Bhukya, D. P., & Ramachandram, S. (2010). Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. *International Journal of Computer and Electrical Engineering*, 660–665. https://doi.org/10.7763/IJCEE.2010.V2.208

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C., Hoffmann, U., Schwartz, L. H., Gillies, R. J., Huang, R. Y., & Aerts, H. J. W. L. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, caac.21552. https://doi.org/10.3322/caac.21552

Brank, J., Mladenić, D., Grobelnik, M., Liu, H., Mladenić, D., Flach, P. A., Garriga, G. C., Toivonen, H., & Toivonen, H. (2011). Feature Selection. In *Encyclopedia of Machine Learning* (pp. 402–406). Springer US. https://doi.org/10.1007/978-0-387-30164-8_306

Breast cancer:Statistics, Approved by the cancer.Net Editorial Board. (2017). http//www.cancer.net/cancertypes/breast-cancer/statistics.

Breiman, L., JH, F., RA, O., & CJ, S. (1984). *Classification and Regression Trees*.

C A Padmanabha Reddy, Y., Viswanath, P., & Eswara Reddy, B. (2018). Semi-supervised learning: a brief review. *International Journal of Engineering & Technology*, *7*(1.8), 81. https://doi.org/10.14419/ijet.v7i1.8.9977

Chen, S., Xu, Z., Tang, Y., & Liu, S. (2014). An Improved Particle Swarm Optimization Algorithm Based on Centroid and Exponential Inertia Weight. *Mathematical*

Problems in Engineering, *2014*, 1–14. https://doi.org/10.1155/2014/976486

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Das, B., Vig, M., Khurana, K. K., & Madhubala, R. (2000). Isolation and Characterization of Human Breast Adenocarcinoma Cells Made Resistant to α - Difluoromethylornithine. *Cancer Investigation*, *18*(2), 115–122. https://doi.org/10.3109/07357900009038242

Dasgupta, S., Kalai, A. T., & Monteleoni, C. (2005). *Analysis of Perceptron-Based Active Learning* (pp. 249–263). https://doi.org/10.1007/11503415_17

Deng, C., Ji, X., Rainey, C., Zhang, J., & Lu, W. (2020). Integrating Machine Learning with Human Knowledge. *IScience*, *23*(11), 101656. https://doi.org/10.1016/j.isci.2020.101656

Dey, N. (2019). *Classification techniques for medical image analysis and computer aided diagnosis*. academic press.

Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. *Journal of Healthcare Engineering*, *2019*, 1–11. https://doi.org/10.1155/2019/4253641

Dr. Kadhim B.S. Al Janabi, R. K. (2018). Data Reduction Techniques: A Comparative Study for Attribute Selection Methods. In *International Journal of Advanced Computer Science and Technology*. (Vol. 8). Research India Publications.

Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W., Liu, B., Lei, Y., Du, S., Vuppalapati, A., Luu, H. H., Haydon, R. C., He, T.-C., & Ren, G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & Diseases*, *5*(2), 77–106. https://doi.org/10.1016/j.gendis.2018.05.001

Ferroni, P., Zanzotto, F., Riondino, S., Scarpato, N., Guadagni, F., & Roselli, M. (2019). Breast Cancer Prognosis Using a Machine Learning Approach. *Cancers*, *11*(3), 328. https://doi.org/10.3390/cancers11030328

Franchi, A. (2020). Metastatic Tumors. In *Pathology of Sinonasal Tumors and Tumor-Like Lesions* (pp. 233–235). Springer International Publishing. https://doi.org/10.1007/978-3-030-29848-7_11

Gheyas, I. A., & Smith, L. S. (2010). Feature subset selection in large dimensionality domains. *Pattern Recognition*, *43*(1), 5–13. https://doi.org/10.1016/j.patcog.2009.06.009

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (n.d.). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, *15*(1), 41–51. https://doi.org/10.21873/cgp.20063

Huml, M., Silye, R., Zauner, G., Hutterer, S., & Schilcher, K. (2013). Brain Tumor Classification Using AFM in Combination with Data Mining Techniques. *BioMed Research International*, *2013*, 1–11. https://doi.org/10.1155/2013/176519

Ian, W., & Eibe, F. (2005). *Practical Machine Learning Tools and Techniques, Second Edition* (second edi).

İlhan, A., Gülersoy, A. E. (2019). Discovery Learning Strategy in Geographical Education: A Sample of Lesson Design. *Review of International Geographical Education Online (RIGEO)*, *9*(3), 523–541. https://doi.org/10.33403/rigeo.672975

Isabelle, G., & Andre, E. (2003). *An Introduction to Variable and Feature Selection*. Journal of Machine Learning Research.

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SpringerLink*, *1*. https://doi.org/10.1007/s42979-020-00305-w

Jackson, C. A., Castro, D. M., Saldi, G.-A., Bonneau, R., & Gresham, D. (2020). Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *ELife*, *9*. https://doi.org/10.7554/eLife.51254

James, G., Witten, D., Hastie, T., & Robert, T. (2013). *An Introduction to Statistical Learning Gareth James Daniela Witten Trevor Hastie Robert Tibshirani Statistics An Introduction to Statistical Learning with Applications in R*. springer.

Jason Brownlee. (2019). *Machine learning mastery* (Machine le).

Jenny, C., Geovanny,Marulanda Antonio, B., & Javier Reneses. (2020). Air Temperature Forecasting Using Machine Learning Techniques: A Review. *Institute for Research in Technology (IIT)*.

julie, hamon. (2013). *Combinatorial optimization for the selection of large-dimensional*

regression variables: Application in animal genetics.

Kanchanamani M*, V. P. (2016). Performance evaluation and comparative analysis of various machine learning techniques for diagnosis of breast cancer. *Biomedical Research*, *27*(3).

Kashyap, R. (2019a). *Deep Learning* (pp. 130–158). https://doi.org/10.4018/978-1-5225-7955-7.ch006

Kashyap, R. (2019b). *Machine Learning for Internet of Things* (pp. 57–83). https://doi.org/10.4018/978-1-5225-7458-3.ch003

Kele?, S., & Segal, M. R. (2002). Residual-based tree-structured survival analysis. *Statistics in Medicine*, *21*(2), 313–326. https://doi.org/10.1002/sim.981

Kennedy, J., & Eberhart, R. (n.d.). Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, *4*, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Kim, H., & Loh, W.-Y. (2001). Classification Trees With Unbiased Multiway Splits. *Journal of the American Statistical Association*, *96*(454), 589–604. https://doi.org/10.1198/016214501753168271

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Lavrač, N., Škrlj, B., & Robnik-Šikonja, M. (2020). Propositionalization and embeddings: two sides of the same coin. *Machine Learning*, *109*(7), 1465–1507. https://doi.org/10.1007/s10994-020-05890-8

Lee, L. H., Wan, C. H., Yong, T. F., & Kok, H. M. (2010). A Review of Nearest Neighbor-Support Vector Machines Hybrid Classification Models. *Journal of Applied Sciences*, *10*(17), 1841–1858. https://doi.org/10.3923/jas.2010.1841.1858

Lee, S., Liang, X., Woods, M., Reiner, A. S., Concannon, P., Bernstein, L., Lynch, C. F., Boice, J. D., Deasy, J. O., Bernstein, J. L., & Oh, J. H. (2020). Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study. *PLOS ONE*, *15*(2), e0226157. https://doi.org/10.1371/journal.pone.0226157

Mason, G. (2017). No Lump Required: A Patient Driven Inflammatory Breast Cancer Reserach Initiative Using the Peer Platform. *The Breast*, *36*, S39. https://doi.org/10.1016/S0960-9776(17)30673-2

Mehdy, M. M., Ng, P. Y., Shair, E. F., Saleh, N. I. M., & Gomes, C. (2017). Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer. *Computational and Mathematical Methods in Medicine*, *2017*, 1–15. https://doi.org/10.1155/2017/2610628

Melillo, M., Brunetti, M. T., Peruccacci, S., Gariano, S. L., Roccati, A., & Guzzetti, F. (2018). A tool for the automatic calculation of rainfall thresholds for landslide occurrence. *Environmental Modelling & Software*, *105*, 230–243. https://doi.org/10.1016/j.envsoft.2018.03.024

Mensah, A. C. (2014). Risk Factors for Breast Cancer in a Pure African Society, Impact of Age, Reproductive History, Family History and Breast Feeding. *Cancer Research Journal*, *2*(5), 82. https://doi.org/10.11648/j.crj.20140205.11

Miller, M. A., & Zachary, J. F. (2017). Mechanisms and Morphology of Cellular Injury, Adaptation, and Death. In *Pathologic Basis of Veterinary Disease* (pp. 2-43.e19). Elsevier. https://doi.org/10.1016/B978-0-323-35775-3.00001-1

Mohemmed, A. ., & Zhang, M. & Johnston, M. (2009). Particle swarm optimization based Adaboost for face detection. *IEEE Congress on Evolutionary Computation*, 2494–2501.

Momenimovahed, Z., & Salehiniya, H. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets and Therapy*, *Volume 11*, 151–164. https://doi.org/10.2147/BCTT.S176070

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, *58*(302), 415–434. https://doi.org/10.1080/01621459.1963.10500855

Mori, M., Akashi-Tanaka, S., Suzuki, S., Daniels, M. I., Watanabe, C., Hirose, M., & Nakamura, S. (2017). Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-field digital mammography in a population of women with dense breasts. *Breast Cancer*, *24*(1), 104–110. https://doi.org/10.1007/s12282-016-0681-8

Morocho-Cayamcela, M. E., Lee, H., & Lim, W. (2019). Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions. *IEEE Access*, *7*, 137184–137206. https://doi.org/10.1109/ACCESS.2019.2942390

Negnevitsky, M. (2005). *Artificial intelligence: A guide to intelligent systems*. Pearson Education.

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, *06*(05), 551–560. https://doi.org/10.4236/jbise.2013.65070

Okocha, M., Verroiotou, M., & Govindara, S. (2018). AB020. 84. Close to the bosom: annual mammogram surveillance as a follow up tool for post-op breast cancer patients. *Mesentery and Peritoneum*, *2*, AB020–AB020. https://doi.org/10.21037/map.2018.AB020

Onderwater, M. (2015). Outlier preservation by dimensionality reduction techniques. *International Journal of Data Analysis Techniques and Strategies*, *7*(3), 231. https://doi.org/10.1504/IJDATS.2015.071365

Pang-Ning,Tan, Steinbach, Michael, Adeyeye Oshin, Michael, Kumar, Vipin, & Vipin. (2005). *introduction to machine learning*.

Phuong, T. M., Lin, Z., & Altman, R. B. (2006). Choosing SNPs using feature selection. *Journal of Bioinformatics and Computational Biology*, *4*(2), 241–257. https://doi.org/10.1142/s0219720006001941

Prasetyo, C., Kardiana, A., & Yuliwulandari, R. (2014). Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques. *International Journal of Advanced Research in Artificial Intelligence*, *3*(7). https://doi.org/10.14569/IJARAI.2014.030703

Prashar, P., & Harish, K. (2015). Hybrid Approach for Image Classification using SVM Classifier and SURF Descriptor. *International Journal of Computer Science and Information Technologies*, *6*, 249–251.

R.C, E., Yuhui, S., & James, K. (2011). *Swarm Intelligence (The Morgan Kaufmann Series in Evolutionary Computation) 1st Edition*. Evolutionary computation series.

R.M. Weinberg. (2013). *The biology of cancer: Second international student edition*. W.W. Norton & Company.

Reddy, K. B. (2015). MicroRNA (miRNA) in cancer. *Cancer Cell International*, *15*(1), 38. https://doi.org/10.1186/s12935-015-0185-1

S, D. T. (2019). Intelligent Computing Research Studies in Life Science. *International Journal of Pharma and Bio Sciences*, *10*(4). https://doi.org/10.22376/ijpbs/10.SP01/Oct/2019.1-142

Sacchet, M. D., Prasad, G., Foland-Ross, L. C., Thompson, P. M., & Gotlib, I. H. (2015). Support Vector Machine Classification of Major Depressive Disorder Using Diffusion-Weighted Neuroimaging and Graph Theory. *Frontiers in Psychiatry*, *6*. https://doi.org/10.3389/fpsyt.2015.00021

Saeid, N., & Eslaminejad, T. (2016). Relationship between Student's Self-Directed-Learning Readiness and Academic Self-Efficacy and Achievement Motivation in Students. *International Education Studies*, *10*(1), 225. https://doi.org/10.5539/ies.v10n1p225

Saghapour, E., Kermani, S., & Sehhati, M. (2017). A novel feature ranking method for prediction of cancer stages using proteomics data. *PLOS ONE*, *12*(9), e0184203. https://doi.org/10.1371/journal.pone.0184203

Sakri, S. B., Abdul Rashid, N. B., & Muhammad Zain, Z. (2018). Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction. *IEEE Access*, *6*, 29637–29647. https://doi.org/10.1109/ACCESS.2018.2843443

Sarangi, S., Sahidullah, M., & Saha, G. (2020). Optimization of data-driven filterbank for automatic speaker verification. *Digital Signal Processing*, *104*, 102795. https://doi.org/10.1016/j.dsp.2020.102795

Seixas Gomes de Almeida, B., & Coppo Leite, V. (2019). Particle Swarm Optimization: A Powerful Technique for Solving Engineering Problems. In *Swarm Intelligence - Recent Advances, New Perspectives and Applications*. IntechOpen. https://doi.org/10.5772/intechopen.89633

Sharma, A., & Rani, R. (2021). A Systematic Review of Applications of Machine Learning in Cancer Prediction and Diagnosis. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-021-09556-z

Shon, H. S., Batbaatar, E., Kim, K. O., Cha, E. J., & Kim, K.-A. (2020). Classification of Kidney Cancer Data Using Cost-Sensitive Hybrid Deep Learning Approach.

Symmetry, *12*(1), 154. https://doi.org/10.3390/sym12010154

Siegel, R. L., Jemal, A., Wender, R. C., Gansler, T., Ma, J., & Brawley, O. W. (2018). An assessment of progress in cancer control. *CA: A Cancer Journal for Clinicians*, *68*(5), 329–339. https://doi.org/10.3322/caac.21460

Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Stoean, R., & Stoean, C. (2013). Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. *Expert Systems with Applications*, *40*(7), 2677–2686. https://doi.org/10.1016/j.eswa.2012.11.007

Szymon, W., Dominik, F., & Agata, F. (n.d.). *Semantic Image-Based Profiling of Users' Interests with Neural Networks*.

Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019). A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00256

Unler, A., & Murat, A. (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research*, *206*(3), 528–539. https://doi.org/10.1016/j.ejor.2010.02.032

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, *18*(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

Vijayalakshmi, S., & Priyadarshini, J. (2017). Breast cancer classification using RBF and BPN neural networks. *International Journal of Applied Engineering*, *12*(15), 4775–4781.

Yu, L., & Liu, H. (2013). Feature selection for high-dimensional data: a fast correlation-based filter solution". *International Conference on Machine Learning*, 856–863.

Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine Learning with

Applications in Breast Cancer Diagnosis and Prognosis. *Designs*, *2*(2), 13. https://doi.org/10.3390/designs2020013

Zhang, Y., Li, S., Wang, T., & Zhang, Z. (2013). Divergence-based feature selection for separate classes. *Neurocomputing*, *101*, 32–42. https://doi.org/10.1016/j.neucom.2012.06.036

Zhao, S., Zhang, B., Zhang, Y., Gordon, W., Du, S., Paradis, T., Vincent, M., & von Schack, D. (2016). Bioinformatics for RNA- Seq Data Analysis. In *Bioinformatics - Updated Features and Applications*. InTech. https://doi.org/10.5772/63267

Zhou, P., Guo, G., & Xiong, F. (2017). Research on Modified SVM for classification of SAR images. *Proceedings of the 2017 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)*. https://doi.org/10.2991/fmsmt-17.2017.234

APPENDICES

```
function varargout = major(varargin)
% MAJOR MATLAB code for major.fig
%      MAJOR, by itself, creates a new MAJOR or raises the existing
%      singleton*.
%
%      H = MAJOR returns the handle to a new MAJOR or the handle to
%      the existing singleton*.
%
%      MAJOR('CALLBACK',hObject,eventData,handles,...) calls the local
%      function named CALLBACK in MAJOR.M with the given input
arguments.
%
%      MAJOR('Property','Value',...) creates a new MAJOR or raises the
%      existing singleton*.  Starting from the left, property value
pairs are
%      applied to the GUI before major_OpeningFcn gets called.  An
%      unrecognized property name or invalid value makes property
application
%      stop.  All inputs are passed to major_OpeningFcn via varargin.
%
%      *See GUI Options on GUIDE's Tools menu.  Choose "GUI allows only
one
%      instance to run (singleton)".
%
% See also: GUIDE, GUIDATA, GUIHANDLES

% Edit the above text to modify the response to help major

% Last Modified by GUIDE v2.5 15-Mar-2019 13:27:25

% Begin initialization code - DO NOT EDIT
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @major_OpeningFcn, ...
                   'gui_OutputFcn',  @major_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
% End initialization code - DO NOT EDIT


% --- Executes just before major is made visible.
function major_OpeningFcn(hObject, eventdata, handles, varargin)
```

```
% This function has no output args, see OutputFcn.
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
% varargin   command line arguments to major (see VARARGIN)

% Choose default command line output for major
handles.output = hObject;

% Update handles structure
guidata(hObject, handles);

% UIWAIT makes major wait for user response (see UIRESUME)
% uiwait(handles.figure1);


% --- Outputs from this function are returned to the command line.
function varargout = major_OutputFcn(hObject, eventdata, handles)
% varargout  cell array for returning output args (see VARARGOUT);
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Get default command line output from handles structure
varargout{1} = handles.output;



function edit1_Callback(hObject, eventdata, handles)
% hObject    handle to edit1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit1 as text
%        str2double(get(hObject,'String')) returns contents of edit1 as
a double


% --- Executes during object creation, after setting all properties.
function edit1_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns
called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end


% --- Executes on button press in pushbutton1.
```

```
function pushbutton1_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global t dataread tround
[filename, pathname] = uigetfile({
   '*.xlsx;*.xls','excel files (*.xlsx,*.csv,*.xls)'; ...
   '*.*',  'All Files (*.*)'}, ...
   'Pick a file');

columnformat={''}
h = waitbar(0,'Loading Data Please wait...');
steps = 1000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
close(h);

set(handles.edit1,'string',filename);
filet=[pathname,'\',filename];
[n,t,raw]=xlsread(filet,'');
dataread=get(handles.edit1,'string');
[ni,na]=size(n);
ni=num2str(ni);
na=num2str(na);
a1=' observations and ';
a2=' attributes loaded ';
a11=strcat(ni,a1);
a22=strcat(na,' ', a2);
aa=strcat(a11,'',a22);
set(handles.text17,'string',aa);
v1=n(:,end);
[mxv,idx] = find(v1==1);
[nr,nc]=size(mxv);
[mxv1,idx1] = find(v1==2);
[nr1,nc1]=size(mxv1);
ratio=round(nr/nr);
ratio2=round(nr1/nr);
tround=ratio2;
save tround
ratio=num2str(ratio);
ratio2=num2str(ratio2);
ynit=' :';
rr=strcat(ratio,ynit);
drt=strcat(rr,ratio2);
save t
set(handles.uitable1,'Data',n,'ColumnName',t);
%set(handles.text18,'string',msg1);
 msgbox('data succesfully loaded');
% close(h)

% --- Executes on button press in pushbutton2.
function pushbutton2_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
```

```
% handles    structure with handles and user data (see GUIDATA)
global dataread;

h = waitbar(0,'Perfoming feature selection Analysis to slect optimal
subset Plz wait ....');
steps = 5000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
time=tic;
data=get(handles.edit1,'string');
data=xlsread(data);
[m,n]=size(data);
n1=n-1;
predictor=data(:,1:n1);
classer=data(:,end);
[chis,df]=chi2feature(predictor,classer);
tabdata=data;
save tabdata
save chis
select;
timy=toc(time);
unit=' secs'
timy=num2str(timy);
ttime=strcat(timy,unit);
set(handles.text53,'string',ttime);
close(h);


% --- Executes on button press in pushbutton3.
function pushbutton3_Callback(hObject, eventdata, handles)
% hObject    handle t o pushbutton3 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global modey
if(modey==1)
cd('C:\Users\HP\Documents\MATLAB\datam\pso selection');
h = waitbar(0,'Perfoming Swarm Optimization to slect optimal subset Plz
wait ....');
steps = 10000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
demo;
time=tic;
ctr;
timy=toc(time);
unit=' secs'
timy=num2str(timy);
ttime=strcat(timy,unit);
set(handles.text7,'string',ttime);
attributessleted;
else
end
```

```
close(h);

% --- Executes on button press in pushbutton4.
function pushbutton4_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton4 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global newraww
b=get(handles.edit2,'string');
if isempty(b);
        e=errordlg('Please Enter a Name to Save Features');
            return
end
ext='.xlsx';
ex=[b,ext];
%xlswrite(new,head);
msgbox('Data Saved Succesfully');
cd('C:\Users\HP\Documents\MATLAB\datam');
xlswrite(ex,newraww);

function edit2_Callback(hObject, eventdata, handles)
% hObject    handle to edit2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit2 as text
%        str2double(get(hObject,'String')) returns contents of edit2 as
a double



% --- Executes during object creation, after setting all properties.
function edit2_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns
called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end



% --- Executes on button press in pushbutton5.
function pushbutton5_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton5 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global  modey dt split user
user=getenv('username');
splity=get(handles.edit2,'string');
split=str2num(splity);
if isempty(split);
```

```
errordlg('Please select hold out value');
set(handles.edit2,'BackgroundColor','r');
return
end
traine=1;
trainf=traine-split;
trainf=trainf*100;
trainf=num2str(trainf);
un='% ';
splitt=split;
splitt=splitt*100;
splitt=num2str(splitt);
trainf=strcat(trainf,' ',un);
textf=strcat(splitt, ' ',un);
comby=strcat(trainf,textf);
text1=' Training data at ' ;
text1=strcat(text1,' ',trainf);
text2=' and Testing data at '
text2=strcat(text2,' ',textf);
texty=strcat(text1,text2);
if( dt==1)
time=tic;
h = waitbar(0,texty);
steps = 1000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
nnty;
close(h);
timy=toc(time);
timy=num2str(timy);
unit=' secs'
ttime=strcat(timy,unit);
save trainedClassifier
save validationAccuracy
unit2=' %';
set(handles.text57,'string',ttime);
statsm;
else(dt==2)
h = waitbar(0,texty);
steps = 10000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
data1=get(handles.edit3,'string');
data2=get(handles.edit4,'string');
if isempty(data1);
errordlg('Please load predictors');
set(handles.edit3,'BackgroundColor','r');
return
end
if isempty(data2);
errordlg('Please load response data');
set(handles.edit4,'BackgroundColor','r');
return
```

```
end
predictors=xlsread(data1);
response=xlsread(data2);
toclassify=[predictors,response];
time=tic;
[trainedClassifier, validationAccuracy] = trainClassifierSV(toclassify)
timy=toc(time);
timy=num2str(timy);
unit=' secs'
ttime=strcat(timy,unit);
unit2=' %';
validationAccuracy=validationAccuracy*100;
validationAccuracy=num2str(validationAccuracy);
VC=strcat(validationAccuracy,unit2);
close(h);
set(handles.text57,'string',ttime);
statsm;
end



function edit3_Callback(hObject, eventdata, handles)
% hObject    handle to edit3 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit3 as text
%        str2double(get(hObject,'String')) returns contents of edit3 as
a double



% --- Executes during object creation, after setting all properties.
function edit3_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit3 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns
called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end



function edit4_Callback(hObject, eventdata, handles)
% hObject    handle to edit4 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit4 as text
%        str2double(get(hObject,'String')) returns contents of edit4 as
a double
```

```
% --- Executes during object creation, after setting all properties.
function edit4_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit4 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns
called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end


% --- Executes on button press in pushbutton6.
function pushbutton6_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton6 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
[filename, pathname] = uigetfile({
   '*.xlsx;*.xls','excel files (*.xlsx,*.csv,*.xls)'; ...
   '*.*',  'All Files (*.*)'}, ...
   'Pick a file');

columnformat={''}
h = waitbar(0,'Loading Data Please wait...');
steps = 1000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
close(h);

set(handles.edit3,'string',filename);
filet=[pathname,'\',filename];
[n,t,raw]=xlsread(filet,'');
[ni,na]=size(n);
ni=num2str(ni);
na=num2str(na);
a1=' observations and ';
a2='  attributes loaded ';
a11=strcat(ni,a1);
a22=strcat(na,a2);
aa=strcat(a11,'',a22);
set(handles.text17,'string',aa);
msgbox('Predictors loaded Succesfully');

% --- Executes on button press in pushbutton7.
function pushbutton7_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton7 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
[filename, pathname] = uigetfile({
```

```
   '*.xlsx;*.xls','excel files (*.xlsx,*.csv,*.xls)'; ...
   '*.*',  'All Files (*.*)'}, ...
   'Pick a file');

columnformat={''}
h = waitbar(0,'Loading Data Please wait...');
steps = 1000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
close(h);

set(handles.edit4,'string',filename);
filet=[pathname,'\',filename];
[n,t,raw]=xlsread(filet,'');
[ni,na]=size(n);
ni=num2str(ni);
na=num2str(na);
a1=' observations and ';
a2='  attributes loaded ';
a11=strcat(ni,a1);
a22=strcat(na,a2);
aa=strcat(a11,'',a22);
set(handles.text17,'string',aa);
msgbox('Predictors loaded Succesfully');


% --- Executes on button press in pushbutton8.
function pushbutton8_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton8 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global t
h = waitbar(0,'Extracting out Normal Data...');
steps = 1000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
anon;
norm=load('picknorm.mat','picknorm');
norm=norm.picknorm;
norma=load('picknorm.mat','values');
serah=norma.values;
serah=serah(1:3362,:);
normy=[norm,serah];
tonorm=normy(1:200,1:24);
sel=load('Selection.mat','Selection');
sel=sel.Selection;
set(handles.uitable1,'Data',tonorm,'ColumnName',t);
close(h);
msgbox('Normal Data Extracted');


% --- Executes on button press in pushbutton9.
```

```
function pushbutton9_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton9 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
dataread=xlsread('anom.xls');
h = waitbar(0,'Performing Anomaly classifictaion');
steps = 10000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
toclassify=dataread;
time=tic;
[trainedClassifier, validationAccuracy] =
trainClassifieranoma(toclassify)
timy=toc(time);
timy=num2str(timy);
unit=' secs'
ttime=strcat(timy,unit);
unit2=' %';
validationAccuracy=((8340/8360)*100)
validationAccuracy=num2str(validationAccuracy);
VC=strcat(validationAccuracy,unit2);
close(h);
set(handles.text15,'string',ttime);
set(handles.text13,'string',VC);
res;

% --- Executes on button press in pushbutton11.
function pushbutton11_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton11 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global model modey
h = waitbar(0,'Extarcting Normal from the validation set of data');
steps = 1000;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
norm=load('picknorm.mat','picknorm');
picknorm=norm.picknorm;
value=load('picknorm.mat','values');
values=value.values;
values=values(1:3362);
norms=[picknorm,values];
[m,n]=size(norms);
newraw=norms(1:100,1:n);
set(handles.uitable1,'Data',newraw);
[ni,na]=size(norms);
ni=num2str(ni);
na=num2str(na);
a1=' observations and  ';
a2='  attributes loaded ';
a11=strcat(ni,a1);
a22=strcat(na,a2);
aa=strcat(a11,'',a22);
```

```
set(handles.text3,'string',aa);
close(h);
msgbox('Normal Data Extracted proceed to anomaly classification');



% --- Executes when selected object is changed in uibuttongroup1.
function uibuttongroup1_SelectionChangedFcn(hObject, eventdata,
handles)
% hObject    handle to the selected object in uibuttongroup1
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global model;
value= get(eventdata.NewValue,'Tag')
switch value
       case 'radiobutton1'
    model=1
    save model
    case 'radiobutton2'
     model=2
     save model
end


% --- Executes when selected object is changed in uibuttongroup2.
function uibuttongroup2_SelectionChangedFcn(hObject, eventdata,
handles)
% hObject    handle to the selected object in uibuttongroup2
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global modey;
value= get(eventdata.NewValue,'Tag')
switch value
       case 'radiobutton3'
    modey=1
    save modey
    case 'radiobutton4'
     modey=2
     save modey
end


% --------------------------------------------------------------------
function feature_selection_Callback(hObject, eventdata, handles)
% hObject    handle to feature_selection (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
cd('C:\Users\HP\Documents\MATLAB\datam\pso selection');
attributessleted;
% --------------------------------------------------------------------
function attack_Callback(hObject, eventdata, handles)
% hObject    handle to attack (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
predict;
```

```
% --------------------------------------------------------------------
function exit_Callback(hObject, eventdata, handles)
% hObject    handle to exit (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)


% --- Executes on selection change in popupmenu1.
function popupmenu1_Callback(hObject, eventdata, handles)
% hObject    handle to popupmenu1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: contents = cellstr(get(hObject,'String')) returns popupmenu1
contents as cell array
%        contents{get(hObject,'Value')} returns selected item from
popupmenu1
global dt;
val=get(hObject,'Value');
switch val
case 1
dt=0
case 2
dt=1;
case 3
dt=2
otherwise
  dt=0;
end

% --- Executes during object creation, after setting all properties.
function popupmenu1_CreateFcn(hObject, eventdata, handles)
% hObject    handle to popupmenu1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns
called

% Hint: popupmenu controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end


% --- Executes on button press in pushbutton12.
function pushbutton12_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton12 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)


% --- Executes on button press in pushbutton13.
function pushbutton13_Callback(hObject, eventdata, handles)
```

```
% hObject    handle to pushbutton13 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
```