iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VIII **Month of publication:** August 2020

DOI: https://doi.org/10.22214/ijraset.2020.30979

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

Predictive Analysis of Lung Cancer, Breast Cancer and Prostate Cancer using Machine Learning

Hritik Bhardwaj[1], Rudraksh Saxena[2], Shubham Tyagi[3]
[1, 2, 3] *Student, Computer Science, Raj Kumar Goel Institute of Technology And Management, Uttar Pradesh, India*

Abstract: ***Machine learning is modern and highly sophisticated technological applications became a huge trend in the health care industry. It provides methods, techniques and tools that can help in solving diagnostic problems in a variety of medical domains e.g. prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning and support, and for the overall patient management. In this paper, various machine learning algorithms has been used for developing efficient decision support that can be used in healthcare applications. Cancer has been the leading cause of death worldwide for many years. The objective of this study is to diagnose and predict the possibility of occurrence of three most common types of cancer including Lung Cancer, Breast Cancer and Prostate Cancer using various Machine Learning supervised algorithms like Logistic regression, Naive Bayes classification, Kernel support vector classification , Decision Tree Classifier, Random Forest Classifier and XG Boost Classifier.***

Keywords**: *Machine Learning (ML), Naïve Bayes Classification, Kernel Support Vector Classification, Decision Tree Classifier, Random Forest Classification, Logistic Regression Model, XG Boost Classifier, Lung Cancer, Breast Cancer, Prostate Cancer.***

I. INTRODUCTION

Cancer is a leading cause of death worldwide, accounting for an estimated 9.6 million deaths in 2018. According to World Health Organization (WHO), lung cancer, breast cancer and prostate cancer are among the top most common type of cancer [1]. Lung cancer remains the most commonly diagnosed cancer and the leading cause of cancer death worldwide. [2] As per WHO says, there were 2.09 million cases of Lung cancer cases in the year 2018.[1] Breast cancer is the most frequent cancer among women, impacting 2.1 million women each year, and also causes the greatest number of cancer-related deaths among women. In 2018, it is estimated that 627,000 women died from breast cancer – that is approximately 15% of all cancer deaths among women. [1] Prostate cancer is the second most frequent cancer diagnosis made in men and the fifth leading cause of death worldwide.[3] Based on WHO 2018 estimates, 1.28 million cases new cases prostate cancer were reported worldwide in 2018, with higher prevalence in the developed countries. The cancer burden can be reduced through early detection of cancer and management of patients who develop cancer. Many cancers have a high chance of cure if diagnosed early and treated adequately [1].

The objective of this briefing is to is to assess the efficiency and accuracy of the used ML models for the early detection of these three most common types of cancer.

II. MACHINE LEARNING

Machine learning is a method of data analysis that automates analytical model building. [4] It is an application of artificial intelligence that provides system the ability to learn and improve from experience without being explicitly programmed. There are mainly two phases of learning process (i) On the basis of dataset provided, the unknown dependencies are to be estimated for the system and (ii) New output of the system is to predict if estimated dependencies are known.

A formal definition of machine learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.[4]

The types of Machine Learning are describe below:

A. Supervised Learning

Supervised machine learning is a type of machine learning algorithms which can apply what has been learned in the past to new information with the use of labelled examples to predict future events precisely. For example, a data set of apartments of specific measurement with real costs is given, then the supervised algorithm can provide answers such as for new apartment what would be the approximate price. [5]

B. Unsupervised Learning

In Unsupervised machine learning method, data with no labels are given to the learning algorithm and leaving it on its own to find structure and patterns in its input. [5] In this, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels. [4]

C. Reinforcement Learning

Reinforcement learning is a method based on output with how an intelligent agent ought to take actions in an environment to maximize some kind of long-term reward like humans do in their life. Reinforcement learning is different from supervised learning in the sense that a correct input and output pairs are by no means presented, not all moves explicitly corrected. In short, Reinforcement learning is the type of machine learning model to make a sequence of decisions.[5]

III.ALGORITHMS

To automatically learn and enhance from experience without being explicitly programmed, Machine Learning algorithms are programmed that adjust themselves to perform better as they are exposed to more data. Machine learning uses algorithms in a distinct way to accommodate its own parameter, given feedback on its previous program making prediction about the dataset. An algorithm contains combine knowledge of statistics, probability, calculus, vector algebra, matrices, optimization techniques etc. There are six classification algorithms has been used it this system which are as follows:

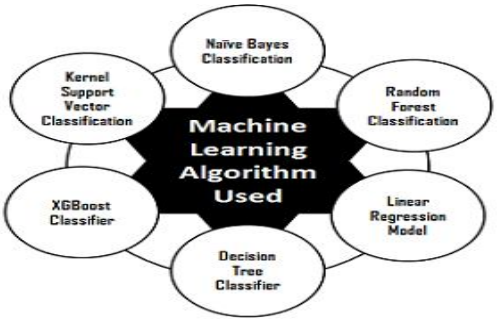


Fig. 1 Machine Learning Algorithms Used

A. Naive Bayes Classification

Naïve Bayes calculations are an arrangement method dependent on applying Bayes' hypothesis with a solid supposition that every one of the indicators is autonomous to one another. [6] Naïve Bayes mainly targets the text classification industry. It is mainly used for clustering and classification purpose. [7] The underlying architecture of Naïve Bayes depends on the conditional probability. It creates trees based on their probability of happening. This model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Bayes theorem provides a way of calculating the posterior probability P(A/B) of class from P(A) is the prior probability of class.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

Fig. 2 Equation for Naïve Bayes Classification

P(B) is the prior probability of predictor and P(B/A) is the likelihood which is the probability of predictor given class. Naïve Bayes classifier assumes that the effect of the value of a predictor (A) on a given class (A) is independent of the values of other predictors called conditional independence. [8]

B. Kernel Support Vector Classification

SVM is a concept that is used to classify the labelled data and to apply regression analysis on the given dataset. There is a hyperplane to have sets of input data where SVM divides the dataset into two classes as the classification is done on the basis of labelled data in the best possible way. [9] The closest point to hyperplane is referred as Support Vector. Each dataset has a support vector point. The gap between dataset is known as margin. Greater margin will affect better computation result. [4]

C. Decision Tree Classifier

It has flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch denotes an outcome of test, and each leaf node holds a class label. The topmost node in a tree is the root node. Decision tree is useful because construction of decision tree classifiers does not require any domain knowledge. It can handle n dimensional data. The learning and classification steps of decision tree induction are simple and fast. The representation of acquired knowledge in tree form is easy to assimilate by users. Decision tree classifiers have good accuracy. [10]

D. Random Forest Classifier

A Random Forest Algorithm takes the decision tree concept a step further by producing a big number of decision trees to make a forest. These trees are reformed on the basis of selection of data and variables randomly. [9] As we realize that a forest is comprised of trees and more trees implies progressively robust forest. So also, arbitrary random forest algorithm makes choice trees on data samples and afterward gets the forecast from every one of them lastly chooses the best solution by methods for casting a vote. It is an outfit strategy which is superior to anything a solitary choice tree since it decreases the over-fitting by averaging the outcome. [6]

E. Logistic Regression

It is one of a classification technique in which if the decision threshold is dependent on the classification problem then the setting of threshold is an important aspect in classification core. [11] It gives statistical data. It is basically multi-class binary classification. Its concept is mostly similar to the concept of probability i.e., modelling of an event is occurring versus event is not occurring. For e.g., for a hospital dataset, whether the patient is diabetic or not needs to be predicted. Logistic Regression makes use of sigmoid function which takes solution of linear regression and output value between 0 and 1. It has S-shaped curve known as logistic curves.

Sigmoid Function = 1/(1 + e –values)

F. XG Boost Classifier

It is short for eXtreme Gradient Boosting package. It is an efficient and scalable implementation of gradient boosting framework by (Friedman, 2001) (Friedman et al., 2000). The package includes efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking. The package is made to be extendible, so that users are also allowed to define their own objectives easily. [12]

IV. IMPLEMENTATION

The implementation phase in machine learning contains these steps as shown in Fig 3:

Fig.3 Steps of Implementation

A. Data Collection

All the dataset presented here have been collected from Kaggle community.

1) *Lung Cancer*: Lung cancer typically doesn't cause signs and symptoms in its earliest stage. The dataset includes diagnosis and result section which investigates whether patient is having any symptoms of Lung cancer.

	Name	Member_ID	Diagnosis	Age	Smokes	Smokes (years)	Smokes (packs/year)	AreaQ	Alkhol	family history	Result
0	Wick	91550	M	35	3	0.0	0.0	5	4	0	1
1	Constantine	915664	M	27	20	0.0	0.0	2	5	0	1
2	Anderson	915691	M	30	0	0.0	0.0	5	2	0	0

Fig. 4 Lung Cancer Patients Dataset

2) *Breast Cancer:* This dataset is taken from UCI machine learning repository. The dataset contain various perimeters like mean of radius, texture, perimeter, area, smoothness which diagnose whether there is any possibility of breast cancer or not.

	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
0	17.99	10.38	122.80	1001.0	0.11840	0
1	20.57	17.77	132.90	1326.0	0.08474	0
2	19.69	21.25	130.00	1203.0	0.10960	0

Fig. 5 Breast Cancer Patients Dataset

3) *Prostate Cancer:* This is a recent dataset of patients to implement the machine learning algorithm and thereby interpreting the result whether patient is building any symptoms of prostate cancer or not.

	id	diagnosis_result	radius	texture	perimeter	area	smoothness	compactness	symmetry	fractal_dimension
0	1	M	23	12	151	954	0.143	0.278	0.242	0.079
1	2	B	9	13	133	1326	0.143	0.079	0.181	0.057
2	3	M	21	27	130	1203	0.125	0.160	0.207	0.060

Fig. 6 Prostate Cancer patients Dataset

B. Data Preparation

Data preparation includes cleaning the data and in machine learning, it is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. As this paper deals with medical data, any changes have not been made in order to preserve the patient's data. Data correlation is the way in which one section of data correlates or connects with another section of same data. The set of correlation values between pairs of attributes form a matrix which is called a correlation matrix is shown in Fig.7, Fig.8 and Fig.9 with respect to datasets.

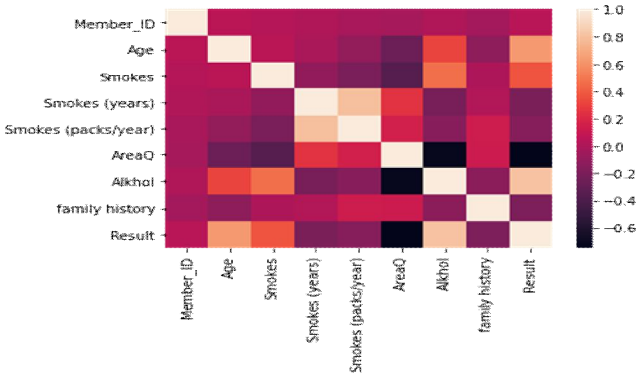


Fig. 7 Correlation Heatmap of Lung Cancer Dataset

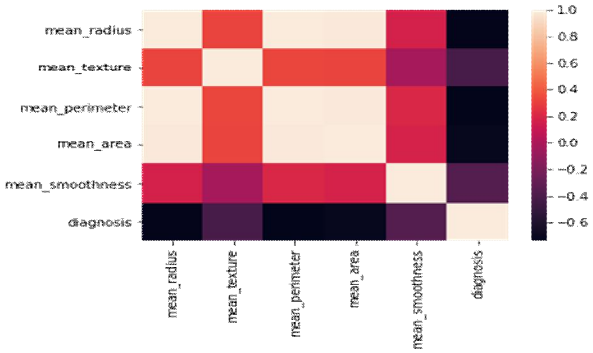


Fig. 8 Correlation Heatmap of Breast Cancer Dataset

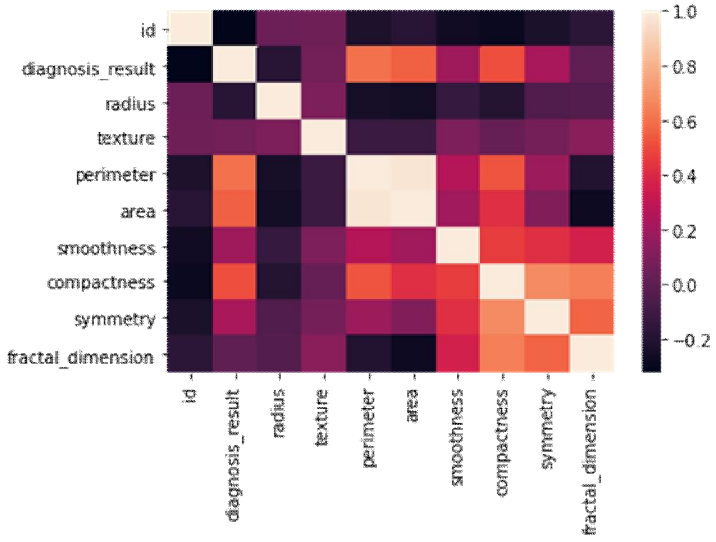


Fig. 9 Correlation Heatmap of Prostate Cancer Dataset

C. Training Model

The above datasets are trained using different machine learning algorithm. Model get trained on high amount of data can generalise themselves better. Generalisation is the ability of model to give generalised prediction across varied or diverse data. Training is basically the procedure of giving the machine efficiency to make further predictions after learning from the training dataset.

```
#Random Forest Classification
from sklearn.ensemble import RandomForestClassifier
classifier_rfc = RandomForestClassifier(n_estimators=20, random_state=0)
classifier_rfc.fit(x_train, y_train)
y_rfc_pred = classifier_rfc.predict(x_test)
```

```
#Logistic Regression Model
from sklearn.linear_model import LogisticRegression
log_model = LogisticRegression()
log_model.fit(x_train, y_train)
y_pred_log = log_model.predict(x_test)
```

```
#Decision Tree Classifier
from sklearn.tree import DecisionTreeClassifier
classifier_dtc = DecisionTreeClassifier()
classifier_dtc.fit(x_train,y_train)
y_dtc_pred = classifier_dtc.predict(x_test)
```

```
# XG Boost Classifier
from xgboost import XGBClassifier
xgboost_model = XGBClassifier()
xgboost_model.fit(x_train, y_train)
y_pred_xg = xgboost_model.predict(x_test)
```

```
#Kernel Support Vector Classification
from sklearn.svm import SVC
classifier_svc = SVC(kernel='linear',random_state=0)
classifier_svc.fit(x_train,y_train)
y_svc_pred = classifier_svc.predict(x_test)
```

```
#Naive Bayes Classification
from sklearn.naive_bayes import GaussianNB
classifier_naive = GaussianNB()
classifier_naive.fit(x_train, y_train)
y_naive_pred = classifier_naive.predict(x_test)
```

Fig.10 Training dataset with different algorithms

D. Testing Model

Testing of model is done to analyze the performance of the algorithms in term of accuracy, precision, F1 scores etc. During testing to make sure whether the prediction is correct or not is checked by using already pre-defined dataset. More the accuracy, higher the result.

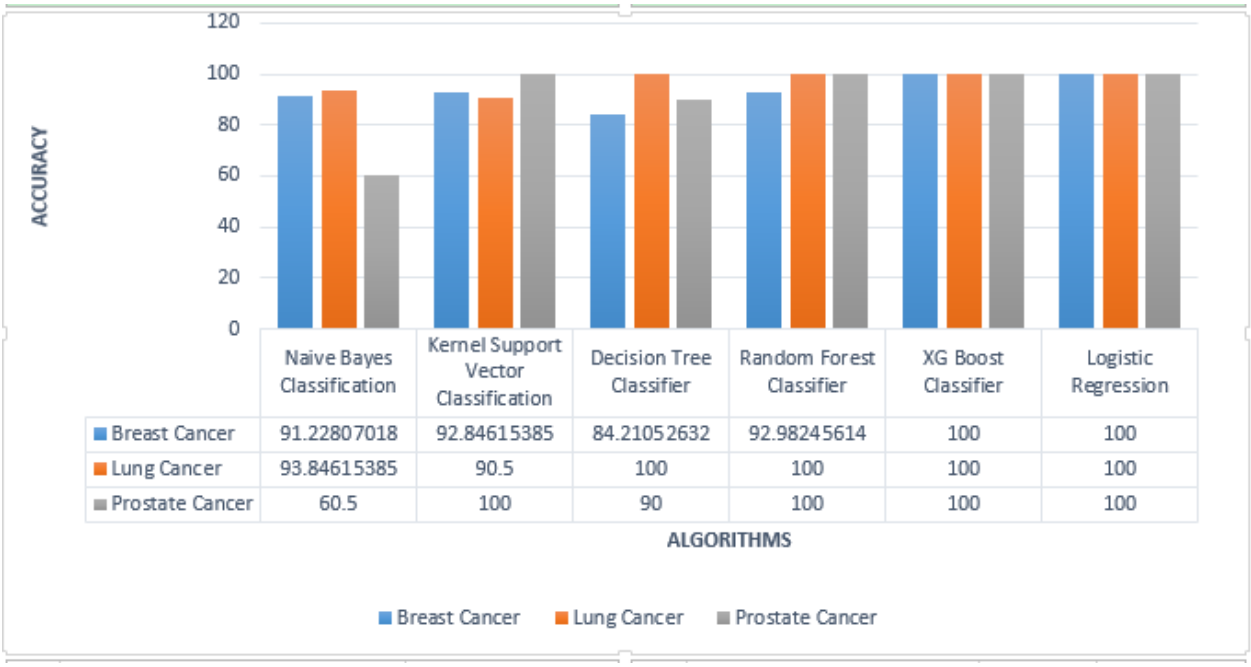


Fig.11 Algorithms with respect to their accuracy scores in respective disease

E. Prediction

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome. The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be. [4]

In Fig.12, Fig.13 and Fig.14 actual values and predicted values of the datasets have been compared:

	Result	Predicted Result
0	1	1
1	1	1
2	0	1

Fig.12 Actual Vs Predicted Values of Lung Cancer Dataset

	diagnosis	Predicted diagnosis
0	0	0
1	0	0
2	0	0

Fig.13 Actual Vs Predicted Values of Breast Cancer Dataset

	diagnosis_result	Predicted diagnosis_result
0	1	1
1	0	1
2	1	1

Fig.14 Actual Vs Predicted Values of Prostate Cancer Dataset

V. CONCLUSIONS

Machine learning is swiftly infiltrating many areas within the healthcare industry, from diagnosis and prognosis of cancer and epidemiology, with significant potential to transform the overall medical landscape. This study is a condensed snapshot of applications of machine learning for diagnostically predicting the possibility of three common types of cancer including lung cancer, breast cancer and prostate cancer using various ML algorithms like Random Forest, Decision Tree, Naive Bayes Classification, XG Boost Classifier, Kernel Support Vector Classification and Logistic Regression and corresponding results have been analyzed. Fusion of disparate multimodal and multi-scale biomedical data continues to be a challenge For improvements, more features in the dataset can be added. Substantial improvement in Python-based workflow for better data visualization and optimization will also be an option.

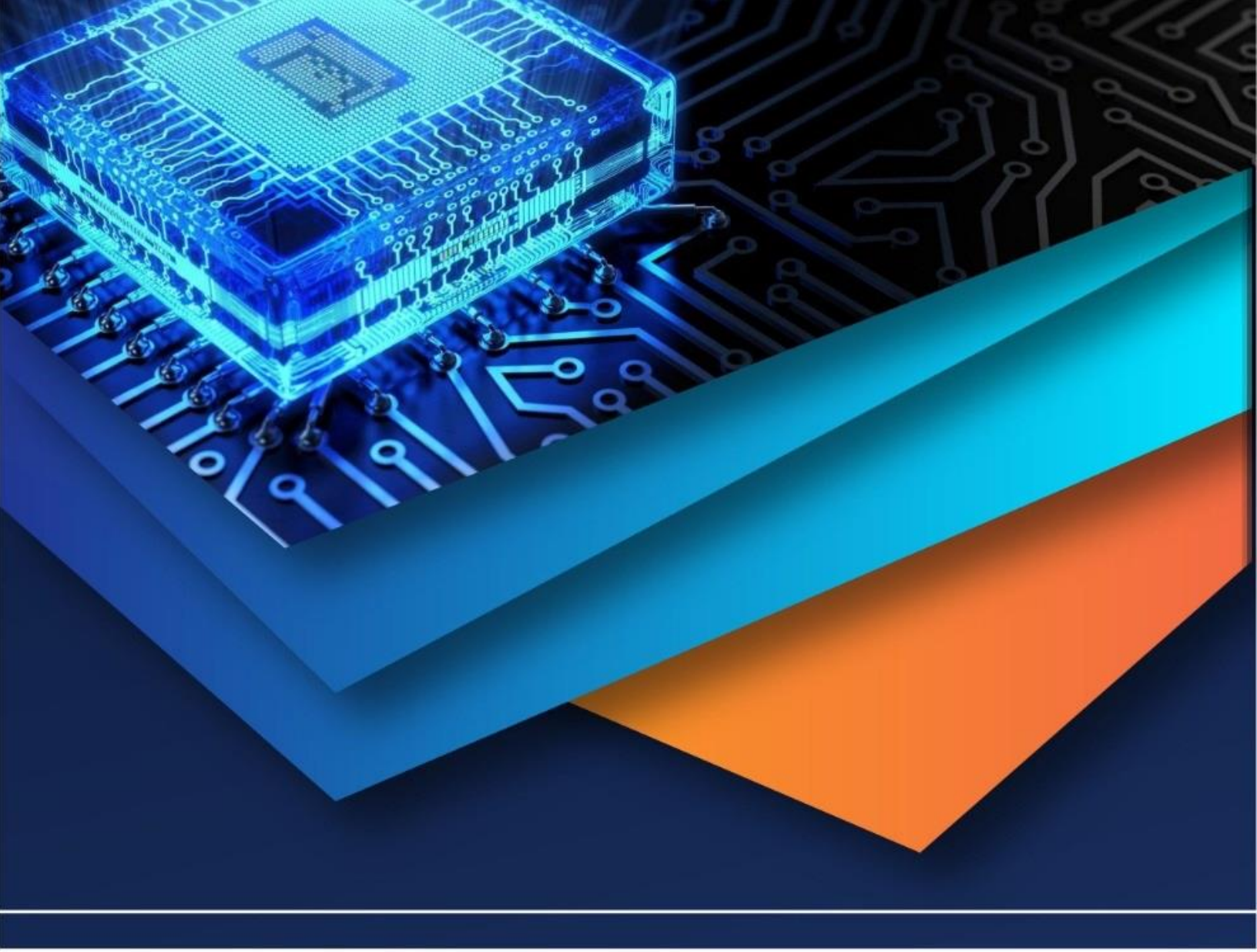
VI. FUTURE SCOPE

The value of machine learning in healthcare industry is its ability to grow huge datasets above the scope of human efficiency, and then accurately convert analysis of that corresponding data into clinical acumen that assist physicians in planning and providing care, ultimately leading to better outcomes, lower costs of care, and increase in patient's satisfaction.

A potential future development of the presented work is to apply more advanced machine learning models to other data with different features, concerning the survival diagnosis and prognosis of the patients and early detection of the cancer with recommendation system for test and medicines according to symptoms. Further using advanced analytics techniques, one can try to develop system making it easier for doctors to visualize patient's data for diagnoses, treatment and other purpose. Moreover, it can also be developed in web-based application with additional services.

REFERENCES

[1] [World Health Organization (WHO)] [News-Room] [Fact-Sheets] [Detail] [Cancer] [Online]. Available at: https://www.who.int/news-room/fact-sheets/detail/cancer
[2] [Lung Cancer][The Cancer Atlas][Online]. Available at: https://canceratlas.cancer.org/the-burden/lung-cancer/
[3] [Epidemiology of Prostate Cancer] [Online]. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6497009/#R01
[4] Shruti Katiyar, Shruti Jain."Predictive Analysis on Diabetes, Liver and Kidney Diseases using Machine Learning", Volume 8, Issue V, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 2285-2292, ISSN : 2321-9653.
[5] Mr. ShismohammadMulla, Dr. Mahesh Chavan."Application of Machine Learning in Computer Vision: A Brief Review", Volume 8, Issue VII, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 508-516, ISSN : 2321-9653.
[6] Muktevi Srivenkatesh. "Prediction of Prostate Cancer using Machine Learning Algorithms", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5, January 2020*FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
[7] D. Lowd, P. Domingos, "Naïve Bayes Models for Probability Estimation
[8] S. Kanchana."Statistical Analysis Using Machine Learning Approach for Multiple Imputation of Missing Data ", Volume 6, Issue II, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 2090-2095, ISSN : 2321-9653.
[9] Kumar, Ajay and Sushil, Rama and Tiwari, Arvind Kumar, Machine Learning Based Approaches for Cancer Prediction: A Survey (March 11, 2019). Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019. Available at SSRN: https://ssrn.com/abstract=3350294 or http://dx.doi.org/10.2139/ssrn.3350294
[10] Bhaskar N. Patel, Satish G. Prajapati and Dr. Kamaljit I. Lakhtaria."Efficient Classification of Data Using Decision Tree", Bonfring International Journal of Data Mining, Vol. 2, No. 1, March 2012.
[11] RavinthiranPartheepan."Breast Cancer L and R Classification and Analysis using Machine Learning Techniques", Volume 8, Issue III, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 329-334, ISSN : 2321-9653.
[12] T. Chen and T. He, "XGBoost: extreme gradient boosting", R Package. Version 0.4-2, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY