Medical Diagnosis of Breast Tumor Using Kernel Machines

A Thesis Submitted to

The Maharaja Sayajirao University of Baroda

For The Degree of

Doctor of Philosophy

in

Applied Mathematics

By

Shah Pooja Jitendra

Reg. No.: FOTE/945, Reg. Date: July 21, 2017

Under the guidance of

Dr Trupti P. Shah

Department of Applied Mathematics
Faculty of Technology and Engineering
The Maharaja Sayajirao University of Baroda
Vadodara 390 001, Gujarat, India

20th April, 2023

Certificate

This is to certify that Ms Pooja Jitendra Shah has worked under my guidance to prepare the thesis entitled ***"Medical Diagnosis of Breast Tumor Using Kernel Machines"*** which is being submitted herewith towards the requirement for the degree of Doctor of Philosophy in Applied Mathematics.

20/4/2023

Dr. Trupti P. Shah
Associate Professor
Department of Applied Mathematics
Faculty of Technology & Engineering
The Maharaja Sayajirao University of Baroda Vadodara - 390 001.

Dr. B. M. Shah
Head of the Department
Department of Applied Mathematics
Faculty of Engineering and Technology
The Maharaja University of Baroda,
Vadodara – 390 001

Head
Applied Mathematics Department
Faculty of Technology & Engineering
The Maharaja Sayajirao University of Baroda
Vadodara

Prof.(Dr.) Dhanesh Patel
Dean
Faculty of Engineering and Technology
The Maharaja University of Baroda,
Vadodara – 390 001

Dean
Faculty of Tech. & Engg.
The M. S. University of Baroda.
Vadodara.

Declaration

I hereby declare that the thesis entitled ***"Medical Diagnosis of Breast Tumor Using Kernel Machines"***, which is being submitted for the fulfillment of the requirements for the award of the Degree of Doctor of Philosophy in the Department of Applied Mathematics, to The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat is a record of my own, original research work and has not been submitted in part or full to any other University or Institute for the award of any degree or diploma. Due acknowledgment has been cited for the literature that was referred for carrying out the present work.

Shah Pooja Jitendra

Department of Applied Mathematics
Faculty of Technology and Engineering
The Maharaja Sayajirao University of Baroda
Vadodara, 390 001, Gujarat, India.

Dedicated to

The Lord Shree Krishna

Lovely Parents

&

Beloved Family

Acknowledgement

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

It is a great pleasure for me to express my deep sense of indebtedness and heartfelt gratitude to my guide Dr. Trupti P. Shah, Associate Professor in Department of Applied Mathematics, Faculty of Technology and Engineering, The M. S. University of Baroda, Vadodara. She has provided her valuable guidance at each and every step of this journey. Her enlightened approach, keen interest and constant encouragement made it possible for me to complete this work. Her persistent advice came a long way in researching and writing of this thesis. She is not just supervisor but also a parental figure for me. Words are not enough to express my gratitude towards her.

Moreover, I would like to express my heartfelt gratitude to the Head of the department of Applied Mathematics, The Maharaja Sayajirao University of Baroda, Dr. B. M. Shah, for his academic support and the facilities provided by him to carry out my research work. I express my gratitude to the Board of Research Committee for their suggestions during the various phases of the program.

I would also like to thank the countless experts who have offered guidance, feedback and suggestions along the way - Prof. Dhanesh Patel, Prof. Praghna Kantawala, Dr. Purnima Pandit, Dr. Jaita Sharma, Dr. Rajesh Shah, Dr. Nilima Shah, Dr. Nimisha Pathak and Dr. Bharat Ratanpal.

I am thankful to our non-teaching staff members, Pooja ma'am, Madhuri ma'am, Shankarbhai and Ravibhai for their kind support.

Besides my advisor, I would also like to thank Dr. Anjali Jiwani from Computer Science Engineering department for their insightful comments and valuable suggestions.

I thankful to Rupalim Burman, Hostel Superintendent, for providing constant support and encouragement throughout the journey. I also thankful to hostel mates Chetna, Shivani and Urvi for their good wishes.

I wish to place on record my wholehearted thanks to Shardav Bhatt, Anil Chawada, Devang Pathak, Shivam Suthar for their kind help in my research whenever I required. I also thank my dear friends and colleagues, Bhavesh, Drashti, Prani, Amar sir, Vishant, Bhavyata, Gargi, Kalpna, Gayatri for their moral support and help during this

work. I would also like to thank Dr. Nakul Soni for providing guidance during the thesis writing in *LaTeX* format.

I thankfully acknowledge the university authorities for providing the financial support under 'DST-PURSE Program Phase-II' for the year 2018–19.

Words are not enough to express my gratitude for my parents. I would like to convey my heartfelt thanks for helping me understand what I was and am capable of, for giving me the strength I needed to follow my dreams, and for believing that I have the skills and potential to achieve my goals. There were moments when I gave up; it was only you who walked with me to provide me with the support that I needed. Without their blessings, I could not have achieved this academic milestone and reached my ambitions. Their goodwill and inspiration far exceeded what the words of appreciation can express. I would like to special thank Dr. Vijay Desai for always supporting me and giving motivations to achieve my goal. He is not only my uncle but also father figure for me. Words are not enough to express my gratitude towards him. It requires luck to have an elder and younger brother like you. My two beloved siblings - Jaymin and Sapan - greatly appreciate and feeling blessed to have both of you for providing me with the continuous motivation, appreciation, strength, understanding and help. I got so much more from the both of you than what I asked for. I am also thankful to Nirali, Devanshi, Nidhi, Reyansh, Prayag and Vaidehi for their good wishes.

I am also thankful to father-in-law and mother-in-law for supporting me to achieve my goal. Last but not the least, I would like to thank my husband, Dr. Kevil Shah for being my source of strength during this entire journey. By offering me a helping hand in everything from daily household chores to key milestones in the thesis journey. Kevil has shown equal dedication to this work as I have. Without his support and understanding, it would be difficult to complete the work. To all my sisters and family, your encouragement, warmth, and support has been instrumental at every stage of my personal and academic life.

Above all, I bow down to The Almighty God for granting me the wisdom, health and strength to undertake this research work even at the later stage of my academic life and enabling me to its completion.

Contents

List of Figures

List of Tables

Abstract

When it comes to medical diagnosis and care, artificial intelligence may be a dependable support system for busy medical professionals and institutions, reducing stress and maximizing productivity. AI-aided medical diagnosis improves clinical decision-making, management, automation, administration, and workflows. It can aid in the diagnosis of cancer, the triage of critical findings in medical imaging, the identification of acute abnormalities, the provision of assistance to radiologists in the prioritisation of cases involving life-threatening conditions, the diagnosis of cardiac arrhythmias, the prediction of the outcomes of strokes, and the assistance in the management of chronic diseases. Artificial intelligence (AI) is an ever-expanding, constantly-evolving, and patient-centric ecosystem of data, algorithms, analytics, deep learning, neural networks, and insights. Artificial intelligence in medical diagnostics has showed significant promise in recent years for both improving medical treatment overall and relieving some of the industry's tremendous stresses.

Machine learning-based autonomous detection and diagnosis systems have demonstrated impressive accuracy and speed. In the recent decade, analytical techniques for breast cancer have made significant progress. Many different automatic classification methods have been used in the past few years. Different strategies produce varying results. Problems still need to be solved, such as the need for innovative approaches.

In this study, we proposed several predictive models using various methodologies and soft computing techniques like Support Vector Machines, Deep Neural Networks, Adaptive Neuro Fuzzy Inference Systems and Radial Basis Function Networks and implemented this models on WDBC and WBC data sets. In the construction and implementation of models different optimization strategies and network topologies are used. The comparative analysis of the performance of the classification accuracy obtained by the various researchers in their models with our respective proposed models is carried out. The proposed predictive models attained the greatest classification accuracy in a remarkably short amount of time.

Chapter 1

Introduction

1.1 Introduction

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Normally, human cell grow and multiply (through a process called cell division) to form new cells as the body needs them. When cells grow old or become damaged, they die and new cells take their place. Sometimes this orderly process breaks down, and abnormal or damaged cells grow and multiply when they shouldn't. These cells may form tumors, which are lumps of tissues. Tumors can be cancerous (Malignant) or non-cancerous (Benign). Cancerous tumors spread into; or invade nearby tissues and can travel to distant places in the body to form new tumors and it is called malignant tumors. Benign tumors don't spread into, or invade; nearby tissues. When removed, benign tumors usually don't grow back, whereas cancerous tumors sometimes do. There are different types of cancer like skin cancer , breast cancer , lung cancer, prostate cancer, kidney (renal) cancer. The most prevalent and serious cancer that affect women is the breast cancer. A large number of women succumb to breast cancer each year. Recently after skin cancer, breast cancer is the second most hazardous cancer diagnosed in women worldwide and becomes the reason for death. Breast Cancer is a cancer that develops from breast tissues. The first symptom of breast cancer is usually an area of thickened tissue in the breast or as lump in the breast or an armpit or breast pain does not change with the monthly cycle, pitting like the surface of an orange or color changes such as redness in the skin of the breast, a rash around or an one nipple, discharge from a nipple which may contain blood, a sunken or inverted nipple, a change in the size or shape of the breast, pealing or flaking or scaling of the skin of the breast or nipple [ref: medicalnewstoday.com/articles/37136]. The several risk factors that cause breast cancer are age, genetics, a history of breast cancer or breast lumps, dense breast tissue, estrogen exposure and breast feeding, body weight, alcohol consumption, radiation exposure, hormone treatments [ref: https://www.medicalnewstoday.com/articles/37136.php]. As per the survey report of GLOBOCAN 2012, Ferlay et al. (2014) found that, 1067 million women were detected with breast cancer [38]. Also, the authors discovered that the breast cancer has highest proportion of 25% out of all cancers within women. In the research of Global Cancer Statistics 2018, Bray et al. (2018) stated that, 2.1 million new cases of breast cancer were identified and out of all registered cases of breast cancer, 53% were diagnosed as malignant [21]. As per the survey report of WHO, each year 2.1 million women are impacting with breast cancer and also large number of

women die due to deficiency in early diagnosis and early treatment [21]. In 2018, 627,000 women were died due to breast cancer. As per the survey report of WHO in 2020, 2.3 million women were diagnosed with breast cancer and 685000 were died worldwide [who.int/news-room/fact-sheets/detail/breast-cancer].

Early detection of breast cancer is important as it is associated with an increased number of available treatment options, increased survival and improved quality of life. Majority of the all affected women are middle-aged that is in the age of 50's & 60's [109]. Early detection provides the best chance of effective treatment. The earlier the stage of breast cancer the better the chance of survival. The very early detection of breast cancer have a 97 to 100% chances of cure but once it spreads to the lymph nodes or elsewhere, the chance of cure goes down significantly. If an early diagnosis is made, patients can avoid the cost of different tests such as mammograms, ultrasounds, other imaging tests, biopsies, well as at the same time they can reduce the number of frequent visits to the doctor which can help them mentally and financially. Early diagnosis can also save that doctor's time and they can reach to the more patients.

It is most essential to identify and cure breast cancer in its early stage. While successful treatment depends on early detection, the diagnosis of breast cancer is difficult due to the dense breast tissues with the detection being subject to human error, the doctors looked for a way to improve the accuracy of the diagnosis. With the help of computer aided technologies and Artificial Intelligence (AI), it is possible to make early diagnosis of breast cancer. Development of such a tool or system is required to make early diagnosis of breast cancer using soft computing techniques.

In medical science, an extensive and diverse spectrum of applied mathematics research is being conducted. AI or ML is all about mathematics, which in turn helps in creating algorithm that can learn data to make an accurate predication. Machine Learning (ML) is an emerging technique which provides an efficient way to enhance the knowledge in data in order to improve the performance of the disease predictive models. There are server ML algorithm like Support Vector Machine (SVM), Artificial Neural Network (ANN), Deep Learning (DL), etc.. Using these algorithms AI is built into machines. The basic requirements for any intelligent behavior is learning. Soft computing approaches are being used in medical science by researches worldwide. The thesis is concerned with the diagnosis of breast cancer through the application of various soft-computing approaches, with a particular emphasis on Kernel-based methodologies. Classification plays an important role in medical science where data mining techniques are used to diagnose and analyse disease at

an early stage.

1.1.1 Literature survey

Many researchers have worked in diagnosis of breast cancer using various soft computing techniques. Liu et. al. used SVM for classification of breast cancer data and achieved 96.71% accuracy with polynomial kernel and 97.07% accuracy with radial basis function kernel [70]. Chen et. al. and Keerthi have classified the breast cancer after applying various feature selection techniques like rule extraction, roughest based feature selection, Genetic Algorithm (GA) etc and obtained good classification accuracy [27, 41]. Polat and Salih developed the least square SVM Classifier and Obtained 98.53% accuracy [95, 103]. Also Akay, Maglogiannis et. al. and Osareh et. al. had built SVM and compared with other classifiers like Bayesian, ANN , K-nearest neighbors probabilistic neural network and obtained nearby 97% of classification accuracy [7, 74, 90].

For different data set like Wisconsin Breast Cancer, Wisconsin Diagnostic Breast Cancer, Wisconsin Prognostic Breast Cancer, Aalaei et. al. employed ANN with GA based feature selection and a Particle Swarm optimization algorithm based classifier (PS- classifier) to diagnosis of breast cancer [1].Abdel-Zaher et. al. employed Deep Neural Network as a classifier with recursive feature elimination technique [3].Karabatak et. al. developed ANN classifier based on association rule and implemented on WBC data set [57]. Agarap Abien Fred M. experimented Six ML method on WBC data set namely Gated Recurrent Unit with SVM; Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbor (NN) search [6].P.R. Innocent et. al. conducted a study of fuzzy methods for medical diagnosis in nursing assessment using Type-II fuzzy sets (2007) [51]. Many authors namely Baig et. al., Awotunde et. al. and Madkour et. al. have developed a control system using fuzzy logic in diagnosis of various disease like brain tumor, malaria, wherping cough, chickenpox etc [16, 15, 73]. Elif Derya Übeyli proposed an integrated view of ANFIS to detect breast cancer and tested on WBC [122]. Seyedesh S.N. et al. designed a hierarchical fuzzy neural system with Extended Kalman Filter (EKF) [86]. M. Ashraf et. Al. introduced an information gain technique with ANFIS for breast cancer classification [14]. Chakravarthy and Ghosh demonstrated scale based clustering with Radial Basis function network [26]. Kiyan and Ypldrim proposed statistical neural network topology in RBFN and Compared with various classifiers like ANFIS, ANN, RBFN and evaluated on WBC data set and they achieved 97.55% success rate [64].

1.2 Methodology for classification

The literature survey depicts that most of the research have been focused on the diagnosis of the breast cancer and the researchers have proposed various predictive models using the benchmark data sets namely Breast Cancer Wisconsin (Diagnostic) Data sets. This data sets are widely available on the University of California at Irvine (UCI) Machine Learning repository. Our aim is to diagnose the breast cancer based on Breast Cancer Wisconsin (Diagnostic) Data sets. Features of these data sets are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. All features are assigned numeric values with four significant digits. In the data set, diagnosis is also specified by "2" for malignant and "4" for benign breast tumors. Hence, this is classification problem.

Using the same data sets, we have proposed various predictive models along with the time analysis for classification of breast tumor using various soft computing techniques like Artificial Neural Network (ANN), Support Vector Machine (SVM), Deep Learning (DL), Radial Basis Function Network (RBFN) and Adaptive Neuro Fuzzy Inference System (ANFIS). We employed various optimization techniques like Stochastic Gradient Descent (SGD), Adaptive Moment Estimation (Adam), Limited-memory Broyden fletcher Goldfarb Shanno (L-BFGS), Particle Swarm Optimization (PSO) in training the proposed models. We have also proposed various feature reduction techniques like Principal Component Analysis (PCA), Independent Component Analysis (ICA), Relief Based algorithm in data pre-processing. Proposed predictive models are implemented on WDBC and WBC data sets for classification of tumor into benign or malignant. Details of the data sets are mentioned in Appendix. The comparative analysis of our predictive models with predictive models of other researchers is carried out in detail and we found that the predictive models proposed by us gives highest classification accuracy in just few seconds.

1.3 Organization of the thesis

The layout of the thesis along with the proposed methodologies used in constructing the predictive models for classification of breast cancer (described in chapters 3 to 6) is as follows.

1.3.1 Chapter 1: Introduction

This chapter mainly deals with the motivation as well as literature survey of the breast cancer.

1.3.2 Chapter 2: Mathematical Preliminaries

This chapter concerns with the mathematical concepts used throughout the study.

1.3.3 Chapter 3: The Ultimate kernel machine based on Support Vector Machines

In machine learning, Support Vector Machines are supervised learning models with associated learning algorithms that analyse data for classification, regression analysis and outliers detection. This algorithm has a good generalisation ability, better performance and a robust mathematical theory. Machine learning, optimization techniques from operations research, and kernel functions from functional analysis are all combined in this approach. It is often referred to as a large margin classifier. When it comes to diagnose breast cancer, SVM has proven to be extremely effective. To build the cost-effective kernel machine for breast cancer diagnosis, the tools of PCA and k-fold Cross-Validation (CV) techniques are employed. The model is implemented on WDBC and WBC data sets to check the condition of the tumor for its malignancy. Classification accuracy and time computation are obtained and comparative experimental results are analysed under different conditions. For WBC data set, 100% accuracy is obtained using Polynomial kernel in just 0.03 second.

1.3.4 Chapter 4: Regularized Deep Neural Network with hybrid approach of Independent Component Analysis

Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised. One of the main advantages of deep learning lies in being able to solve complex problems that require discovering hidden patterns in the data and/or a deep understanding of complex relationships between a large number of interdependent variables. It not only has the ability to tackle nonlinear programming problems with the restrictions of equality and inequality, but it also

has a greater overall performance.

In this study, we investigate the use of Regularized Deep Neural Network (R-DNN) for the prediction of breast cancer. A variety of optimization techniques, such as Limited-memory Broyden Fletcher Goldfarb Shanno (L-BFGS), Stochastic Gradient Descant (SGD), Adaptive Moment Estimation (Adam), and activation functions like as Tanh, Sigmoid, and Rectified Linear Unit (ReLu) are used in the simulation of R-DNN. The Independent Component Analysis (ICA) approach is used to identify the most effective features to be used in the study. To measure the efficacy of the model, training and testing of the proposed network is carried out using the WDBC and WBC data sets. The detailed analysis of the accuracy is carried out and compared to the accuracy of other author's model. We find that the proposed network attains the highest accuracy.

1.3.5 Chapter 5: A Hybrid Approach of Adaptive Neuro Fuzzy Inference System and Novel Relief Algorithm

An Adaptive Neuro-Fuzzy Inference System or Adaptive Network-based Fuzzy Inference System (ANFIS) is a kind of artificial neural network that is based on Takagi–Sugeno fuzzy inference system. The technique was developed in the early 1990s [53, 55]. Since it integrates both neural networks and fuzzy logic principles, it has potential to capture the benefits of both in a single framework. Its inference system corresponds to a set of fuzzy IF–THEN rules that have learning capability to approximate nonlinear functions [4]. Hence, ANFIS is considered to be a universal estimator [52].

ANFIS provides accelerated learning capacity and adaptive interpretation capabilities to model complex patterns and apprehends nonlinear relationships. It is possible to identify two parts in the network structure, namely premise and consequence parts. In more details, the architecture is composed by five layers. The first layer takes the input values and determines the membership functions belonging to them. It is commonly called fuzzification layer. The second layer is responsible of generating the firing strengths for the rules. The role of the third layer is to normalize the computed firing strengths, by dividing each value for the total firing strength. The fourth layer takes as input the normalized values and the consequence parameter set. The values returned by this layer are the defuzzificated ones and those values are passed to the last layer to return the final output.