iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** I **Month of publication:** January 2022

DOI: https://doi.org/10.22214/ijraset.2022.39743

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: 7.538
Volume 10 Issue I Jan 2022- Available at www.ijraset.com

Breast Cancer Prediction Using Classification Techniques of Machine Learning

Angela More[1], Siddhi Mhatre[2], Varsha Kamble[3], Vanshika Patil[4], Sujata Bhairnallykar[5]
[1, 2, 3, 4]*Computer Engineering Department,* [5]*Assistant Professor, Saraswati College of Engineering, Kharghar, Navi Mumbai, Maharashtra, India - 410210.*

Abstract: ***Data analytics play vital roles in diagnosis and treatment in the health care sector. To enable practitioner decision-making, huge volumes of data should be processed with machine learning techniques to produce tools for prediction and classification Breast Cancer reports 1 million cases per year. We have proposed a prediction model, which is specifically designed for prediction of Breast Cancer using Machine learning algorithms Decision tree classifier, Naïve Bayes, SVM and K-Nearest Neighbour algorithms. The model predicts the type of tumour, the tumour can be benign (noncancerous) or malignant (cancerous) . The model uses supervised learning which is a machine learning concept where we provide dependent and independent columns to machine. It uses classification technique which predicts the type of tumour.***
Keywords: Cancer, Machine learning, Prediction, Data Visualization, SVM, Naïve Bayes, Classification.

I. INTRODUCTION

Breast cancer is among the most pervasive cancers found in women. It is a cancer in which the breast cells grow abnormally.
Female breast cancer has surpassed lung cancer as the most commonly diagnosed cancer worldwide, with an estimated 2.3 million new cases (11.7%)[1]. 10% of the breast cancer cases are hereditary and the other 90% are related to lifestyle factors. A significant increase in the incidence rates of breast cancer was observed in 15 PBCRs in females. Majority of patients underwent multi-modality treatment and 97.7% were epithelial tumours. Israel (84.6) had the highest incidence of breast cancer in Asia. In India, Hyderabad district (48.0) had the highest incidence rate[2]. According to a report published by National Cancer Registry Programme (NCRP), cancer cases are expected to increase from 13.9 lakh in 2020 to 15.7 lakh by 2025, assuming a 20 percent increase overall [3]. Common cancers can be prevented to be fatal if treated early. A breast cancer diagnosis made early can lead to effective treatment.
The goal of the research is to classify the patients in Malignant and Benign types of tumours by classification techniques hence achieving higher accuracy. The dataset is from the Kaggle website. We have used supervised learning which is a machine learning concept where we provide dependent and independent columns to the machine for learning and after the learning process is completed the machine will predict the value for the dependent variable for a given input in the form of an independent variable. The classification techniques used for detecting the tumour are Decision tree, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB) classification in Jupyter notebook along with data visualization.

II. LITERATURE REVIEW

The following is a brief summary of work done in the following domain:

Author	Dataset Used	Techniques Used	Tools Used	Result	Error Rate
Muktevi et al. [4]	Wisconsin breast cancer-Kaggle	SVM, Random Forest, KNN, LR, NB	Python	Random Forest had the best accuracy of all with 98.24%	RF had mean absolute error of 0.01%
Ramik Rawal [5]	Wisconsin -Kaggle	SVM, Logistic Regression, Random Forest, K-NN	Jupyter Notebook-Python	SVM- 97.13% highest efficiency and accuracy	
Gaurav Singh [6]	UCI repository	K-NN, SVM, LR, NB	Python	K-NN- 99% SVM- 96% LR-97% NB-95%	

Min-Wei et al. [7]	UCI Repository, ACM SIGKDD Cup 2008	SVM classifiers, SVM ensembles	Weka	SVM ensembles perform slightly better than single SVM classifiers	
Deepika et al. [8]	UCI repository	Naïve Bayes, MLP	Weka	Naïve Bayes had better accuracy	
Ch. Shravya et al. [9]	UCI repository	SVM, K-NN, LR	Spyder Platform	SVM predicted the best accuracy of 92.78% followed by KNN-92.23%	
Wang et al. [10]	Wisconsin Breast Cancer Database (1991),Wisconsin Diagnostic Breast Cancer (1995)	SVM, ANN, Adaboost, PCA	WEKA	8 PCs -92.6% correlation, 10 PCS- 95%	

This project aims to improve the accuracy and error rate of classification using different classification algorithms.

III. BACKGROUND STUDY: MACHINE LEARNING ALGORITHMS

The following four machine learning classification techniques are used:

A. Decision Tree Classifier

In 1980, J. Ross Quinlan developed ID3(Iterative Dichotomiser) which is a decision tree algorithm. The decision tree classifier is an example of supervised machine learning A decision tree works on possible solutions to a decision based on certain conditions. It classifies conditions at every node to find a solution.

Algorithm

1) Starts at root node
2) Root value compared with record real dataset attribute
3) Jump to next node based on comparison
4) Compare attribute value with sub-node value and jump accordingly
5) Repeat till left node of tree is reached.

B. Naïve Bayes

Naïve Bayes classification technique is based on the Bayes' Theorem. It is named 'naïve' because this algorithm assumes each input variable to be independent. It is a quick and simple ML algorithm that is used to predict a class of that which are usually large. When a class variable is given, Naïve Bayes classifier assumes the presence or absence of a feature to be unrelated to the presence or absence of other features. It is quite useful and effective when complex problems are involved.

Algorithm

1) Separate the training data by class
2) Calculate mean and standard deviation for each attribute
3) Summarize and organize dataset by class
4) Calculate the Gaussian Probability Density function
5) Lastly calculate the class probabilities

C. *Support Vector Machine*

Support Vector Machines (SVMs) are machine learning algorithms which are used to deal with both classification and regression problems. SVM linear classifier is built around the margin maximization principle. SVM algorithm is used to create a decision boundary, called hyperplane that segregates n-dimensional space into different classes so that new data points can be added easily. This linear classifier plans to broaden the space between the decision hyperplane and the closest data points by finding the most suitable hyperplane.

Algorithm

1) Correctly classify the training dataset on the basis of lines/boundaries
2) Selects the one having maximum distance from the closest data point out of the lines/boundaries

D. *K-Nearest Neighbour*

K-Nearest Neighbour is a Machine Learning algorithm which belongs to the Supervised Learning technique. K-NN algorithm can solve both classification and regression problems but is mainly used for classification problems. It is a non-parametric classification method, which doesn't make assumptions based on underlying data. It is also known as a lazy learner algorithm because the training set is not learned immediately, instead, it stores the dataset and then at classification time, it performs an action on it. In the K-NN algorithm, similarity is assumed between a new case and available data, and the new case is placed in the category with the most similarity to the available cases. All the available data is stored and the classification of new data point based is done based on similarity. As a result, when new data appears, the K-NN algorithm can be easily used to classify it into an appropriate category. K-NN stores datasets during the training phase, and when it gets new datasets, it categorizes them into a category that is similar to the new data.

Algorithm

1) Load training and testing data from dataset
2) Choose the value of nearest data point (K)
3) For each data point, calculate Euclidean (distance between testing and rows of training data) distance and sort them ascendingly
4) Choose top K row from sorted array and allot a class to test point based on the most frequent class

IV. PROPOSED METHODOLOGY

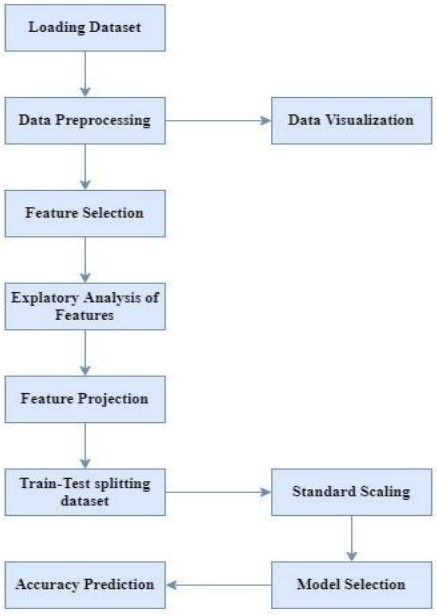


Fig (1) Different phases of breast cancer classification prediction

A. Phase 1: Loading Dataset

In this phase the data from the Breast Cancer Wisconsin (Diagnostic) Data Set is loaded into the JUPYTER notebook for further analysis.

B. Phase 2: Data Preprocessing

The raw data is transformed into an understandable format for future analysis. Data pre-processing refers to manipulation of datasets to ensure that only necessary data is used further for classification. The dataset which will be used may contain attributes which are not useful for classification and can cause errors in the model. Hence data pre-processing is necessary as it sorts this raw data into constructive data. This is a very important step as the performance of the model depends on the data used.

C. Phase 3: Data Visualization

This data that has been pre-processed is now visualized using seaborn and matplotlib libraries of Python. Data visualization helps in better understanding as it translates the large datasets into graphic and metric models which are easy for human mind to comprehend. In the following analysis, matrix is used to check whether there are any null values in the dataset. Also a co-relation between the different attributes is plotted for visualization purposes.

D. Phase 4: Feature Selection

Feature Selection is the process where the features which contribute most to your prediction variable or output are selected manually.

In the dataset, there are around 31 columns which contain different features which may or may not be necessary for analysis. So certain features which do not help in predicting accuracy are dropped and not used.

E. Phase 5: Exploratory Analysis Of Features

Exploratory Data Analysis is the process in which data is inspected to find patterns and anomalies, a hypothesis is done to study and analyse assumptions using different graphical methods and summary statistics. In this paper, the number of benign and malignant tumours in the diagnosis column of the dataset are counted and are visually represented using elements like charts and graphs.

F. Phase 6: Feature Projection

In order to improve data transformation and representative learning, in this paper feature projection technique is used. Feature projection or feature extraction refers to the linear combination of new features which is created using existing features. Feature projections transform the high dimensional data to low dimensional data with fewer attributes.

G. Phase 7: Train- Test Splitting Dataset

The train-test splitting technique is used to evaluate and analyse the performance of any machine learning model. This technique is used for both classification and regression problems. In this model, classification techniques have been evaluated using this technique where the data is split into two subsets- training and testing, and then the relevant data is tested to get the necessary results.

H. Phase 8: Standard Scaling

Standardization is a scaling technique that rescales the distribution values of a dataset. It can be useful when data follows Gaussian distribution. In the model, the scaler is fitted on training data and is then used to transform trained data. This helps to avoid any data leakage throughout the model training process.

I. Phase 9: Model Selection

In model selection, fitting machine learning models are selected for training the dataset to get relevant results. In our paper, the model selection process is applied across different types of models like SVM, K-NN, Naïve Bayes, and Decision Tree Classifier. In our dataset, we have a dependent variable Y having only two sets of values which are Malignant(M) and Benign(B). Our model works on this set of values for predicting the outcome based on the different classification algorithms.

J. Phase 10: Accuracy Prediction

In this paper, accuracy is predicted for classifying the type of tumour in the dataset. Accuracy is the percentage of correctly classified instances namely TN, TP, FN, and TP for the trained dataset. It is the ratio of the number of correct predictions to the number of total predictions. The misclassifications in the dataset are calculated which helps in understanding why such accuracy has been predicted.

V. PERFORMANCE EVALUATION

In performance evaluation, we check the different accuracies which are predicted using the different classification algorithms. This work was implemented in JUPYTER notebook of ANACONDA navigator in 1.6 GHz Dual-Core Intel Core i5 with 8GB memory. We used four algorithms namely SVM, K-NN, Naïve Bayes and Decision Tree for making the prediction. The data was split using train-test splitting method into 70-30% model for training and testing. The data was then standard scaled for analysis. Further pipeline method was used to make initial predictions on the dataset. It had the following predictions:

Algorithm	Accuracy
SVM	96.808
KNN	95.744
Naïve Bayes	93.617
Decision Tree	95.213

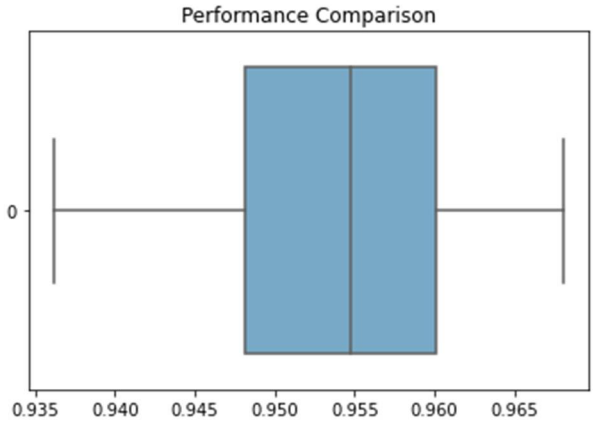


Fig (2) boxplot of performance comparison

The different algorithms were carried out and prediction of training and testing dataset was made.

Later a classification report was prepared which consisted of accuracy of data for both malignant and benign tumours. Precision, support and F1-score for both 0-Benign and 1-Malignant were calculated. F1 score is the weighted average score of precision and recall for both malignant and benign tumours. Both false positives (FP) and false negatives (FN) are taken into account when calculating this score. The table below represents the output gathered from different classification algorithms. It mentions the weighted average.

Table 1

Algorithm	Accuracy	Precision	Recall	Support	F1-score
SVM	97.8723	0.98	0.98	188	0.98
KNN	97.3753	0.97	0.97	188	0.97
Naïve Bayes	93.617	0.94	0.94	188	0.94
Decision Tree	96.808	0.97	0.97	188	0.97

- Accuracy: Determines best model for recognizing patterns in a dataset
- Precision: Number of TP by total number of positive predictions
- Recall: Measure to correctly identify TP
- Support: Number of correct samples in each class of target values
- F1-score: Weighted average score of precision and recall

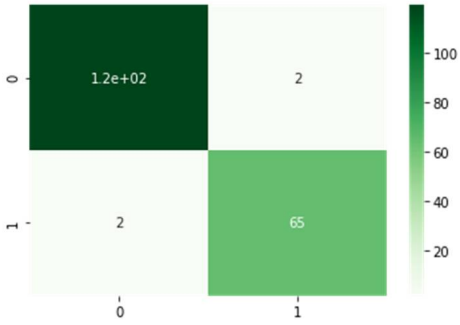
A confusion matrix was plotted to calculate the miscalculations in each model. A confusion matrix is used to evaluate and recount the performance of a classifier.

Table 2

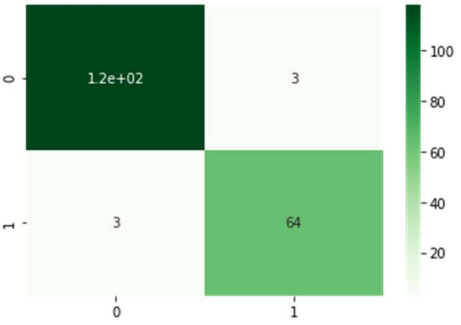
	Predicted Class		
Actual Class		Yes	No
	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

- True Positive (TP): Correctly predicts positive class
- False Positive (FP): Incorrectly predicts positive class
- False Negative (FN): Incorrectly predicts false class
- True Negative (TN): Correctly predicts false class

It was seen that SVM algorithm showed the least miscalculations of about 4 as 2 of them were FP and other 2 were FN. We can see the miscalculations in the below mentioned diagrams of confusion matrix.



Fig(3) SVM



Fig(4) K-NN

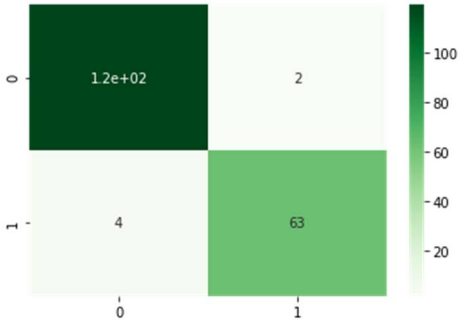


Fig (5) Decision Tree

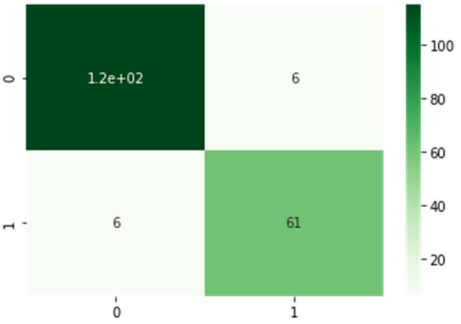


Fig (6) Naïve Bayes

VI. CONCLUSION

This paper analyses a Breast Cancer Wisconsin (Diagnostic) Data Set with 32 attributes and makes a prediction using different classifiers whether the tumour is benign or malignant. It can be seen that the highest accuracy is obtained using the SVM model with an accuracy of 97.87% for the linear kernel. K-NN model showed an accuracy of 97.37% while Naïve Bayes showed 93.61% accuracy and the Decision Tree showed 96.81% accuracy. We can summarize by saying that the SVM model showed the most efficient and effective accuracy out of all the 4 models used for the classification of benign and malignant tumours.

VII. FUTURE SCOPE

The future scope for this project is very wide. We can consider predicting whether the gene is mutating or not and if so, which one is mutating and what are the chances of one having breast cancer. In addition to it, we can try predicting which type of breast cancer is the patient prone to and what are other types of tumours and cancers might develop so that further diagnosis and tests can help prevent it, as prevention is always better than reduction.

REFERENCES

[1] Hyuna Sung, PhD[1]; Jacques Ferlay, MSc, ME[2]; Rebecca L. Siegel, MPH[1]; Mathieu Laversanne, MSc[2]; Isabelle Soerjomataram, MD, MSc, PhD[2]; Ahmedin Jemal, DMV, PhD[1]; Freddie Bray, BSc, MSc, PhD[2], "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", CA CANCER J CLIN 2021;71:209–249, Volume 71, Number 3 May/June 2021, p. 209

[2] Report of National Cancer Registry Programme (ICMR-NCDIR), Bengaluru, India 2020, Chp. 7, p. xvi

[3] Ritu Madhan, Shivani Kalariya, "Design and Development of Prosthetic Brassieres for Breast Cancer Patients", Journal of Scientific Research Institute of Science, Banaras Hindu University, Varanasi, India, Volume 65, Issue 4, 2021

[4] Muktevi Srivenkatesh, "Prediction of Breast Cancer Disease using Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-4, February 2020

[5] Ramik Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING", Journal of Emerging Technologies and Innovative Research, May 2020, Volume 7, Issue 5

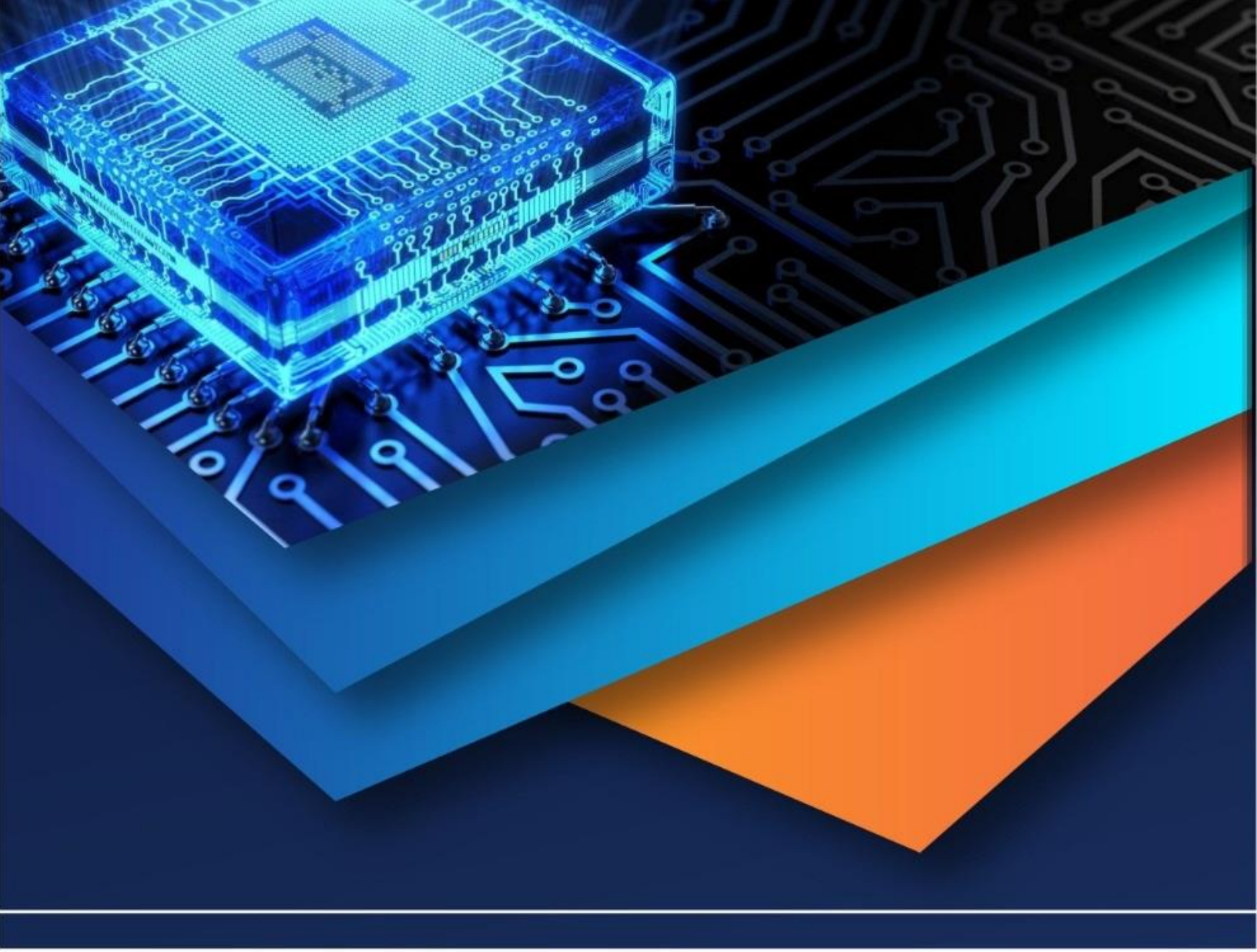
[6] Gaurav Singh ,"Breast Cancer Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 6, Issue 4 Page Number: 278-284, July-August 2020

[7] Min-Wei Huang,Chih-Wen Chen ,Wei-Chao Lin,Shih-Wen Ke,Chih-Fong Tsai, "SVM and SVM Ensembles in Breast Cancer Prediction", PLoS ONE 12(1): e0161501.

[8] Deepika Verma, Nidhi Mishra, "Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques", 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)

[9] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.

[10] Wang Haifeng; Yoon Sang Won, "Breast Cancer Prediction Using Data Mining Method", IIE Annual Conference. Proceedings; Norcross (2015): 818-828



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY