iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.47953

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 XII Dec 2022- Available at www.ijraset.com

Comparative Study of Different Machine Learning Classifiers Using Multiple Feature Selection Techniques for Breast Cancer Classification

Anurag Deyol[1], Astu[2], Ojas Shandilya[3], Yash Nayak[4]
[1, 2, 3, 4] *Department of Electrical Engineering Netaji Subhas University of Technology, New Delhi, India*

Abstract: ***This research investigates use of several Machine Learning classifiers under feature selection methods: Without Dimensionality reduction, using Correlation Coefficient Score, using Voting Classifier, and using Tree Based Feature Selection. The different ML Classifiers used in this research are: Logistic Regression, Decision Trees, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbours (KNN) and Naïve Bayes Classifier. These classification models are run on data generated from processing mammography scans to extract shape, texture, size and other spatial features from the tumour contour. The performance of these ML classifiers is evaluated by performance metrics like: Precision Score, Recall Score, F1 Score, and Accuracy Score. The dataset used for the purpose of our study was The Wisconsin Breast Cancer Dataset for both training and testing. The comparison of these results helps us better understand the nature of these classifiers for such classification problems, give us more insights on feature engineering and selection, and their potential use in clinical trials. After computing the results, we were able to get accuracy levels as high as 97.9% and were able to reach accuracy between 90-95% in general.***

Keywords: Machine Learning, Classifier, Correlation, Logistic Regression, Decision Trees, SVM, Random Forest, KNN, Naïve Bayes.

I. INTRODUCTION

Breast cancer is a form of cancer that originates in breast tissue. Signs of breast cancer may include a breast lump, a change in its shape, dimpling of the skin, fluid leaking from the nipple, and a red or scaly skin patch. Patients with distant disease spread may experience bone pain, swollen lymph nodes, shortness of breath, and yellow skin.

Breast cancer is one of the most common malignancies in India, affecting one in every 28 women, and is the main cause of cancer-related fatalities in women. Most commonly, breast cancer presents as a distinct lump within the breast. When a lump is felt with the fingertips, more than 80 percent of cases are detected. Mammograms, however, detect breast cancers at their earliest stages. According to medical professionals from various healthcare facilities, the primary factor contributing to women dying from cancer in its advanced stages is a lack of early identification of the disease. Women have been urged to maintain a vigilant watch over their health to avoid ignoring any signs of cancer. Early discovery helps to manage a lot of cases and lower the fatality rate. The best time to detect breast cancer is before it develops large enough to feel or cause symptoms, when it is easier to cure. The risk factors for breast cancer include, but are not limited to:

1) *Family history:* Breast cancer may be passed on genetically from parent to child.
2) *Formation of lumps in breast:* Later in age, Women with benign breast lumps have an increased risk of developing breast cancer.
3) *Dense breast tissue:* Women with denser breast tissue have a higher risk of developing breast cancer
4) *Age:* Likelihood of developing breast cancer increases with age, especially after 50 years.
5) *Diet and lifestyle choices:* Consumption of tobacco, high fat diet and alcohol is directly associated with increased chances of breast cancer.
6) *Radiation exposure:* Regular X-ray and CT scan exposure may raise the chance of developing breast cancer.
7) *Obesity:* Overweight women are more likely to develop breast cancer.
8) *Oestrogen exposure:* Due to longer exposure to oestrogen, women who began menstruation sooner or women who experience menopause later than average have an increased risk of breast cancer. Most Hormone Replacement Therapy forms raise the risk of breast cancer in women.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Depending on how quickly the disease is detected, treatment for breast cancer can be exceedingly effective, with 90 percent or greater survival rates. Systemic therapy (antitumor drugs administered orally or intravenously) is used to treat and/or reduce the risk of cancer metastasis (metastasis). Endocrine (hormone) therapy, chemotherapy, and, in rare cases, targeted biologic therapy are anti-cancer medications (antibodies). Overall, breast cancer has a relative 5-year survival rate of 90%. This suggests that 90% of women diagnosed with breast cancer live at least 5 years after diagnosis. 84 percent is the relative 10-year survival rate, while the invasive 15-year survival rate is 80%. Thus, early detection and treatment can improve breast cancer patients' survival rates.

II. DATA SET

For training and testing, the Wisconsin Breast Cancer Data Set available from the UCI Machine Learning Repository, an online open-source repository, was utilised.. The data set includes 569 incidents collected over a three-year period by Dr William H Wolberg, W. Nick Street and Olvi L. Mangasarian from the University of Wisconsin Hospitals. 357 of the given 569 instances were benign, while 212 were malignant. Among the 32 total features for each instance, ten are real-valued:

1) Radius (average distance between the centre and perimeter)
2) Texture (standard deviation of grey-scale values)
3) Perimeter
4) Area
5) Smoothness (local variation in radius lengths)
6) Compactness (perimeter^2 / area – 1)
7) Concavity (severity of concave portions of the contour)
8) Concave Points (number of concave portions of the contour)
9) Symmetry
10) Fractal Dimensions (coastline approximations -1)

The mean, standard error and largest (mean of the three largest values) of these ten features were computed for each image, resulting in a total of 30 features. The remaining two features are patient's id number and diagnosis where M stands for Malignant and B stands for Benign. All the features were recorded with four significant digits with no missing values.

III.MACHINE LEARNING CLASSIFIERS

Machine Learning is commonly divided into three main learning paradigms, supervised, unsupervised, & reinforcement learning. In supervised learning, the machine is trained on a set of data instances that have been labelled to yield the desired outcome. However, unsupervised learning lacks predetermined data sets and any idea of the expected outcome, making the goal more difficult to attain but has the added advantage of needing no labelled data. In case of reinforcement learning the machine learns to take correct actions in an environment in order to maximize the notion of a cumulative reward.

Classification is one of the most prevalent methods utilized in supervised learning. It employs labelled historical data to develop a predictive model for the future. In the field of medicine, large databases holding patient records with their symptoms and diagnosis are maintained by clinics and hospitals. Therefore, researchers utilize this information to develop classification models capable of drawing inferences from historical cases. With machine-based aid and the large amount of medical data available today, medical inference has gotten simpler.

A. Logistic Regression

It is a technique use to find out the relationship between independent and dependent. Also, logistic regression is used as a method of predictive modelling in machine learning. If there is a relationship between the input or output variables, then regression algorithm is used in it

B. Support Vector Machine (SVM)

SVM is a commonly used supervised learning algorithm for evaluating and identifying trends. Incorporating the learning model and algorithm into this method enables the solution of classification and regression analysis problems. The objective of SVM is to obtain a hyperplane that precisely separates d-dimensional data into its groups. A SVM model depicts the data as points in space, mapped so that the data of different categories are separated by a gap that is as large as possible. Then, new data are mapped into the same space, and their anticipated category membership is predicated on which side of the separation they exist.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. K-Nearest Neighbours

A supervised machine learning approach. Both classification and regression issues can be solved with this approach. The output of a categorization task is always discrete. A regression problem's output is an actual number (a value including a decimal point). KNN is a learning algorithm, but it is also referred to as a lazy algorithm because it does not use the training data for generalisation, i.e., there is no explicit training phase or it is very minimal, and all the training data is required during the testing phase. It analyses one instance of the testing data at a time, examines the class of the 'K' nearest training instances, and predicts the class of the testing instance as the one with the highest occurrences in the neighbourhood. K Nearest Neighbours is employed for statistical estimation and pattern recognition. It is already considered a non-parametric method.

D. Naïve Bayes

Based on Bayes' theorem, it is a classification approach. We assume predictor independence between classes. By employing the Naive Bayes theorem, we calculate the probability of a particular tuple belonging to a class attribute by multiplying the probabilities of the attribute values of the tuple with the probability of the class. The anticipated class for a tuple is the class with the highest probability. The Bayesian mathematical theorem underlies a number of mathematical classification techniques, including Bayesian algorithms. The Bayes Classifier is an easier-to-use statistical algorithm. The classification algorithm Naive Bayes illustrates the relationship between the independent variables and the target variable.

Let X = $\{X_1, X_2, X_3 \ldots\ldots\ldots\ldots X_n\}$ is the sample set, and $Y = Y_1, Y_2, Y_3 \ldots\ldots Y_m$ is the class set.

$$P(X|Y_i) = \frac{\mathrm{P(X|}Y_i)\mathrm{P}(Y_i)}{P(X)} = \frac{\left(P(X_1|Y_i) \times P(X_2|Y_i) \times P(X_3|Y_i) \ldots\ldots P(X_n|Y_i)\right) \times P(Y_i)}{P(X_1) \times P(X_2) \times P(X_3) \ldots\ldots P(X_n)}$$

E. Decision Tree

The decision tree is a function that receives a vector of attribute-value pairs as input and produces a single result, the decision, as output.

A decision tree divides the space of a function into axis-parallel rectangles or hyperplanes. Decision Trees are a nonparametric supervised learning technique applicable to classification and regression issues. The goal is to develop a model capable of forecasting the value of the target variable by discovering fundamental decision-making principles from data features. Inputs may be discrete or continuous in general; however, all inputs in this case are discrete and have Boolean values. They are optimal for several classification and recognition issues requiring big data sets and sophisticated information representation [8,9]. Here, the decision tree classifier selects an attribute as the dividing attribute, splits the tree recursively using other attributes, and organises it hierarchically with the root nodes indicating the class mark attribute. At each level, the dividing property yields the highest information gain.

F. Random Forrest

Random forests classifier (RFC) is one of the most efficient learning methods for high-dimensional classification and skewed situations. It has proven to be quite effective in pattern recognition and machine learning. The high variance of tree classifiers is a drawback. In practise, it is not unusual for a little change to the training data set to result in the formation of a tree. that is radically different. Consequently, a simple decision tree approach rarely generalises the results. The random forest classifier is comprised of a vast number of individual decision trees that operate as an ensemble. Each individual tree generates a result, a final vote is conducted, and the class that receives the most votes become the model's prediction. To create an RFC, bootstrapping and feature selection are performed. Bootstrapping is a resampling technique involving random sampling with replacement of a dataset, and feature selection is the process of generating random feature subsets for each decision tree. The combination of bootstrapping and feature selection for each decision tree improves the generalizability of the results.

IV. MODEL ANALYSIS

This section explains the performance evaluation metrics that are used in our investigation. An evaluation of the classifiers performance typically involved the utilization of a confusion matrix. A confusion matrix is a grid of 2 rows and 2 columns. We represent the actual labels and predicted labels on the x and y axes respectively. Positive and Negative refers to our predicted labels whereas True and False refers if our prediction actually correct or incorrect.

	Prediction	Outcome
True Positive (TP)	Positive	True
False Positive (FP)	Positive	False
True Negative (TN)	Negative	True
False Negative (FN)	Negative	False

A. Accuracy

Accuracy of a classifier is a measure of the classifier's ability to accurately forecast classification of instances. It is the proportion of predictions of the total number of instances in the dataset that are accurate. Notably, the accuracy is highly dependent on the threshold chosen by the classifier and, as a result, can vary between testing sets. Therefore, it is not the best method for comparing various classifiers, but it can provide a general overview of the class. Consequently, accuracy can be computed using the following formula:

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \times 100$$

B. Precision

Precision, usually referred to as confidence, is the proportion of true positives and true negatives identified as true positives. This indicates how effectively the classifier handles positive observations but offers little about negative observations.

$$Percision = \left(\frac{TP}{TP + TN}\right) \times 100$$

C. Recall

Recall, also known as sensitivity, is the proportion of positive observations that are accurately predicted to be positive. This metric is desirable, particularly in the medical field, due to the number of correctly diagnosed observations. In this study, correctly identifying a malignant tumour is more important than incorrectly identifying a benign one.

$$Recall = \left(\frac{TP}{TP + FP}\right) \times 100$$

D. F1 Score

The F1 score is the harmonious combination of accuracy and recall. Since this is the average of precision and recall, a model's F1 Score will be high if both precision and recall are high, low if both are low, and medium if one is low and the other is high.

$$F1\,Score = \left(\frac{Percision \times Recall}{Percision + Recall}\right) \times 2$$

V. RESULTS AND DISCUSSION

A. Without Using Dimensionality Reduction

Reduced dimensionality refers to a dataset's reduced number of features. In machine learning problems like classification or regression, there are often too many variables to work with. Due to the higher number of variables, it is difficult to model them. Additionally, some of these features may be redundant, adding noise to the training dataset and making it difficult to reach generalization. However, in this approach we use no dimensionality reduction, hence all 30 features are used to train and test the models. In the dataset used for this research, since there were effectively only 10 real valued features, and there mean, standard error and largest (mean of the three largest values) were computed, thus the correlation among these features is high. However, using no feature reduction helps us getting a good baseline score for the classifiers used.

Model Used	Accuracy	Precision	Recall	F1 Score
Logistic Regression	95.2%	95.5%	94.5%	95.1%
Support Vector Machine	92.0%	94.0%	89.7%	91.6%
Decision Trees	93.9%	94.2%	93.3%	93.8%
Random Forest	95.6%	96.0%	95.2%	95.7%
Naïve Bayes	93.6%	94.4%	92.5%	93.4%
K Nearest Neighbours	93.8%	94.1%	93.2%	93.8%

Fig 1: Performance metrics for model trained without dimensionality reduction.

B. Using Feature Selection using Correlation Score

Correlation is an analysis that measures the strength of association between two features. The correlation coefficient might vary from +1 to -1. Value ± 1 represents the ideal degree of connection.

1) *Positive correlation coefficient score:* The relation between the two variables will mutually related i.e., while one increases the other also increases.
2) *Negative correlation coefficient score:* The relation between the two variables will be inversely related, while one increases the other decreases.
3) *As correlation value gets closer to 0:* The relation between the two variables will be weaker.

For two features with high correlation score, both being fed to the machine only increases noise and increases training time, using only one of these features will suffice. As discussed earlier, since 30 features were achieved from 10 real valued features, thus the degree of correlation between features is very high. The heatmap for correlation between the features is shown in fig 1 is calculated by Pearson correlation coefficient.

Some classifiers like Naïve Bayes actually directly benefit from positive correlated features while random forest may indirectly benefit from them. Removing correlated features can also tackle the problem of Multicollinearity. One of Wisconsin datasets hallmark is its relatively high correlation score, thus we only selected features with score lesser than 0.6. So, the selected 9 features are smoothness_mean, radius_se, texture_se, smoothness_se, symmetry_se, fractal_dimension_se, texture_worst, symmetry_worst and fractal_dimension_worst.

Model Used	Accuracy	Precision	Recall	F1 Score
Logistic Regression	93.6%	93.9%	92.6%	93.6%
Support Vector Machine	90.4%	90.8%	89.7%	90.3%
Decision Trees	90.4%	90.8%	89.7%	90.3%
Random Forest	96%	96.4%	95.5%	96.0%
Naïve Bayes	93.9%	94.6%	94.5%	94.6%
K Nearest Neighbours	94.8%	94.1%	93.2%	93.8%

Fig 2: Performance metrics for model trained with features selected from correlation score.

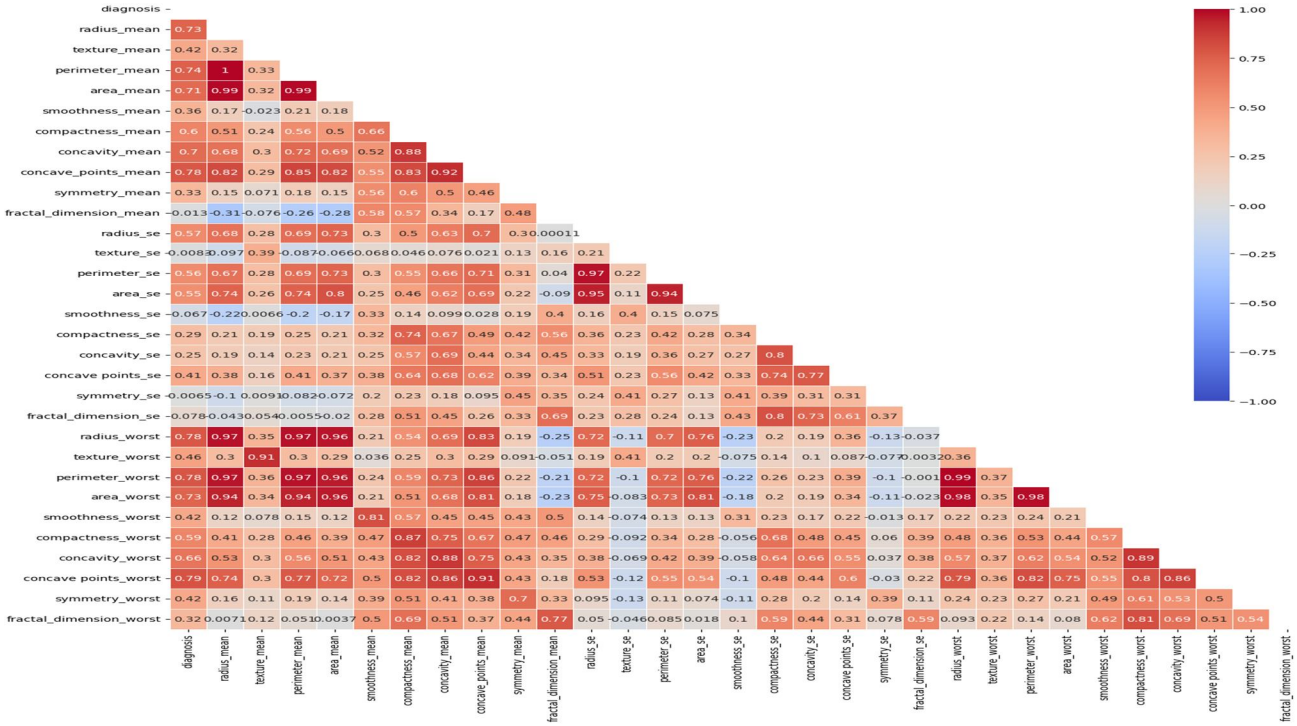


Fig 3: Pearson Correlation Score Heatmap for all 31 features.

C. Using Voting Classifier

A voting classifier is a machine learning model that is trained on an ensemble of many models and predicts an output label based on the models with the highest probability of the selected label. It basically combines each classifier's findings and votes for the final output. The idea behind this approach is that different models may learn different patterns in the data and using not one, but all these models may mask each other's shortcomings to produce better results. There are two possible forms of voting criteria:

1) *Hard Voting:* Votes are based on the output label that is anticipated.
2) *Soft Voting:* Voting is determined by the output label's predicted likelihood.

Model Used	Accuracy	Precision	Recall	F1 Score
Soft Voting	97.9%	98.5%	95.2%	96.7%
Hard Voting	97.4%	98.4%	94.9%	96.2%

Fig 4: Performance metrics for voting classifier-based model.

The voting classifier used for this investigation consisted of all the previously used classifiers to produce its final output.

D. Using Extra Trees Classifier Based Feature Selection

In this approach we use EXRTA (Extremely Randomized) Trees Classifier to compute feature importance's, which is then used to discard the irrelevant features. Extra Trees Classifier is a sort of ensemble learning that integrates the classification results of several de-correlated decision trees gathered in a "forest". Conceptually, it is quite like a Random Forest Classifier, but there are significant differences. to Random Forest classifier in two aspects, using the whole original sample rather than using bootstrapped replicas and by randomly selecting feature rather than choosing optimum feature split points. These differences motivate the reduction of both bias and variance, which makes Extra Trees algorithm much faster (almost 20-40%) however leads to slightly worse or better performance depending on the application (almost ±5%).

Model Used	Accuracy	Precision	Recall	F1 Score
Logistic Regression	94.9%	95.2%	94.4%	94.9%
Support Vector Machine	91.2%	93.4%	90.0%	90.9%
Decision Trees	92.6%	92.9%	91.7%	92.5%
Random Forest	96.0%	96.0%	95.7%	96.0%
Naïve Bayes	94.1%	94.5%	93.3%	94.1%
K Nearest Neighbours	93.3%	93.3%	92.4%	93.3%

Fig 5: Performance metrics for model trained with tree-based feature selection.

VI. ANALYSIS OF RESULTS

The highest performance was achieved by voting classifier, this can be explained by the fact that using different classifiers and then combining their results makes different patterns in the data to be understood by the model. Some classifiers may work better than others, but combining them and voting for the final results can help in masking of shortcoming for pattern recognition amongst the classifiers. In general, random forest approach works better than other classifiers with or without feature selection.

Slightly worse performance can be seen in non-feature selected models than feature selected models in logistic regression and SVM. This is due to availability of lesser amount of data to obtain the hyperplane in SVM. For KNN due to removal of correlated data that introduced noise into the model, when removed caused slight improvement in the overall performance. Extra Trees Classifier based feature selection produces better results than Pearson Correlation Score based feature selection as it does not depend on a fixed formula and randomization in feature selection process for forest construction leads to better generalization.

The reason why feature selection is important is to improve upon the training times of these models. For very large datasets using redundant features increases the training time without improving on performance all that much. We can observe that performance of feature selected models is only about 1-2% lesser than non-feature selected models while being trained faster.

VII. CONCLUSION

In this analysis, we have illustrated six machine learning classifiers with different forms of feature selection. Afterwards we used two different feature selection paradigms and used voting classifier for the distinction between benign and malignant breast cancer tumors. This application is capable to classify Breast Cancer into benign and malignant in a few minutes. The performance metrics for all classifiers are upwards of 90% and reaches upwards of 97% for voting classifier.

REFERENCES

[1] N. Karankar, P. Shukla and N. Agrawal, "Comparative study of various machine learning classifiers on medical data," 2017 7th International Conference on Communication Systems and Network Technologies (CSNT), 2017, pp. 267-271, doi: 10.1109/CSNT.2017.8418550.J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[2] Dana Bazazeh, Raed Shubair. "Comparative study of machine learning algorithms for breast cancer detection and diagnosis" , 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016

[3] Priyank Jain, Shriya Sahu, “Analysis of different machine learning classifiers on MP election commission and breast cancer big dataset", Materials Today: Proceedings, 2021,

[4] Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.

[5] Y. Xiao, W. Huang and J. Wang, "A Random Forest Classification Algorithm Based on Dichotomy Rule Fusion," 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2020, pp. 182-185, doi: 10.1109/ICEIEC49280.2020.9152236.

[6] Ahmad Taher Azar, Hanaa Ismail Elshazly, Aboul Ella Hassanien, Abeer Mohamed Elkorany, “A random forest classifier for lymph diseases,Computer Methods and Programs in Biomedicine”,Volume 113, Issue 2,2014,

[7] S. Harika, T. Yamini, T. Nagasaikamesh, S. H. Basha, S. Santhosh Kumar, Mrs. S. Sri DurgaKameswari, “Alzheimers Disease Detection Using Different Machine Learning Algorithms”, doi: 10.22214/ijraset.2022.46937, IJRASET, 2022, vol. 10.

[8] Amreen Khanum D, Prof. Kavitha G, Prof. Mamatha H S, “Parkinson’s Disease Detection using Machine Learning Algorithms,” doi: 10.22214/ijraset.2022.46272, IJRASET, 2022, vol. 10

[9] Diksha, Monika, Palvi, Pankaj Verma, “A review on Machine Learning: Application and Algorithms”, International Journal of Research in Engineering and Science (IJRES), vol. 10.10, 2022

[10] Patil, Vaishnavi, Shravani Burud, Goutami Pawar, Tanaya Rayajadhav, and Sunil B. Hebbale. "Breast Cancer Detection using MATLAB Functions." Advancement in Image Processing and Pattern Recognition vol. 3, no. 2 (2020).

[11] Deepti Sharma, Rajneesh Kumar, Anurag Jain. "Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning" , Measurement: Sensors, 2022

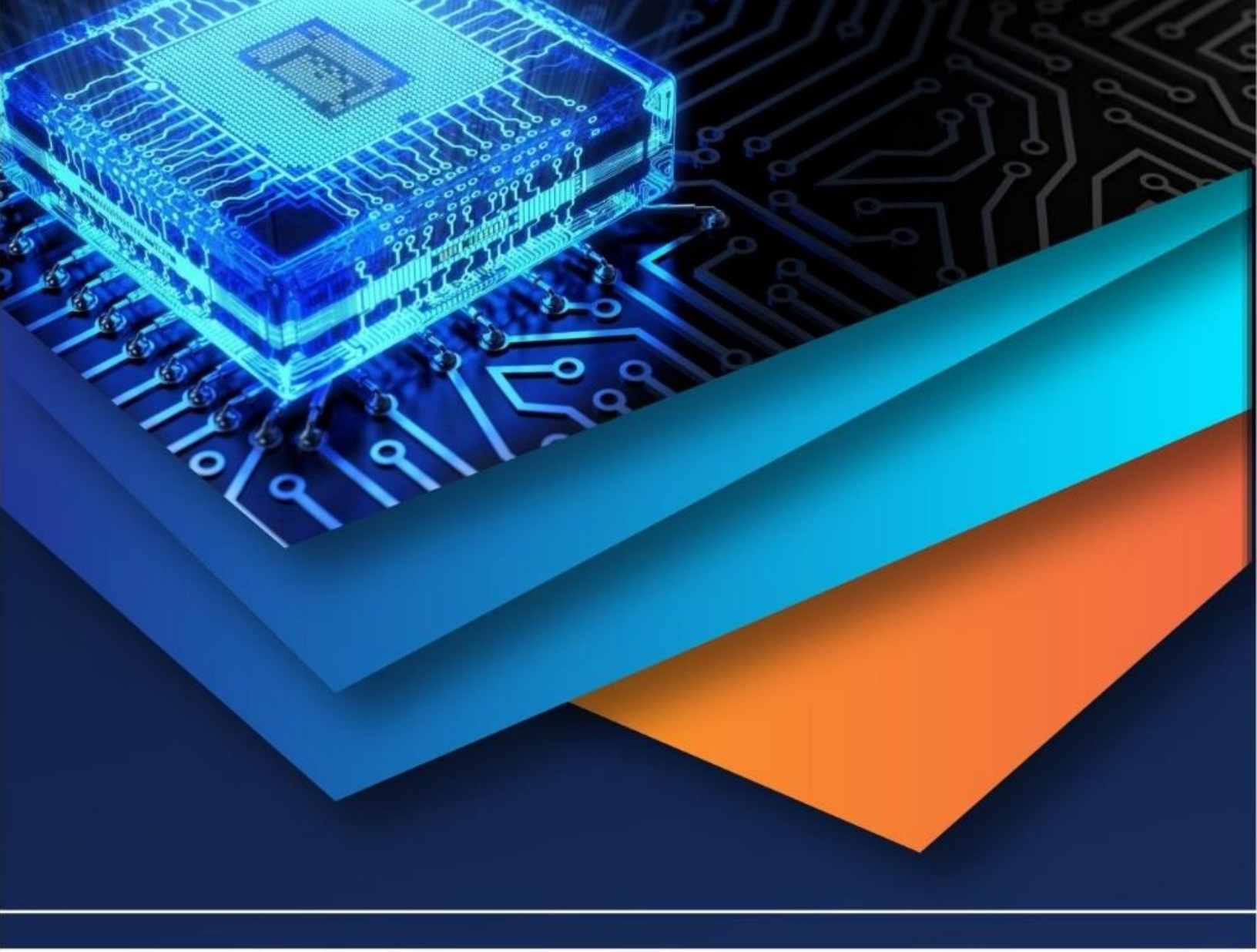
[12] Yunfeng Wu. "Breast Cancer Diagnosis Using Neural-Based Linear Fusion Strategies" , Lecture Notes in Computer Science, 2006

[13] Khorshid, Shler Farhad, Adnan Mohsin Abdulazeez, and Amira Bibo Sallow. "A comparative analysis and predicting for breast cancer detection based on data mining models." Asian Journal of Research in Computer Science vol. 8, no. 4 (2021): 45-59.

[14] Abdallah, Yousif MY, Sami Elgak, Hosam Zain, Mohammed Rafiq, Elabbas A. Ebaid, and Alaeldein A. Elnaema. "Breast cancer detection using image enhancement and segmentation algorithms." Biomedical Research 29, no. 20 (2018): 3732-3736.

[15] Nassif, Ali Bou, Manar Abu Talib, Qassim Nasir, Yaman Afadar, and Omar Elgendy. "Breast cancer detection using artificial intelligence techniques: A systematic literature review." Artificial Intelligence in Medicine (2022): 102276.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)