Comparative Study of Machine Learning Algorithms in Breast Cancer Prognosis and Prediction

Majid Khawar[1†]
majidkhawar10406@gmail.com
NFC-IET, Multan , Pakistan

Dr. Naeem Aslam [2††]
naeemaslam@nfciet.edu.pk
NFC-IET, Multan , Pakistan

Rao Muhammad Mahtab Mahboob [3†††]
raomahtab55@gmail.com
University of Agriculture,Faisalabad, Pakistan

Mueed Ahmed Mirza [4††††]
mueedahmed92@gmail.com
NFC-IET, Multan , Pakistan

Hassan Jahangir [5†††††]
hassanjahagir111@gmail.com
NFC-IET, Multan , Pakistan

Muhammad Awais Mughal [6††††††]
awais.ashraf12@gmail.com
Riphah I U, Lahore , Pakistan

Summary
Machine learning is a classification of artificial intelligence that apply collection of analytical and development approach which enable computer to determine the former pattern. It means that it's severely source desire in to the medical applications, such those based on large or multiplex data values. Machine learning is also concerned many times in cancer detection and diagnosis. In the cancer research the early prognosis and diagnosis of cancer is essential. A collection of machine learning approaches such as naïve Bayes, support vector machine (SVMs), artificial neural network (ANN) and Decision Trees (DT's) are used in medical research for progress of anticipate model following in successful and precise decision. It's a challenge to extract the meaningful information from the large stored dataset. In this study we will provide the performance of machine learning tools by using dataset of cancer related to Breast Cancer and predict the cancer susceptibility, cancer recurrence and cancer survival also we will tell you which tool is better for in term of accuracy and efficiency with respect to CPU time and memory consumption

Key words:
Accuracy, Efficiency, Prediction, Cancer Susceptibility, Cancer Recurrence, Cancer Survival, Precise Decision, Multiplex Data, Performance.

1. Introduction

Nowadays generation are growing, systematically lessen human work we're presenting an idea of machine learning and records technological know-how. A Process of Machine Learning in which the system attain accordingly with training. Learning is elimination but recall and ingestion of absorption records and selection totally depended on completed facts. It's strenuous to take decision based on achievable inputs. To conquer this issue, certain algorithms had been developed. To solve a particular project feed the set of rules with greater specifically facts. Most cases we will work with data as data source, examine output that can be used for the prediction [1].

1.1 Background scope and motivation:

Over the last ten years, a continuous development concerned to cancer research has been accomplished. Genetic disease cancer, this brought on by modifications gene, which control cells functions specifically division and developed. Adjustments genetics, reason cancer may come from guardian. Also get up at same time someone life-time because mistake that comes due to cell division or DNA damaged. The most cancers inflicting environmental exposure together with substance which include radiations, along with ultraviolet rays from sun, chemical tobacco smoke, massive use of alcohol, enough weight of body, bodily laziness and terrible substance Every person most cancers has a completely uniquely combination genetic adjustments. When cancer continues to grow the additional many changes occur in spite of in same tumour distinct cells might also have unique genetic modifications. Malignity also said cancer that is unexpected expansion of cells. Most cancer from tumours but not all tumours are cancerous. In order to find the early stage of cancer, there are different methods such as screening are applied by the scientists before the cancer indication. [2]. There are many new strategies for a beforehand prophecy of cancer and treatment results. In medical research, due to the new technology of medication multiple types of data have been together. For the specific guess about the disease is the challenging tasks for the physicians [3]. Cancer is a main root cause of disease

Manuscript received August 5, 2020
Manuscript revised August 20, 2020

among human deaths in lots of evolved countries. Cancer class in medical practice depended on clinical and pathological data may additionally produce incomplete or deceptive results. Nowadays, 100 types cancers based on effected cells which classified. At present, around the world death major cause is cancer. Many factors involved spreading cancer, like living area, person married or not, age, genes, living style etc. The early prognosis of cancers increases the risk of their treatment .For this reason, there had been several studies tries to create intelligent structures to assist doctors for early diagnosis of cancers. Currently maximum physicians figuring out the sort of cancers prefer to make a surgical biopsy. But most of them do not forget that biopsy is very critical project and must be avoided as lots as possible. Therefore, providing an intelligent machine which can assist physicians pick out the sort of most cancers and avoid needless surgical biopsy might be beneficial for each patients and physicians [5].

1.2 Machine Learning and cancer prognosis:

Now a day, machine learning applications widely used prognosis and diseases prediction. In cancer research machine learning is not contemporary. The purpose of machine learning approaches in the medical research is to model which can be used to for the prognosis or many other similar activities. Support Vector Machine (SVMs) and Decision Trees (DTs), artificial neural network, naïve Bayes techniques have been used in cancer diseasedisclosure and projection. In cancer prognostication the primarily goal is specify to predict cancer awareness, cancer repetition, cancer survival. Cancer awareness (susceptibility) is the estimation assessment of risk or cancer sensitivity. In the cancer recurrence (cancer repetition) the prediction of redevelopment chances of cancer after apparent resolution of the disease, local control to repetition. Cancer survival predict the life expectancy, survivability, cancer durability. Different type of data like genomic data, microarray, clinical, mutations, tumour size are used with machine learning application like support vector machine, decision trees, for cancer prognosis or prediction [6].

Machine learning in fitness care is intended to enhance clinical records analysis. Record in image format sets copied computed using x ray, imaging technique picture of anatomy, screening process low energy x rays for breast cancer identification ,or mental disorder studies to understand genes have been used for cancer prognosis or stage identification. Machine getting to know have large impact cancer diagnosis with purpose of beneficial for patient care the development of the computational tools for disease prediction is important [7]. Almost 20 years the Artificial intelligence, machine learning have used to diagnose and cancer classification, but only some studies have investigated relevance cancer prognosis. To develop models for cancer progression, cancer recurrence, and survivability in particular, Machine Learning or semi-supervised learning techniques were recently implemented. [8].

Major cause secondly, in women death due to breast most cancers (after lung-cancer). In woman breast cancer represents about twelve percent of all new cancer instances and twenty five percent of all cancers. The main role play in caner care are the information and communication technologies. In life any time breast cancer can be appeared in ladies and affect around ten percent of the women. In latest years, the incidence rate keeps growing and data display that survival price is eighty-eight percent after five years from diagnosis and eighty percent after ten years from analysis [9]. Big data was superior not only size of information however also creating price from, Big statistics, turns into a synonymous of records mining, business analytics and commercial enterprise intelligence, has made a large trade in BI from reporting and choice to prediction results. High performance outcomes, medicine price reducing, health care fee enhancing, and important thing keep people alive, machine learning procedures applied in medical science rising rapidly. Algorithms applied for the prediction of many cancers. Machine learning approaches like Support vector machine, decision trees, artificial-neural-networks can use for most cancers prediction. Our aim is to evaluate efficiency and effectiveness of those algorithms on various tools in terms accuracy, confusion matrix, sensitivity, specificity and precision [10].

2. REVIEW OF LITERATURE

Beginning of death for both man and woman is cancer. Cancer is specified to group of connected diseases. Anywhere in the body the disease can be developed and the cancer is group of more than hundred different type of diseases. The every type of cancer, body cells start to disassociate and these cells circulate inclosing the tissues except quit. Cancer develop a lot of cells anywhere in the

body. Most commonly the human cells developed and disassociated from the new cells. Cells become damaged after the cells rear old they expire and new cells place. Cancer starts when genetic adjustments intervene with this orderly process. Cells begin to grow uncontrollably. These cells may shape a mass called a tumour. A tumour may be cancerous or benign. A cancerous tumour is malignant, meaning it is able to grow and spread to other elements of the body. A benign tumour approach the tumour can grow but will now not spread. Some sorts of cancer do no longer shape a tumour. These include leukemia, most kinds of lymphoma, and melanoma. By 2020, cancer rate can be increase to 15 million, using the effective tools for cancer identifications its important step in medical field. For this reason and achieving the goal researcher from different fields are working on it to improve medical disease diagnosis process [11]. Anywhere in the body cancer can occur. Like commonly breast cancer is common in women. Prostate cancer commonly in men. Colorectal, lung cancer are common in both men, women. Cancer have many types. Normally named of cancer are the organism where the cancers consist in human body. Like, lung cancer is in cells of lungs, brain cancer is in cells of brain. Types of cancer are consist lung, breast, skin, head and neck, gynecological, hematological, CNS (central nervous system), genitourinary, colorectal, upper gastrointestinal cancer [12]. List of cancers childhood cancers, cancer in adolescents and young adults, we can concern cancer surgeon and discuss on particular type of cancer and location of cancer. Cancer treatment depends on types, its stages, and physical health. Kill many cancer cells and reducing damage to normal cells is the medication treatment in cancer treatment. New technologies make possibility for that treatments may be directly tumour removing, using some chemical on affected cells and may use x rays to remove cancer cells.

Commonplace cancer in women is breast cancer, affecting approximately 10 percentage of all girls at few levels of their life. In modern times, the rate keeps increasing and data show that rate of survive is eighty eight percentage after five year from analysis, 80 percentage after ten years from prognosis. Early prediction of breast cancer so far have made heaps of improvement, death rate of breast cancer by 39 percent, starting from 1989. Due to varying nature of breast cancers symptoms, patients are often subjected to a barrage of tests, including but not limited mammography, ultrasound and biopsy, to check their likelihoods of being diagnosed breast cancer. Biopsy, is most indicative amongst these procedures, which entails extraction of pattern cells or tissues for examination. Breast most cancers is a form of cancer that occurs frequently in women and is the leading purpose of women's deaths. These deaths may be reduced through early detection of the cancerous cells. Cancerous cells are detected by means of performing numerous exams like biopsy, MRI, ultrasound. Dataset used in this undertaking contains features which can be computed from a digitized picture of a first-rate needle aspiration, biopsy of a breast mass. They describe traits of cellular nuclei present within the image. Diagnosis of breast most cancers is done with the aid of classifying the tumours. Tumours may be both malignant and benignant. Malignant tumours greater dangerous than benignant [13]. Breast most cancers is a disease in which cells within the breast develop out of control. There are one-of-a-kind varieties of breast cancer. The sort of breast most cancers relies upon on which cells inside the breast become cancer. Breast most cancers can begin in distinctive parts of the breast.. Breast cancer can unfold outdoor the breast through blood vessels and lymph vessels. When breast most cancers spreads to other parts of the body, it is said to have metastasized. The common sorts of breast cancer are the cancer cells grow out of doors the ducts into other parts of the breast tissue. Invasive most cancers cells also can unfold, or metastasize, to other components of the body called Invasive ductal carcinoma. The Cancer cells spread from lobules to the breast tissues which can be close by. These invasive cancer cells also can spread to other elements of the body called Invasive lobular carcinoma.

Machine learning is a class of artificial intelligence which associated the issues of learning from sample data to general theory of assumption. Machine learning models are increasingly used in the field of science. The principal objective of those version is to decide the effective variables and the relationship between them. Machine gaining knowledge of belongs to a technology and engineering of making intelligence appliance. Automated information acquisition centered by way of gadget learning through layout and implementation of algorithms had empirical facts is required with the aid of algorithms. Basically the techniques for studying of the gadget is taught via system gaining knowledge of depending on use probability. Also, in medical, gadget learning techniques are extensively used for the analysis of

most cancers and also to differentiate among benign and malignant cancers [14]. There are two phases in every learning process (i) From given datasets, the assessment of unknown dependencies (ii)To estimate the up to date conclusion of system by using the estimated dependencies. Machine learning have some common types.

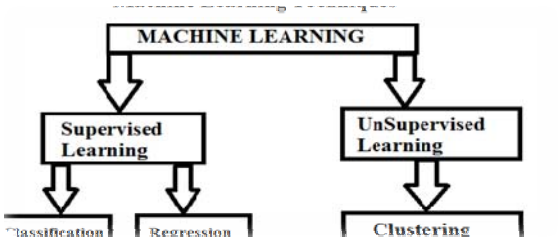

Fig. 1. Machine Learning

Supervised-Learning: Fixed training data is used to approximate input data to the desired result. Supervised-learning is used in classification problems. In the classification learning process there is categorization of finite classes of the data. It does it by using running statistics thru a gaining knowledge of algorithm. The intention of supervised learning is, efficiently perceive the new statistics given to it via supervised getting to know and use of previous facts set and getting to know algorithms can study the approach to become aware of facts. The algorithms operating below supervised gaining knowledge of takes inputs that the output is already recognized for the purpose just so the algorithms will create device to discover via holding it examine the precise output with the already regarded output to test for to any extent further errors. The system is then organized consequently. The well-known supervised studies are the classification, boosting, regression, predictions. Then the model is modified with the aid of it consequently.

Unsupervised-Learning: During the learning process no categorized data are provided and no annotation of the result. Unsupervised learning used in clustering that attempt to find class or cluster that expressed the data values. Unsupervised learning studies however systems will learn to represent specific input patterns in a manner that reflects the applied math structure of the assortment of input patterns. By contrast with supervised learning or reinforcement learning, there aren't any express target outputs or environmental evaluations related to every input; rather the unattended learner brings in contact previous biases on what aspects of the structure of the input ought to be captured within the output. A specific output is not having by unsupervised learning. Finding the structures and patterns in the data is aimed by the learning agent.

Semi-Supervised Learning: Association of supervised and unsupervised learning. To generate precise model semi-supervised learning combines ordered and unordered data. Semi-supervised learning is used where much unordered data than order dataset.

When we place a machine learning mechanism, sample data initiate the basic element. Every sample data is relate with particular characteristics and consists different type of principles. Also particular class of data used to permit the preferred choice of approaches and procedures that are used for investigation or prediction. Some issues related to data may the data nature and to make valuable data so data preparation steps is necessary. Data quality issues may be data duplication and missing values. If the data of poor quality then prediction quality also poor. Prediction quality can be improved after improving data sets. To make data more relevant for investigation or prediction, data preparation steps used that concern data improvement. There are different approaches used in data preparation that concerned on data improvement for sophisticated connection in particular Machine learning methods [3].

The main purpose of machine learning approaches is create model that is normally given for the auguring , estimation and many more related activities. The common activity of machine learning algorithm is classification. Learning function that classifies the data values in given classes known as classification. When classification model is created in machine learning techniques, the training errors can be produced. The best classification model suit the training set all over and precisely categorize all the situations. Classification model is acquire using one or many machine learning applications, it's needed to investigate the performance of classification. The accomplished anatomy of every suggested model to uniform in term to responsiveness, special features, precision. When the data is pre-processed machine learning methods includes Artificial Neural Networks, Decision Trees, Support Vector Machines and Bayesian Network is available. We have to specify the two kinds of machine learning approaches that are Support Vector Machine, Decision Trees and type of data unified as well as the estimation process that are used to evaluate the long term achievement of approaches that are used for

cancer prediction or outcomes [3].There has been a lot of research on cancer diagnosis by using machine learning techniques. To evaluate prostate cancer survivability Decision tree, logistic Regression, Artificial Neural Network are employed [15]. In medical field disease prediction play an important role. There are different types of diseases predicted in research namely Hepatitis, Liver, mental disorder, heart disease, Diabetes, Breast Cancer, Thyroid disease, etc. [16].

Support vector machine (SVMs): Supervised-learning model techniques Support Vector machine which commonly concerned to classification and regression. Support vector machine (SVMs) is latest technique of machine learning used study of cancer prognosis, prediction. Primarily Support vector machine (SVMs) design input data values, quality space of greater scale and determine hyper plane that divide data values two classes. Output classifier accomplished significance conception and can used to dependable classification new sample [3]. Each form of classifier needs a metric to measure the similarity or distance among patterns. SVM classifier uses inner product as metric. If there are established relationships among pattern's attributes, such records can be accommodated thru extra dimensions, and this could be found out via a mapping. SVMs have been effectively carried out to a number of actual world problems. It find packages information mining, bioinformatics and pattern recognition. [17].

Decision Trees (DTs): Supervised-machine learning techniques Decision Tree which commonly used for classification and regression. Decision tree is efficient training algorithm [6]. Decision tree (DTs) is tree structure classification mechanism where nodes show the input value and leaves compare to decision results. Focus on architecture of decision trees are simple and easy to interpret and learn. In the decision tree there would be a tree like structure that node show the variable based on the input and the tree leaves that resembles to the opinion results. New sample classification able to predict data categorization in specific architecture with suitable reasons [18]. It's a predictive modelling approaches used to create model that predict a value of selected variable for decision making. In tree model selected variable use discrete values called classification trees and if selected variable use continuous values called regression trees [3]. The records sample is divided into homogeneous subsets primarily based on the most first rate splitter in input attributes [19]. Decision tree generated using J48 may use for classification. It makes use of the truth that every characteristic of the information may be used to determine via splitting the records into smaller subsets. J48 examines the normalized information advantage that results from selecting a characteristic for splitting the records. [17].

There are many different machine learning approaches and quality selected algorithms have been commonly used to disease prediction and prognosis. Mostly of these use machine learning methods to model that advancement of cancer and determine significant mediator which used in classification. The achievement in disease prediction is certainly under the nature of medical diagnosis and prognostic prediction employ for diagnostic decision. In the cancer prediction machine learning techniques used to predict cancer risk, cancer control and cancer survival. Therefore cancer prediction is involved with three Prognostication tasks. (i) Prognostication of cancer sensitivity or risk assessment) (ii) Prognostication of cancer return or cancer local control to repetition) (iii) Prognostication of cancer durability or cancer survived) [3].

Predication of cancer susceptibility: Cancer susceptibility a term used to describe the mutations (changes occur) in a specific genes that grow a person risk of any type of cancer. Cancer susceptibility gene mutation are commonly inherited (Parent to Child) and may be seen inside families. Person has cancer susceptibility, gene mutation may help to prevent, diagnoses and treat cancer. It's not a necessary person who have a cancer susceptibility gene mutation will develop cancer. Different type of data (genomic data, mutations) are used with machine learning techniques like naïve Bayes model, lot of decision trees, and support vector machine used for cancer susceptibility prediction [6].

Prediction of cancer recurrence: Recurrence means that come back cancer after some time. This happen after treatment of months or may be years after the original cancer was treated. The chances of cancer recurrence is depend on the type of primary cancer. Cancer reoccur due to tiny area of cancer developing cells can remains in human body after the treatment, after sometime these cells can multiply and produce many enough to cause symptoms or for test to find them. It's depend on the primary cancer, when and where cancer recurs. There are three types of cancer recurrence that are local, regional and distant. Support vector machine, artificial neural

network algorithm and clinical information, molecular data and microarray data used for cancer recurrence prediction [6].

Prediction of cancer survival: Researchers collect the past data of same people and have same type of cancer after applying many type of statistical or analytics to estimate the projection. Analytics tells three types of survival that are particular, relative, overall. Percentage of patient with particular survival who have not expire from cancer after diagnosis. We can also said disease specific cancer survival cause of death in medical record. Percentage of cancer patient who survived a specific time after diagnosis is relative survival. And after diagnosis percentage of people who have not expire known as overall survival. There is recurrence free survival which consist have no sign of cancer in human body after treatment. For prediction and modelling of the cancer (cancer survival) the micro array data and gene expressions are used with support vector machine and naïve Bayes machine learning algorithms [20].

3. METHODOLOGY

In the medical research machine learning applications are used for disease diagnosis or prediction. Tools regarding machine learning may use to analyze data values that may help us in disease prediction. Early prediction helpful for better
treatment.Using computerized computing implement
and machine learning processes to promote and enhance in medical field inquiry or diagnosis is encouraging and essential operations. In cancer prediction there is prediction of cancer awareness, repetition and cancer survival. Data, in the machine learning, in most common structure is that to splitting into the training dataset and testing data. Also can be to three types of dataset: trained, validate and test data, and fit our model on the training data, and the using training data in order to make predictions on the test data. Training dataset is the actual dataset that we may use to train our model. Model learns from dataset which used. Test data is a sample data which used to provide a neutral analysis of a final model fit on the training dataset. The Test dataset provides standard used for the model evaluation and used when model completely trained according to data. Splitting the dataset into training, validation testing sets can be determined on two categories. Firstly, it depends on how much the finally numbers samples in data and secondly the actual model the user is training. Data may be consisted collection of observations, Disease, pain, altered body functions, DNA data (SNPs), codded format of the data, pathological reports, and textual format of data like report results, numerical formats, records of clinician. The data which may use for the prediction will be Breast cancer dataset. Before using the dataset with machine learning tools there may be steps of data preprocessing. In this step we may treat the missing values or irrelevant data removing. Data have been imported in CSV format and leaned before used. Values that are disappeared and null, are treated and all known valid values are used for investigation. Based on the desired output we may use the breast cancer data and lung cancer data with the machine learning classification algorithm techniques using support vector machine (SVMs) and decision trees (DTs) algorithms. Tools that may use for this analysis are the WEKA, R studio (based on R language), Spyder (Python), Jupyter Notebook (Python) for the data analysis. Result included statistical and visualization. We analyze these result and predict the cancer disease (cancer susceptibility, recurrence, survival). We present the performance evaluation methods use to evaluate the proposed methods. Performance evaluation methods of cancer prediction or prognosis are the classification accuracy, analysis of sensitivity, specificity, TPR, FPR, TNR, FNR, confusion matrix. We may analyze the accuracy and efficiency of the tools with respect to the CPU time and memory consumption and predicted that which tool perform better performance on Support vector machine and decision tree algorithms.

4. RESULTS AND DISCUSSIONS

Confusion Matrix:

Performance of the classification algorithms as summarized referred to a confusion matrix. It is a simple way to alter achievement of classification class by differentiation of how many effective occurrence are correct or incorrect class and how many are the negative times are correct or incorrect classified.

TABLE 1: CONFUSION MATRIX

		Actual	
		Positive	**Negative**
Predicted	Predicted Positive	True Positive **(TP)**	False Positive **(FP)**
	Predicted Negative	False Negative **(FN)**	True Negative **(TN)**

Positive (P): Inspection positive
Negative (N): Inspection not positive
True Positive (TP): Inspection positive, and prediction positive.
True Negative (TN): Inspection negative, and prediction negative.
False Positive (FP): Inspection negative, but the prediction positive.
False Negative (FN): Inspection positive, but the prediction negative.
Accuracy: Evaluation of classification models done by one the metrics called accuracy. Accuracy is the fraction of prediction. It determines the total quantity of precise forecast over the total quantity of prognosis made by model. Calculated as the:
Accuracy = (TN) + (TP) / (TN) + (TP) + (FN) + (FP)
Recall:
Recall compute number of the sufferer that prognosis to the difficulty among those patient that actually have the difficulty. Calculated as the:
Recall= (TP) / (FN) + (TP)
Precision:
Precision measured number of patients, actually difficulty which those classified have difficulty by model. Precision calculated as:
Precision = (TP)/ (FP) + (TP)
F1 Score:
Burdened common of the precision and take into account is called F1 rating. Therefore FP and FN are extract via that outcome into the deliberation. Unconscious it isn't always as comprehend as the validity, but F_1 is mostly additional useful, than the validity. F1 Score can be computed as follows:
F1 score =2* (recall) * (precision)/(recall)+(precision)
Sensitivity:
Sensitivity is portion corrected positive investigations over the total-positive values. We can also called the sensitivity as the Recall and well known as the true positive rate. Best case of sensitivity is the 1.0, but the worst case sensitivity is the 0.00. Sensitivity can be computed as the:
Sensitivity = (TP) / (FN) + (TP)
Specificity:
Specificity is the results of performance to negative outcomes. Specificity is totally negative result of the recall. Specificity is calculated as:
Specificity = (TN) / (FP) + (TN)

CPU time:
The time to impose for processing by the CPU for the processing for a particular software or process. Programs and applications commonly do now not use the processor 100% of the time that they're walking; some of that time is spent on I/O operations and fetching and storing facts at the RAM or storage device. The CPU time is only when this system actually uses the CPU to carry out tasks such as doing arithmetic and good judgment operations. CPU time is also called processing time. The CPU time required by packages and strategies are regularly minuscule, fractions of a second, that's why a lot of programs can be jogging at the identical time, but nonetheless get their activate the CPU.
Memory Usage:
The physical device which have a capability to manage and store the information for specific time is the computer memory. Devices used in memory devices utilize integration circuits and used utility program, hardware, operating systems and software. The memory available on your system is the sum of all physical memory installed on your system and the page file on your hard disk, which is used to complement the physical memory.

(A) BREAST CANCER DATASET RESULTS COMPARISONS:

(i) Support Vector Machine Results

Accuracy, CPU Time and Memory Usage Comparisons of Different tools regarding Support Vector Machine Algorithms:

TABLE 2 : Tools performance comparisons using SVM algorithm with BREAST cancer

List of Tools	CPU Time (S)	Memory Usage (MB)	Accuracy
WEKA	0.092	4.521	0.78
R Studio	0.051	0.049	0.22
Spyder (Python)	0.064	0.040	0.68
Jupyter Notebook (Python)	0.021	0.043	0.72

Accuracy, CPU Time and Memory Usage Comparisons of Different tools regarding Support Vector Machine Algorithms:

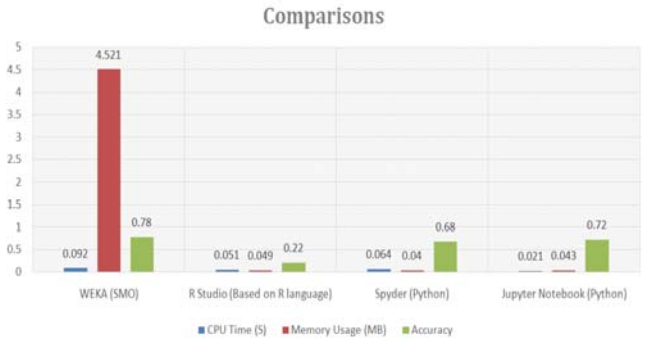

Fig. 2. Accuracy, CPU Time and Memory Usage Comparisons of Different tools regarding Support Vector Machine Algorithms with BREAST cancer data

ROC Parameters (SVM):

TABLE 3: ROC PARAMETERS OF TOOLS USING SVM ALGORITHM WITH BREAST CANCER DATA

List of Tools	F1 Score	Sensitivity	Specificity	TPR	FPR	TNR	FNR
WEKA	0.20	0.76	0.80	0.87	0.10	0.80	0.41
R Studio	0.20	0.45	0.15	0.13	3.00	0.15	0.15
Spyder (Python)	0.80	0.68	0.00	1.00	0.00	0.00	0.00
Jupyter Notebook (Python)	0.83	0.72	0.00	1.00	0.00	0.00	0.00

Tools ROC Parameters:

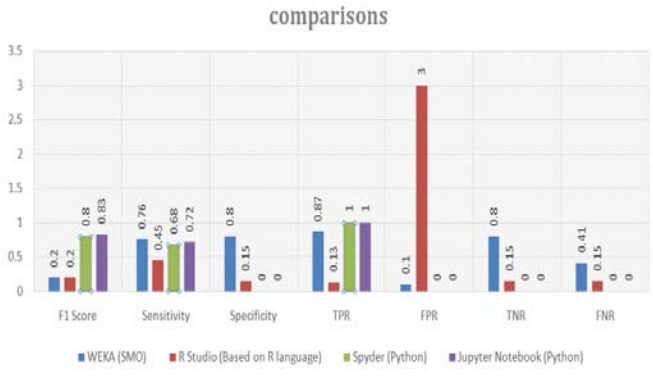

Fig. 3. ROC Parameter Comparisons of Different tools regarding Support Vector Machine Algorithms with BREAST cancer data

(ii) Decision Tree Results:

Accuracy, CPU Time and Memory Usage Comparisons of Different tools regarding Decision Tree Algorithms:

TABLE 4: TOOLS PERFORMANCE COMPARISONS USING DECISION TREE ALGORITHM WITH BREAST CANCER DATA

List of Tools	CPU Time (S)	Memory Usage (MB)	Accuracy
WEKA	0.057	1.582	0.77
R Studio	0.061	0.011	0.65
Spyder (Python)	0.066	0.040	0.72
Jupyter Notebook (Python)	0.050	0.042	0.77

Accuracy, CPU Time and Memory Usage Comparisons of Different tools regarding Decision Tree Algorithms:

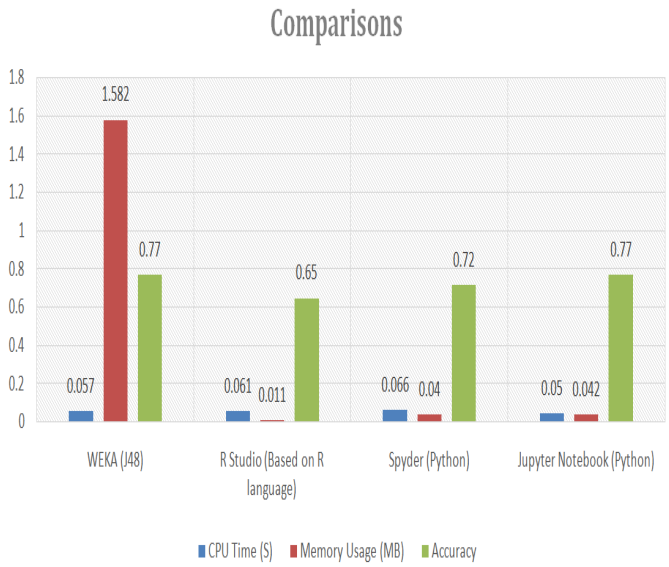

Fig. 4. Accuracy, CPU Time and Memory Usage Comparisons of Different tools regarding Decision Tree Algorithms with BREAST cancer data

ROC Parameters (DTs):

TABLE 5: ROC Parameters of tools using Decision Tree algorithm with BREAST cancer data

List of Tools	F1 Score	Sensitivity	Specificity	TPR	FPR	TNR	FNR
WEKA (J48)	0.78	0.80	0.73	0.78	0.21	0.73	0.24
R Studio	0.76	0.78	0.29	0.75	0.26	0.29	0.59
Spyder (Python)	0.80	0.75	0.60	0.88	0.10	0.60	1.00
Jupyter Notebook (Python)	0.83	0.77	0.72	0.95	0.03	0.72	1.81

Tools ROC Parameters:

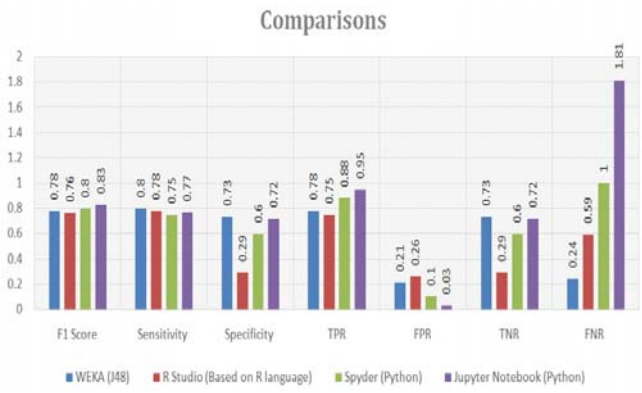

Fig. 5. ROC Parameter Comparisons of Different tools regarding Decision Tree Algorithms with BREAST cancer data

3. CONCLUSION

In that study we will provide presentation of machine learning tools by using dataset of cancer related and predict the cancer susceptibility, cancer recurrence and cancer survival.

Using confusion matrix and ROC Parameters if the TPR and TNR could be very high then it indicate there is no cancer. And the FNR is very low then predict the cancer but you are not aware. When you are not aware then there is no treatment for cancer then FNR could be very low and cancer will be extreme stage. FPR should be little bit high, predict that there is no cancer but you are in doubt. After medical checkup it predict that there is no cancer then FPR will be down. We have used the machine learning algorithms that are support vector machine and decision tree with the two type's cancer datasets that are the lung cancer, breast cancer datasets.

By using different tools (WEKA, R Studio, Spyder, Jupyter Notebook) using different algorithms according to the Table 3 for the ROC parameters using Support vector machine algorithm and Table 5 for the ROC parameters using decision tree algorithm with breast cancer data the WEKA tool gives the better result for the cancer prediction. According to the Table 2 performance comparison based on the CPU time (S) and memory usage (MB) of tools (WEKA, R Studio, Spyder, Jupyter Notebook) with support vector machine algorithm using breast cancer data. WEKA tool gives the better results which are the CPU time is 0.092 seconds and Memory usage of WEKA tool is 4.521 and accuracy according to confusion matrix is 0.78 which is higher than other tools. And According to the Table 4 performance comparison based on the CPU time (S) and memory usage (MB) of tools (WEKA, R Studio, Spyder, Jupyter Notebook) using the Decision Tree algorithm with breast cancer data. Jupyter Notebook tool gives the better results which are the CPU time is 0.050 seconds and Memory usage of Jupyter Notebook tool is 0.042 and accuracy according to confusion matrix is 0.77 which is higher than other tools.

References

[1]. M. P. M. S. S. M. M. Mr.P.SATHIYANARAYANAN, "Identification of Breast Cancer Using The Decision," in Preceeding of International Conference on Systems Computation Automation and Networking, 2019.

[2]. Z. W. H. C. A. P. a. X. L. Wembin Yue, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis," www.mdpi.com/journal/design, pp. 2-13, 2018.

[3]. T. P. E. K. P. E. V. K. I. F. Konstantina Kourou, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, pp. 8-17, 2015.

[4]. J. A. Shikha Agrawal, "Neural Network Techniques for Cacer Predition: A Survey," in 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, 2015.

[5]. R. &. S. Oskouei, "Data Mining and medical world: breast cancers diagnosis, treatment, prognosis and challenges," American journal of cancer research, 2017.

[6]. D. S. W. Joseph A. Cruz, "Applications of Machine Learning in Cancer Prediction and Prognosis," in Cancer Informatics, Canada, 2006.

[7]. J. D. M. W. K. J. R. B. I. Blaz Zupan, "Machine

Learning for survival analysis: a case study on recurrence of prostate cancer," Artificial Intelligence in Medicine, pp. 59-75, 2000.

[8]. F. M. Z. S. R. N. S. F. G. M. R. Patrizia Ferroni, "Breast Cancer Prognosis Using a Machine Learning Approach," Cancers, pp. 1-9, 2019.

[9]. E. A. P. A. E. M. a. R. A. Ahmad LG, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," Journal of Health & Medical Informatics, vol. IV, no. 2, 2013.

[10]. H. M. H. A. M. T. N. Hiba Asri, "Using Machine Learning ALgorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Computer Science, pp. 1064-1069, 2016.

[11]. T. Turki, "An Empirical Study of Machine Learning Algorithms for Cancer Identification," IEEE , 2018.

[12]. T. T. W. L. D. P. R. L. K. A. B. D. C. D. K. M. S. M. K. L. M. D. M. A. S. V. Sunil Gupta, "Machine-learning prediction of cancer survival: a retrospective study using electronic adminstrative record and a cancer registry," group.bmj.com, 2015.

[13]. P. N. R. A. R. Anusha Bharat, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," IEEE Third International Conference on Circuits, Control, Communication and Computing, 2018.

[14]. M. M. M. M. A. B. Mitra Montazeri, "Machine learning models in Breast cancer survival prediction," IOS Press, pp. 31-42, 2016.

[15]. S. L. W. L. J. C. Y. HUAIYU WEN, "COMPARISION OF FOUR MACHINE LEARNING TECHNIQUES FOR THE PREDICTION OF PROSTATE CANCER SURVIVABILITY," IEEE, pp. 112-116, 2018.

[16]. S. S. S.Vijiyarani, "Disease Prediction in Data Mining Technique- A Survey," International Journal of Computer Applications & Information Technology, pp. 17-21, January, 2013.

[17]. D. S. R. L. N. S.Aruna, "KNOWLEDGE BASED ANALYSIS OF VARIOUS STATISTICAL TOOLS IN DETECTING BREAST CANCER," CCSEA, CS&IT, pp. 37-45, 2011.

[18]. J. W. Z. L. X. Z. Yawen Xiao, "A Deep Learning-based Multi-model Ensemble Method for Cancer Prediction," Computer Methods and Programs in Biomedicine, pp. 1-16, 2017.

[19]. S. A. Pahulpreet Singh Kohli, "Application of Machine Learning in Disease Prediction," in 4th International Conference on Computing communication and Automation, India, 2018.

[20]. M. G. R. S. L. S. H. E. Azadeh BASHIRI, "Improving the Prediction of Survival in Cancer Patients using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review," Iran J Public Health, pp. 165-172, 2017.

[21]. T. A. H Bharathi, "A Review of Lung cancer Prediction System using Data Mining Techniques and Self Organizing Map (SOM)," Internation Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, pp. 2190-2195, 2017.

[22]. T. A. H Bharathi, "A Review of Lung cancer Prediction System using Data Mining Techniques and Self Organizing Map (SOM)," Research India Publications, India, 2017.

[23]. T. A. H Bharathi, "A Review of Lung Cancer Prediction System using Data Mining Techniques and Self Organizaing Map (SOM)," India, 2017.

MAJID KHAWAR received the BS Software Engineering degree from Riphah International University Islamabad, Pakistan, in 2017, and Perusing the M.S. in computer science from NFC, Institute of Engineering & Technology Multan, Pakistan. He is currently working as a Lecturer at University College of Management & Sciences, Khanewal. He is interested in Artificial Intelligence, Software Engineering, Machine Learning, and Bio-Informatics.

DR NAEEM ASLAM received the Ph.D. degree in Bio-Informatics from the University of Veterinary and Animal Science (UVAS), Lahore. He is currently working as the Head of Computer Science Department, NFC Institute of Engineering and Technology, Multan, Pakistan. He has 17 Years of Professional experience in education and industry worked on versatile projects. He is interested in Data Science, Artificial Intelligence, Machine Learning and Bio-Informatics.

RAO MUHAMMAD MAHTAB MAHBOOB received the software engineering degree from GC University Faisalabad, Pakistan, in 2015, and the M.S. degree in computer science from the University Of Agriculture Faisalabad Pakistan, in 2017. He is currently working as a Lecturer at University College of Management &Sciences, Khanewal. His research interests include Software Engineering, Data Mining, Artificial Intelligence, Machine Learning, and Bio-Informatics.

MUEED AHMAD MIRZA received the Master in Computer Science degree from GC University Faisalabad, Pakistan, in 2014, and Perusing the M.S. in computer science from the NFC, Institute of Engineering &Technology Multan, Pakistan. He is currently working as a lecturer at University College of Management & Sciences, Khanewal. His research interests include Networking, Data Mining, Artificial Intelligence, and Machine Learning.

HASSAN JAHANGIR received BS Software Engineering from COMSATS University Islamabad, Pakistan, in 2018, and Perusing the M.S. in computer science from the NFC Institute of Engineering &Technology, Multan, Pakistan. He is currently working as a lecturer at University College of Management & Sciences, Khanewal. His research interests include Image processing, Artificial intelligence and Machine learning.

M AWAIS MUGHAL received BS Computer Science from COMSATS University Lahore, Pakistan, in 2017, and Perusing the M.S. in computer science from the Riphah International University, Lahore, Pakistan. He is currently working as a lecturer at University College of Management & Sciences, Khanewal. His research interests include Data Science, Graph Algorithms, and Machine Learning.