**BREAST CANCER DETECTION**

# **Absract**

Breast cancer is one of the leading cause for the death of women. In women Breast cancer is treated as the most significant issue. According to statistics released by the International Agency for Research on Cancer (IARC) in December 2020, Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide. Early diagnosis of this helps to prevent the cancer. If breast cancer is detected in early stage, then Survival rate is very high. Machine Learning methods are effective ways to classify data. Especially in the medical field, where those methods are widely used in diagnosis and analysis for decision making. In this paper, Data Visualization and performance comparisons between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB), K Nearest Neighbours (k-NN), Adaboost, XGboost and Random Forest conducted on Wisconsin breast cancer Dataset. The main objective is to evaluate the accuracy in the classification of data in terms of efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Our aim is to review various Techniques To detect early, efficiently and accurately Using Machine Learning. Experimental results show that XGboost offers the highest accuracy (98.24%) with the lowest error rate. Keywords: Breast Cancer, Machine Learning, Wisconsin, Algorithms, Detection.

# **1.introduction**

The second major cause of women's death is breast cancer (after lung cancer). 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women’s death is estimated. Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body. The cause of Breast Cancer includes changes and mutations in DNA. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less common. There are many algorithms for classification of breast cancer outcomes. The side effects of Breast Cancer are – Fatigue, Headaches, Pain and numbness (peripheral neuropathy), Bone loss and osteoporosis. There are many algorithms for classification and prediction of breast cancer outcomes. The present paper gives a comparison between the performance of four classifiers: SVM, Logistic Regression, Random Forest and kNN which are among the most influential data mining algorithms. It can be medically detected early during a screening examination through mammography or by portable cancer diagnostic tool. Cancerous breast tissues change with the progression of

the disease, which can be directly linked to cancer staging. The stage of breast cancer (I–IV) describes how far a patient’s cancer has proliferated. Statistical indicators such as tumour size, lymph node metastasis, and distant metastasis and so on are used to determine stages. To prevent cancer from spreading, patients have to undergo breast cancer surgery, chemotherapy, radiotherapy and endocrine. The goal of the research is to identify and classify Malignant and Benign patients and intending how to parametrize our classification

techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy. 10-fold cross validation test which is a Machine Learning Technique is used in JUPYTER to evaluate the data and analyse data in terms of effectiveness and efficiency.

# **1.2 problem solving**

Breast Cancer is one of the leading cancers developed in many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue. This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

# **1.3 proposed system**

In this project we will use Machine Learning Algorithms to detect breast cancer, based off of data. Breast Cancer (BC) is a common cancer for women around the world. Early detection of BC can greatly improve prognosis and survival chances by promoting clinical treatment to patients. We will use the UCI Machine Learning Repository for breast cancer dataset.

# **2.python**

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics developed by Guido van Rossum. It was originally released in 1991. Designed to be easy as well as fun, the name "Python" is a nod to the British comedy group Monty Python. Python has a reputation as a beginner-friendly language, replacing Java as the most widely used introductory language because it handles much of the complexity for the user, allowing beginners to focus on fully grasping programming concepts rather than minute details.

# **3.libraries used**

**Pandas:** Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data.

**Numpy:** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

**Matplotlib:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Twintern, wxPython, Qt, or GTK.

**Seaborn :** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**Sklearn:** scikit-learn is a [free software](https://en.wikipedia.org/wiki/Free_software) [machine learning](https://en.wikipedia.org/wiki/Machine_learning) [library](https://en.wikipedia.org/wiki/Library_(computing)) for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) [programming language](https://en.wikipedia.org/wiki/Programming_language). It features various [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and [clustering](https://en.wikipedia.org/wiki/Cluster_analysis) algorithms including [support-vector machines](https://en.wikipedia.org/wiki/Support_vector_machine), [random forests](https://en.wikipedia.org/wiki/Random_forests), [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting), [*k*-means](https://en.wikipedia.org/wiki/K-means_clustering) and [DBSCAN](https://en.wikipedia.org/wiki/DBSCAN), and is designed to interoperate with the Python numerical and scientific libraries [NumPy](https://en.wikipedia.org/wiki/NumPy) and [SciPy](https://en.wikipedia.org/wiki/SciPy). Scikit-learn is a [NumFOCUS](https://en.wikipedia.org/w/index.php?title=NumFOCUS&action=edit&redlink=1) fiscally sponsored project.

# **4.machine learning**

**K-nearest-neighbours (kNN) algorithm:** It is a simple supervised learning algorithm in pattern recognition. It is one of the most popular neighbourhood classifiers due to its simplicity and efficiency in the field of machine learning. KNN algorithm stores all cases and classifies new cases based on similarity measures; it searches the pattern space for the k training tuples that are closest to the unknown tuples. The performance depends on the optimal number of neighbours (k) chosen, which is different from one data sample to another.

**Support Vector Machine (SVM):** It is a supervised learning method derived from statistical learning theory for the classification of both linear and nonlinear data. SVM classifies data into two classes over a hyperplane at the same time avoiding over-fitting the data by maximizing the margin of hyperplane separating.

**Naïve Bayes (NB) It is a probabilistic classifier:** It is one of the most efficient classification algorithms based on applying Bayes' theorem with strong (naïve) independent assumptions. It assumes the value of the feature is independent of the value of any other features, given the class variable. Based on the maximum probability. It detects the class membership for the given tuple to a particular class.

**Logistic regression:** In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one. Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable (target) is categorical.

**5. CONCLUSION & FUTURE SCOPE**

In this project in python, we learned to build a breast cancer tumour predictor on the Wisconsin dataset and created graphs and results for the same. It has been observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of prediction systems. These systems can assist in proper treatment methods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.