SDM COLLEGE OF ENGINEERING & TECHNOLOGY,
DHARWAD – 580 002

MAJOR PROJECT

ON

BREAST CANCER PREDICTION : BIG DATA AND MACHINE LEARNING APPROACH

Submitted in partial fulfillment of the requirements for the Award of Degree of

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

Submitted by:

Name	USN
❖ **NIKITA JOSHI**	**2SD16IS034**
❖ **ANURAG DEVAGIRI**	**2SD16IS009**
❖ **POOJA GUNAGI**	**2SD16IS035**

Under the Guidance of :
1. Prof. Leena Sakri

Project Guide:
1. Dr. S.R. Biradar

Department of Information Science & Engineering
2019-2020

SDM COLLEGE OF ENGINEERING AND TECHNOLOGY, DHARWAD-580002
(An autonomous Institution affiliated to VTU, Belagavi – 590018)

Department of Information Science and Engineering

CERTIFICATE

This is to certify that the Mini Project work entitled "BREAST CANCER PREDICTION : BIG DATA AND MACHINE LEARNING APPROACH" is a bonafied work carried out by **Nikta Joshi, Anurag Devagiri and Pooja Gunagi** bearing USN **2SD16IS034,2SD16IS009 and 2SD16IS035** respectively, for successfully completing the Major Project for VII Semester B.E. Degree in Information Science and Engineering of SDM College of Engineering and Technology,Autonomous Institution under Vishveshvaraya Technological University, Belagavi during the year 2019–2020. It is certified that all necessary suggestion indicated for internal assessment have been incorporated in the report deposited in the department library. The project work has been approved as it has successfully satisfied the academic requirements as prescribed for the Bachelor of Engineering Degree.

Signature of the Guide	Signature of the HOD	Signature of the Principal
Prof. Leena Sakri	Dr. J. D. Pujari	Dr. S. B. Vanakudre

ACKNOWLEGEMENT

The successful completion of any task would be incomplete, without the mention of people who are responsible for the completion of this work.

First and foremost, we would like to express our sincere gratitude to our research supervisors who has in the literal sense, guided and supervised us. Their guidance deserves much more than the credit on the successful completion of our project and also for giving us an opportunity to implant our skills and for providing the required help and co-operation to complete the report. We are indebted with a deep sense of gratitude for the constant inspiration and valuable guidance throughout the work.

We are humbly thankful to Prof. J.D. Pujari, H.O.D., Dept of Information Science and Engineering for providing us a 24-hour internet access and high bandwidth round the clock.

Team members:

Name	**USN**
Nikita Joshi.	***2SD16IS034***
Anurag Devagiri	***2SD16IS009***
Pooja Gunagi.	***2SD16IS035***

TABLE OF CONTENTS

I. ABSTRACT

Mammograms are used for early detection of breast cancer. The breast cancer diagnostic system extracts features from these mammograms and classifies them as malignant or benign. These systems are very helpful to doctors in detecting and diagnosing the disease faster than any other traditional methods.
In this thesis an attempt has been made to classify the extracted features from Mammograms.
As benign or malignant by using Naive Bayes, K-NN, Support Vector Machine, Decision Tree and Random Forest Classifier approaches. Performance variation of the approaches by varying various parameters is studied. Finally, the results are compared to find the best performing approaches.

II. OBJECTIVE

The objective of this project is to predict the occurrence of tumor cells and classify the class of Breast Cancer (Malignant or Benign). This will reduce the time complexity as it will simplify the process of sending the reports to far off places in order to predict the Breast Cancer. This is an efficient technique which can easily replace the manual work of prediction.

III. INTRODUCTION

Breast cancer is the major occurring cancer among women. It has become a major reason for death in women. When the growth of cells in breast tissue became uncontrollable it forms a mass tissue called tumor. These tumors are mainly classified in to benign and malignant tumors which are cancerous and non-cancerous. Benign tumors are not harmful which do not spread to the other parts of the body. They can be removed completely and they do not grow back again. Malignant tumors are threat to life and they can spread other parts of the body and reappearing of malignant tumors can be seen often even when they are removed. Breaking away from the breast tumors cancer cells can travel through lymph vessels and blood vessels to reach other parts of the body. It may attach to other tissues of the body parts and grow to form new tumors that can cause damage to the entire function. Several tests are included to diagnose the patient, it includes surgical biopsy where patient need to go through operation. Nowadays data mining application has been increased in medical field. There are a few arguments that can support the use of data mining in health sector for breast cancer like early detection, early avoidance, and indication based medication, rectifying hospital data errors [1]. Many machine learning algorithms are used for the better treatment to the patient. Supervised algorithms such as classification and un-supervised such as regression and clustering which are helpful for the diagnosis of the patient.

There are many techniques to predict and classification on breast cancer pattern. This paper compares performance of three classification algorithms and their combination using ensemble approach that are suitable for direct interpretability of their results. This paper presents a new model by combining three classifiers that enhances the accuracy in recognizing breast cancer patients. Vote ensemble technique is used for classification of benign and malignant tumor. In this paper we inspected the generalization performance of J48, Naïve Bayes, and SVM in order to boost the prediction models for decision-making system in the prediction of breast cancer survivability. We are using a new Voting classifier approach where all three classification algorithms are combined for the prediction of breast cancer.

IV. LITERATURE REVIEW

Necessity for Research

Amongst all non-communicable diseases present, Cancer is one of the dangerous disease and there is a urgent and strict need for research in this domain. The symptoms of Cancer may not be clear at the initial stages and hence pose a great challenge for the medical practitioners and pathologists to predict the occurrence of tumour cells. Recently, in the year 2018 the reports from World Health Organisation (WHO) have shown that India has about 1 million cancer cases and more than 50% of these will be diagnosed in Women or precisely around 17000 more women will fall prey to this disease as compared to men. Therefore, because of all these challenges faced at every step of treatment, including the variations in the reports generated by one expert to another, by the manual identification of images, applying appropriate algorithms to detect the cancer can serve helpful in this domain due to its urgent need.

Reference Papers

1. Feature Selection for Breast Cancer Detection using Machine Learning Algorithms
   Authors : Sreyam Dasgupta, Ronit Chaudhuri, Swarnalatha Purushotham

2. Prediction of Breast Cancer Using Big Data Analytics
   Authors : K. Shailaja, B. Seetharamulu , M.A. Jabbar

3. Big data analytics for early detection of breast cancer based on machine learning
   Authors : Desislava Ivanova

4. Prediction of Breast Cancer using Big Data and Machine Learning Approaches
   Authors : B. Seetharamulu , M.A. Jabbar

EXISTING SYSTEM

1.Histology deals with the study of the microscopic structure of cells and tissues of organisms.

The knowledge of biological (microscopic) structures and their functions at the sub-cellular, cellular, tissue and organ levels is central to the study of disease proliferation and prognosis of disease. Also to study and analyse histological image under microscope, pathologists identify the morphological characteristics of tissue which indicates the presence of disease like cancer.

2. Role of Biotechnology in Cancer Control

Biotechnology has helped researchers to understand cancer in different ways such as Gene profiling, Genome analysis, cell culture, culturing transgenic cell lines and specially identification of new biomarkers for detection of risk and progression of cancer.

Available technologies in biotechnology help scientists to understand cause of cancer and the behavioral way of cancerous tissue in different environments.

3. Breast Cancer Histopathology Image Processing

Breast cancer is the most prevalent form of cancers among women, and image analysis methods that target this disease have a huge potential to reduce the workload in a typical pathology lab and to improve the quality of the interpretation.

4. Cancer and Radiation Therapy: Current Advances and Future Directions

Radiation remains an important modality for cancer treatment with ongoing efforts towards designing new radiation treatment modalities and techniques which continue to improve the survival and quality of life of cancer patients. With the improved clinical outcomes of cancer treatment, minimizing radiation therapy related toxicities has also become a priority.

PROPOSED SYSTEM

Cancer detection has always been a major issue for the pathologists and medical practitioners for diagnosis and treatment planning.

The manual identification of cancer from microscopic biopsy images is subjective in nature and may vary from expert to expert depending on their expertise and other factors which include lack of specific and accurate quantitative measures to classify into normal or cancerous one.

The objective is to develop a software that can guide and assist pathologists to detect the type. This software takes text inputs and records from patients and predicts the occurrence of Breast Cancer.

V. ANALYSIS AND REQUIREMENTS

Specific Requirements:

Inputs and Outputs

Input : Breast Cancer Dataset(Wisconsin) from UCI Repository
Output : Prediction of Breast Cancer tumor cells and occurrence of Breast Cancer depending upon the values entered by the user.

Functional Requirements

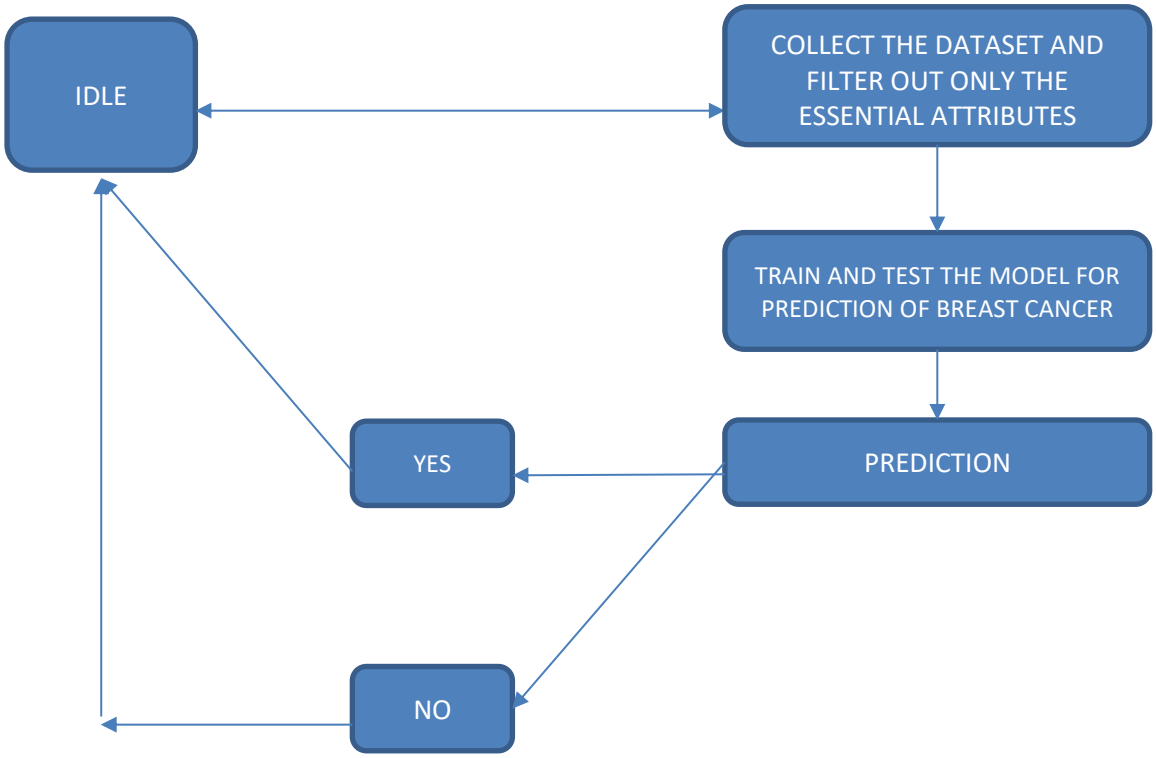
Jupyter Notebook is a platform for accepting(reading) the cancer dataset as input performing several different algorithms and comparing their efficiencies(accuracy) to determine the appropriate algorithm with higher accuracy rate. The output is expected to have a proper and efficient algorithm working in the background and that compares with the entered values by the user.

Hardware Constraints

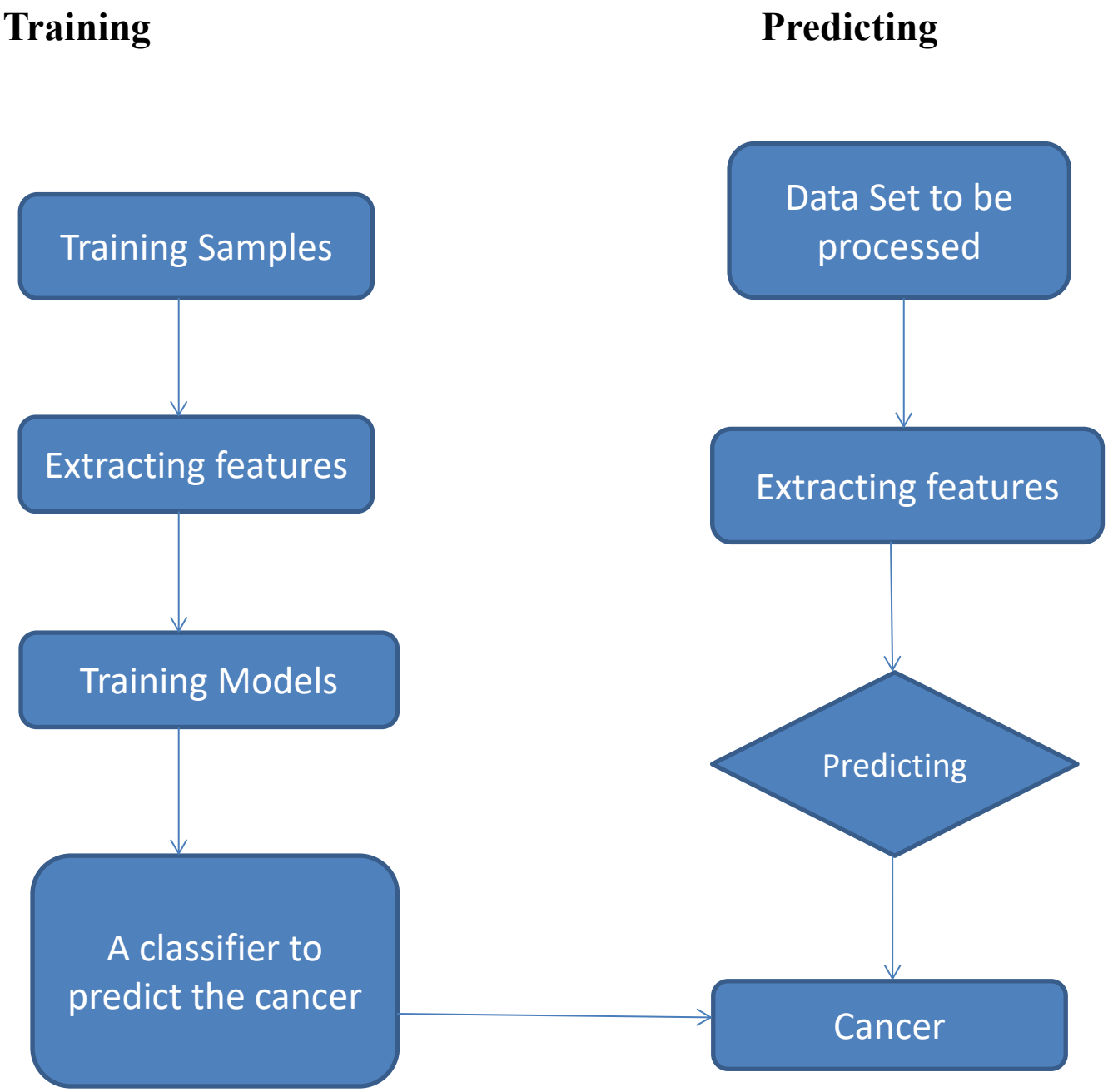
The system must contain a minimum of 8GB RAM to accommodate Anaconda Navigator software to run either Spyder or Jupiter Notebook to run several different algorithms of prediction and finally select the one that provides higher accuracy.

VI. DESIGN PHASE

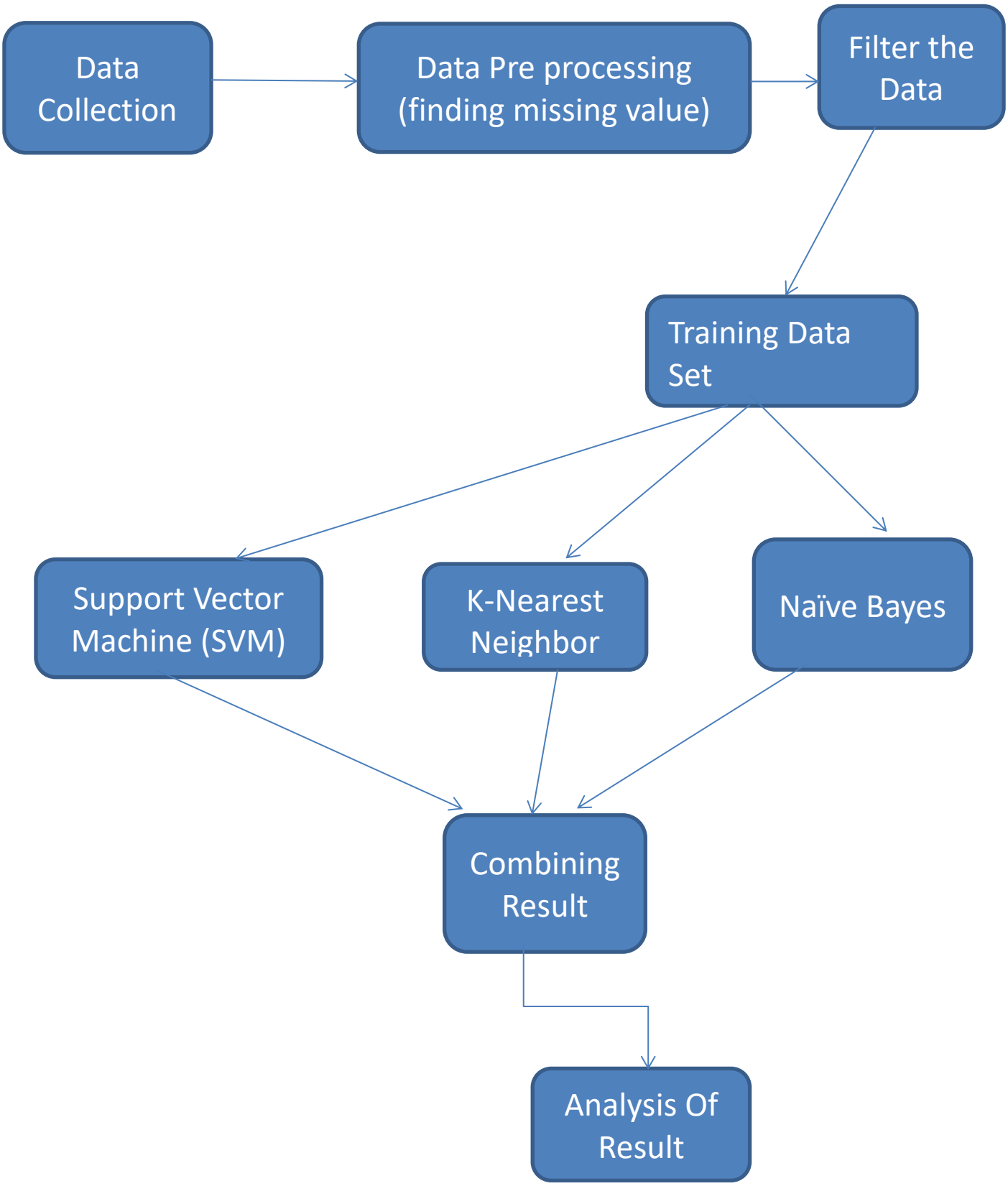
ADVANCED STATE DIAGRAM FOR BREAST CANCER PREDICTION



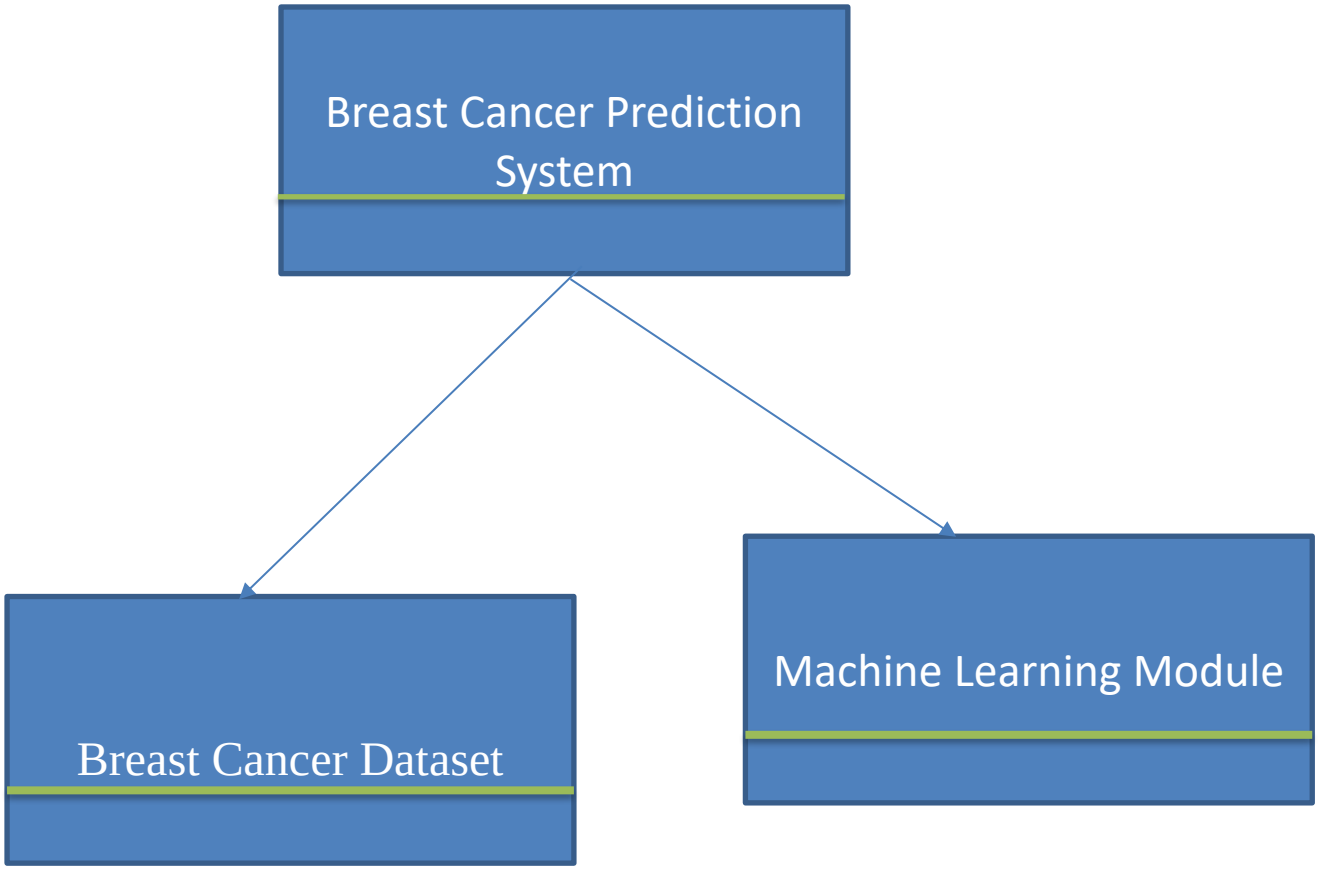
Flow Chart for Breast Cancer Prediction



Architecture Diagram for cancer prediction



Aggregation Concurrency Model for Cancer Detection



VII. IMPLEMENTATION

The implementation of this project is done using various algorithms such as K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest Classifier, Decision Tree, Naïve Bayes. The algorithm that provides highest efficiency is opted for prediction of Breast Cancer.

K-Nearest Neighbour

K-Nearest Neighbour strategy is a non-parametric and indolent learning algorithm. Its classification method is slightly different. It adopts the methodology of classifying the huge data based on the resemblance of its neighbours. It is a simple method that supplies all accessible instances and classifies new instances based on the resemblance measure. Here, a test point is selected and distances from the test point to every point in the training set is calculated. The KNN algorithm involves initialization of K followed by computation of distance between the input sample and the training samples. This step is followed by sorting the distances, taking the K nearest neighbours and finally applying simple majority. *k*-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbours, so that the nearer neighbours contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbour a weight of 1/*d*, where *d* is the distance to the neighbour. =The neighbours are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the *k*-NN algorithm is that it is sensitive to the local structure of the data.

Support Vector Machine

Support Vector Machine is a selective classifier formally defined by the separating hyperploid. When labelled training data is already given, this method results into a optimal hyperploid which further segregates into new instances. By varying the regularization parameter which is the tuning parameter of SVM, we will achieve substantial non-linear classified partition with higher accuracy in quite a reasonable amount of time. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When data are unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The **support-vector clustering** algorithm, created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabelled data, and is one of the most widely used clustering algorithms in industrial applications.

Naive Bayes

The basis for Naïve Bayes technique is Bayes Theorem. This classifier assumes that the existence of a particular characteristic in a class is not related to the existence of any other attribute. It is a method that predicts the probability of different class depending upon numerous attributes. They are non-pareil strategy utilized for the purpose of categorization as well as regression. As and how the tree grows deeper and deeper, the complexities regarding the rules of decision tree increases. A decision node consists of 2 or more subdivisions. Non-parent node will represent a categorization. The decision node at the topmost level represents the best predictor and is called as a root node. These trees are capable of handling categorical as well as numerical data. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression,[5]:718 which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

In the statistics and computer science literature, naive Bayes models are known under a variety of names, including **simple Bayes** and **independence Bayes**. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naïve Bayes is not (necessarily) a Bayesian method.

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. Random forests or random decision **forests** are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Decision Tree

They are non-pareil strategy utilized for the purpose of categorization as well as regression. As and how the tree grows deeper and deeper, the complexities regarding the rules of decision tree increases. A decision node consists of 2 or more subdivisions. Non-parent node will represent a categorization. The decision node at the topmost level represents the best predictor and is called as a root node. These trees are capable of handling categorical as well as numerical data.

VIII. TOOLS AND TECHNOLOGIES

In context with our project of Breast Cancer Prediction System, the tools used are :

Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Spyder

Spyder is an open source cross-platform integrated development environment (IDE) for scientific programming in the Python language. Spyder integrates with a number of prominent packages in the scientific Python stack,
including NumPy, SciPy, Matplotlib, pandas, IPython, SymPy and Cython, as well as other open source software. It is released under the MIT license. Initially created and developed by Pierre Raybaut in 2009, since 2012 Spyder has been maintained and continuously improved by a team of scientific Python developers and the community.

The driving technologies behind these tools are Machine Learning and Big Data Approach.

Big Data

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data was originally associated with three key concepts: *volume*, *variety*, and *velocity*. When we handle big data, we may not sample but simply observe and track what happens. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and *value*.

Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

IX. APPLICATIONS

This software assists pathologists and guides them to detect Breast Cancer using machine learning algorithms. For accurate prediction of cancer, the hospitals can use this software with much greater efficiency as compared to the conventional techniques. It takes the report values as the inputs.This software is very much useful in the cancer hospitals and various cancer testing labs for the purpose of cancer prediction.

X. CONCLUSION

Several Machine Learning techniques are used for the classification of benign and malignant tumor. In this paper we used best of three supervised learning classification algorithms for prediction of breast cancer and compared on different parameters. We mainly concentrate which classifier has the better accuracy for prediction and we combined two or more algorithms for the highest accuracy using one of ensemble approach. Instead of using one classification power we are using combinational power of rest of algorithms. The model we induced by combining three multiple class will be more reliable and it is sophisticated. From the malignant. Above all 235 instances correctly classified as malignant and 6 instances are incorrectly classified as benign. For all the experimental results based on accuracy and other parameters we can conclude that combining all three algorithms using voting approach is the best approach for predicting breast cancer. In future combination of algorithms should be taken wisely without over fitting problem. Stacking approach can be used for the better classification whereas in stacking the predictions by each different model given as input for a Meta level classifier whose output is the final class.

XI. REFERENCE

- [1]. K. Shailaja et al., "Applications of Big Data Analytics: A Systemat- ic Review", International Journal of Engineering Research in Com- puter Science and Engineering, volume 5, 2018.
- [2]. Ms. Shweta Srivastava et al., "A Review Paper on Feature Selec- tion Methodologies and Their Applications", International Journal of Engineering Research and Development, Volume 7, PP. 57-61, 2013.
- [3]. Animesh et al, "Study and analysis of Breast cancer Cell Detec- tion using Naïve Bayes, SVM and Ensemble Algorithms",International Journal of Computer Applications,vol2, 2016.
- [4]. G. Sumalath et al, "A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques",International Journal of Innovative Research in Computer and Communication Engineering,vol 5,2017.
- [5]. Hiba Asri, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", The 6th International Symposium on Frontiers in Ambient and Mobile Systems, pp.1064-1069.
- [6]. American Cancer Society. Breast Cancer Facts & Figures 2005- 2006. Atlanta: American Cancer Society, Inc. http://www.cancer.org/. [7]. https://www.r-project.org/