**A Thesis/Project/Dissertation Report**

on

**Breast Cancer Detection using Machine Learning**

***Submitted in partial fulfillment of the***

***requirement for the award of the degree of***

*Computer Science and Engineering*



**Under The Supervision of**

**Name of Supervisor: Dr. S. Srinivasan**

**Designation: Associate Professor**

Submitted By

Name of Student/s

**MD Talha 19SCSE1010826**

**Kashif Ahmad 19SCSE1180046**

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING / DEPARTMENT OF COMPUTERAPPLICATION**

**GALGOTIAS UNIVERSITY, GREATER NOIDA**

**INDIA**

**MONTH, YEAR**

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**GALGOTIAS UNIVERSITY, GREATER NOIDA CANDIDATE’S DECLARATION**

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **“Breast Cancer Detection using Machine Learning”** in partial fulfillment of the requirements for the award of the Bachelor of Technolog submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of Dr. S. Srinivasan Designation, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

**MD Talha 19SCSE1010826**

**Kashif Ahmad 19SCSE1180046**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor Name

Designation

**CERTIFICATE**

The Final Thesis/Project/ Dissertation Viva-Voce examination of MD Talha 19SCSE1010826 and Kashif Ahmad 19SCSE1180046 has been held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ and his/her work is recommended for the award of Bachelor of Technology.

**Signature of Examiner(s) Signature of Supervisor(s)**

**Signature of Project Coordinator Signature of Dean**

Date: November, 2013

Place: Greater Noida

**Abstract**

Each year number of deaths is increasing extremely because of breast cancer. It is the most frequent type of all cancers and the major cause of death in women worldwide. Any development for prediction and diagnosis of cancer disease is capital important for a healthy life. Consequently, high accuracy in cancer prediction is important to update the treatment aspect and the survivability standard of patients. Machine learning techniques can bring a large contribute on the process of prediction and early diagnosis of breast cancer, became a research hotspot and has been proved as a strong technique. In this study, we applied five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbours (KNN) on the Breast Cancer Wisconsin Diagnostic dataset, after obtaining the results, a performance evaluation and comparison is carried out between these different classifiers. The main objective of this research paper is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective whit respect to confusion matrix, accuracy and precision. It is observed that Support vector Machine outperformed all other classifiers and achieved the highest accuracy (97.2%).All the work is done in the Anaconda environment based on python programming language and Scikit-learn library.

**Table of Contents**

|  |  |  |  |
| --- | --- | --- | --- |
| **Title** | |  | **Page No.** |
| **Candidates Declaration** | | | **3** |
| **Abstract** | | | **4** |
| **Contents** | | | **7** |
| **List of Figures** | | | **5** |
| **Acronyms** | | |  |
| **Chapter 1** | **Introduction** | | **7** |
| **Chapter 2** | **Related Works** | | **5** |
| **Chapter 3** | **Methodology** | | **9** |
| **Chapter 4** | **Implementation** | | **11** |
| **Chapter 5** | **Conclusion** | | **41** |
| **Chapter 6** | **Future Enhancements** | |  |
|  | **References** | |  |

**CHAPTER-1**

**Introduction**

One of the most usual cancers is Breast Cancer, affects women highly worldwide, exemplifying the majority of other cancer types and related deaths according to global statistics, considering it as a significant public health issue in present’s society. The amount of death can be controlled by earlier diagnosis and it save people life significantly. Accurate diagnosing finds the stages of the cancer as benign, malignant, and severe malignant; it leads to apply timely treatment to the cancer, and it avoid patients undergoing improper treatments. Therefore, accurate diagnosis of cancer classes is the matter of several recent researches. Numerous approaches have been proposed for diagnosing the breast cancer data whereas the accuracy of classification is less. Since the breast cancer dataset has various and unique advantages of clinical features, machine learning algorithms have been used. But the machine learning algorithms are not fully automatic. Hence to provide an automatic learning, detection, and classification of breast cancer dataset is applied using deep learning algorithms. Deep learning algorithms are recognized as better method and used for breast cancer classification and prediction. Generally, data analytics and data mining approaches are widely used classifying complex datasets. Particularly in medical industry those approaches are widely applied for diagnosing and decision making. The earlier methods extract the standard features for classifying the data. Though the classification accuracy is less. The main objective of this paper is to analyse and observe the features used to predict the cancer class as benign or malignant. To extract a greater number of features and hidden features deep learning algorithm is used. Convolutional Neural Network is used as the deep learning algorithm for analysing and diagnosing the complex data. Deep learning algorithm learn the data using striding method. Convolution value is calculated from the stride function, where stride function can fetch the features from various sized images.

**CHAPTER-2**

**RELATED WORKS**

To understand the issues and challenges faced by the earlier research works, this section presents a detailed survey on various methods focused on breast cancer analysis. Author in explained breast cancer is one of the most important kind of cancers among various cancers. Breast cancer is considered as serious cancer type, it is a hot research topic with great value . Using data science and machine learning algorithms healthcare industries developing a great assistance for medical practitioners in decision making. Nowadays, it is a big challenge of diagnosing the pattern of breast cancer, since patterns are different is shape, texture and other clinical features. So, the healthcare industry is paying more attention in developing an efficient application using machine learning algorithms . In the earlier works, some of the researchers have focused on detecting breast cancer using image analysis for analysing the cancers have spread beyond the breast, other organs and nearby lymph nodes , and cell biology using selective but small datasets from algorithm evaluation challenges . Some of the earlier research works have focused on machine learning algorithms for diagnosing cancer dataset . The dataset used in the experimenting the machine learning algorithms is Wisconsin Diagnostic Breast Cancer dataset , and obtained significant output. Authors in used Gated Recurrent Unit combined with Support Vector Machine algorithms for diagnosing breast cancer on WDBC dataset. Some of the research works have used multiple kernel learning method for various healthcare problems and provides the way to examine the learned model. One of the authors combined SVM and Random optimization methods for demographic, clinical and biochemical data prediction. One of the authors in utilised MKL method for cancer-based thrombosis risk analysis to predict the progression of cancer disease in an oncology setting of breast cancer patients. Also, the author has provided a proof of concept for assessing the MKL based decision support system is really useful for cancer diagnosis.

**CHAPTER-3**

**LIMITATIONS AND MOTIVATION**

Some of the limitations identified from the above discussion are, the number of cancer classes are more, whereas SVM can classify only two (positive and negative), so it needs a classifier can find more classes. Thinking about Gated Recurrent Unit, it is highly suitable for huge volume of numerical or time series, and dependent data analysis. Also, combining GRU with SVM creates more time and computational complexity. Cell biology is not a common method used in all the medical centers. In the initial stage the breast cancer analysis has done over breast images and clinical data of the breast cancer, not on tissue analysis. In addition to the above said limitations, the detection and classification accuracy is less and all the methods are tested on different datasets.

From the above discussion it is identified that still the medical / healthcare industry requires a fully automatic and efficient algorithm for diagnosing and classifying breast cancer stage. Hence, to improve the accuracy of classification this paper used a fully automatic and effective method as convolution neural network and prove the performance. To do that, the entire contribution of the work is carried out into two different stages such as data acquisition from IoT industry and data analysation using convolution neural network.

The initial stage of the work has concentrated on data acquisition from medical network. The sample network is illustrated in Figure-1. The number of patients 𝑷 = {𝑷𝟏,𝑷𝟐,...,𝑷𝒊,...𝑷𝒏}, are monitored using different IoT devices such as, mammogram, CT scan and other scanning devices interconnected in the healthcare. The set of all data monitored from the devices are directly feed into the internet hub (it is a software model) where it collects and transmit to the corresponding server. From the server various medical experts 𝑴𝑬 = {𝑴𝑬𝟏, 𝑴𝑬𝟐, ... , 𝑴𝑬𝒊, ... , 𝑴𝑬𝒎 } access the data with authorisation and process the data using various data/image analytical methods to predict the health condition of the patients and the medical report/summary is uploaded in the crowd cloud DB. The final analyzation reports are sent to cloud and the volume of data is increased into a high amount, which is unimaginable. Since, the data is coming from various medical industries, IoT devices and medical experts, it become crowd data. The crowd data has two major kinds such as breast cancer data and the appropriate medical report used for enhancing the detection and classification accuracy.