iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42742

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 V May 2022- Available at www.ijraset.com

Machine Learning with Applications in Breast Cancer Detection

Garvit Kaushik[1], Vedant Golash[2], Divyanshu[3], Saksham Garg[4]

***Abstract:** This briefing will give an overview of Machine Learning Techniques (MLT) and their applications. It's used to identify patterns in data, analyze trends and make predictions using algorithms. It evolved from a study in computational theory and pattern recognition to touch industries that range from education to government. It is widely considered a key tool in decision-making. We present a review on the most recent ML methods used in modeling cancer progression. We present the latest works using ML techniques to model patient outcomes or cancer risk, in light of the increasing use of ML methods for cancer research.*

I. BACKGROUND

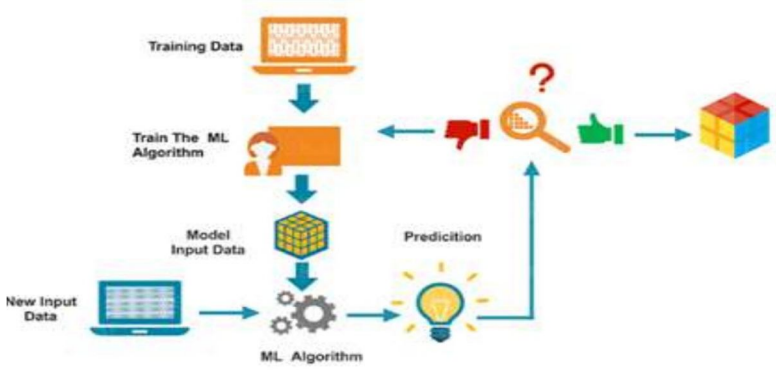
According to the American Cancer Society, the 10-year survival rate for breast cancer is 83%. Breast cancer patients who are only diagnosed with the disease will have a five-year relative survival rate that is 99%. This stage is where 62% of all cases are diagnosed. This study aims to evaluate the accuracy and efficiency of the ML models used for prediction of survival time in breast cancer patients.

II. PROBLEM DESCRIPTION

Due to increasing complexity of breast cancer treatment protocols and patient samples, it is difficult to predict the survival rate. A better personalized treatment and treatment plan could be possible with reliable and validated predictions. This would also help to control the growth of cancer. In good clinical practice, clinicians often use data from multiple sources, such as medical records and clinical laboratory tests. This allows for more accurate diagnosis, treatment, and prognosis. The three main domains of prediction and prognosis for cancer development are risk assessment, prediction or prediction of susceptibility to cancer, and prediction or relapse prediction. The first domain is the prediction of the likelihood of developing certain types of cancer before the patient is diagnosed. The second concern is the prediction of cancer recurrence based on diagnostics and treatment. The third case is concerned with predicting several parameters that characterize cancer development and treatment following diagnosis. These include survival time, drug sensitivity and progression. Both the survivability rate (p) and the likelihood of cancer relapse depend heavily on the quality of the diagnosis (p). Around 40% of all ML studies related to breast cancer prediction focused on patient survival. There are many examples of machine learning-based approaches that have been applied to various datasets. It is obvious that major clinical studies use models related to artificial neural networks (ANN) or support vector machines (SVM), with statistical methods used for validation. This has helped to overcome some of the problems associated with validation and classification. It is evident that there is a need to increase the ML impact on survival time prediction studies for breast cancer in the areas of generality and better accuracy as well as validation.

III. WHAT IS MACHINE LEARNING?

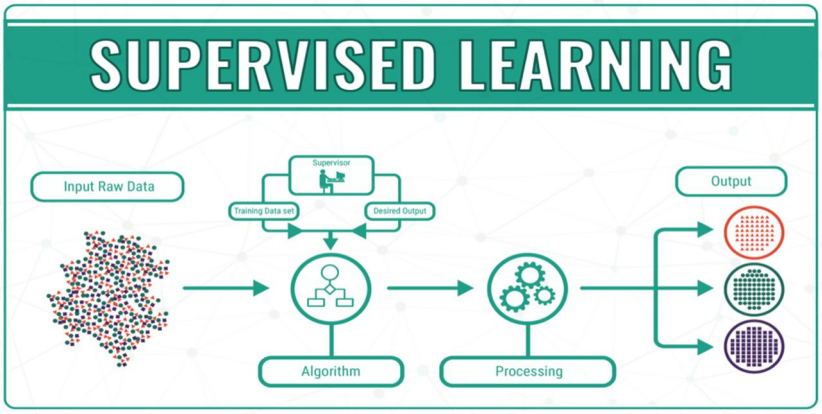
Machine Learning is a subset in artificial intelligence that focuses mainly upon machine learning from its experience and making predictions based on it. There are three types of machine learning:



1) Supervised Learning – Train Me!
2) Unsupervised Learning - I can learn by myself
3) Reinforcement Learning – My life, my rules! (Hit & Trial)

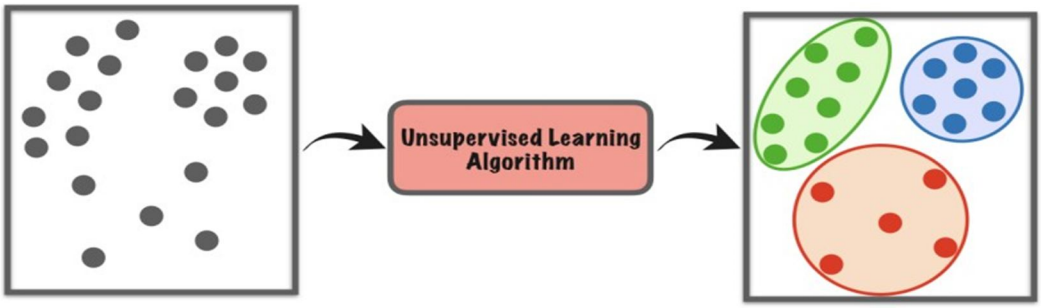
A. What is Supervised Learning?

Supervised learning is where the learning is guided and monitored by a teacher. A dataset acts as a teacher, and it is responsible for training the model or machine. After the model is trained, it can make predictions or make decisions based on new data.



B. What is Unsupervised Learning?

The model learns from observation and discovers patterns in the data. When the model is given a dataset it automatically discovers patterns and relationships by creating clusters. It cannot add labels to the clusters. For example, it can't say that this is a group of mangoes or apples, but it will separate all the mangoes from the apples.



Let's say we present images of mangoes, bananas, and apples to the model. Based on patterns and relationships, it creates clusters from the data and divides it into those clusters. If new data is added to the model, it will add it to one of those created clusters.

C. What is Reinforcement Learning?

This is how an agent interacts with the environment to determine the best outcome. It is based on the hit-and-run method. A point is awarded to the agent for correct answers. The model then trains itself based on the positive reward points. Once it is trained, it can predict new data.

IV. RISE OF MACHINE LEARNING

Modern Machine Learning Algorithms can overcome strict static program instructions to make data-driven predictions that help companies make better decisions.

IDC predicts that Machine Learning spending will increase from $12 billion in 2017 up to $57.6 million by 2021. Machine Learning patents have grown at a 34% CAGR between 2013-2017, making them the third fastest growing category of patents. MemSQL and O'Reilly Media conducted a survey to determine how Machine Learning was being used in the workplace. It found that 61% of respondents cited Machine Learning as the most important data initiative in their company for the next year. 74% of all respondents thought Machine Learning was a game- changer.

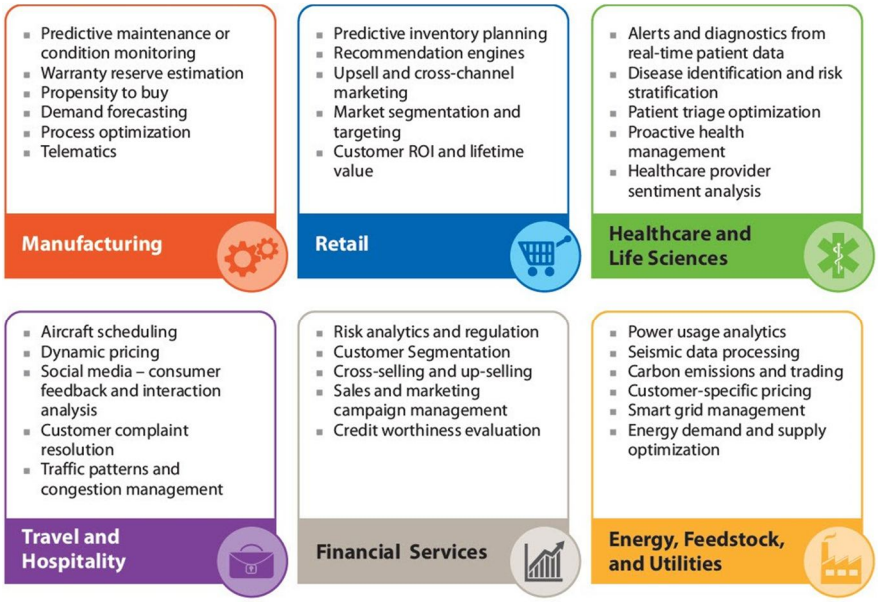


Figure 2: Machine Learning applications across industries

Machine Learning has already helped doctors diagnose, make marketing more personal, spot potential fraud cases across many fields, translate obscure legalese into plain language, predict the movement of stock markets, eliminate false alarms at concerts and stadiums, to name a few.

V. METHODOLOGY

We obtained the UCI breast cancer dataset and used Google Colaboratory for the purposes of coding. Our method uses classification techniques such as Support Vector Machine (SVM), K- Nearest Neighbors (K-NN), and Naive Bayes.

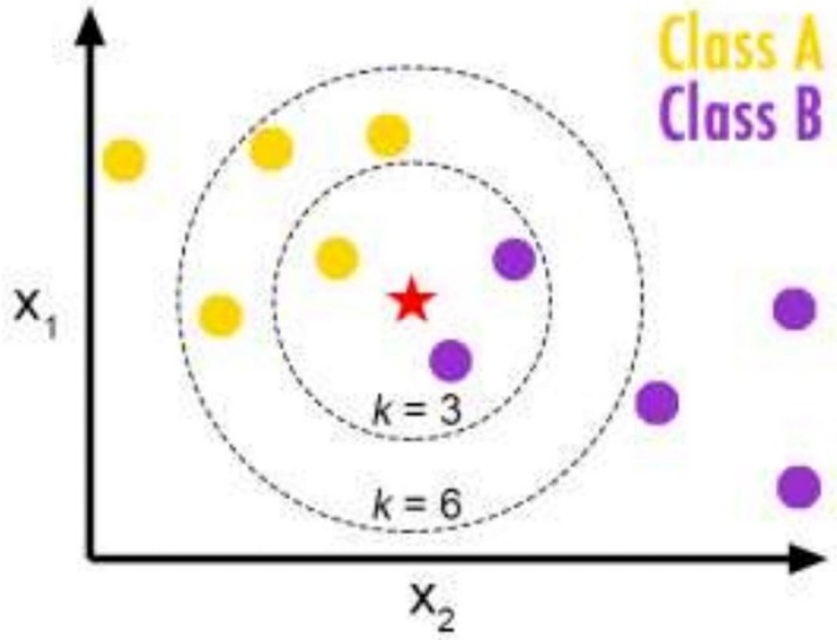
VI. MODEL SELECTION

The selection of an algorithm is the most exciting part of any machine learning model. To analyze large data sets, we can use multiple types of data mining techniques. At a high level, all these different algorithms can be divided into two categories: supervised and unsupervised learning. We have two variables in our dataset: the dependent variable and the outcome variable. Y has only two sets of values: M (Malign), or B (Benign). The Classification algorithm for supervised learning is then applied to it. Three types of Machine Learning classification algorithms have been chosen:

A. K-Nearest Neighbors

K-Nearest Neighbor can be described as a supervised machine-learning algorithm. It labels the data it receives. This method is nonparametric because it uses the closest training data points to classify test data points, rather than the dimensions (parameters). It can be used to solve both regression and classification tasks. It is used in Classification technique to classify objects based upon the k closest examples within the feature space. KNN works on the principle that similar data points are found in the same environment. KNN reduces the effort of creating a model, adapting several parameters or making further assumptions. It captures the concept of proximity using a mathematical formula called Euclidean Distance, which calculates distance between two points on a plane.

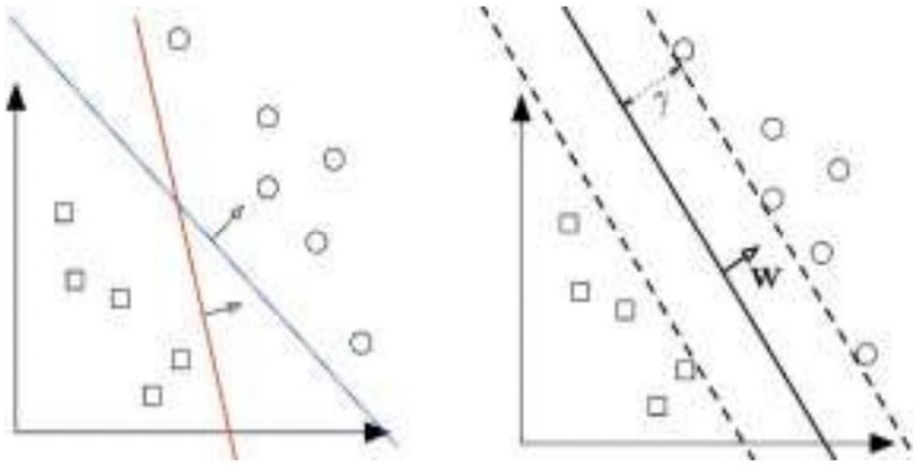
The most important step in the implementation of the KNN algorithm is choosing the value of K. K's value is not fixed. It varies depending on the data set. The prediction's stability will be lower if K is lower. The same goes for K's value. An increase in its value will reduce ambiguity, which leads to smoother boundaries as well as increased stability. K is the value that determines whether a new category will be assigned a data point. K is the number of training data points within a reasonable distance of a test data point. The test data point will be assigned to the class with the highest number of neighbors (i.e. Class with high frequency



B. Support Vector Machines

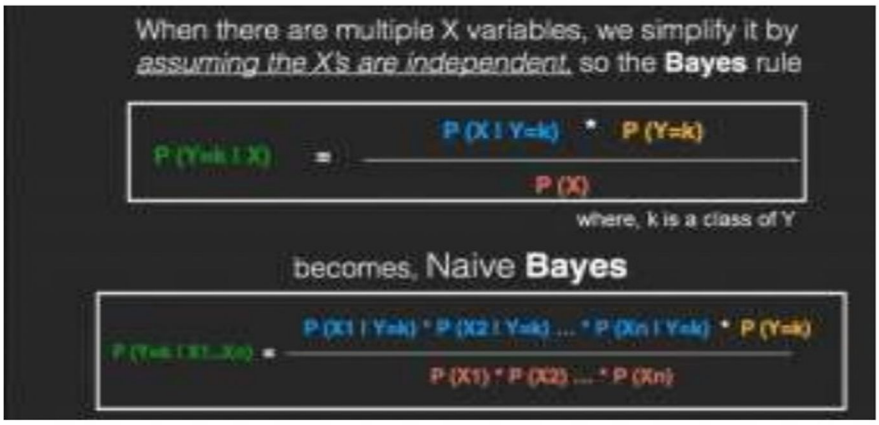
Support Vector Machine (SVM) is a supervised machine-learning algorithm that excels at pattern recognition. It is also used to train algorithms for learning how to extract classification and regression rules from data. SVM is best used when there are many features and lots of instances. It is built using a hyperplane, which is a line that extends beyond 3 dimensions. The hyper plane separates the members into the appropriate class. The mathematical equations that make up the hyperplane of SVM are used to build it. WTX=0 is the equation of the hyperplane. It is very similar to the line equation y=ax + b. W and X are vectors, where W is always the normal vector to the hyperplane. WTX is the dot product vector. SVM is used to deal with datasets that have more features. In this instance, the equation WTX=0 will be used instead of the line equation y=ax + b.

An SVM training algorithm creates a model that assigns each new data item to the appropriate category. Each data item in an SVM model is represented by points in an n dimensional space, where n is how many features are represented. The value of each feature in this space is called the value of that particular coordinate. The hyper-plane divides these two classes and is used to classify them. A new data item can be mapped in the same space, and their category is predicted based upon the side of the hyperplane that they appear.



C. Naïve Bayes

Naive Bayes, a probabilistic machine-learning algorithm that is based on Bayes Theorem and used for a variety of classification tasks, is called "Naive Bayes". Because the X's do not depend on each other, it is called Naive. It's powerful, regardless of its name.



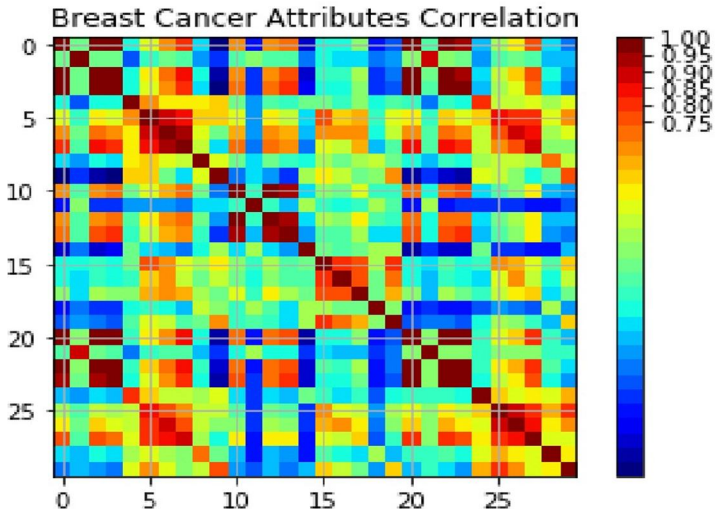
VII. DATA PRE-PROCESSING AND EXPLORATORY DATA ANALYSIS

Data preprocessing plays an important role in any data analysis task. There are many factors that can lead to missing or out-of-range data. Data that hasn't been thoroughly screened for these problems could lead to misleading results. Before any analysis can be performed, it is important to ensure that the data quality and representation are correct.

Data that contains a lot of irrelevant or unreliable information will make it more difficult to create knowledge and model. This problem can be solved by preparing and filtering data. This process can take a lot of time.

Data preprocessing includes binarization, categorization, standardization, cleaning, feature extraction, etc. We used categorization to do our work, as we had a dataset that was already suitable for classification algorithms. Two outputs are possible in our dataset: M for "Malignant" or B for "Benign". Our dataset was transformed to only contain numeric values. Instead of M, we chose 1 and B was 0. Correlation analysis is required to verify the results of these pre- processing steps.

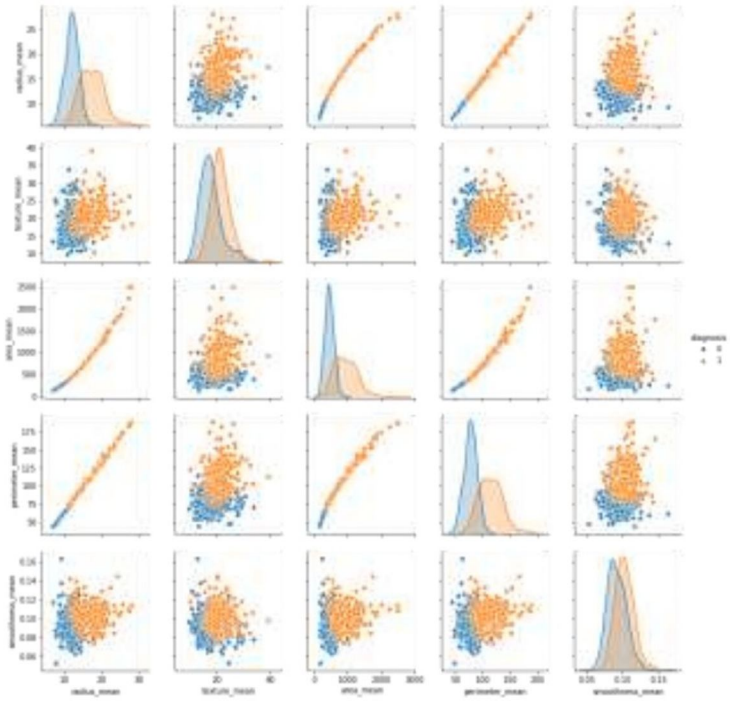
The Pearson correlation coefficient is the most common measure of dependence between quantities. The sample correlation coefficient can then be used to calculate the population Pearson correlation r between X or Y if we have a series containing n measurements of X, Y.

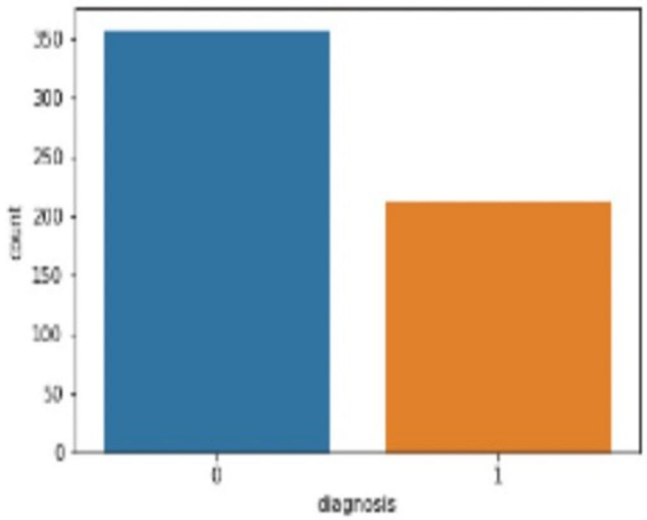


The red area around the diagonal in the output graph suggests that attributes are closely related. The blue boxes indicate negative correlations while the yellow and green areas suggest moderate correlation.

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

The next step after cleaning the data is Exploratory Data Analysis. EDA refers to the use of quantitative and visual methods to analyze and summarize a dataset, without making assumptions about its content. This is an important step before you dive into statistical modeling or machine learning. It provides context that allows you to create the right model for your problem and correctly interpret the results. The pairs plot, also known as a scatterplot matrix, is one of the most powerful tools in EDA. The pair plot can be used to show both the distribution of single variables as well as relationships between variables. Pair plots can be used to identify trends and provide follow-up analysis. Here is the pairplot of 6 features: 'radius_mean, ’texture_mean’, area_mean’, & smoothness_mean’.





VIII. EVALUATION OF THE MODELS

A. *Confusion Matrix*

It displays the difference between the predicted and observed outcomes and lists the correct and incorrect predictions, categorized by type. A confusion matrix C, by definition, is one in which Ci,j equals the number of observations that are known to be in group I and those predicted to be within group j. In binary classification, C0.0 is the count of true negatives, C1,0 is false negatives, C1,0 is true positives, and C0.1 is false positives.

1) True Negative / TN: When a case was negative or predicted to be negative

2) TP / True positive: When a case was positive, predicted positive.

3) FN/False Negative: When a case was positive, but it was predicted to be negative

4) FP/False Positive: When a case was negative, but it was predicted to be positive

Table 1: Confusion Matrix for kernel SVM

63	4
1	46

Table 2: Confusion Matrix for KNN Classifier

63	4
3	44

Table 3: Confusion Matrix for Naïve Bayes Classifier

63	4
4	43

B. *Classification Report*

The Classification report measures the quality of predictions made by a classification algorithm. This report displays the main classification metrics precision and recall on a per-class base. To predict the metrics in a classification report, you can use True Positives, False positives, True Negatives, and True Negatives. In this instance, positive and negative are the generic names of the predicted classes.

SVC RESULTS

	precision	recall	f1-score	support
0	0.98	0.94	0.96	67
1	0.92	0.98	0.95	47
accuracy			0.96	114
macro avg	0.95	0.96	0.96	114
weighted avg	0.96	0.96	0.96	114

NAIVE BAYES RESULTS

	precision	recall	f1-score	support
0	0.94	0.94	0.94	67
1	0.91	0.91	0.91	47
accuracy			0.93	114
macro avg	0.93	0.93	0.93	114
weighted avg	0.93	0.93	0.93	114

KNN RESULTS

	precision	recall	f1-score	support
0	0.95	0.94	0.95	67
1	0.92	0.94	0.93	47
accuracy			0.94	114
macro avg	0.94	0.94	0.94	114
weighted avg	0.94	0.94	0.94	114

1) *Precision:* The classifier's intuitive ability to not label a negative sample as a positive is called precision. It is the ratio of true to false positives for each class.
2) *Recall:* The classifier's intuitive ability to locate all positive samples is called recall. It is the ratio of true negatives to true positives for each class.
3) *F1 Score:* F1 score can be described as a weighted harmonic means of precision and recall. The best score is 1.0, while the worst score is 0. F1 scores are generally lower than accuracy measurements because they incorporate precision and recall into the computation. The weighted average F1 should be used for classifier models comparisons, and not global accuracy.

IX. CONCLUSION

This paper presents an original method of integrating data from two different cancer studies. It includes all details. This method is horizontal and vertical, and uses graph and document- oriented databases. It has no data loss and shows good performance. The paper presents the results of a set of ML-based models that were used to predict survival time in breast cancer.

Different algorithms take into account different aspects and use different mechanisms. While ANNs have dominated BC diagnosis for decades, it is evident that alternative ML methods have been used to improve healthcare systems and offer a wider range of options to doctors.

The future potential development of this work is to apply ML modeling to other data with differing features regarding the survival prognosis for patients. Our Python-based workflow will be improved and made web-based with additional services.

X. FUTURE SCOPE

Even with significant advances in genetics and modern imaging technology the majority of breast cancer patients are still surprised when they receive their diagnosis. [p]

Some people find it too late. A later diagnosis can lead to anxiety, aggressive treatment and uncertain outcomes. Early detection and early diagnosis of breast cancer has been an important pillar of research in the field.

The new deep learning model was created by a team of researchers from MIT's Computer Science and Artificial Intelligence Laboratory CSAIL and Massachusetts General Hospital (MGH). It can be used to predict from a mammogram if a patient will develop breast cancer in the near future. Their model was trained on mammograms from more than 60,000 patients at MGH. The model then learned how to recognize subtle patterns in breast tissue which could indicate malignancy. [7]

Thermalytix, a breakthrough invention by Bengaluru-based Niramai, uses Machine Learning (ML), and Artificial Intelligence(AI) to overcome technological and socio-cultural limitations. Thermalytix, which combines thermal image technique with artificial intelligence, is a revolutionary diagnostic tool that detects breast cancer early.

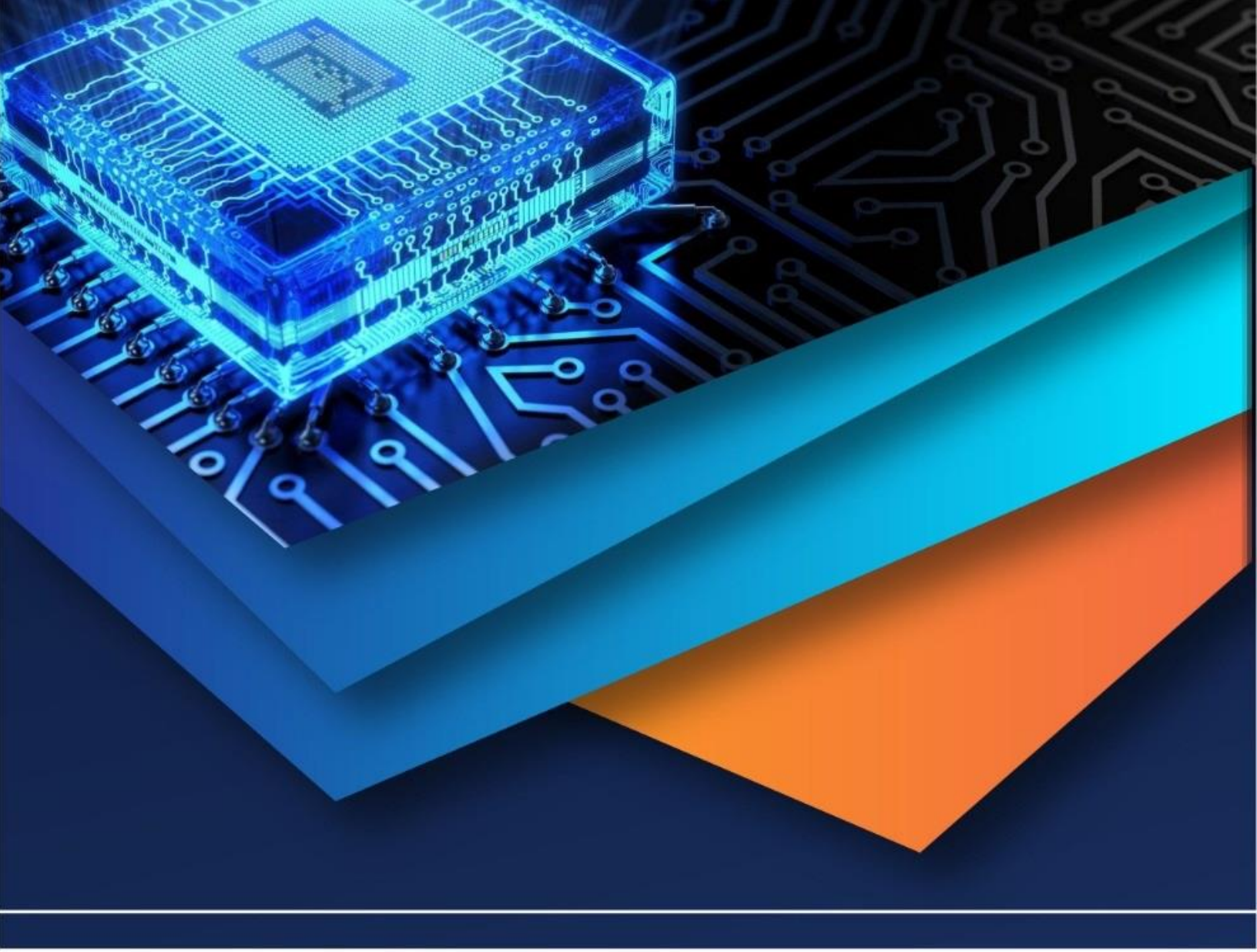
REFERENCES

Journal Article

[1] Cruz, J.A.; Wishart, D.S. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Inform. 2006,
[2] Introduction, Machine Learning Methods] [2] I. Rish; An empirical study of the Naive Bayes classifier [Definitions and Background]
[3] Tattersall, M.H.N.; Ellis, P.M.; Butow, P.N.; Hagerty, R.G.; Dimitry, S. Communicating prognosis in cancer care: A systematic review of the literature. Ann. Oncol. 2005, 16, 1005–1053
[4] Marcelo Gagliano, John Van Pham, Boyang Tang, Hiba Kashif, James Ban Applications of Machine Learning in Medical Diagnosis [History Overview]
[5] Theodoros Evgeniou, Massimiliano Pontil, Support Vector Machines – Theory and Applications [A brief Overview to the SVM theory]
[6] Padraig Cunningham, Sarah jane Delaney, K-nearest neighbor classifiers [Introduction].
[7] A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction, Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, Regina Barzilay, Radiology 2019; 292:60–66.
[8] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[9] Yue, Wenbin & Wang, Zidong & Chen, Hongwei & Payne, Annette & Liu, Xiaohui. (2018). Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. Designs. 2. 13. 10.3390/designs2020013.

Website Article

[1] Silicon Valley Data Science- https://www.svds.com/value-exploratory-data-analysis/.
[2] [Niramai][Thermalytix] https://www.niramai.com/about/thermalytix/.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY