Analyzing the impact of Machine Learning on Cancer treatments

Victor Chang[1], Gunji Srilikhita[2], Qianwen Ariel Xu[2], M.A. Hossain[3] and Mohsen Guizani[4]

1. Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, UK
2. Cybersecurity, Information Systems and AI Research Group, School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK
3. Vice President Office, Cambodia University of Technology and Science, Cambodia
4. Faculty Affairs and Institutional Advancement, Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates.

Emails: victorchang.research@gmail.com/v.chang1@aston.ac.uk; srilikhita99@gmail.com, qianwen.ariel.xu@gmail.com; alamgir@camtech.edu.kh and mguizani@ieee.org

Abstract

The survival rate of breast cancer prediction has been a significant issue for researchers. Nowadays, the health care industry has completely transformed by using modern technologies and their applications for medical services. Among those technologies, machine learning is one of them, which has gained attention by people that its new advanced technology would give accurate results by using modeling methods for prediction. As this is a branch of artificial intelligence, it employs various statics, probabilistic and optimistic tools. This is applied to medical services, primarily based on proteomic and genomic measurements. The aim is to use the dataset of cancer treatment and predict the results of patients by using machine learning with its modeling methods for accurate results. Recently experts have even used this machine learning in cancer for prognosis and forecasting.

Keywords: Machine learning; cancer analysis; breast cancer analysis using R; methods for breast cancer analysis.

1. INTRODUCTION

1.1 Significance of research

Technological advancement in healthcare institutions has gained momentum in current phenomena where the entire world is revolving around modern tools—for example, using smart

applications to connect with various people, opting for home delivery, online shopping, remote diagnosis, online treatment henceforth. Therefore, it becomes essential for experts of the healthcare sector to apply modern technologies while performing several activities like recording patients' history and medical practitioners enforced to reform treatment procedures. Hence, this overall subject will help understand the important role of machine learning in cancer treatmentusing R. In fact, R is known as chemotherapy or R-CHOP, frequently given in cyclesthree weeks apart. Every cycle of R-CHOP lasts for almost three weeks, completely based on size and kind of tumor (Zhang et al., 2017). The treatment is generally incurred in a chemotherapy per day unit. It means that the entire process is a complicated one, in which machine learning helpsin processing various complex concepts or terms. Thus, this research will explore techniques to perform cancer analysis and outline cancer treatment processing. We can present how experts use machine learning in understanding complicated concepts or images of diseases.

1.2 Research aim and questions

Cancer is a very dangerous disease for the human body that destroys the functioning of a person's body organs completely. As a result, practitioners of the healthcare sector are opting for severaladvanced technologies to facilitate cancer patients with creative services or ideas. For example, machine learning aided doctors in identifying cancer symptoms at a prior stage, which aids in curing it as soon as possible (Koutsouleris et al., 2018). Therefore, conducting an in-depth investigationon cancer treatment alternatives is a smart move in this innovative environment. Hence, the foremost aim of this research is "to analyze the impact of machine learning on cancer treatment by using R". In order to collect further information on this subject, several related questions are demonstrated as follows.

- What is the role of machine learning in the healthcare sector?
- What is the significance of machine learning in cancer treatment with the use of R?

1.3 Research problem

According to the designed title, this research revolves around machine learning how this technology has reformed medical services and covers examples of cancer treatment to understand the crucial role of technology. However, the foremost problem of research is to describe therole of machine learning in treating cancer with the help of suitable examples. Cancer is a life-taking disease that is not easy to recover from. Experts of the healthcare sector are trying toconduct an in-depth investigation to identify accurate solutions to cancer. Therefore, it will be very

problematic to outline the role of machine learning in cancer treatment success via chemotherapy (Redlich et al., 2016).

This research investigated the relationship between machine learning techniques with the health sector, especially cancer treatment, and then conducted cancer prediction using several machine learning algorithms, including Random Forest, Naïve Bayes, and Decision Tree. The results show that all of the three algorithms achieved accuracy over 80%, and Random Forest achieved the best result, which is 96.47%. In addition, a computer with machine learning algorithms is far more efficient than human beings, capable of performing thousands of biopsies in just mini-seconds. Thus, the utilization of machine learning algorithms can largely reduce the workload of medical experts and make cancer treatment more efficient.

2: LITERATURE REVIEW

2.1 Machine Learning in Cancer Detection and Diagnosis

Noor and Narwal (2017) cancer is one disease and many associated diseases that indulge in uncontrolled cellular growth and reproduction. This problem is seen as the leading reason behind death in the developed world and second in developing countries, like killing nearly eight million individuals in a single year. The initial diagnosis and prognosis of cancer type have increased the demand for research as it aids successive clinical management of patients. For effective clinical decisions, distinguishing between benign and malignant tumors is indispensable. Conservatively, the statistical method is used for categorizing the high and low risk of cancer. Still, experts have applied machine learning for cancer prognosis and assumptions to overcome the loopholes of traditional statistical techniques. The author has stated that machine learning is a branch of artificial intelligence that provides many statistical, probabilistic, and optimization tools that permit computers to learn or acquire things from previous instances. This potentiality of machine learning is appropriate for medical applications, especially for those departments based on complicated proteomicand genomic calculations. Thus, machine learning is highly used by experts in cancer diagnosis and detection (Bibault et al., 2016).

2.2: Machine learning techniques used in Cancer Detection

Current technologies like microarray and sequencing have covered the style of computational techniques or tools. There are several indispensable issues in cell biology which

requisite the dense nonlinear communications between functional modules to be measured. Now the significance of computer simulation in the cellular process is broadly approved and numerous simulation algorithms are beneficial for studying specific subsystems formed. While using machine learning, data and output are run on a computer for designing an impressive program. One of the new frameworks is created to solve the classification issue of cancer on microarraygene expression facts that seek the benefits of correntropy expense and make it vigorous against distinct noises or outliers. Another new technique was introduced, a knowledge-based system for cancer classification using clustering, noise removal, and techniques classification. EM or expectation maximization was utilized as a clustering technique for clustering the facts in a similar assembly. Additionally, Classification and Regression Trees are preferred forgenerating fuzzy regulations while classifying cancer disease under a knowledge-based system offuzzy rule reasoning tools. Principal Component Analysis was combined with an existing knowledge-based system to reduce the multicollinearity obstacle (Zhou et al., 2018).

According to Ferlay et al. (2018), 18.1 million cancer cases across the world annually have faced uncertainties and challenges. The pathologists have conducted diagnoses of cancer and 96-98% are successful in diagnosing cancer. Machine learning has played an important role in curing and providing treatment to cancer patients in this phase. This is AI, which helps take data, find all patterns, and train by using data and getting outcomes from it. It works fasterthan humans and thousands of biopsies are used in a second to gain outcomes. Machines are repeating the process to gain positive outcomes. ML provides accuracy as all things are performedthrough the help of IoT that is highly used and works faster than others to analyze data patterns.ML is of two types such as supervised as well as unsupervised. In the case of supervised learning,there is the use of an algorithm that is taught through data and making predictions for the future. In the case of unsupervised learning, data are not labeled for the creation of jobs to fit data andinformation with the help of patterns. This can be wrong or right, depending upon the situation. Experts have introduced various ML models that include linear regression, logistic regression, decision tree, SVM, and Random Forest.

2.3: R-CHOP

According to the National Cancer Institute (2012), chemotherapy is an integration of medicines because experts believe that combinations work more effectively than single drugs as distinct drugs can diversely kill cancer cells. Every drug in an amalgamation is approved by Food

and Drug Administration for treating the cancer issue or related circumstances.R-CHOP is utilized to treat non-Hodgkin lymphoma, a mixture of other drugs for treating a different kind of cancer.

According to Joseph and David S. Wishart (2006), machine learning is not new in cancer research because artificial neural networks and decision trees have been utilized for detecting cancer for around 20 years. Machine learning techniques are currently used in a broadrange of applications, from cancer-detecting to tumor classification through X-ray and CRT pictures.

2.4: Breast Cancer Detection using Machine learning techniques

Mohammed et al. (2020) expressed that breast cancer is the second leading reason behind women's death worldwide. In 2019, approximately 268,600 newcases of invasive breast cancer were identified for diagnosis in US women's whereas roughly 62,930 non-invasive new cases of breast cancer were found. In order to maximize the possibility of treatment and survivability, early detection is a must or best method. Mohammed et al. (2020) expressed that breast cancer is the second leading reason behind women's death worldwide. In 2019, approximately 268,600 new cases of invasive breast cancer were identified for diagnosis in US women's whereas roughly 62,930 non-invasive new cases of breast cancer were found. In order to maximize the possibility of treatment and survivability, early detection is a must or best method. A mammogram is an x-ray picture of the breast used to analyze women with breast cancer who do not have any signs or symptoms of this disease. Screening mammography is a kind of mammogram that experts use for checking the symptoms of disease (Kumar, Ramachandra, and Nagamani, 2013). It aids in minimizing the death rate of breast cancer amongst women of 40 to 70. Authors have found that multiple classifier algorithms have been applied on healthcare data clusters to estimate patient and associated diagnosis assessment. For instance, machine learning tools have been useful in analyzing tumor behavior for people with breast cancer. Swan et al. (2013) compared three diverse classifiers such as J48, NB, and SMO in context with accuracy to detect breast cancer.

The rise in new technologies in the medical field, especially cancer, helps cure a patient's suffering. Cancer research has been done to screen it early and kind of cancer before causing any symptoms in the past years. The strategies have been made to identify early predictions of cancer treatment and gain some outcomes. Mohammed et al. (2020) applied data mining algorithms on diverse healthcare datasets for breast cancer classification. All these algorithms have shown appropriate classification outcomes and motivated several investigators to overcome challenging tasks. For example, a convolutional neural network was preferred for assuming and segmenting

invasive ductal carcinoma in histology pictures of the breast with almost 88% accuracy. ML is highly used and creates a relationship among discovery and patterns to gain the outcome of kind of cancer people suffer. New trends in works integrate various data such as genomics and clinical to achieve some outcomes. There is a problem in validating external aspects to know about the performance of the predictive model used in business. ML has assisted and improved cancer accuracy, prediction of survival, and recurrence. The accuracy rose from the previous year and was enhanced by 15-20% using ML tools. With the analysis of several studies, it has been found that ML techniques are highly used in cases of curing cancer patients and improving their health conditions. This helps in the early detection of cancer and its types so that treatment is provided accordingly for the highest results.

2.5: Deep Learning of Wisconsin Breast Cancer diagnosis

Yuefeng Zhang (2019) stated that the machine learning technique had been used in medical diagnosisfrom the previous era. The Wisconsin Breast Cancer Database dataset is broadly utilized by experts during experiments of research. Many publications have concentrated on conventional machine learning techniques like decision trees and based methods. Stahl and Geekette have recently applied this technique to the WBCD dataset in breast cancer diagnosis using values measured from digitized pictures of Fine Needle Aspirate of a breast mass. Stahl has used the WBCD dataset with useful characteristics like mean, standard error henceforth and experimented with three kinds of depth neuron networks; 1,2 and 3 unseen coatings of nearly 30 neurons without having any data processing. These kinds of characteristics can outline the cell nuclei shown in the picture. In contrast, Geekette has utilized only initially determined structures with pre-processing like scale and center but has not offered detailed information on pre-processing or network designing of neurons in use. An example of code for obtaining a statistical summary of the dataset is shown in Figures 1 and 2 (Tseng et al., 2015).

```
import pandas as pd


headers =
["ID","CT","UCSize","UCShape","MA","SECSize","BN","BC","NN","Mitoses"
,"Diagnosis"]
data = pd.read_csv('breast-cancer-wisconsin.csv', na_values='?',
        header=None, index_col=['ID'], names = headers)
data = data.reset_index(drop=True)
data = data.fillna(0)
data.describe()
```

Figure 1 Code to get a statistical summary

	CT	UCSize	UCShape	MA	SECSize	BN	BC	NN	Mitoses	Diagnosis
count	699.000000	699.000000	699.000000	699.000000	699.000000	683.000000	699.000000	699.000000	699.000000	699.000000
mean	4.417740	3.134478	3.207439	2.806867	3.216023	3.544656	3.437768	2.866953	1.589413	2.689557
std	2.815741	3.051459	2.971913	2.855379	2.214300	3.643857	2.438364	3.053634	1.715078	0.951273
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	2.000000
25%	2.000000	1.000000	1.000000	1.000000	2.000000	1.000000	2.000000	1.000000	1.000000	2.000000
50%	4.000000	1.000000	1.000000	1.000000	2.000000	1.000000	3.000000	1.000000	1.000000	2.000000
75%	6.000000	5.000000	5.000000	4.000000	4.000000	6.000000	5.000000	4.000000	1.000000	4.000000
max	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	4.000000

CT - Clump Thickness
UCSize - Uniformity of Cell Size
UCShape - Uniformity of Cell Shape
MA - Marginal Adhesion
SECSize - Single Epithelial Cell Size
BN - Bare Nuclei
BC - Bland Chromatin
NN - Normal Nucleoli

Figure 2: Statistical summary

Some of the data obtained via dataset are missing info, the small size of a dataset, several ranges of data values, unbalanced information, and skewed facts. Significantly, one of the common problems with the dataset is the missing of facts and there is an absence of exception in the WBCD dataset. Particularly as highlighted in Figure 2, there are almost 16 Bare Nuclei entries missing. Furthermore, the WBCD dataset contains almost 699 samples, which are very small for obtaining deep learning after classifying the dataset in training and testing sub-parts.

Specialists use different suitable open source codes to understand the style or complexity of machine learning in breast cancer diagnosis. Consequently, it is an overview of Wisconsin Breast Cancer diagnosis using experiments or examples of codes that experts understand via in-depth learning. It means that topographies of breast cancer are figured from digitalized pictures of fine needle aspirate of a breast mass. This demonstrates the characteristics of cell nuclei available in an image (Shouval et al., 2021). Through this dataset, experts can estimate whether breast cancer in Wisconsin is benign or maligned. See Figure 3.

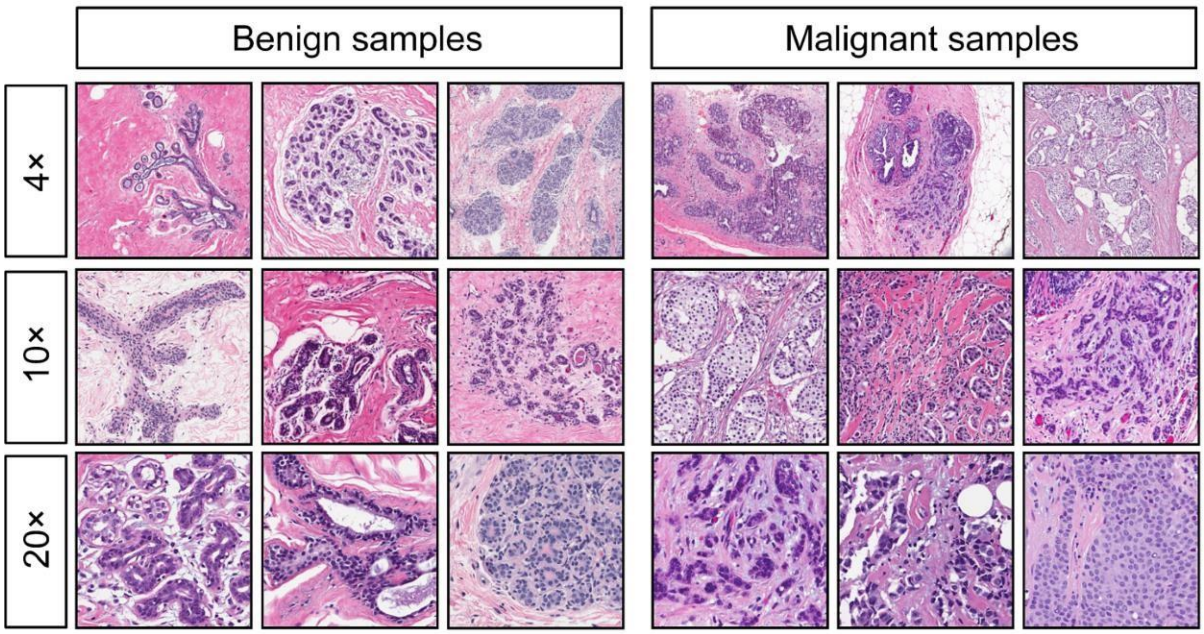

Figure 3 Breast Cancer Diagnostic EDA

This digitalized image depicts the features of cell nuclei. The separating plane, described in the above picture, was acquired using the Multi-surface method-Tree. It is a classification tool that utilizes Linear Programming for constructing a decision tree. Additionally, an actual linear program is preferred to get the extrication plane in three-dimensional space.

3. RESEARCH METHODOLOGY

This section of the paper will reveal the different methods supported in gathering relevant information on cancer treatment and how machine learning is aiding medical practitionersduring emergency situations. The objective is to demonstrate the path followed by the researcher for examining viewpoints of several authors, scholars, and bloggers on machine learning in context with the healthcare industry.

3.1 Research design

Several machine learning (ML) models are applied in cancer treatments as it predicts a recurrence of oral cancer fromthe emission of patients. Various data are collected from patients to reach a final conclusion that ML is better than other types of models used in cancer treatment. Most of the people who have cancer prefer such a model to diagnose and accordingly take treatment easily. Thisis regarded as an accurate model that has used testing to identify low-risk and high-risk groups for cancer recurrence. ML is an AI technique for completing tasks, improving, and

effectively getting iteration. It is regarded as the first model that has been built for discriminating cancers along with their subtypes and provides treatment to patients so that they can recover fast. Therefore, ML has created an impact on cancer treatment by using R coding for patients as it helps in early diagnosis and categorizes subtypes of cancer and provides better treatment to them. In addition to this, it is a highly used tool in the medical industry for cancer treatment and provides accurate and fast results to users.

Decision-tree algorithm is used as classification techniques and supervised variouslearning techniques that prefer facts for improving outcomes. K-nearest neighbors algorithm is another simple supervised learning tool helpful in pattern recognition. Support Vector Machines, Naïve Bayes, and logistic regression are useful tools supported to find the problem of breast cancer. This advanced technology has supported medical practitioners in overcoming the problem of breast cancer by treating it at a prior stage (Arora, Dhawan, and Singh, 2020).

3.2 Data collection methods

Data collection is an appropriate procedure for collecting data and measuring chosen variables. It is a research element in every study incorporating physical and social sciences. The objective is to perform evaluation analysis to respond to research questions. Data gathering aims to seek quantitative evidence that permits assessment to formulate credible outcomes on the subject. Data gathering and validation comprise four steps when considering census, whereas seven stages involve sampling. A researcher can appropriately gather information to understand complex concepts through this process. Different methods are introduced, which supports accumulating information such as interviews, questionnaires or surveys, observations, documents and records, focus groups, historical records, presentations, etc. But all these methods are categorized under two techniques of data gathering such as, primary and secondary. Primary methods direct the researcher to accumulate first-hand data, whereas the secondary technique outlines existing facts and already used information to disclose relevant elements (Elio et al., 2011).

3.3 Data analysis tool

Data assessment signifies a procedure of transforming and modeling information in useful data for using research for understanding complex elements. The main purpose of data analysis is to extract relevant facts from acquired data and make relevant decisions on the basic analysistool. This processing consisted of multiple facets and frameworks, encompassing distinct toolsunder various names and preferred in distinct business, science, or social science regions. Incurrent

business, data analysis aids researchers in making the correct decision using numerous information gathered from many primary and secondary sources (Fernandez- Cavia, 2013). It is a core practice in modern business and selecting the right data analytics tool is very challenging for the researcher because it aids in extracting relevant information from a largebunch. Some of the useful methods of data analysis are described further.

Statistical analysis tools and **textual assessment** are two main categories that consist of various other techniques.

Statistical packages include SPSS, R (Foundation for statistical computing), SAS (Statistical analysis software), GraphPad Prism, Minitab, and many more. On the other hand, a few significanttext analysis tools include, MonkeyLearn, Aylien, IBM Watson, Thematic analysis, Google CloudNLP, Amazon Comprehend, Meaning Cloud, and Lexalytics. All the analysis tools are suitable fortransforming plenty of data smartly and identifying the best information amongst all to conclude the subject. For instance, thematic analysis or analyzing graphs is used in this research to obtain essential information on machine learning and how medical practitioners are using smart information technology in examining cancer problems (Uri, 2015).

3.4 Ethical consideration

Ethics means principles and morals, which create a difference between good or bad. It means moral values need to be followed by researchers while gathering significant information on the subject to control the chance of internal loopholes. Ethical considerations of research are based on right and wrong, and if ethical considerations are absent, then the researcher mightencounter any allegations and negative claims. Thus, the investigator should gather important information on the chosen subject by considering essential ethical norms and elements (Tungprapa, 2015). General Data Protection Regulation (GDPR) has been carefully used and exercised throughout the entire data analysis.

4. DATA ANALYSIS

4.1 Data Visualization

4.1.1 Bar Graph

The Wisconsin Breast Cancer Database (WBCD) was employed in our study for breast cancer prediction. M means that the diagnosis result is malignant and B refers to benign. In addition, ten real-value characteristics are measured for every cell nucleus, including Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, and Fractal

dimension. Bar graphs are used for the process of analysis of data using bar functions. These bars presentdifferent heights. The bar chart is presented in vertical form.

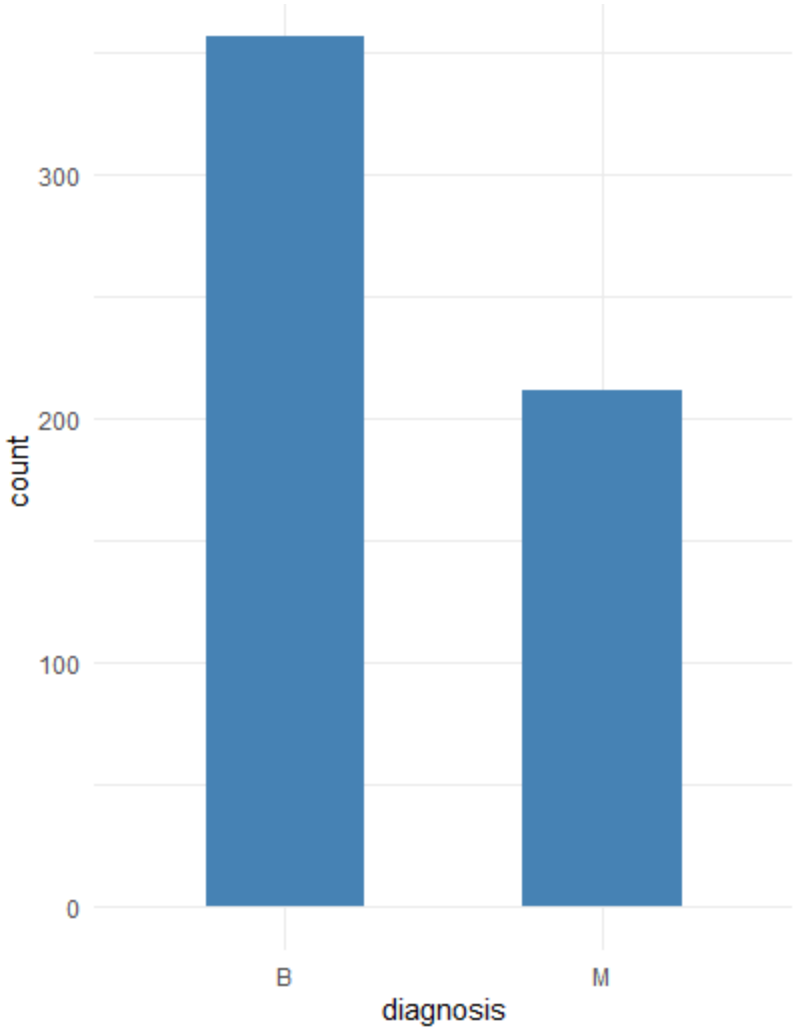

Figure 4: Graph for breast cancer count (Benign vs. Malignant)

The above graph presents a clear picture of the number of breast cancer cases found correct after complete diagnosis. The graph shows that more than 300 individualshad contracted the benign form of the disease. According to the graph, the diagnosis shows more than 200 individuals contracted the disease but in its malignant form. Breast cancer is a very seriousdisease that, if undiagnosed, can have serious implications on the person's health. Breastcancer, if undiagnosed, may cause death in the long run. The benign form of the cancer is derived from the tumor, which occurs in the human breast. The malignant form also occurs eventually from the tumor. However, the basic difference lies in the fact that the tumor, when malignant, starts to spread over the whole of the body. This can damage other cells of the body. This type of tumor causes what is known as cancer. The above graph correlates with the fact that the number of benign cases is far more than the number of malignant cases. This is good since benigncases can be treated more easily. The graph shows that the diagnosis process is generally correct when used with the help of machine learning algorithms. The following breast cancer analysis uses a machine-learning algorithm for the process of diagnosis. The benign form has a difference of more than 100 from the malignant form.

4.1.2 Violin Plots of Breast Cancer

As the name suggests, violin plots take the shape of a violin. These plots were developed through the mixture of both kernel density plots along with the box plot. The data represented by these plots are the same represented by the latter and include the median of the data. The other values represented by the plot include both the upper and lower values of the adjacent and the quintessential interquartile range.

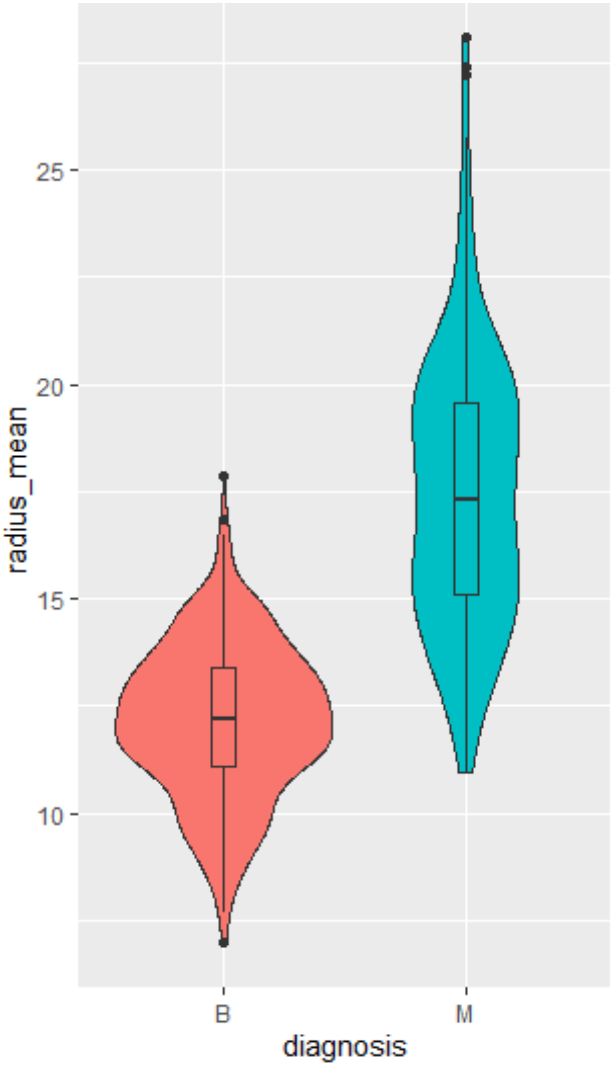

Figure 5: Violin Plot (Radius mean vs. diagnosis)

The above violin plot depicts the radius mean of diagnosed cancer of both forms, i.e., benign and malignant. The benign form of cancer has an equal distribution in terms of its radius. However,the distribution in the case of a malignant form of breast cancer is far more stretched. The radiusesof few values of malignant cancers are more than 22.5. The interquartile range of radius means fora malignant form of the cancer is from 15 to 20 units, whereas for a benign form of the virus, it lies in the range of 11.5 to 13.5 units. In other words, this shows that the benign form of cancer has a much shorter radius than the malignant form. On the other hand, the benign condition had radius values from about 7.5 to 17.5 units. Moreover, in the case of the violin plots of malignant, the curve flattens a bit in between, which shows that the values of radius in the middle portion are a bit less.

However, the curve of distribution is long in the case of a malignant form of breast cancer. Also,the malignant form has more outliers in the data, as represented by the violin plot. The black dots in the violin plot have small black dots, which show the outliers in the data. On the other hand, the benign form has a much thinner distribution before the value of the radius hits ten units, after which the graph expands, showing much more values of radius of the breast cancer tumor.

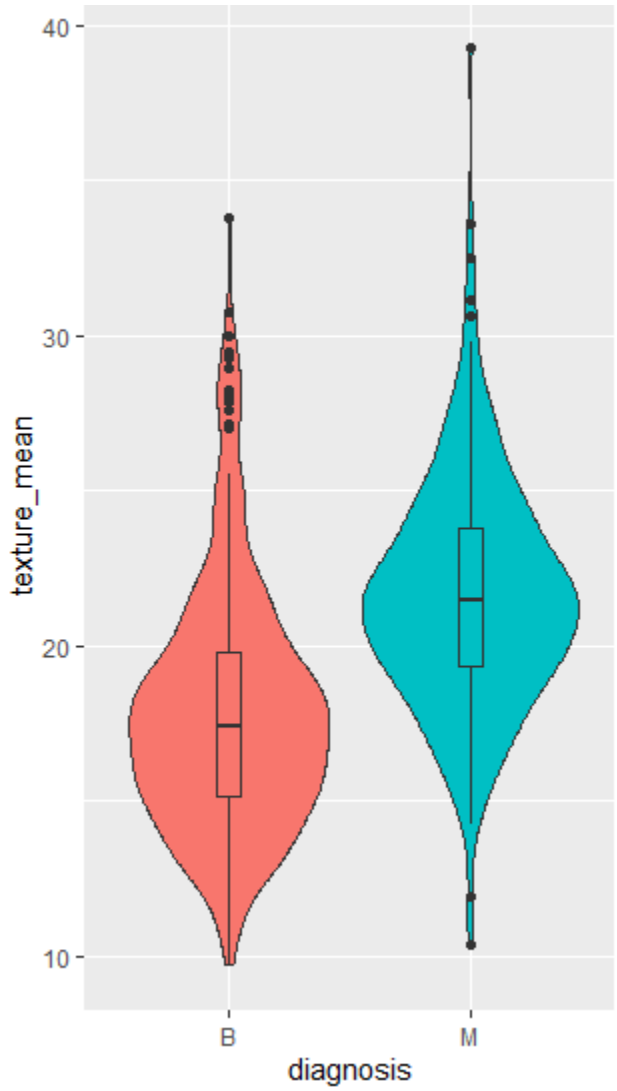

Figure 6: Violin Plot (Texture mean vs. diagnosis)

The above plot is a violin plot of texture and diagnosis of two types of breast cancer: benign and malignant. The benign data of texture starts from 10 and continues till 35. While on the other even though the value of malignant starts from 10 but continues nearly till 40. The interquartile value of a benign type of cancer starts from 15 up to 20, whereas the value of the

interquartile range for the malignant type of breast cancer starts from 20 and continues till nearly 25. The distribution of benign and malignant types of breast cancer shows the nearly same type of distribution, with their ends being slightly wider than the top portion. Thetop of both the benign and malignant plots have outliers near their top portion. Such outliers in thedata have been marked with black dots. This graph is different from other graphs because both the benign and malignant types of cancer show rather the same type of distribution. The distribution value is much similar when compared with the graphs of other attributes. This showsthat texture in breast cancer is quite similar in benign and malignant cases.

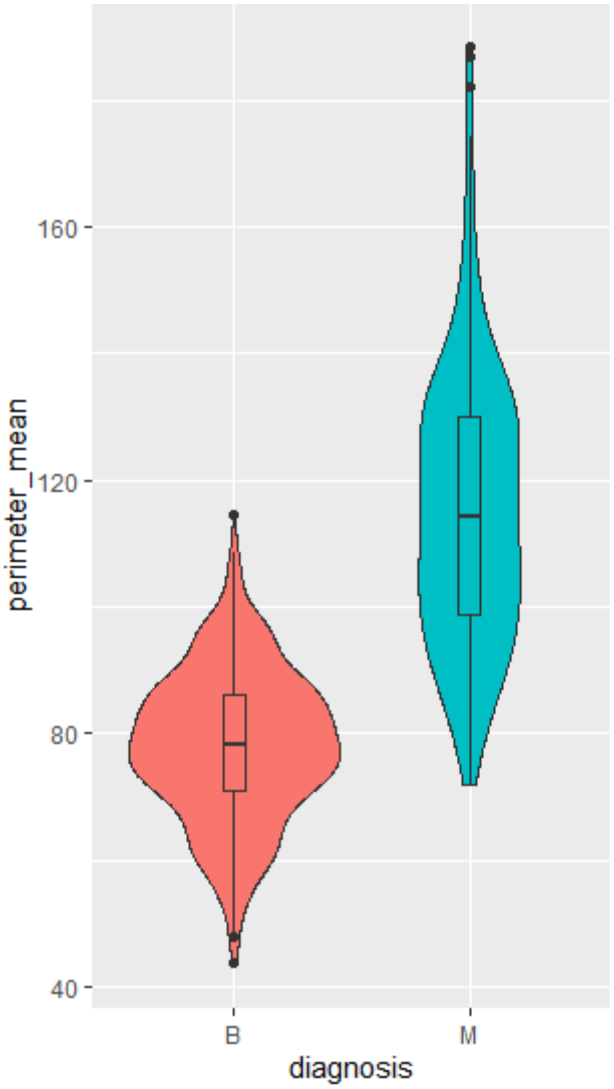

Figure 7: Violin Plot (Perimeter mean vs. diagnosis)

The above violin plot shows the data between perimeter and diagnosis of two types of cancer, i.e., benign and malignant. The above graph shows that the violin plot for a benign type of breast cancer starts from the value of slightly above 40 and continues until slightly below the value of 120. The interquartile range for the plot starts from nearly 70 and continues to 110. The distribution does not have any outliers of data. However, the graph shows a much wider distribution in the middle andgenerally tapers very quickly towards both ends. This shows that the perimeter of benign cancer most frequently has a perimeter of around 80. On the other hand, the malignant form ofcancer shows that the distribution value starts from slightly below 80 and continues until much above 160. The data for the benign type of breast cancer do not have any significant outliers. However, in the case of a malignant form of breast cancer, the data for perimeter shows few outliersabove 160. The interquartile value in the case of a malignant type of breast cancer ranges from 110to 130. The plot for malignant is much flatter and shows a much more equitable distribution throughout the range of values. However, in the case of benign forms of breast cancer, the values show a less equitable distribution and are fatter in the middle section.

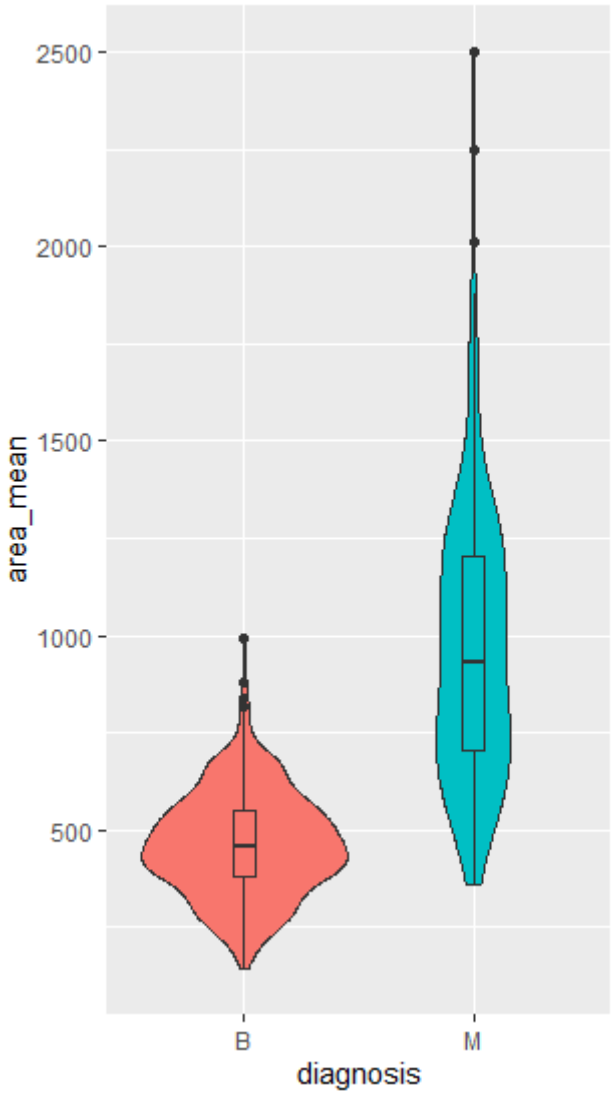

Figure 8: Violin Plot (Area mean vs. diagnosis)

The above plot shows the violin graph between area means and diagnosis of cancer of two types , i.e., benign and malignant. The graph shows that for the type of benign type of breast cancer, the area lies in the range of 0 to 1000. The graph shows a very wide distribution in the middle section near the value of 500. Then the distribution value decreases steeply and ends with a few outliersin the data. The interquartile range is from near the value of 500 and is concentrated near it. In the case of a malignant type of cancer, the violin graph distribution starts from slightly below 500 and continues up to 2500, showing a very steep decrease. The interquartile range for the malignant type of breast cancer is from 750 and continues up to 1250. The graph shows a rather flatter distribution when compared against its benign counterpart. However, the data has many outliers, and the plot is extremely flat above the value of 1500 and continues up to 2500. The outliers lie near the value of 2500.

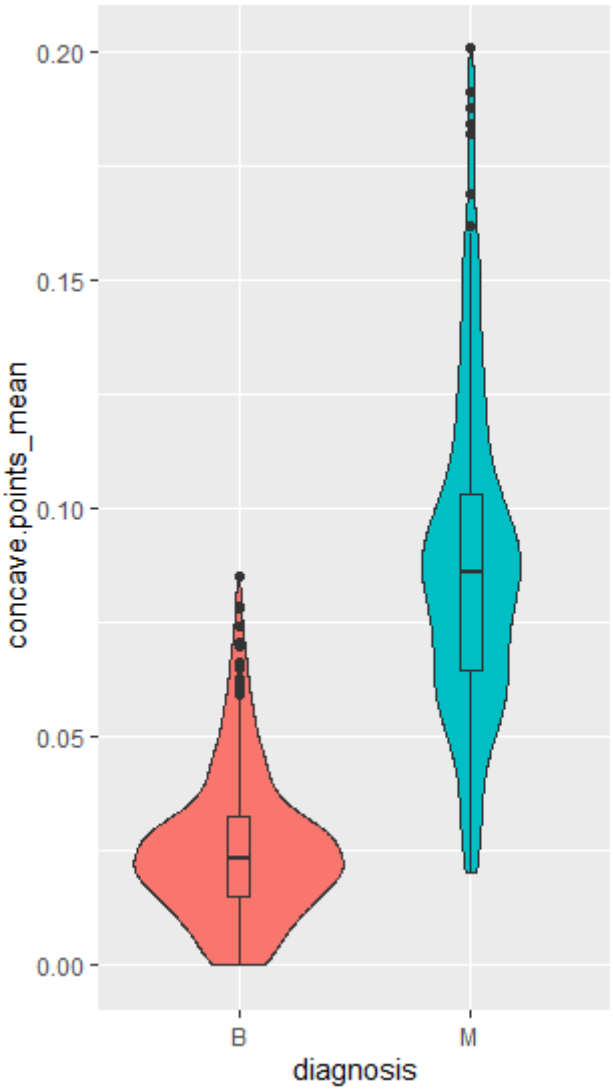

Figure 9: Violin Plot (Concave points mean vs. diagnosis)

The above violin graph shows concave points in the diagnosis of breast cancer. The benign form of cancer or tumor has concave points in the range between 0.00 to nearly 0.10. The interquartile range is concentrated in the very small region between 0.02 and 0.04. The data for benign cancer has much more outliers in the concave range of 0.05 to 0.10. In the case of cancer of malignant type, the distribution of concave points starts from 0.02 and continues up to 0.20. The data distribution here to have many outliers, as pointed by the black dots in the graph. The interquartile for this malignant type of cancer is from slightly above the value of 0.05 and continues to very slightly above the value of 0.10. The distribution is much flatter than that for benign, where the middle shows a very heavy distribution, whereas the distribution here is nearly smooth from 0.05 to the value of 0.10.

4.1.3 Correlation chart

A correlation chart is the graphical representation of the different variables. This is used for determining the correlation between the variables, where correlation refers to a change that occurs in a variable's value when there is a change in the value of a different variable. The correlation value ranges from +1 to -1. This shows that when the value is positive, it offers a positive correlation, i.e., both variables show a change in the same direction. In contrast, a negative correlationshows that both the values work in inverse ways. In contrast, the zero values show no correlation. The correlation graph paints the correlation between variables in shades of three different colors.The blue shade stands for positive correlation, the white shade stands for little correlation, and the red stands for negative correlation, as evidenced by the presented scale.

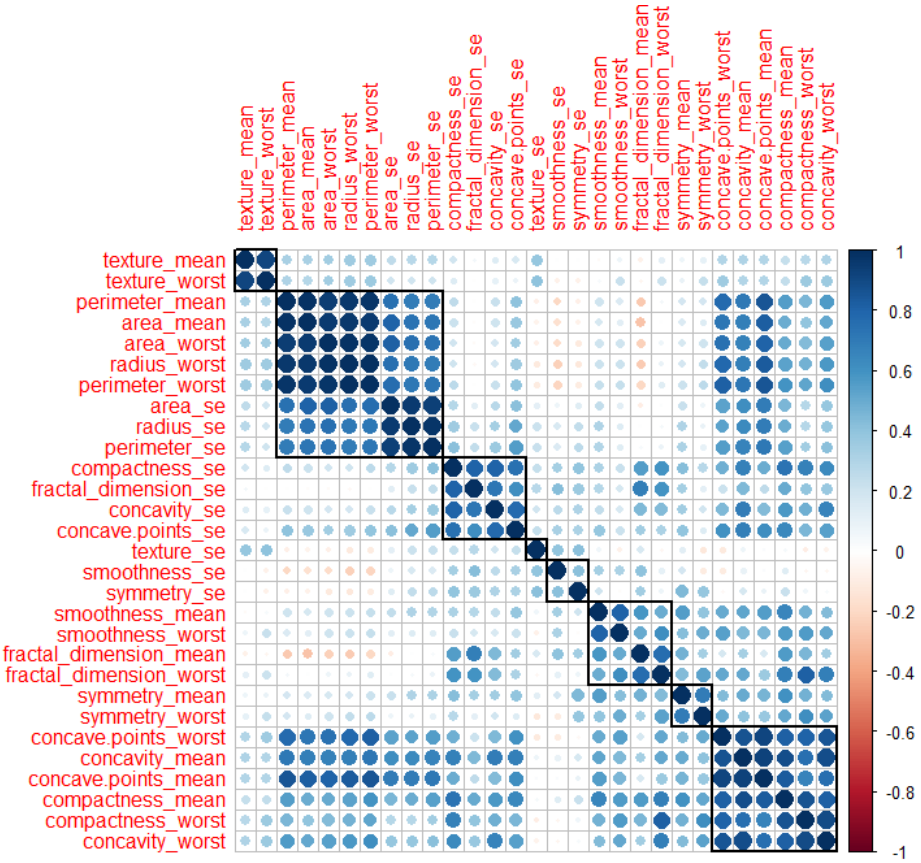

Figure 10: Correlation chart of cancer

The above chart shows that the concavity worst is correlated to texture mean and the correlation value is quite high in this case. The perimeter means and the area means have a very high positive value of correlation. The area also has a high positive correlation with perimeter worst, radius worst and area worst. However, the fractal dimension attribute does not correlate with the area mean attribute and the correlation value stands at zero. The correlation value of area means has a

negative correlation factor with fractal dimension mean.

The radius attribute has a very strong correlation value with area and perimeter with a correlation factor nearing 0.8 to 1. The value of the correlation of radius is in the range of 0.4 to 0.2 for texture mean and texture worst. The correlation value of perimeter means, perimeter worst, areameans and area worst with radius is between 0.4 and 0.6. This shows that these factors have a goodcorrelation with radius. This means that as the breast cancer radius increases, so does the perimeter and area. The fractal dimension means and fractal dimension worst has zero correlation value with radius. The zero-correlation value shows that the radius of cancer does not affect the fractal dimension and fractal dimension worst.

4.2 Results of Machine Learning Algorithms

4.2.1 Random Forest

Random Forest is a ML algorithm that is widely known for its flexibility and simple approach. Classified as a supervised machine learning algorithm, it belongs to a rather special class of algorithm known as ensemble learning algorithms. The algorithm was first developed by a researcher at Bell Labs. Tin Kam Ho was the researcher who initially presented a paper that numerous other scientists and researchers would research in the coming years to develop the algorithm to its current state. That paper was developed by the scholar after initially studying the Random Subspace Method. The algorithm eventually reached its pinnacle when it was further worked upon by Leo Breiman and Adele Cutler (Noshad et al., 2019). The algorithms are known for their ability to deliver fairly good results. The algorithm has been used in almost all the sectors, which depend on machine learning since the configuration requirements of the algorithm are not much required while possessing the ability to fit to operate on almost any data. The algorithm thus earned itself the nickname Blackbox.

Since the algorithm is in the category of ensemble learning that refers to those algorithms, which consist of other algorithms to improve their performance, these algorithms are always considered weak learners. The algorithm in the case of Random Forest (RF) is a decision tree as the random forest is a collection of numerous of the latter. The decision trees are put in subsets. This allows the decision trees in a particular subset to operate independently and arrive at a result. All the results from the decision trees are then taken to the mode of the results used for the final results.

This method is used for classification problems; however, when faced with regression problems, the algorithm uses the mean value to deliver results. Hence, from simple reasoning, it is clear that random forest works better when the number of decision trees is more, thus eliminating the major issue with the decision tree algorithm, i.e., overfitting the model (Magidi et al., 2021). However, the RF algorithm's performance is less good than Gradient boost algorithms, which is another way of enhancing the ability of the weak learner algorithms. The algorithm can mostly handle any dataset even if the data contained in the dataset is huge and still deliver optimal performance. Even though algorithms like SVM, i.e., Support Vector Machine and Neural Networks, are better at dealing with more complicated data, the time taken to train those models is greater than the random forest model. Hence, this remains a top choice for data and machine learning engineers in recent times.

```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 106   5
         M   1  58

               Accuracy : 0.9647
                 95% CI : (0.9248, 0.9869)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9233

 Mcnemar's Test P-Value : 0.2207

            Sensitivity : 0.9206
            Specificity : 0.9907
         Pos Pred Value : 0.9831
         Neg Pred Value : 0.9550
             Prevalence : 0.3706
         Detection Rate : 0.3412
   Detection Prevalence : 0.3471
      Balanced Accuracy : 0.9556

       'Positive' Class : M
```

Figure 11: Random Forest Algorithm Result

The RF algorithm has an accuracy of 96.47%, which shows the model is quite efficient in predicting the results. The model developed presents good results in the classification problem.

4.2.2 Naïve Bayes

The algorithm is considered simple, yet it delivers consistent results. It is a type of supervised algorithm. The machine learning algorithm works on the formula of the Bayes Theorem. The algorithm delivers excellent results when faced with problems of text classification. Since the algorithm is based on Bayes Theorem, the objects in the algorithm are each assigned with a specific

probability. When performing classification problems, the algorithm makes two crucial assumptions that distinguish the algorithm from other machine learning algorithms (Abellán & Castellano, 2017). The features or, in other terms, the variables are considered independent. The other assumption is the equal contribution the variable makes to the result of the problem. However, the algorithm is not any single entity, but many algorithms that follow the BayesTheorem are referred to as the Naïve Bayes algorithm. Due to its independent consideration of features, the algorithm has conferred the name Naïve since there is rarely any situation where theindependent assumption holds (Safri, Arifudin & Muslim, 2018). However, the algorithm has certain advantages over other algorithms, which may be listed as follows: -

- The algorithm is generally used with a dataset consisting of categorical data since the results delivered are better than other algorithms.
- When true for the features, the independence makes the algorithm deliver excellentresults even in the case of the dataset being large and consisting of many features.

There are three types of machine learning algorithms of the Naïve Bayes algorithm, which are listed as follows: -

1. Multinomial: - All the differences in the algorithms are based on the type of the data and in this case, this is multinomial.
2. Bernoulli: -When the variables belong to Boolean type, the algorithms are Bernoulli.
3. Gaussian: -This algorithm is implemented when the algorithm has a normal distribution of data.

The algorithm finds almost exclusive use in the field of text classification and hence is widely implemented in the filtering of spam, sentiment analysis, and many other such cases of text classification.

```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 102  10
         M   5  53

               Accuracy : 0.9118
                 95% CI : (0.8586, 0.9498)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8077

 Mcnemar's Test P-Value : 0.3017

            Sensitivity : 0.8413
            Specificity : 0.9533
         Pos Pred Value : 0.9138
         Neg Pred Value : 0.9107
             Prevalence : 0.3706
         Detection Rate : 0.3118
   Detection Prevalence : 0.3412
      Balanced Accuracy : 0.8973

       'Positive' Class : M
```

Figure 12: Naive Bayes Algorithm Result

The Naïve Bayes algorithm has an accuracy of 91.18%, which is quite a good model.

4.2.2 Decision Tree

The algorithm is known as weak learners since the results delivered by the algorithm are generally above the value of random guess. They come under the category of the supervised machine learning algorithm, which has two sub-types listed as follows: -

1. When the data is of categorical type, the categorical decision tree is utilized. Categorical data generally does not have any value which any form of the naturalorder can sort.
2. When the data is in continuous type, a decision tree is utilized for classification results.

The algorithm is named so because its structure is similar to an inverted tree. The tree starts with a root node which has the entire dataset subsumed in itself. The branching process takes place after that. This is also referred to as the splitting process when a node is divided into more branches. The nodes apart from the root node are of two types: decision and leaf node. In the case of the decision node, the decision-making process is done in this node, while in the case of theleaf node, the final result is delivered.

In the algorithm, some decision tree nodes often need to be removed since these are not required in the result, making the pruning process very crucial since suchnodes are removed from the decision tree (Ramya, Teekaraman & Kumar, 2019). The attribute isselected for making the root which makes it a rather complicated process resolved through the useof methods like entropy gain, Gini index, and other such methods.

```
399 samples
 30 predictor
  2 classes: 'B', 'M'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 399, 399, 399, 399, 399, 399, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.02013423  0.9151113  0.8169275
  0.06040268  0.9081137  0.8017457
  0.79194631  0.8695800  0.6913667

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.02013423.
```

Figure 13: Decision Tree Result

The decision tree presents an accuracy of 81.69% and the results are quite good since the algorithm is a weak learner.

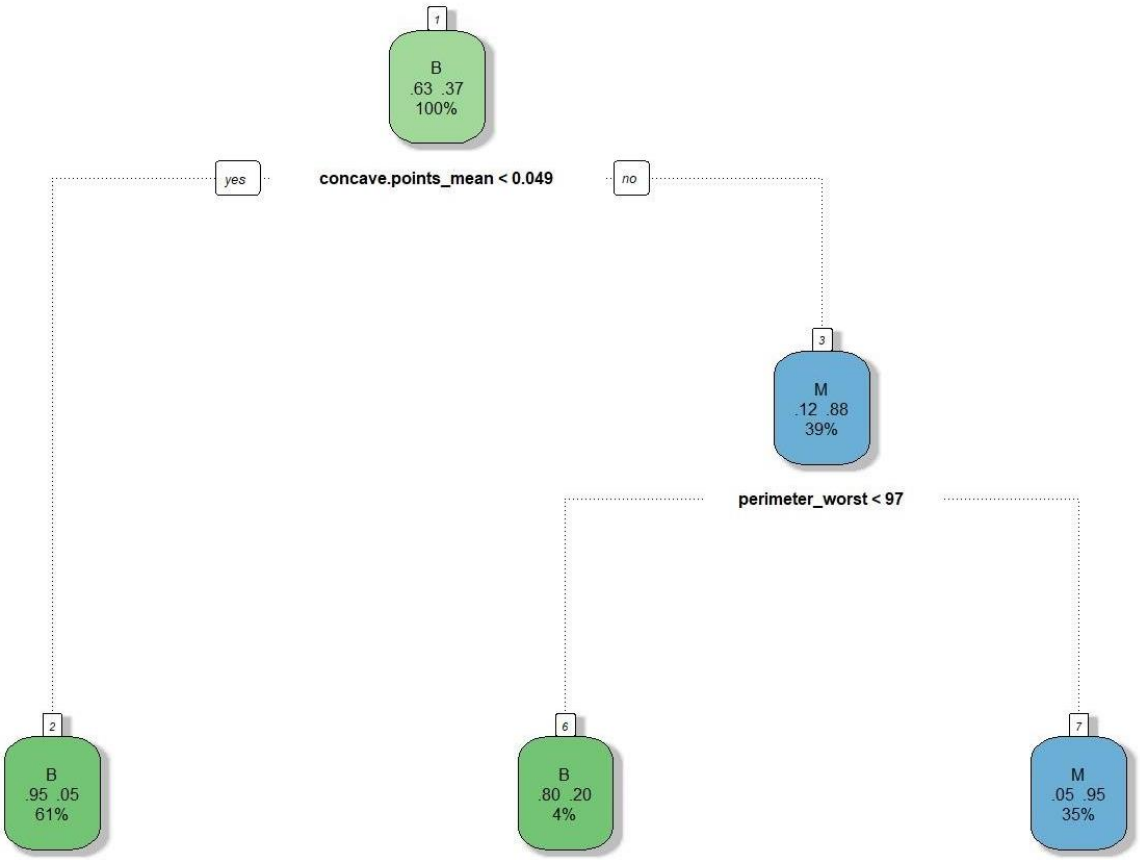

Figure 14: Decision Tree Model

The above is a model of the decision tree algorithm developed for the classificationproblem.

Table 1. Summary of the Algorithm Performance

Algorithm	**Accuracy**
Naïve Bayes Algorithm	91.18%
Random Forest Algorithm	96.47%
Decision Tree	81.69%

From the above table, it can be inferred that the algorithm which delivers the best results is the Random Forest algorithm which shows a stunning accuracy of 96.47%. This is followed by the Naïve Bayes with 91.18%, while the decision tree has an accuracy of only 81.69%, least in the group.

5. DISCUSSION AND FINDINGS

5.1 Discussion

As per the research title, the authors assess the influence of Machine learning on cancer treatment under R technology, which is a foremost element to disclose the important role of modern tools in the healthcare sector. This title consists of one independent variable and two dependent variablesfor disclosing its relationships. Machine learning is considered an independent variable in the designed title, whereas cancer treatment and R technique are two dependent variables. Thus, the main discussion aspects of this research are machine learning, specifically in the healthcare industry. Instead of outlining the entire medical sector, cancer treatment is the main focus of the study. The effectiveness of advanced technology in overcoming cancer problems is investigated. For example, the chemotherapy technique is successfully implemented by medical practitioners in the healthcare department with the help of machine learning to reduce the fast-growing cells from a body for overcoming cancer problems. Machine learning plays a crucial role in diagnosis processing, identifying internal issues ofthe human body, and many more.

By investigating the application of machine learning algorithms in the healthcare industry through literature review and conducting the breast cancer prediction through the use of several classical machine learning algorithms, this paper disclosed numerous findings. First of all, machine learning has gained the attentionof every medical practitioner by facilitating experts in diverse ways. Although cancer treatment isthe main focus of research, data has revealed the importance and dependency of the healthcare industry on machine learning or IT technology. For example, machine learning algorithms can detect patterns linked with life-taking disease and dangerous health status by studying several records of healthcare and patient facts. In fact, current improvement in machine learning has maximized the access of the medical industry by innovating cancer diagnosisand treatment alternatives. After knowing this, various domestic, national, and international institutions have invested heavily in making digitalized setups for treating cancer patientsand facilitating patients with smart techniques.

In addition, this research evaluated the effectiveness of machine learning techniques in

cancer prediction. To be specific, we employed three classical ML algorithms, Naïve Bayes, Random Forest, and Decision Tree, on the Wisconsin Breast Cancer Dataset to predict breast cancer. The comparison result shows that they all perform effectively in prediction as their accuracy values are greater than 80% and Random Forest performs best with an accuracy of 96.47%. The results indicate that machine learning is a beneficial tool in diagnosing cancer problems.

6. CONCLUSION AND RECOMMENDATION

6.1 Conclusion

From the above research, the effectiveness of machine learning gets disclosed, especially in enhancing cancer treatment procedures by focusing on distinct examples. It is understood that artificial intelligence is identified everywhere, such as personal digital assistants answering or questioning, robot-advisors trade, driverless cars, and many more. In fact, AI has penetrated the overall lives of human beings. Its utilization has exploded the biomedical investigation and health care incorporating over every dimension of cancer investigation wherever AI applications are very vast. Machine learning is known as computers performing tasks that are completely linked with human intelligence because humans are responsible for coding and programming to make programs work smartly. There is an algorithm or model known as code that direct computer for their actions, reason, and showing learning. In summary, machine learning is a category of AI that is clearly programmed to perform a particular activity and learn iteratively to make further forecasting or judgments. More data is exposed to Machine learning, which improves its performance more or more. Deep learning is another subset of machine learning that utilizes an artificial neural network model for processing the human brain and its learning from an excessive number of facts.

All the above data reveals, around 14 million new patients with cancer are diagnosed by pathologists annually across the globe. Several pathologists have performed cancer diagnoses and prognoses for decades. In this, a maximum pathologist has attained nearly 96 to 98% of success rate while diagnosing cancer. Machine learning is working too fast than humans, such as a biopsy that generally consumes pathologists at least ten days. In contrast, a computer with machine learning algorithms can have the capability to complete thousands of biopsies in just mini seconds. It means that machines can accomplish anything which humans and experts do not easily do can repeat thousands of times without getting tired. After thousands of iterations or more, they also let the machine repeat a similar procedure for doing it better. Another usefulness of machine learning

in the healthcare sector is its great preciseness while treating cancer issues. For example, breast cancer is also a dangerous health issue that easily identifies a specialist via ultrasound. In this ultrasound, the machine utilizes sound waves to make detailed pictures known as sonograms that cover distinct areas of the breast.

6.2 Potential research contribution

Researchers disclose MammoScreen, an AI machine learning tool designedto determine several regions that are suspicious of breast cancer on 2D digital mammograms. This process also identified probabilities of malignancy. Research also found that this system can produce many picture positions with distinctscores for doubt of malignancy,which is mined from four perspectives of a standard mammogram.Hence, complete data of artificial intelligence had shown how much machine learning of AI had improved efficiencies of healthcare activities and directed humans or medical experts to do things smartly. When AI is consuming repetitive or dangerous activities, it frees up the workforce team from work pressure. Instead of this, researchers also tried a lot to support or enhance computers to predict future activities that might happen next. For instance,AI is guessing the future or predicting technologies without learning anything or having in-depth knowledge.

Typically, numerous empirical studies are conducted by a researcher to show the appropriateness of machine learning in breast cancer and how it immediately gets treated by soft computing techniques or tools. Although breast cancer is a major reason behind death, it isone type of cancer that increases broadly across women. Thus, many imaging tools have been designed for primary detection and treatment to minimize death levels due to breast cancer.In a nutshell, a complete thesis has proven the benefits of machine learning or IT technologies in solving the problem of cancer. This improved the healthcare sector significantly and provided many growth opportunities to many medical practitioners.

6.3 Recommendation

Throughout the discussion and analysis, it is disclosed that the main objective of the research is to focus on cancer treatment via machine learning instead of outlining any other health issues. This research is too narrow with a specified area of the healthcare industry. There are still many options for a researcher to continue further research on a similar topicbut cover a broad area. According to the title, machine learning focuses on cancer treatment, but now there is an option to focus on other health problems. Apart from this, a researcher can use any specific technology like the role of cryptocurrency or encryption keys in protecting confidential patient data, biometric in

scanning humans for enhancing data privacy, andmany more. If a researcher wants to cover a broad area, then there is an option to focus on machine learning in improving the healthcare industry completely. Under this topic, researchers get an opportunity to outline the wider use of machine learning in the healthcare sector,especially by medical practitioners. These alternatives are creative and exciting elements that can reveal multiple machine learning capabilities using suitable examples and related cases.

Conflict-of-Interest Statement

Authors have no conflict of interests with anyone and any organization.

Funding information

This work is partly supported by VC Research (VCR 0000159) for Prof Chang and any overheads

REFERENCES

Anoy Chowdhury, 2020. Breast Cancer Detection and Prediction using Machine Learning. Available at< https://www.researchgate.net/publication/342303246_Breast_Cancer_Detection_and_Prediction_using_Machine_Learning >.

Arora, M., Dhawan, S. and Singh, K., 2020. Data Driven Prognosis of Cervical Cancer Using Class Balancing and Machine Learning Techniques. *EAI Endorsed Transactions on Energy Web*, *7*(30), p.e2.

Bibault, J.E., Giraud, P. and Burgun, A., 2016. Big data and machine learning in radiation oncology: state of the art and future prospects. *Cancer letters*, *382*(1), pp.110-117.

Chaplot, N., Dhyani, P. and Rishi, OP, 2013. A Review on Machine Learning concepts for prediction based application. *International Journal of Computational Science, Engineering & Technology, 1*(2), pp.12-18.

De Silva, D., Ranasinghe, W., Bandaragoda, T., Adikari, A., Mills, N., Iddamalgoda, L., Alahakoon, D., Lawrentschuk, N., Persad, R., Osipov, E. and Gray, R., 2018. Machine learning to support social media empowered patients in cancer care and cancer treatment decisions. *PloS one*. 13(10). p.e0205855.

Deist, T. M., Dankers, F. J., Valdes, G., Wijsman, R., Hsu, I. C., Oberije, C., Lustberg, T., van Soest, J., Hoebers, F., Jochems, A. and El Naqa, I., 2018. Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers. *Medical physics*. 45(7). pp.3449-3459.

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., ... & Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. International journal of cancer, 144(8), 1941-1953.

Jiang, F and et. Al., 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, *2*(4).

Koutsouleris, N and et. Al., 2018. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: a multimodal, multisite machine learning analysis. *JAMA psychiatry*, *75*(11), pp.1156-1172.

Kumar, G.R., Ramachandra, G.A. and Nagamani, K., 2013. An efficient prediction of breast cancer data using data mining techniques. *International Journal of Innovations in Engineering and Technology (IJIET)*, *2*(4), p.139.

Li, Y. and Chen, Z., 2018. Performance evaluation of machine learning methods for breast cancer prediction. *Appl Comput Math*. 7(4). pp.212-216.

Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020, July). Analysis of breast cancer detection using different machine learning techniques. In *International Conference on Data Mining and Big Data* (pp. 108-117). Springer, Singapore.

Noor, M. M., & Narwal, V. (2017). Machine learning approaches in cancer detection and diagnosis: mini review. IJ Mutil Re App St, 1(1), 1-8.

National Cancer Institute, 2019. R-CHOP. Available at< https://www.cancer.gov/about-cancer/treatment/drugs/r-chop>.

Redlich, R and et. Al., 2016. Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. *JAMA psychiatry*, *73*(6), pp.557-564.

Shouval, R and et. Al., 2021. Machine learning and artificial intelligence in haematology. *British journal of haematology*, *192*(2), pp.239-250.

Swan, A.L and et. Al., 2013. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics: a journal of integrativebiology*, *17*(12), pp.595-610.

Tseng, W.T and et. Al., 2015. The application of data mining techniques to oral cancer prognosis. *Journal of medical systems*, *39*(5), pp.1-7.

Wang, P., Li, Y. and Reddy, C.K., 2019. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, *51*(6), pp.1-36.

Yuefeng Zhang, 2020. Deep Learning in Wisconsin Breast Cancer Diagnosis. Available at< https://towardsdatascience.com/deep-learning-in-winonsin-breast-cancer-diagnosis-6bab13838abd>.

Zhang, B and et. Al., 2017. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer letters*, *403*, pp.21-27.

Zhou, M and et. Al., 2018. Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. *American Journal of Neuroradiology*, *39*(2), pp.208-216.

Emami Sigaroodi, A and et. Al., 2012. Qualitative research methodology: phenomenology. *Journal of Holistic Nursing and Midwifery*, *22*(2), pp.56-63.

Pickerd, J and et. Al., 2011. Individual accounting faculty research rankings by topical area and methodology. *Issues in Accounting Education*, *26*(3), pp.471-505.

Fernandez-Cavia, J., 2013. Destination brands and website evaluation: A research methodology.

Uri, T., 2015. The Strengths and Limitations of Using Situational Analysis Grounded Theory as Research Methodology. *Journal of Ethnographic & Qualitative Research*, *10*(2).

Tungprapa, T., 2015. Effect of using the electronic mind map in the educational research methodology course for Master-degree students in the faculty of education. *International Journal of Information and Education Technology*, *5*(11), p.803.

Abellán, J., & Castellano, J. G. (2017). Improving the Naive Bayes classifier via a quick variable selection method using maximum of entropy. *Entropy*, *19*(6), 247.

Magidi, J., Nhamo, L., Mpandeli, S., Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28. Mabhaudhi, T. (2021). Application of the Random Forest Classifier to Map Irrigated Areas Using Google Earth Engine. *Remote Sensing*, *13*(5), 876.

Noshad, Z., Javaid, N., Saba, T., Wadud, Z., Saleem, M. Q., Alzahrani, M. E., & Sheta, O. E. (2019). Fault detection in wireless sensor networks through the random forest classifier. *Sensors*, *19*(7), 1568.

Ramya, K., Teekaraman, Y., & Kumar, K. R. (2019). Fuzzy-Based Energy Management System

With Decision Tree Algorithm for Power Security System. *International Journal of Computational Intelligence Systems*, *12*(2), 1173-1178.

Safri, Y. F., Arifudin, R., & Muslim, M. A. (2018). K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor. *Sci. J. Informatics*, *5*(1), 18.