Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Breast Cancer and Prostate Cancer Detection using Classification Algorithms

Priyanshi Agrawal, Sharmista Deb,
Shilpa A V, Shirisha N Raju
Dept. of Computer Science,
Sir M. Visvesvaraya Institute of Technology,
VTU, Bengaluru, India

Sreenivasa B C
Associate Professor,
Dept. of Computer Science,
Sir M. Visvesvaraya Institute of Technology,
VTU, Bengaluru, India,

***Abstract* - It is a known fact that, cancer rates have increased to great heights in the recent times. The only way to completely cure cancer is to detect its presence at an early stage, for which appropriate diagnosis is available. The main plot of this paper is the detection, and classification of such cancerous cells in the patient's genome expression, on detection of which he/she can be provided rightful treatment. Modern day techniques have evolved that can help to detect the presence of cancer, such as Deep Learning, Artificial Neural Networks, Deep Convolution Networks, and Data Mining etc. In this paper we have dealt mainly with two types of cancer. Breast cancer and Prostate cancer in females and males respectively. We have implemented Machine Learning to find out signs of cancer, and what type of cancer, if seen. The reason being, physicians are capable of diagnosing a patient with cancer, with an accuracy of 71%, according to a latest research, on the other hand Machine Learning techniques can show up to 91% accuracy for rightful classification. Since, the primary focus is to detect and classify the type of cancer in the patient, we have used classification techniques/algorithms under Machine Learning such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) and Naïve Bayes (NB). Effort has been made, to identify the best techniques providing the highest accuracy for both the cancers, and enhancing them with stratified K-fold and dimensionality reduction.**

Keywords: Machine Learning, Classification algorithms, Cancer, Prediction and classification, Naïve Bayes, SVM, Logistic Regression, Random Forest, Decision Tree.

I. INTRODUCTION

Breast cancer affects about 10% of all women, at some or the other stages of their lives, making it the most common cancer, among women. Fig 1, represents the cases and rates of Breast cancer in women based on age. Based on the growth patterns, the expression of oestrogen, progesterone, human epidermal and growth factor receptor, including Ki-67 proliferation index, we can classify the invasive breast cancer into a heterogenous category of disease. The statistics portrays that, the survival rates of breast cancer after 5 years post diagnosis is 88%, and after 10 years post diagnosis is 80%. Breast cancer in men is notably very rare, which accounts to less than 1% of the total breast cancer cases. On the positive side, the early prediction of breast cancer, can increase the survival rate and hence, it has turned out to be the most important step, in this process.

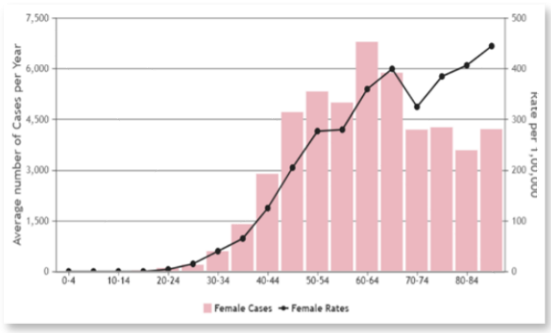

Fig 1: Cases and rates of Breast Cancer in women, based on age

Besides, Prostate cancer affects about 7%, close to 1.3 million men across the globe. Fig 2, represents the rates of Prostate Cancer based on country. Prostate cancer is the next significant cancer in men, after skin cancer, caused in the prostate gland which may lead to death. The rates of prostate cancer have been observed to increase with factors such as, age, sex hormones and steroid hormones. Can Prostate cancer be detected early? Yes, Screening can be used to find out cancers at an early stage, which can help the medical practitioners, to provide a better treatment.

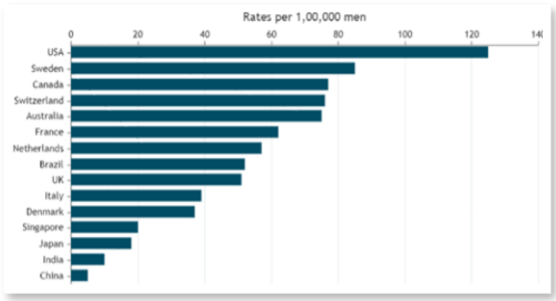

Fig 2: Rates of Prostate Cancer, based on country

Manual techniques for prediction and treatment of the same exists, but Machine Learning can act as a helping hand thereby reducing the number of wrong predictions, i.e., false positive (FP) and false negative (FN) decisions in the generated confusion matrix. Consequently, new methods such as Deep Learning, Artificial Neural Networks, Deep Convolution Networks, and Data Mining have become popular research tools for medical researchers who are trying to predict the outcome of a disease with the help of datasets, by identifying and exploiting patterns and interrelations among various huge number of attributes. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

use of machine learning in medicine is nowadays becoming more and more important, researchers are now using Machine Learning (ML) in applications in ECG analysis and cancer detection.

In the decision-making process of medical practitioners, machine learning and deep learning approaches, can prove out to be of great assistance. An unfortunate increase in breast cancer and prostate cancer diagnosis cases is accompanied with massive datasets, which are notably very relevant to carry forward the ongoing medical practices and researches. It is indeed much more significant to the application of machine learning and deep learning as mentioned above. Precedent studies have witnessed how important the same research topic is, where they solemnly declared how the machine learning algorithms have been used for the detection, classification of breast cancer and prostate cancer, eventually showing significant results.

This paper implements an approach for analysing early cancer detection and prediction with the machine learning techniques and cancer specific knowledge to retrieve conclusions about gene patterns. Among the many supervised machine learning classification algorithms, upon analysis five of them were chosen. Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest Classifier (RF), Naïve Bayes (NB) were developed and evaluated based on their performances. Stratified K-fold and dimensionality reduction are applied on all the above-mentioned models, to enhance the accuracy, and it was observed that the accuracy of most of the models, was actually improved.

II. MOTIVATION

Cancer is a disease that rapidly spreads all through the patient's body and not give a single sign of its presence. This leads to many deaths, due to treatment un-available at that stage. Breast cancer develops on the breast tissues. In America, after Skin cancer, the second most common cancer seen in women is breast cancer. Prostate cancer is the second most common cancers found in men, in USA and worldwide. Both Breast cancer and Prostate cancer have definitely turned out to be a problem, and a topic to think about. So, this motivated a number of researchers to study about the causes of cancer, and how to nullify its presence. Our effort will be a contribution in the domain of medicine, with respect to detection and classification of cancer in the patient more efficiently. As already mentioned, among the modern technologies, machine learning does provide a significantly high rate of accuracy in the classification process. Thus, this does justice to the topic chosen and the implementation techniques.

III. RELATED WORK

Prediction of Breast Cancer with the help of various machine learning algorithms, more precisely classification models and the analysis of their performances:

The paper depicts the importance of data-mining techniques, how data-mining can extract appropriate insights from raw data available, and their useful application in the field of medical science. The prediction domain has been worked upon. It is further stated that, robust data-analytical techniques using distinct data-mining approaches such as Decision tree, K-star algorithm, Bayes theorem and decision table were used, to predict the risk factor the patient. Comparison, had been done among the various models applied, and as to which model provides the maximum accuracy. The maximum efficiency obtained was observed to be approximately close to ~ 76.23% [1].

An application of machine learning classification models on Wisconsin diagnostic dataset for breast cancer prediction:

The paper states that a comparison of various classification algorithms such as Multilayer Perceptron (MLP), Support Vector Machine (SVM), SoftMax Regression, Linear Regression and K-nearest neighbour (KNN) had been conducted, on the Wisconsin diagnostic dataset. The paper describes the dataset, says that it consists of digitized images from the FNA tests of the breast masses. The paper also describes the partition of the dataset into 70% and 30%, classified as training and testing set respectively. The sensitivity and specificity of all the classification algorithms used were found and recorded. It was concluded that Multilayer Perceptron stood high with an accuracy close to ~ 99.04% [2].

Applications of Machine Learning in cancer prediction and related prospects:

The domain of this paper is focused on Machine Learning. The importance of early cancer detection and classification of the detected cancer into high and low risk groups had been expressed. The basic ability of a machine learning model is to detect the key features from complex datasets, and this motivated the authors of the paper to experiment with the various supervised techniques, more specifically classification algorithms such as, Logistic Regression (LR), Support Vector Machines (SVM), Decision tree (DT), Artificial Neural Networks (ANN) and Random Forest (RF). Among the accuracies that were recorded, it was observed that Support Vector Machine resulted maximum with ~89%. This paper also discusses about a survey, an extensive search that was relevant to the use of Machine Learning technologies, in the prediction of susceptibility of cancer, its recurrence and chances of survivability [3].

Study of Breast Cancer: How can machine learning classification models be used for cancer outlook?

The main objective of this paper is to predict accurately, the survival rate/time of the patient suffering from breast cancer. Two datasets, from the breast cancer studies were united with the help of data integration approach. Apart from data normalization, applied machine learning techniques such as Support Vector Machine (SVM), Kernel Ridge Regression, Lasso Regression, Decision Tree Regression (DT), K-Neighbourhood Regression, were also applied. These algorithms were chosen because, they achieved the most accurate results. Furthermore, the authors mention the scope of the project, by discussing about a python workflow developed as a support [4].

Use of machine learning techniques for predicting the breast cancer recurrence:

This paper expresses the concern about the number and size of most of the medical databases, and how a large amount of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

data is not analysed, for finding valuable knowledge. It states that the advanced data mining techniques can be used to analyse this data. The main purpose of this paper, is to make use of data mining as it is capable of discovering hidden patterns and relationships, in a way yielding higher accuracies alongside highly accurate results. The models deployed were Support vector machine (SVM), Decision Tree (DT) and Artificial Neural Network (ANN), out of which SVM stood out with approximately 95.09% accuracy and the least error rate. The 10-fold cross validation resulted in the observed accuracy [5].

Magnetic Resonance Imaging (MRI): An application of Machine Learning in prostate cancer prediction:

The authors of this paper have aimed at providing a synopsis on the possible applications of machine learning in the field of radiology, and screening mainly focused on the prostate cancer magnetic resonance imaging. Furthermore, the difference between the deep learning pipeline and the machine learning pipeline has also been highlighted. Various potential clinical applications are outlined and the results were promising too, but this applicability still needs better and robust validation. However, Machine learning is justified to have huge potential in improving the diagnosis and the best accuracy obtained was observed to be ~ 92.2% [6].

Use of Machine Learning classification models to predict invasive prostate cancer:

The authors of this paper have segmented the data set taken from the National Cancer Institute, into patients without cancer, with non-invasive i.e. low risk cancer and patients with invasive i.e. high-risk cancer. The current screening of the prostate cancer at that time based on prostate-specific antigen (PSA), gave both false positive and false negative cases which had negative impact on the result obtained, but they applied Machine learning algorithms to detect the presence of the invasive prostate cancer, and the proposed machine learning pipeline achieved an accuracy of ~91.5%. They concluded that machine learning when combined with feature engineering and various other advanced features can enhance prediction models to a greater extent [7].

Role of Machine Learning and Artificial Intelligence in Prostate cancer detection:

This paper puts light on Artificial Intelligence and how it can help to achieve a particular goal according to the data provided. It mentions that AI has characteristics to reshape our health care systems. It specifies a computerized decision-aided system which is based on the domain of machine-learning that has the capability of improving the diagnostic accuracy. ML has been applied, to perform low-level image analysis. The ML algorithms applied, helped in improving the prostate cancer treatment [8].

Regression concept vectors for bi-directional explanations in histopathology:

In this paper the thought of expressing deep neural networks predictions, in the field of medicine has been highlighted and how they can be very valuable in the field of medical sciences. A methodology named Regression Concept Vectors (RCVs) has been proposed in this paper. When applied to cancer, it has emerged out to play a significant role in the detection of the tumor tissue in the breast tissue samples. They have performed statistical analysis to evaluate robustness score and the consistency of the model. They thus provided a proof of this conception on the breast cancer dataset and also proved that RCVs can have many extended applications in various domains [9].

Screening Mammography: Improving Breast Cancer Detection with the help of Deep Learning Methods:

This paper has expressed how deep learning has come off spurring more engrossment in its application to medicine and its scanning problems using imaging tests. The author has stated the use of end-to-end training approach that helps in leveraging the training datasets efficiently. This approach requires lesion annotations in the beginning of the training phase itself and the labels on image-levels are required in the subsequent phases which eliminates the reliance on the lesion annotations those are rarely available. They used the four-model averaging that improved the accuracy to a notable 0.98, with specificity – 96.1% and sensitivity – 86.7%. They thus concluded that deep learning gives very promising results and reduces the number of false positives and false negatives [10].

IV. PROPOSED SYSTEM

1. Existing System

The existing system of cancer detection, include practices such as physical tests conducted by a doctor, laboratory tests performed, scanning or imaging tests and biopsy.

- **Personal examination of patient carried out by a doctor**: This physical examination by the doctor includes checking different areas of the body to feel the presence of unwanted lumps or tissues which may be part of a tumour.

- **Performing Laboratory Tests**: Various laboratory tests can be performed which might indicate the presence of any abnormalities or unwanted growth of tissues due to cancer. Such tests can be blood test, urine test etc.
- **Scanning and Testing**: The scanning or imaging tests help scrutinize the scanned parts of the patient body in order to detect some irregularities if present. Such a test can be performed on bones or other internal organs of the human body in a non-intrusive manner.

- **Biopsy**: In a biopsy a sample of cells or tissues are collected in order to uncover the presence of some abnormal growth or distortion of the same which might be the cause of a disease.

2. Proposed System

The main objective of our project is to use the concepts of machine learning to make predictions, on which type of cancer he/she has. Looking through a doctor's perspective, the first step towards the treatment is to detect the type of cancer the patient has, so that the right diagnosis can be provided. So, our main propaganda is to deal with two different categories of cancer in both male and female.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

This project has been developed on the below mentioned classification algorithms. Accuracies of each algorithm, has been recorded and the best among all is used for testing the cancer samples.

- Support Vector Machine
- Decision Tree
- Logistic Regression
- Random Forest
- Naïve Bayes

The implementation consists of 2 major parts.

2.1 Testing of cancer, using classification algorithms

Depending upon the patient, one can select the type of test that needs to be performed. The following datasets were used for prostate cancer and breast cancer respectively.

Classes	2
Samples per data	62(M), 38(B)
Samples total	100
Dimensionality	10
Features	Real, Positive

Table 1: Data-set Specifications for Prostate Cancer

Classes	2
Samples per data	210(M), 357(B)
Samples total	567
Dimensionality	32
Features	Real, Positive

Table 2: Data-set Specifications for Breast Cancer

An electronic report is generated consisting of a detailed overview of the dimensions of cancer cells, if found and the result of the test alongside.

2.2 Monitoring of the Accuracies

Pandas is used, to import the .csv file of the above-mentioned datasets. The data frame has been split into two different datasets: one comprising only the data to be trained and tested and one comprising of the class labels for the respective samples. Accuracies of all the above-mentioned classification algorithms, have been recorded, alongside the application of stratified K-Fold to uniformly distribute the data into trained and the test set. Dimensionality reduction, also has been applied on the classification algorithms, to remove the redundant and irrelevant features, so as to increase the accuracy of the models.

2.3 System Architecture

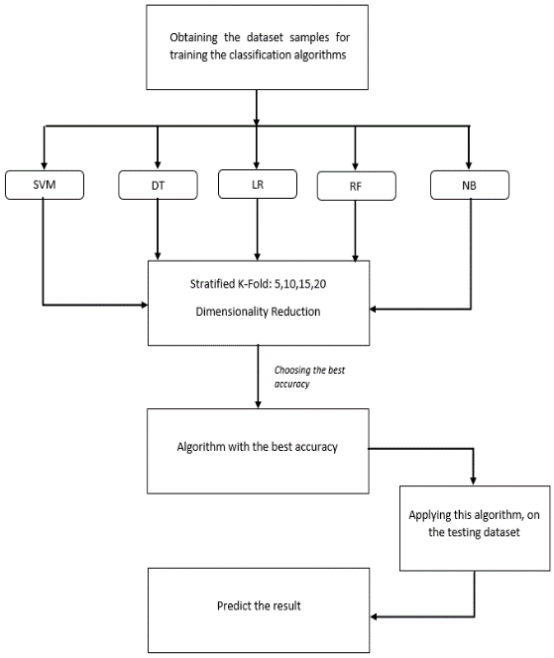

Fig 3: Architectural outline of the classification of cancer

Several steps are maintained to build the classification model that analyse and predict the target class.

The data set samples were obtained, to train the classification algorithms: Support Vector Machine, Decision Tree, Logistic Regression, Random Forest and Naïve Bayes. Stratified K-Fold between the ranges 0-20, with step size being 5, was applied on the above-mentioned algorithms. Dimensionality Reduction was also implemented, to improve the accuracy. The classification model with the best accuracy was chosen and on which the data set was trained. Thus, testing was done on the trained model to predict the cancer class.

V. METHODOLOGY

The system is built atop Python3, and has 3 main steps:

1. Collection of Data Samples

The foremost step, is to collect the data samples. It is a known fact that machine learning models are trained upon the data samples i.e. the examples for the same attributes. Therefore, having a good clench of the sample data is very essential.

A data set can be divided into two subsets:

- Training Dataset – It is a batch of data samples which is used to train the model.
- Test Dataset – It is a batch of data samples which is used to test the trained model.

A qualified test dataset is bound to fulfil the following two conditions:

- It must be sufficiently large enough to capitulate statistically meaningful results.
- It should contain all the important features of the dataset as a whole i.e. it should prototypically represent the data. Simply put, the test dataset

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

should retain the characteristics of its training dataset.

The main objective is to generate a model that extrapolates quite well to new data samples, provided that the above two conditions hold true. Precisely, this test dataset acts as a substitute for new data. The datasets, used in this implementation, is described further in the sub-sections.

1.1 Prostate Cancer Dataset

This dataset was taken from Kaggle and it contains the data of 100 patients for implementing the machine learning algorithms and thereby detecting and predicting results. This data set constitutes of 100 observations and 10 attributes, out of which 8 of them are numerical attributes, one of them is a categorical attribute representing the class of cancer and other being ID, uniquely representing each patient. The 8 numerical attributes are tabulated below (Table 3) along with their ranges.

Attribute	Range		Attribute	Range	
	Min	Max		Min	Max
Radius	9	25	Smoothness	0.07	0.14
Texture	11	27	Compactness	0.04	0.34
Perimeter	52	172	Symmetry	0.14	0.3
Area	202	1878	Fractal Dimension	0.05	0.1

Table 3: Representation of Numerical Attributes of Prostate Cancer Dataset

1.2 Breast Cancer Dataset

This dataset was taken from Kaggle and it contains the data of 567 patients for implementing the machine learning algorithms and thereby detecting and predicting results. The features or say attributes in the dataset are enumerated from a computerized digital image of a fine needle aspirate (FNA) of a breast mass. These features give a description of the cell nuclei present in the image.

This data set constitutes of 567 observations and 10 attributes, out of which 8 of them are numerical real-valued attributes, one of them is a categorical attribute representing the class of cancer and other being ID, uniquely representing each patient. The 8 numerical real-valued attributes are tabulated below (Table 4) along with their ranges.

Attribute	Range		Attribute	Range	
	Min	Max		Min	Max
Radius	6.98	28.1	Smoothness	0.05	0.16
Texture	9.71	39.3	Compactness	0.02	0.35
Perimeter	43.8	189	Symmetry	0.11	0.3
Area	144	2500	Fractal Dimension	0.05	0.1

Table 4: Representation of Numerical Attributes of Breast Cancer Dataset

2. Building Classification Models

Classification is a machine learning method that is used to group the input data into various classes depending on which class it can most relate to. The main purpose of any classification problem is to predict that class under which the new data can fit. So, we have used various classification models, to categorize the cancer as Malignant or Benign. A conclusion is inferred by the classification model from the input values provided to it for training i.e. the training dataset. This conclusion helps the classification model to predict the class labels/categories for the new data i.e. the test dataset.

The classification models built for categorization are:

Support vector machine: It splits the training data points in space on the basis of some parameter values such that the partition is as wide as possible creating a clear gap among them. The test dataset which contains the new examples are basically put across that space and hence predicts the class they are part of depending on which side of the space they are falling.

Decision tree: It assembles various rules that have a particular sequence which is decided by some factors namely, Information gain, Gini index etc. These sequences of rules can thus help to classify the data provided there is a training dataset available with some attributes and classes.

Logistic regression: It can be described as a stabilized machine learning technique that can be used for binary as well as multinomial classification. In this algorithmic design, a logistic function is used to model the probabilities that illustrate the possible outcomes of a single trial.

Random forest classifier: Very precisely it can be described as a model that is a combination of many decision trees on various sub-divisions of the samples of the dataset which provides an average prediction result in order to improve the accuracy and avoid the over-fitting of the data. The name forest as it suggests, describes the combination of various decision trees that by far provides a better accuracy compared to any individual decision tree.

Naive Bayes: It is an intuitive algorithm which derives its basis from the Bayes' theorem. It works on the principle that the presence of a particular feature of the class should not be related to any other feature of that class, i.e. there exists independence between every pair of features.

3. Improvement of Classification Models

Stratified K-Folds and dimensionality reduction are used to improve the classification models, so as to give a better accuracy.

3.1 Stratified K-Folds

Stratified K-Folds cross-validation technique is a commonly used statistical method used to evaluate the dexterity of machine learning models. It is one of the main applications of machine learning which is used in selection and comparison of various models for a particular predictive modelling scenario. It is highly recommended as it is easy to implement, easy to understand and mostly gives the best prediction as it is less biased than other methods.

The Stratified K-Folds cross-validator computes indices of train or test data samples to split the data into train or test sets. The folds are made by preserving the percentage of data samples for each category. In our implementation, K-folds from 0 to 20 with a step-size of 5 has been applied.

3.2 Dimensionality Reduction

Dimensionality reduction is basically a technique where the attributes of the input dataset i.e. random variables can be reduced. This is useful because most of the times, a large number of attributes can lead to the curse of dimensionality

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

which means that more input features can make the prediction task exigent, resulting in poor performance of the machine learning model and hence reduced accuracy.

The various methods used for dimensionality reduction include:

- Auto Encoders
- Algebra Methods
- Feature Selection
- Feature Extraction
- Projection Methods

Dimensionality reduction can be either linear or non-linear, based on the technique used. The basic linear technique, which is called Principal Component Analysis (PCA) has been used in this implementation.

VI. RESULTS

It can be observed, that the Logistic Regression classification model, worked best for prostate cancer detection. After applying Stratified K-Fold, between the ranges 0-20, with the step size being 5, the final average accuracy recorded was 91.25%. It was also observed, that dimensionality reduction had a declining effect when tested on prostate cancer dataset. Hence, Logistic Regression produced an accuracy of 88%. Figure 4 represents the accuracy comparison for prostate cancer. Table 5 depicts the average accuracies of different models on prostate cancer dataset.

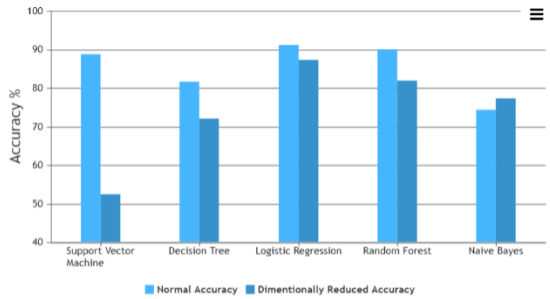

Fig 4: Accuracy comparison for Prostate Cancer

CLASSIFICATION ALGORITHMS	ACCURACY %		ACCURACY % WITH DIMENSIONALITY REDUCTION	
LOGISTIC REGRESSION (BEST OBSERVED)	Fold-0	90.48%	Fold-0	81.5%
	Fold-5	90.19%	Fold-5	87%
	Fold-10	90.18%	Fold-10	90%
	Fold-15	92.38%	Fold-15	90%
	Fold-20	93%	Fold-20	90%
	Final Average	91.25%	**Final Average**	88%
SUPPORT VECTOR MACHINE	88.83%		52%	
DECISION TREE	81.80%		72%	
RANDOM FOREST	90.08%		82%	
NAÏVE BAYES	74.39%		77%	

Table 5: Average accuracies of different models on prostate cancer dataset

On the other hand, for Breast Cancer it was observed that, the Random Forest classification model gave the best accuracy. After applying Stratified K-Fold, between the ranges 0-20 with the step size being 5, the final average accuracy was observed to be 96.05%. Dimensionality reduction worked well for the breast cancer dataset especially in the case of Support vector Machine where an exponential increase in the accuracy was observed. But, for Random Forest a very slight decrease in the accuracy was observed due to randomization of train-test split, resulting up to 95%. Figure 5 represents the accuracy comparison for breast cancer. Table 6 depicts the average accuracies of different models on breast cancer dataset.

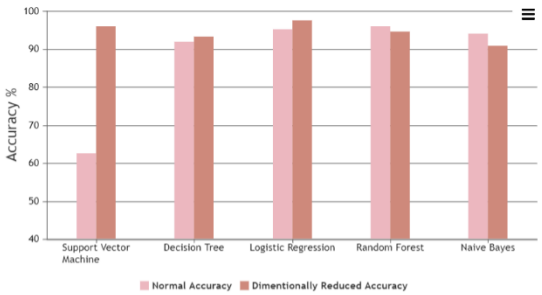

Fig 5: Accuracy comparison for Breast Cancer

CLASSIFICATION ALGORITHMS	ACCURACY %		ACCURACY % WITH DIMENSIONALITY REDUCTION	
RANDOM FOREST (BEST OBSERVED)	Fold-0	95.86%	Fold-0	95.09%
	Fold-5	95.97%	Fold-5	94.98%
	Fold-10	96%	Fold-10	95.01%
	Fold-15	96.07%	Fold-15	94.19%
	Fold-20	96.34%	Fold-20	93.92%
	Final Average	96.05%	**Final Average**	95%
SUPPORT VECTOR MACHINE	62.57%		95.5%	
DECISION TREE	91.98%		93%	
LOGISTIC REGRESSION	95.25%		96%	
NAÏVE BAYES	94.10%		91%	

Table 6: Average accuracies of different models on breast cancer dataset

CONCLUSION AND SCOPE

The most significant part of a complete cancer control planning includes the diagnosis and the treatment for the same. To make the above planning successful, it needs to be incorporated with an early cancer detection application, so that the cancer can be predicted at an early stage. This ensures a very effective treatment yielding up to a greater chance of cure. This implementation is an application of the Machine Learning Techniques in the field of Medicine.

The application developed is delivering results, with high accuracy, as high as **91.25%** for prostate cancer, and **96.05%** for breast cancer. These accuracies are delivered by Logistic Regression and Random Forest respectively. It was also observed that there was a slight increase in the accuracy for most of the models, exceptionally in the case of Support Vector Machine for Breast Cancer.

Although this application classifies the cancer data samples quite accurately for both breast cancer and prostate cancer, it still faces some issues. The application currently classifies the cancer as Malignant (M) or Benign (B). For instance, the classification label contains a high number of classes, this application may fail to classify accurately. The application can be improved in the following ways:

Bigger Dataset: Having a bigger dataset can never be a bad idea as it ensures more precise and enhanced classification models. A machine tends to have a better learning in the presence of bigger dataset, i.e. more the data, better the training and hence better the result. Larger size of data can be of great help as it gives accurate mean result, in turn providing minimum error. It also identifies outlier values that could have skewed the data if the sample was small. Deep Learning being a recent non-linear machine learning technique, continues to yield improved performance with bigger dataset.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Handling outliers and missing values: The accuracy of a classification model tends to reduce due to presence of faulty data such as unnecessary outliers or various missing values. This might lead to an inaccurate or biased model which might not give the expected result. This mostly happens due to the fact that one doesn't inspect the behaviour and correlations among variables accurately. The above explanation justifies why it is a basic need to handle the outliers and missing values effectively.

Algorithm Tuning: For any given parameter, a certain range can be defined that provides the optimum value for the same. This process of finding the optimum value for the parameter is known as parameter tuning, and hence this process can be employed to improve the accuracy of the machine learning model. The prerequisite to find this optimum value for the parameter comes with having a better understanding of their meaning and how this parameter individually impacts the model. The above method can be repeatedly applied with various well performing models.

Grouping multiple Models (Ensembles): The prediction result of multiple models can be amalgamated to give improved and accurate result. In fact, this has turned out to be the next substantial area for amelioration after the application of algorithm tuning. Although, one can obtain rather high accuracy with fragile (highly-tuned) models but combining the predictions from various good-enough models can also deliver satisfactory performance which can in turn be estimated as a better method.

Application of Deep Learning: Deep Learning can be defined as a sub-division of machine learning that attains immense power and flexibility by getting trained in such a way that it is able to delineate the world as nested hierarchy of hypotheses, with each notion of hypotheses described in relation to elementary hypothesis, and more abstract descriptions computed in terms of less abridged ones. In a more intricate view, a deep learning method uses its hidden layer architecture in order to train itself about various level of categories incrementally, such as letters representing low-level, words representing high-level and sentences representing much higher level. A substantial advantage with deep learning is that it works well with huge amounts of data as already mentioned above. This can be understood by considering the metaphor that deep learning models are rocket engines and on the other hand, the massive amount of data fed to it is the fuel to that rocket engine.

ACKNOWLEDGEMENT

We are extremely delighted to manifest our sincere gratitude towards the management of **Sir M. Visvesvaraya Institute of Technology,** Bangalore for lending us the opportunity and the assets to accomplish our paper in their premises.

On the path of learning, the presence of an experienced guide is imperative and this paper would have been incomplete without his invaluable help and guidance, so we would like to extend special thanks to our guide **Mr Sreenivasa B C,** Associate Professor, Department of Computer Science and Engineering.

We would also like to convey our regards and sincere thanks to **Prof Dr. G.C. Bhanu Prakash,** HOD, Department of Computer Science and Engineering for his suggestions, constant support and valuable guidance from time to time. Heartfelt and sincere thanks to **Prof Dr. V.R. Manjunath,** Principal, Sir M. Visvesvaraya Institute of Technology, for his constant encouragement.

REFERENCES

[1] Howlader, Koushik Chandra, Urmi Das, and Mahmudur Rahman. "Breast Cancer Prediction using Different Classification Algorithm and their Performance Analysis."

[2] Agarap, A.F.M., 2018, February. On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (pp. 5-9). ACM.

[3] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., 2015. Machine Learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, pp.8-17.

[4] Mihaylov, I., Nisheva, M. and Vassilev, D., 2019. Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies. Information, 10(3), p.93.

[5] Ahmad, L.G., Eshlaghy A.T., Poorebrahimi, A.,Ebrahimi, M. and Razavi, A.R., 2013. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4(124), p.3.

[6] Cuocolo, R., Cipullo, M.B., Stanzione, A., Ugga, L., Romeo, V., Radice, L., Brunetti, A. and Imbriaco, M., 2019. Machine learning applications in prostate cancer magnetic resonance imaging. European radiology experimental, 3(1), pp.1-8.

[7] Barlow, H., Mao, S. and Khushi, M., 2019. Predicting High-Risk Prostate Cancer Using Machine Learning Methods. Data, 4(3), p.129.

[8] Goldenberg, S.L., Nir, G. and Salcudean, S.E., 2019. A new era: artificial intelligence and machine learning in prostate cancer. Nature Reviews Urology, 16(7), pp.391-43.

[9] Graziani, M., Andrearczyk, V. and Müller, H., 2018. Regression concept vectors for bidirectional explanations in histopathology. In Understanding and Interpreting Machine Learning in Medical Image Computing Applications (pp. 124-132). Springer, Cham.

[10] Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R. and Sieh, W., 2019. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Scientific reports, 9(1), pp.1-12.