# ABSTRACT

Over two decades, the most observed deadliest and dreadful disease is considered to be Cancer. The positive rate of cancer and the death rate over the years is rapidly increasing at an alarming rate. Among women, Breast cancer is the most diagnosed cancer. More than the treatment, the initial clinical examination of breast cancer itself is a very painful process for many patients. Even a small mistake can lead to false-negative and false-positive results that will be a burden on the life of a patient. To make this task easier and accurate we will be dealing with a novel approach by using technology. However, the present technology can make this painful clinical examination comparatively easier. This paper presents data mining algorithms like Decision tree, SVM which will be executed in the Spyder(anaconda) platform where the input is taken as values that differentiate the patient's record whether the cancer tissues are benign or malignant with higher probability based on the training dataset. This project also summarizes the cons of a traditional breast cancer diagnosis.



**Keywords** - Benign, Cancer, Data Mining, Decision tree, Malignant, SVM, Spyder.

# CONTENT

# LIST OF FIGURES

# 1 INTRODUCTION

## 1.1 INTRODUCTION

Thousands of females fall victim to breast cancer every year. The human body comprises millions of cells each with its unique function. When there is the unregulated growth of the cells it is termed as cancer. In this, cells divide and grow uncontrollably, forming an abnormal swelling tissue part called a tumor. Tumor cells grow and invade digestive, nervous and circulatory systems disrupting the bodies' normal functioning. Though every single tumor is not cancerous. Cancer is classified by the type of cell that is affected and more than 200 types of cancers are known. This paper is focused on Breast cancer. Breast cancer is the most common type of cancer among females across the world.

As per the National Breast Cancer Foundation, "Breast cancer is the most commonly diagnosed cancer in women". Breast cancer is the second largest cause of cancer death among women. Women generally approach clinics with a mild to serious pain in their breasts. After the examination of the breasts, the doctors usually suggest an ultrasound scan. The proceedings after the scan are more painful. Some women feel that the pain is intensely increased only after clinical proceedings. It is because of the painful process that is followed to detect whether the lymph node is malignant or benign. Malignant tumors are harmful or cancerous and benign tumors are harmless and can be removed through surgery. The treatment is then decided after thoroughly examining the state of the tumor. With the help of this paper, the detection of the state of the tumor can be decided with the help of data mining algorithms.

## 1.2 PROBLEM STATEMENT

In today's scenario, we have too much delay and inaccuracy in the diagnosis results provided by clinical centers. There are many cases where patients are given wrong inputs regarding their diagnosis and face false-positive or false-negative results which results in either poring the money unnecessarily or even pays for the death due to delay of treatment. In order to achieve these disadvantages of traditional methodology, Our system proposes an easy approach for clinical examination for breast cancer diagnosis

prediction. Using machine learning algorithms the classifier will be training by the dataset (ultrasound scan values) and the classifier will classify the new record values to either benign and malignant tumors. Higher accuracy results will be obtained than traditional methods, which also reduces the patient's waiting time and increases the life rate of people.

## 1.3 ORGANISATION THESIS

The organization thesis of this paper is the problems faced by the cancer patients especially about breas cancer patients, the cancer causes,risks and symptoms. It also discusses about the traditional procedure for diagnosis and treatment of the cancer and some drawbacks of traditional procedure treatment. In order to overcome those minor disadvantages, this paper contains one of the way for executing the detection of cancer using the present technology i.e. machine learning, data mining techniques and the previous dataset record. Using this technology it classifies whether the patient is suffering from benign or malignant tumor.

## 2.2 Yael Ben-Haim, Elad Tom-Tov: A STREAMING PARALLEL DECISION TREE ALGORITHM May 2010

In this journal, it proposes a new algorithm for building decision tree classifiers for classifying both large data sets and streaming data.

As recently noted (Bottou and Bousquet, 2008), the challenge which distinguishes large-scale learning from small-scale learning is that training time is limited compared to the amount of available data.

Thus, in our algorithm both training and testing are executed in a distributed environment, using only one pass on the data. We refer to the new algorithm as the Streaming Parallel Decision Tree (SPDT).

Decision trees are simple yet effective classification algorithms. One of their main advantages is that they provide human-readable rules of classification. Decision trees have several drawbacks, one of which is the need to sort all numerical attributes in order to decide where to split a node. This becomes costly in terms of running time and memory size, especially when decision trees are trained on large data. The various techniques for handling large data can be roughly grouped into two approaches: performing pre-sorting of the data or replacing sorting with approximate representations of the data such as sampling and/or histogram building.

While pre-sorting techniques are more accurate, they cannot accommodate very large data sets or streaming data. Faced with the challenge of handling large data, a large body of work has been dedicated to parallel decision tree algorithms Horizontal parallelism partitions the data so that different processors see different examples. Vertical parallelism enables different processors to see different attributes. Task parallelism distributes the tree nodes among the processors. Finally, hybrid parallelism combines horizontal or vertical parallelism in the first stages of tree construction with task parallelism towards the end. Like their serial counterparts, parallel decision trees overcome the sorting obstacle by applying pre-sorting, distributed sorting, and approximations. Following our interest in streaming data, we focus on approximate algorithms.

Our proposed algorithm builds the decision tree in a breadth-first mode, using horizontal parallelism. The core of our algorithm is an on-line method for building histograms from streaming data at the processors. The histograms are essentially compressed representations of the data, so that each processor can transmit an approximate description of the data that it sees to a master processor, with low communication complexity. The master processor integrates the information received from all the processors and determines which terminal nodes to split and how.

## 2.3 "Early Detection of Lung Cancer Risk Using Data Mining"- Article in Asian Pacific journal of cancer prevention: January 2013

Lung cancer is the leading cause of cancer death worldwide Therefore, identification of genetic as well as environmental factors are very important in developing novel methods of lung cancer prevention. However, this is a multi-layered problem. Therefore, a lung cancer risk prediction system is here proposed which is easy, cost-effective and time-saving. Materials and methods: Initially 400 cancer and non-cancer patients' data were collected from different diagnostic centers, pre-processed and clustered using a K-means clustering algorithm for identifying relevant and non-relevant data.

Next significant frequent patterns are discovered using AprioriTid and a decision tree algorithm. Results: Finally using the significant pattern prediction tools for a lung cancer prediction system was developed. This lung cancer risk prediction system should prove helpful in the detection of a person's predisposition for lung cancer.

In this project, the data mining algorithm used is k-means and decision tree where the collected lung cancer dataset is passed through k-means clustering in order to separate the significant data into groups of relevant and non-relevant data and it also group the unique by their attributes connected records into groups, which helps to build a training machine and also used for recognizing cancer to non-cancerous data for predicting the results of lung cancer. They also used the decision tree algorithm to develop significant patterns in order to obtain an accurate lung cancer predictor.

## 2.4 “An Ad Hoc Random Initialization Deep Neural Network Architecture for Discriminating Malignant Breast Cancer Lesions in Mammographic Images”- Andrea Duggento, Marco Aiello and Carlo Cavaliere, In WILEY (Hindawi), (AUG 2018)

Breast cancer is one of the most common cancers in women, with more than 1,300,000 cases and 450,000 deaths each year worldwide. In this context, recent studies showed that early breast cancer detection, along with suitable treatment, could significantly reduce breast cancer death rates in the long term. X-ray mammography is still the instrument of choice in breast cancer screening. In this context, the false-positive and false-negative rates commonly achieved by radiologists are extremely arduous to estimate and control although some authors have estimated figures of up to 20% of total diagnoses or more.

The introduction of novel artificial intelligence (AI) technologies applied to the diagnosis and, possibly, the prognosis of breast cancer could revolutionize the current status of the management of the breast cancer patient by assisting the radiologist in clinical image interpretation.

Lately, a breakthrough in the AI field has been brought about by the introduction of deep learning techniques in general and of convolutional neural networks in particular. Such techniques require no a priori feature space definition from the operator and are able to achieve classification performances which can even surpass human experts. In this paper, we design and validate an ad hoc CNN architecture specialized in breast lesion classification from imaging data only.

We explore a total of 260 model architectures in a train-validation-test split in order to propose a model selection criterion that can pose the emphasis on reducing false negatives while still retaining acceptable accuracy. We achieve an area under the receiver operatic characteristics curve of 0.785 (accuracy 71.19%) on the test set, demonstrating how an ad hoc random initialization architecture can and should be fine-tuned to a specific problem, especially in biomedical applications.

# 3 EXISTING SYSTEM

## 3.1 TRADITIONAL CLINICAL PROCEDURE

The traditional clinical proceedings for breast cancer are painful. Lymph nodes found in women's breasts might not be always a malignant tumor or cancer. A patient must undergo clinical examination by a breast surgeon then the result like x-ray mammography(Figure 1) an ultrasound scan(Figure 2) or MR mammography will be evaluated by a radiologist. If the report was suspicious, a needle biopsy is recommended followed by surgical removal of a lesion. The above procedure is very painful for women as they take a blood sample or a part of the lesion from the affected part. As it is most painful the pain stays for days to months. As per reports, diagnostic errors play a role in around 10% of patient deaths and breast cancer is no exception.

Research says, "Overall, screening mammograms miss about 20% of breast cancers that are present at the time of screening because mammograms cannot find the affected area for all skin types. False-negative results can delay treatment and a false sense of security for affected women". On the other hand, false-positive results would let the patient go through unwanted painful and expensive procedures.

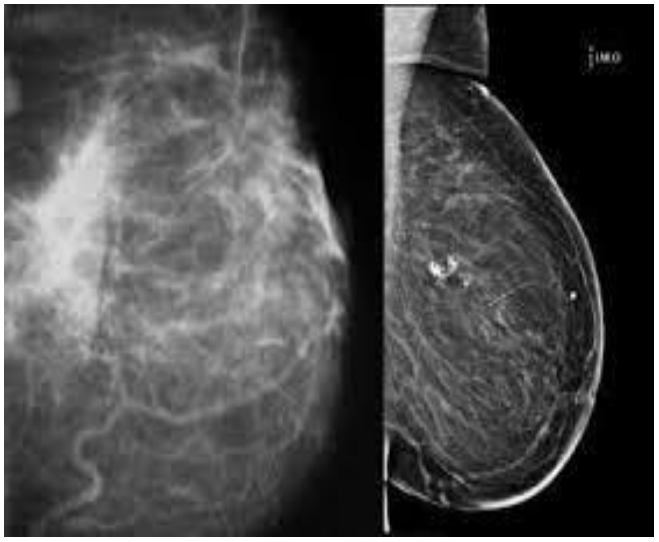

**Figure 1: X-ray mammography**