Ministry of Higher Education and
Scientific Research
University of Diyala
College of Science
Department of Computer Science

Breast Cancer Classification Model Based on Machine Learning Algorithms

A Dissertation
Submitted to the Department of Computer Science\ College of Sciences\ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of master's in computer science

By

Ismail Miteb Hamid

Supervised By

Prof. Dr. Dhahir Abdulhade Abdulah **Ass.Prof.Dr. Salam Abdulkhaleq Noaman**

2020 A.C **1442 A.H**

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

﴿يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنكُمْ
وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ﴾

صدق الله العظيم

سورة المجادلة :الآية (١١)

Dedication

This thesis is dedicated to someone,
I will remember his kindness and support in his life

and after…

My parent
And *My Brothers and Sisters*
And My family

Ismael

Acknowledgements

All my thanks first of all are addressed to Almighty ***Allah***, who has guided my steps towards the path of knowledge and without His help and blessing; this thesis would not have progressed or have seen the light.

My sincere appreciation is expressed to my supervisors ***Prof. Dr. Dhahir Abdulhade Abdulah*** *and* ***Ass.prof.Dr.*** **Salam Abdulkhaleq Noaman** for providing me with ideas, inspiration and continuous support me during the period of my study.

I am extremely grateful to all members of Computer Science Department of Diyala University for their general support.

Finally, I would never have been able to finish my thesis without the help from ***friends***, and support from ***my family***.

Thank you all!

ismael

(Supervisor's Certification)

We certify that this research entitled "***Breast Cancer Classification using Developed Techniques"*** was prepared by ***Ismail Miteb Hamid*** Under our supervisions at the University of Diyala Faculty of Science Department of Computer Science, as a partial fulfillment of the requirement needed to award the degree of Master of Science in Computer Science.

(Supervisor)

Signature: 

Name: **Dr. Dhahir Abdulhadi Abdulah**

Date:2/7/2020

(Supervisor)

Signature: 

Name: **Ass.Prof.Dr. Salam bdulkhaleq Noaman**

Date:2/7/2020

Approved by University of Diyala Faculty of Science Department of Computer Science**.**

Signature: 

Name: **Dr. Taha Mohammad Hassan**

Date: 2/7/2020

(Head of Computer Science Department)

Approved by University of Diyala Faculty of Education for Pure sciences Department of Computer Science.

Linguistic Certification

This is to certify that this thesis entitled " Breast Cancer Classification using Developed Techniques " was prepared by "Ismail Miteb Hamid" under my linguistic supervision. Its language was amended to meet the English style.

Signature :

Name :

Date : / / 20120

Scientific certification

This is to certify that this thesis entitled "Breast Cancer Classification using Developed Techniques" was prepared by "Ismail Miteb Hamid" under my Scientific supervision. It has been evaluated scientifically, therefore, it is suitable for debate by examining committee.

Signature :

Name :

Date : / /2020

Abstract

Breast cancer is one of the leading causes of death among women worldwide. Accurate and early detection of breast cancer can ensure long-term survival for the patients. However, traditional classification algorithms usually aim only to maximize the classification accuracy, and cost failing to take into consideration the misclassification costs between different categories. Furthermore, the costs associated with missing a cancer case (false negative) are much higher than those of mislabeling a benign one (false positive).

To overcome this drawback and further improving the classification accuracy of the breast cancer diagnosis, in this work, present several machine learning algorithms such as Decision Tree (DR) , Random Forest (RF) , Logistic Regression (LR), and Support Vector Machine (SVM) . For all the phases of the work that required data treatment and machine learning techniques are going to use this tool. In technical terms ,the intended output of the work that enables the achievement of the business objectives described before is find the algorithms that can classify more efficiently the different types of breast cancer.

The result of the machine learning by calculate the accuracy of each model obtain , the random forest achieved 0.9857 % accuracy , decision tree achieved 0.9571% accuracy, SVM achieved 0.9714% accuracy, Logistic Regression achieved 0.9643 % accuracy , in an other word the Robust forest archive high accuracy.

List of contents

List of Tables

List of Figure

List of Abbreviations

Abbreviations	Description
2D	Two-dimensional
3D	Three- dimensional
4D	Four- dimensional
ANN	Artificial Neural Network
CAD	Computer-aided Detection
CADX	Computer-aided Detection/Diagnosis
CT	Computed-tomography
DM	Data Mining
DT	Decision Tree
FN	False Negative
FNA	Fine Needle Aspiration
FP	False Positive
KKN	K-Nearest Neighbors
LR	Logistic Regression
MIAS	Mammographic Image Analysis Society
ML	Machine Learning
MRI	Magnetic-resonance-imaging
NB	Naïve Bayesian
NCI	National Cancer Institute
PET	Positron-emission-tomography

RBF	Radial Basis Function
RF	Random Forest
SEE	Span Error Estimate
SOMs	Self-organizing Maps
SRG	Seeded Region Growing
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
US	United State
WBC	Wisconsin Breast Cancer Dataset
WDBC	Wisconsin Diagnosis Breast Cancer
WDBC	Wisconsin Breast Cancer Diagnosis
WPBC	Wisconsin Breast Cancer prognosis

Chapter one

Introduction

Chapter One

Introduction

1.1 Overview

Major public health problem is a cancer, that mortality and cancer incidence has been increased over the past three decades at an accelerated pace globally, [1]. Breast cancer is known as the burden of non-communicable diseases and it is the main reason in deaths in developing countries. The disease GLOBOCAN 2018 diagnosed have affected around (2.2) million new cases in breast cancer around the world, [2]. There are highly chance of causes disease like breast cancer which are still unknown, [3][4].

The distinguish between malignant breast tumors and benign ones by using some of diagnosis techniques. The well-known procedure, Fine Needle Aspiration (FNA). In addition, inexperience or exhaustion cause higher possibility to rise errors, that panic patients when incorrect-positive result happens when incorrect-negative result appears. To help doctors' diagnosis of breast cancer, the evolving well organized diagnosis support system, The research explains that proposing machine learning and data mining approaches for breast tumor diagnosis can obtain lots advantage including high degree of diagnosis accuracy, reducing medical and resource driving down costs, [4][5]. Breast tumor is the popular diagnosed growth among the methods for growth diagnosis, that is treated as breast growth from malignant ones, [6]. However, although, how to obtain better result for common classification issues is still not easy up to now, [7].

Many Machine learning process, like Logistic Regression (LR), Artificial Neural Network (ANN), Support Vector Machine (SVM), Naïve Bayesian (NB) and K-Nearest Neighbors (KNN), are work on

principle black box system which are incapable to explain the forecast results. When knowing into a learning model is more important compared with classification outcomes, it is more efficient to employ a system of white box. Although there are numerous assistance systems of diagnosis in the United State (US), that had explained gaining in scientific research clinical experiments and labs , that have not available in clinical practice, or even widely used, [8].

In this thesis , a statistical analysis of the data available in the dataset WBC, Also there are four machine learning algorithms to use classification data and comparison between them.

1.2 Literature Review

Many of researches have been accomplished in the area of classification of Wisconsin Brest Cancer Dataset (WBCD) by use of smart computing types. These can be described briefly below:

- In Martin J. Yaffe, Earlier [9], 2016. Authors have applied the use of the 2D median filtering, image contrast enhancement, algorithm, Seeded Region Growing (SRG) to disconnect the noise, omit radiopaque artefacts and remove the prediction of the Breast from a digitalized mammogram. In addition, the technicality of Artificial Neural Network is used to classify a mammogram like representing a malignant tumor or normal, which is indicating the existence of a benign aggregate.
- Roselina Sallehuddin [10], 2016. Authors have introduced a specific procedure for feature uprooting to distinguish between digital mammograms through using quick limited shear let transform. To enlarge the differences between type delegates, a thresholding method could be put in place as the last scene of feature uprooting.

The categories have been calculated through the optimal property group.

- In Emina Aličković [11] (2017). Applying backup Self-organizing Maps (SOMs), Support Vector Machine (SVM), Radial Basis Function (RBF) networks for breast tumour detection. The (WDBC) dataset can be utilized in the categories experiment. The following classifiers: 1-norm C-SVM (L1-SVM), 2-norm C-SVM (L2-SVM), and u-SVM are utilized, for which the connection search up on span error estimate (GSSEE), gradient descent based on validation error estimate. The SOM–RBF classifier is developed to improve the performance of only the SOM learning procedure, which is based on distance comparison .The SOM–RBF classifier could be effective materials to detect breast tumour with the high detection precision.
- In Peiguang Lin [12] (2016). presented a hybrid system named GRA-SVM that constitute SVM classifier and filter attribute option. Grey Relational Analysis has been suggested and tested against BUPA Disorder the group of data and Wisconsin Breast Cancer Dataset (WBCD). Experiments results show that GRA-SVM get better the SVM precision around 0.48% through use limited two advantages for the WBCD dataset. For BUPA dataset, GRA-SVM gets better the SVM accuracy around 0.97% through utilizing four advantages.
- In Nilashi, Mehrbakhsh., Ibrahim, ..et [13] (2017). Introduced many dissimilar information mining procedures such as: (Support Vector Machine (SMO SVM), Bayesian Network, Multilayer Perceptron and Decision tree (J48)) these methods were put into practice for Wisconsin Diagnostic Breast Cancer (WDBC). The

SVM SMO procedure has accomplished an accuracy around 97.72%.

- In O.N. Oyelade, [14] (2017) . Shows a procedure about the user interface that depends on the ontology. OWL language can be used to describe the context data and the inquiry form data, that in turn gives a better firm for the latter classification and integration of the inquiry interface style. This step can have a great employ of constructing and submitting the inquiry demand. Moreover, the [22] also include the automatically extraction procedure about the impersonation, while the tests display that the procedure can perform greatly to excerpt the attribute information, relationship data and context data of the inquiry interface.
- In S. Wang, Y. Wang, et al [15] (2017) . Introduced Classification and Regression Trees (CART) which create the fuzzy principles to be employed in the knowledge system. The check results on Pima Indian Diabetes, Stat Log, WDBC, Mesothelioma, Parkinson, and Cleveland's tele monitoring collection of data presents suggested method extraordinarily increases the diseases forecast accuracy. The outcomes displayed that the series of fuzzy principle, CART with limit noise and gathering techniques can have more impacts in diseases forecast from medical datasets.
- In Dalwinder Singh, [16] 2018. Worked on diagnosing breast cancer base on Wisconsin information, firstly, set up an efficient input mechanism that will make it able for the outline to filter through reading and cleaning input from datasets. Secondly, semantic web languages were utilized to set up an ordered principle set and know the outline framework were therefore formed to try and help the causing algorithm. In conclusion, modification of the

structures of Select and Test (ST) can be accommodated to this enhancement

- In S., Birmohan S. [17] 2019. Introduced recovered the method of the Random Forest-Based Rule Extraction Method for Breast Cancer Diagnosis. At beginning, availability of the abundant decision rules could be created by using Random Forest through numbers of decision tree models. Then, the process of a principle of extraction is inventive to detach decision principles which is integrated of trained trees. In the end, multi objective evolutionary algorithm (MOEA) is improved and put in place to find for an optimal principle foreteller that the constituent principle group is the better trade-off in between interpretability and accuracy.
- In Na Liu, Er-Shi Qi, [18] (2019). Introduced merit weighting is applied to evolve an efficient computer as aided diagnosis outline for breast cancer. Merit weighting is appointed because of the enhances the classification achievement more as contrast to merit branch selection. This mostly works when a wrapper procedure employs the Ant Lion Optimization algorithm is showed that finds for better merit weights and values of Multilayer Neural Network at the same time. The option of unobserved backpropagation and neurons training algorithms are applied as variables of nervous networks. The presentation of the suggested method is assessed on three types of breast cancer information. The information of datasets is at first standardized utilizing tanh procedure to expel the impacts of predominant aspect. The outcomes display that suggested wrapper strategies has superior capacity to achieve best precision when contrasted with the existent techniques. The acquired more characterization achievements approves the work

which has the prospect for turning into an option in contrast to another known technicality.

1.3 Statement of the problem

During a literature review of tissue classification, it was noted that breast tissue cancer is the most common malignancy among women and the cause of many endings of life. Early detection of tumours is the best solution to avoid mastectomy, reduces the chances of it coming back and reduces the life death rate. There is no effective way to get rid of this cancer. All the methods used did not carry the problem at all. There are many classification methods used in this field It is not easy to answer the question of the classification approach appropriate for a specific study. Different classification results can be obtained according to the selected works. Different classification methods have their own advantages.

1.4 Aim of Thesis

The aim of this thesis is to solve the problem in the paragraph (1.3) and classification model based on the intelligent computing model to help clinicians or radiologists identify areas of suspicion of the databases used. The main idea to develop a model for classification of Malignant and benign cases in order to reduce the number of erroneous cases in this thesis. And use four algorithms to develop the work which is support vector machine (SVM) , Random Forest (RF) , Logistic Recreation (LR) and Decision Tree (DT) .

1.5 Thesis Outline

This thesis consists four chapters and adding to the chapter one as the following:

Chapter 2: Theoretical and background, it includes the theoretical background of concept that were adopted in the thesis it was reviewed Medical Image and machine learning

Chapter 3: Breast Cancer Classification Model Based on Machine Learning Algorithms: The suggests algorithms and procedures that describe the statistical analysis of the database and explain in detail the tools to use in classification.

Chapter 4: Results and Analysis, this chapter includes a review of the results obtained through the implementation of the proposed system. The results are discussed and analysed with comparisons made with related work.

Chapter 5: Conclusions and Future Work, it includes reviewing the results reached through designing and implementing the proposal as well as the recommendations.