Explainable ML and DL models on Breast Cancer Data

Project Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Technology in the field of Computer Science and Engineering

BY

ABHILASH BANERJEE (123190803001)

ABHISHEK CHAKRABORTY (123190803002)

AMIT KUMAR DUBEY (123190803010)

ANITABHA DAS (123190803015)

ARYA RAJ (123190803028)

Under the supervision

Of

PROF. APURBA PAUL

Department of Computer Science and Engineering

JIS College of Engineering

Block-A, Phase-III, Kalyani, Nadia, Pin-741235

West Bengal, India

May,2023

CERTIFICATE

This is to certify that **Abhilash Banerjee(123190803001) , Abhishek Chakraborty (123190803002),Amit Kumar Dubey(123190803010),Anitabha Das(123190803015) ,Arya Raj (123190803028)** has completed his/her project entitled **Explainable ML and DL models on Breast Cancer data,** under the guidance of **Prof. Apurba Paul** in partial fulfillment of the requirements for the award of the **Bachelor of Technology in Computer Science and Engineering** from JIS college of Engineering (An Autonomous Institute) is an authentic record of their own work carried out during the academic year 2022-23 and to the best of our knowledge, this work has not been submitted elsewhere as part of the process of obtaining a degree, diploma, fellowship or any other similar title.

--------------------------------	------------------------------	-----------------------------
Signature of the Supervisor	**Signature of the HOD**	**Signature of the Principal**

Place: Kalyani

Date: 17/05/2023

ACKNOWLEDGEMENT

The analysis of the project work wishes to express our gratitude to Apurba Paul for allowing the degree attitude and providing effective guidance in development of this project work. His conscription of the topic and all the helpful hints he provided contributed greatly to the successful development of this work, without being pedagogic and overbearing influence.

We also express our sincere gratitude to, Dr. BikramJit Sarkar Head of the Department of Computer Science and Engineering of JIS College of Engineering and all the respected faculty members of Department of CSE for giving the scope of successfully carrying out the project work.

Finally, we take this opportunity to thank Prof. **(Dr.) Partha Sarkar**, Principal of JIS College of Engineering for giving us the scope of carrying out the project work.

Date: 17/05/2023

……………………………………………………

Abhilash Banerjee, B. TECH in Computer Science and Engineering, 4thYEAR/8th SEMESTER

Univ Roll—123190803001

……………………………………………………

Abhishek Chakraborty, B. TECH in Computer Science and Engineering,4thYEAR/8th SEMESTER

Univ Roll--123190803002

…………………………………………………

Amit Kumar Dubey, B. TECH in Computer Science and Engineering,4thYEAR/8th SEMESTER

Univ Roll—1231908030010

……………………………………………………………………

Anitabha Das, B. TECH in Computer Science and Engineering,4thYEAR/8th SEMESTER

Univ Roll—123190803015

……………………………………………………………………

Arya Raj, B. TECH in Computer Science and Engineering,4thYEAR/8th SEMESTER

Univ Roll--123190803028

List of Figures

List of Tables

CONTENTS

ABSTRACT

In recent years, breast cancer has grown in importance. Women seem to be developing breast cancer at a considerably higher rate. If the illness is not detected at all, it has already become fatal, and in many cases, limb amputation is the only method to stop it if it is discovered too late. Therefore, a reliable indicator of this problem can aid in accurate diagnosis. This paper's major objective is to apply various machine learning classification algorithms to accurately forecast the target class and enhance it by evaluating the usefulness of specific aspects of the original Wisconsin Breast Cancer dataset (WDBC) for breast cancer diagnosis prediction. The highest performing algorithm was identified after classifiers were run on the dataset, and then useful dataset features were examined to boost performance even further. We employed the Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest, CNN, RNN, FFNN algorithms in this paper. Performance metrics were employed to compare the results in this case, including accuracy, precision, recall, F-measure and ROC, AUC Curve. Among the algorithms utilised, convolutional neural network produced the best results based on the values of the performance indicators. On the same dataset, we also sought to optimise our suggested model and compare it to other cutting-edge methods given by other researchers. We utilised Explainable AI (XAI) to visualise the models. Explainable AI refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts.

Keywords

- Breast Cancer, Dataset, FFNN, CNN, RNN, Gaussian Naïve Bayes, Random Forest, SVM, Logistic Regression, Decision Tree, Explainable AI(XAI), LIME, SHAP.

1. INTRODUCTION

1.1 Breast Cancer

Breast-cancer is among the most serious illnesses/diseases in India, causing many deaths in the current situation. Due to changes in food and lifestyle, the number of cancer cases in women is increasing day by day. It is the second most common cause of death in women in the world. This uses concepts of Deep learning (DL) and Machine learning (ML) to predict breast cancer based on the data obtained. This cancer is produced by abnormal growth of fatty and fibrous tissues, and the different phases of cancer are caused by cancer cells spreading throughout the tissue. This is one of the most common cancers that affects women, but other types of cancer and those who are affected by them can be treated greatly, according to a government survey, when compared to breast cancer. The various phases of breast cancer are identified via proper treatment and detailing. If we do not provide proper therapy to our patients, it will result in their death. A number of methods for establishing an accurate diagnosis of breast cancer have been presented. Because the dataset contains a variety of distinct report attributes, machine learning may be easily applied to the dataset for prediction. Even by using Technology which is not fully automatically designed to give the output. Hence here we propose the fully automatic classification and prediction of breast cancer based on dataset. Using deep learning technique. This learning technique is recognized as the best method to predict and classify for image dataset.

The second major cause of women's death is breast cancer (after lung cancer). 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women’s death is estimated. Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumor that can often be seen on an x-ray or felt as a lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body. The cause of Breast Cancer includes changes and mutations in DNA. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumors and angiosarcoma are less common. There are many algorithms for classification of breast cancer outcomes. The side effects of Breast Cancer are – Fatigue, Headaches, Pain and numbness (peripheral neuropathy), Bone loss and osteoporosis. There are many algorithms for classification and prediction of breast cancer outcomes. The present paper gives a comparison between the performance of eight algorithms: Logistic Regression, Random Forest, SVM, FFNN, CNN, RNN, which are among the most influential data mining algorithms. It can be medically detected early during a screening examination through mammography or by portable cancer diagnostic tool. Cancerous breast tissues change with the progression of the disease, which can be directly linked to cancer staging. The stage of breast cancer (I–IV) describes how far a patient's cancer has proliferated. Statistical indicators such as tumor size, lymph node metastasis, and distant metastasis and so on are used to determine stages. To prevent cancer from spreading, patients have to undergo breast cancer surgery, chemotherapy, radiotherapy and endocrine. The goal of the research is to identify and classify Malignant and Benign patients and intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms & Deep Learning can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy. 10-fold cross validation test which is a Machine Learning Technique is used in JUPYTER to evaluate the data and analyze data in terms of effectiveness and efficiency.

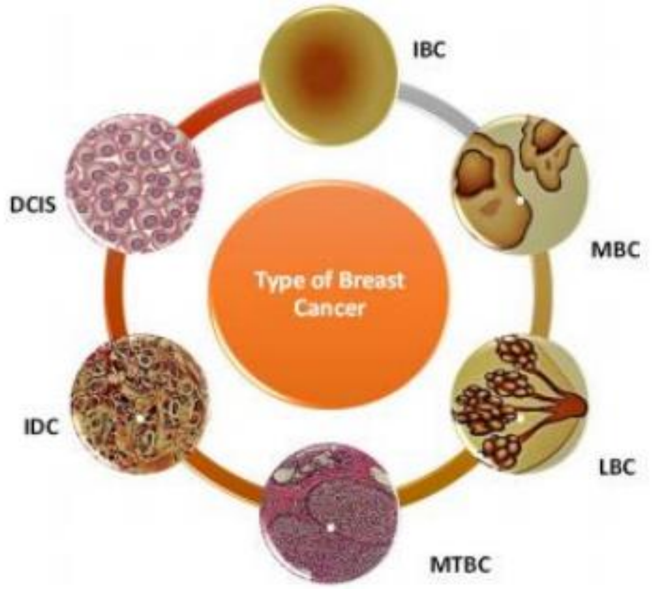


Fig 1: - The various kinds of breast cancer. [Apollo hospitals.]

Cancer is a disease that occurs when there are changes or mutations that take place in genes that help in cell growth. These mutations allow the cells to divide and multiply in a very uncontrolled and chaotic manner. These cells keep increasing and start making replicas which end up becoming more and more abnormal. These abnormal cells later form a tumor. Tumors, unlike other cells, don’t die even though the body doesn’t need them. The cancer that develops in the breast cells is called breast cancer. This type of cancer can be seen in the breast ducts or the lobules. Cancer can also occur in the fatty tissue or the fibrous connective tissue within the breast. These cancer cells become uncontrollable and end up invading other healthy breast tissues and can travel to the lymph nodes under the arms. There are two types of cancers. Malignant and Benign. Malignant cancers are cancerous. These cells keep dividing uncontrollably and start affecting other cells and tissues in the body. They spread to all other parts of the body, and it is hard to cure this type of cancer. Chemotherapy, radiation therapy and immunotherapy are types of treatments that can be given for these types of tumors. Benign cancer is non-cancerous. Unlike malignant, this tumor does not spread to other parts of the body and hence is much less risky than malignant. In many cases, such tumors don’t really require any treatment. Breast cancer is most diagnosed in women of ages above 40. But this disease can affect men and woman of any age. It can also occur when there’s a family history of breast cancer. Breast Cancer has always had a high mortality rate and according to statistics, it alone accounts for about 25% of all new cancer diagnoses and 15% of all cancer deaths among women worldwide. Scientists know about the dangers of it from very early on, and hence there’s been a lot of research put into finding the right treatment for it. Breast cancer detection is done with the help of mammograms, which are basically X-rays of the breasts. It’s a tool which can help detect and diagnose breast cancer. But detection is not easy due to different kinds of uncertainties in using these mammograms. The result of a mammogram are images that can show any calcifications or deposits of calcium in the breasts. These don’t always have to be cancerous.

1.2 Machine Learning

Machine Learning is a subcategory of Artificial Intelligence which allows systems to automatically learn and understand data from experience without the system being programmed to do so. It helps software applications become better at predicting outcomes for various types of problems. The basic idea of ML is to take in input data and use different algorithms to help it predict outcomes and update these outcomes when new data is available as input. The procedures used with machine learning are like that of data mining and predictive modeling. Both require scanning through huge amounts of data to search for any type of pattern in the data and then modify the program accordingly. Machine Learning has been seen many a times by individuals while shopping on the internet. They are then shown ads based on what they were searching for earlier on. This happens because many of these websites use machine learning to customize the ads based on user searches and this is done in real time. Machine learning has also been used in other various places like detecting fraud, filtering of spam, network security threat detection, predictive maintenance and building news feeds.

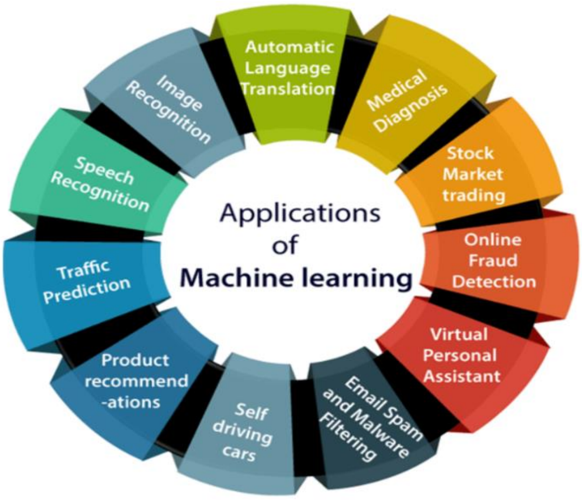

Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

Fig 2: - Application of ML [JavaTpoint]

Machine Learning methods:

- **Supervised learning** – Here both the input and output is known. The training dataset also contains the answer the algorithm should come up with on its own. So, a labeled dataset of fruit images would tell the model which photos were of apples, bananas and oranges. When a new image is given to the model, it compares it to the training set to predict the correct outcome.
  i) Classification ii) Regression

- **Unsupervised learning** – Here input dataset is known but output is not known. A deep learning model is given a dataset without any instructions on what to do with it. The training data contains information without any correct result. The network tries to automatically understand the structure of the model.
  i)Clustering ii) Association

- **Semi-supervised learning** – This type comes somewhere between supervised and unsupervised learning. It contains both labelled and un-labelled data.
- **Reinforcement learning** – In this type, AI agents are trying to find the best way to accomplish a particular goal. It tries to predict the next step which could possibly give the model the best result at the end.

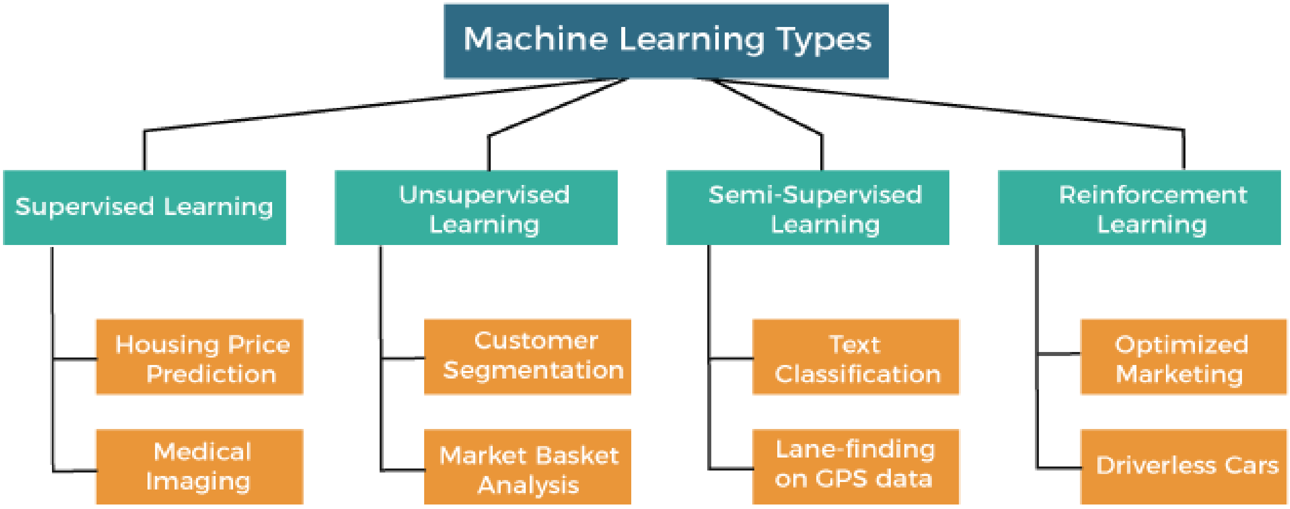


Fig 3: - Real Life Use of Machine Learning [JavaTpoint]

Machine Learning Architecture-

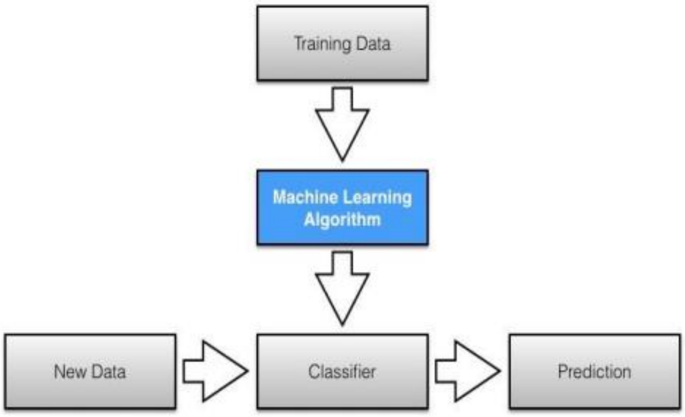


Fig 4: - ML architecture [JavaTpoint]

Advantages of Machine Learning-

- **Automation-**Machine Learning is one of the driving forces behind automation, and it is cutting down time and human workload. Automation can now be seen everywhere, and the complex algorithm does the hard work for the user. Automation is more reliable, efficient, and quick. With the help of machine learning, now advanced computers are being designed. Now this advanced computer can handle several machine-learning models and complex algorithms. However, automation is spreading faster in the industry but, a lot of research and innovation are required in this field.
- **Scope of Improvement-**Machine Learning is a field where things keep evolving. It gives many opportunities for improvement and can become the leading technology in the future. A lot of research and innovation is happening in this technology, which helps improve software and hardware.
- **Enhanced Experience in Online Shopping and Quality Education-**Machine Learning is going to be used in the education sector extensively, and it will be going to enhance the quality of education and student experience. It has emerged in China; machine learning has improved student focus. In the e-commerce field, Machine Learning studies your search feed and give suggestion based on them. Depending upon search and browsing history, it pushes targeted advertisements and notifications to users.

- **Wide Range of Applicability-**This technology has a very wide range of applications. Machine learning plays a role in almost every field, like hospitality, ed-tech, medicine, science, banking, and business. It creates more opportunities.

Disadvantages of the Machine Learning-

- **Data Acquisition-**The whole concept of machine learning is about identifying useful data. The outcome will be incorrect if a credible data source is not provided. The quality of the data is also significant. If the user or institution needs more quality data, wait for it. It will cause delays in providing the output. So, machine learning significantly depends on the data and its quality.
- **Time and Resources-**The data that machines process remains huge in quantity and differs greatly. Machines require time so that their algorithm can adjust to the environment and

learn it. Trials runs are held to check the accuracy and reliability of the machine. It requires massive and expensive resources and high-quality expertise to set up that quality infrastructure. Trials runs are costly as they would cost in terms of time and expenses.

- **Results Interpretations-**One of the biggest advantages of Machine learning is that interpreted data that we get from them cannot be hundred percent accurate. It will have some degree of inaccuracy. For a high degree of accuracy, algorithms should be developed so that they give reliable results.
- **High Error Chances-**The error committed during the initial stages is huge, and if not corrected at that time, it creates havoc. Biasness and wrongness have to be dealt with separately; they are not interconnected. Machine learning depends on two factors, i.e., data and algorithm. All the errors are dependent on the two variables. Any incorrectness in any variables would have huge repercussions on the output.
- **Social Changes-**Machine learning is bringing numerous social changes in society. The role of machine learning-based technology in society has increased multifold. It is influencing the thought process of society and creating unwanted problems in society. Character assassination and sensitive details are disturbing the social fabric of society.
- **Elimination of Human Interface-**Automation, Artificial Intelligence, and Machine Learning have eliminated human interface from some work. It has eliminated employment opportunities. Now, all those works are conducted with the help of artificial intelligence and machine learning.
- **Changing Nature of Jobs-**With the advancement of machine learning, the nature of the job is changing. Now, all the work are done by machine, and it is eating up the jobs for human which were done earlier by them. It is difficult for those without technical education to adjust to these changes.
- **Highly Expensive-**This software is highly expensive, and not everybody can own it. Government agencies, big private firms, and enterprises mostly own it. It needs to be made accessible to everybody for wide use.
- **Privacy Concern-**As we know that one of the pillars of machine learning is data. The collection of data has raised the fundamental question of privacy. The way data is collected and used for commercial purposes has always been a contentious issue. In India, the Supreme court of India has declared privacy a fundamental right of Indians. Without the user's permission, data cannot be collected, used, or stored. However, many cases have

come up that big firms collect the data without the user's knowledge and use it for their commercial gains.

- **Research and Innovations-**Machine learning is an evolving concept. This area has not seen any major developments yet that fully revolutionized any economic sector. The area requires continuous research and innovation.

Overfitting in Machine Learning- Overfitting occurs when our machine learning model tries to cover all the data points, or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model. The overfitted model has low bias and high variance.

How to avoid the Overfitting in Model- Both overfitting and underfitting cause the degraded performance of the machine learning model. But the main cause is overfitting, so there are some ways by which we can reduce the occurrence of overfitting in our model.

- Cross-Validation
- Training with more data
- Removing features
- Early stopping the training
- Regularization
- Ensembling

Underfitting in Machine Learning- Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data.

How to avoid the Underfitting in Model-

- By increasing the training time of the model.
- By increasing the number of features.

Epoch- Epochs are defined as the total number of iterations for training the machine learning model with all the training data in one cycle. In the Epoch, all training data is used exactly once. Further, in other words, Epoch can also be understood as the total number of passes an algorithm has completed around the training dataset. A forward and a backward pass together counted as one pass in training.

Usually, when a machine learning model is trained, then it requires a little number of Epochs. An Epoch is often mixed up with iteration.

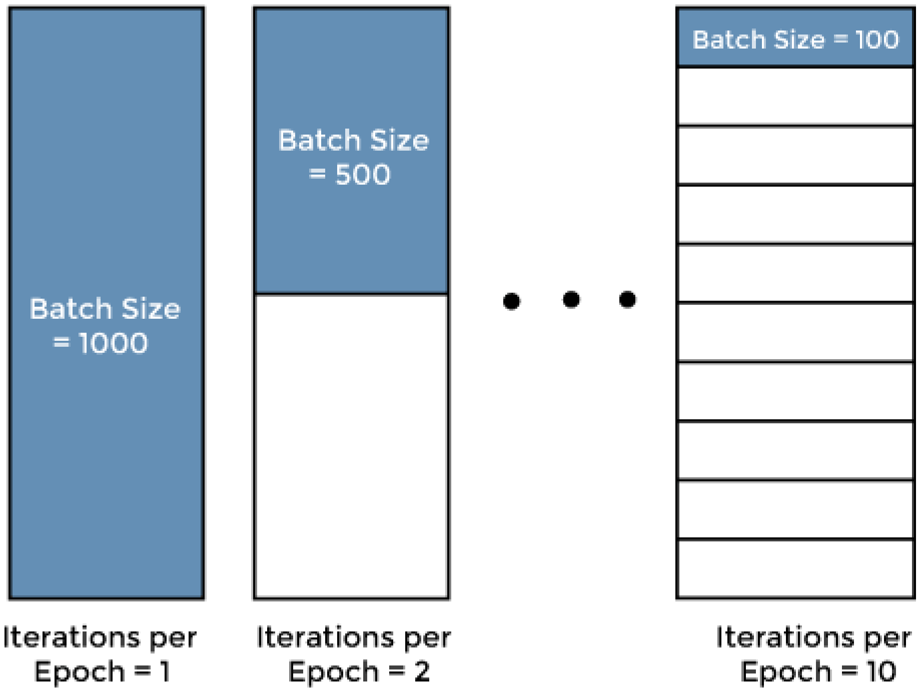


Fig 5:- Description of Epoch [JavaTpoint]

Bias and Variance- Machine learning is a branch of Artificial Intelligence, which allows machines to perform data analysis and make predictions. However, if the machine learning model is not accurate, it can make predictions errors, and these prediction errors are usually known as Bias and Variance. In machine learning, these errors will always be present as there is always a slight difference between the model predictions and actual predictions.

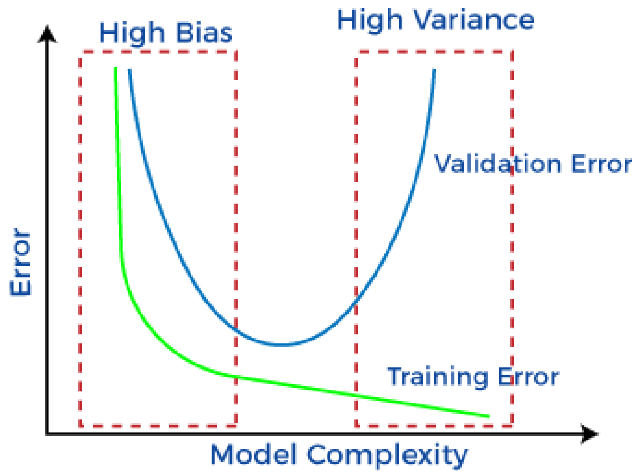


Fig 6: -Graph between bias and Variance [JavaTpoint]

How to choose the right machine learning algorithm-

- **Categorize the problem**- If it is labeled data, it's a supervised learning problem. If it's unlabeled data with the purpose of finding structure, it's an unsupervised learning problem. If the solution implies optimizing an objective function by interacting with an environment, it's a reinforcement learning problem. If the output of the model is a number, it's a regression problem. If the output of the model is a class, it's a classification problem. If the output of the model is a set of input groups, it's a clustering problem.
- **Understand Your Data/ Analyze the Data/ Process the data/ Transform the data-** Data itself is not the end game, but rather the raw material in the whole analysis process. Successful companies not only capture and have access to data, but they're also able to derive insights that drive better decisions, which result in better customer service, competitive differentiation, and higher revenue growth. The process of understanding the data plays a key role in the process of choosing the right algorithm for the right problem. Some algorithms can work with smaller sample sets while others require tons and tons of samples. Certain algorithms work with categorical data while others like to work with numerical input.
- **Find the available algorithms-** After categorizing the problem and understanding the data, the next milestone is identifying the algorithms that are applicable and practical to implement in a reasonable time.

- **Optimize hyperparameters-** There are three options for optimizing hyperparameters, grid search, random search, and Bayesian optimization.

Feature Selection: - This module is used for feature selection/dimensionality reduction on given datasets. This is done either to improve estimators' accuracy scores or to boost their performance on very high-dimensional datasets.

Feature Extraction: - This module is used to extract features in a format supported by machine learning algorithms from the given datasets consisting of formats such as text and image.

The Main Difference Between Feature Selection and Feature Extraction: -

Feature Extraction transforms arbitrary data, such as text or images, into numerical features that is understood by machine learning algorithms. Feature Selection on the other hand is a machine learning technique applied on these (numerical) features.

1.3 Deep Learning

Deep learning is a branch of machine learning which is based on artificial neural networks. It is capable of learning complex patterns and relationships within data. In deep learning, we don't need to explicitly program everything. It has become increasingly popular in recent years due to the advances in processing power and the availability of large datasets. Because it is based on artificial neural networks (ANNs) also known as deep neural networks (DNNs). These neural networks are inspired by the structure and function of the human brain's biological neurons, and they are designed to learn from large amounts of data.

1. Deep Learning is a subfield of Machine Learning that involves the use of neural networks to model and solve complex problems. Neural networks are modeled after the structure and function of the human brain and consist of layers of interconnected nodes that process and transform data.

2. The key characteristic of Deep Learning is the use of deep neural networks, which have multiple layers of interconnected nodes. These networks can learn complex representations of data by discovering hierarchical patterns and features in the data. Deep Learning algorithms can automatically learn and improve from data without the need for manual feature engineering.

3. Deep Learning has achieved significant success in various fields, including image recognition, natural language processing, speech recognition, and recommendation systems. Some of the popular Deep Learning architectures include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Deep Belief Networks (DBNs).

4. Training deep neural networks typically requires a large amount of data and computational resources. However, the availability of cloud computing and the development of specialized hardware, such as Graphics Processing Units (GPUs), has made it easier to train deep neural networks.

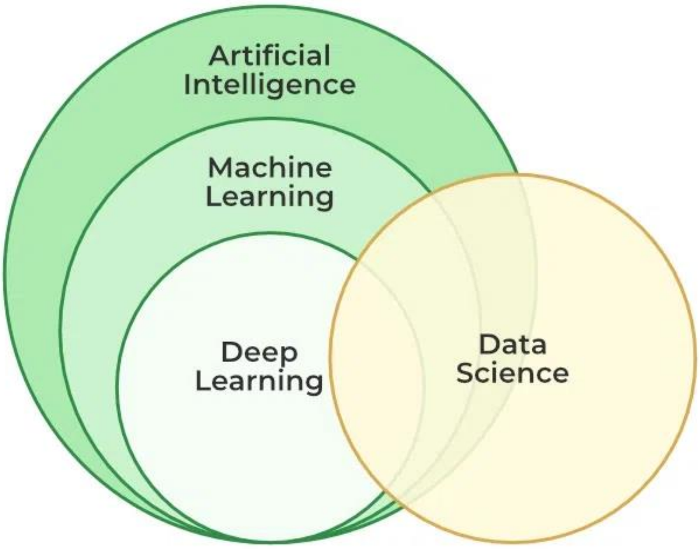


Fig 7: - Venn diagram of ml and dl [GeeksforGeeks]

Advantages of Deep Learning:

- High accuracy: Deep Learning algorithms can achieve state-of-the-art performance in various tasks, such as image recognition and natural language processing.
- Automated feature engineering: Deep Learning algorithms can automatically discover and learn relevant features from data without the need for manual feature engineering.
- Scalability: Deep Learning models can scale to handle large and complex datasets and can learn from massive amounts of data.

- Flexibility: Deep Learning models can be applied to a wide range of tasks and can handle various types of data, such as images, text, and speech.
- Continual improvement: Deep Learning models can continually improve their performance as more data becomes available.

Disadvantages of Deep Learning:

- High computational requirements: Deep Learning models require large amounts of data and computational resources to train and optimize.
- Requires large amounts of labeled data: Deep Learning models often require a large amount of labeled data for training, which can be expensive and time- consuming to acquire.
- Interpretability: Deep Learning models can be challenging to interpret, making it difficult to understand how they make decisions.
- Overfitting: Deep Learning models can sometimes be overfit to the training data, resulting in poor performance on new and unseen data.
- Black-box nature: Deep Learning models are often treated as black boxes, making it difficult to understand how they work and how they arrive at their predictions.

Artificial Neural Networks

Artificial Neural Networks contain artificial neurons which are called units. These units are arranged in a series of layers that together constitute the whole Artificial Neural Network in a system. A layer can have only a dozen units or millions of units as this depends on how the complex neural networks will be required to learn the hidden patterns in the dataset. Commonly, Artificial Neural Network has an input layer, an output layer as well as hidden layers. The input layer receives data from the outside world which the neural network needs to analyze or learn about. Then this data passes through one or multiple hidden layers that transform the input into data that is valuable for the output layer. Finally, the output layer provides an output in the form of a response of the Artificial Neural Networks to input data provided.

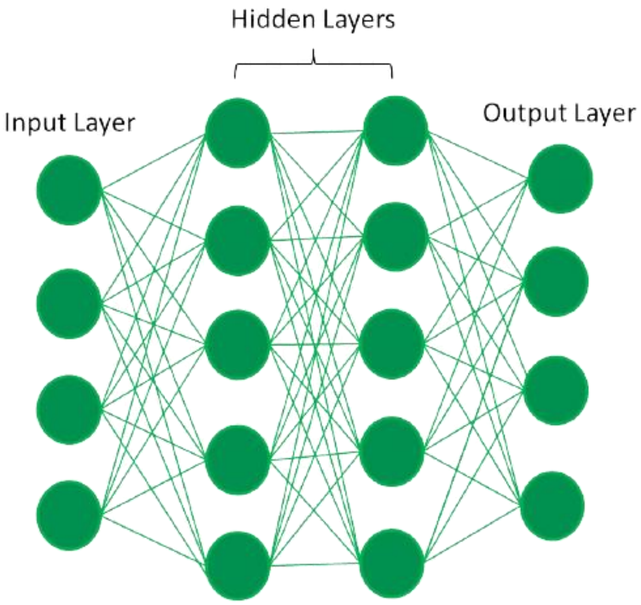


Fig 8: - Basic structure of ANN [GeeksforGeeks]

The structures and operations of human neurons serve as the basis for artificial neural networks. It is also known as neural networks or neural nets. The input layer of an artificial neural network is the first layer, and it receives input from external sources and releases it to the hidden layer, which is the second layer. In the hidden layer, each neuron receives input from the previous layer neurons, computes the weighted sum, and sends it to the neurons in the next layer. These connections are weighted means effects of the inputs from the previous layer are optimized more or less by assigning different-different weights to each input and it is adjusted during the training process by optimizing these weights for improved model performance.

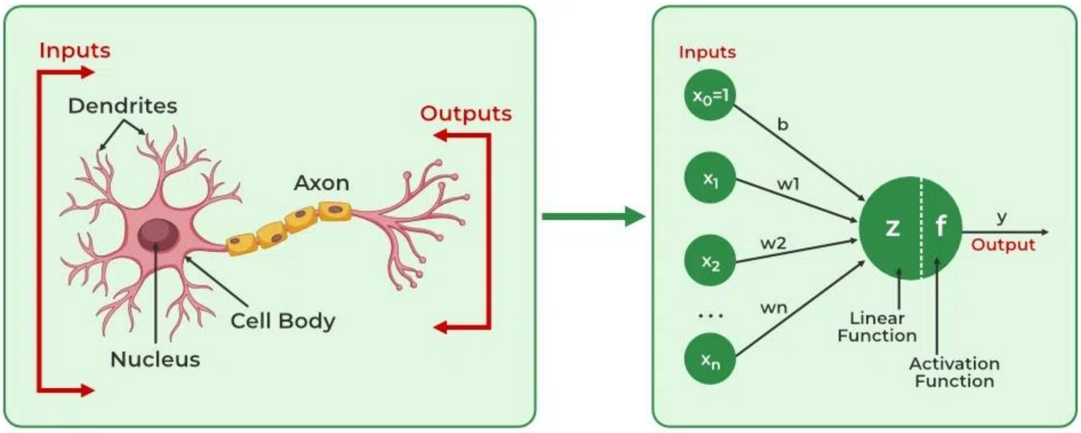


Fig 9: -Understanding the ANN [GeeksforGeeks]

Biological Neuron	**Artificial Neuron**
Dendrite	Inputs
Cell nucleus or Soma	Nodes
Synapses	Weights
Axon	Output
Synaptic plasticity	Backpropagations

Table:3- Difference between Biological Neuron and Artificial Neuron

Applications of Artificial Neural Networks-

- Social Media: Artificial Neural Networks are used heavily in social media. For example, let's take the 'People you may know' feature on Facebook that suggests people that you might know in real life so that you can send them friend requests. Well, this magical effect is achieved by using Artificial Neural Networks that analyze your profile, your interests, your current friends, and also their friends and various other factors to calculate the people you might potentially know. Another common application of Machine Learning in social media is facial recognition. This is done by finding around 100 reference points on the person's face and then matching them with those already available in the database using convolutional neural networks.
- Marketing and Sales: When you log onto E-commerce sites like Amazon and Flipkart, they will recommend your products to buy based on your previous browsing history. Similarly, suppose you love Pasta, then Zomato, Swiggy, etc. will show you restaurant recommendations based on your tastes and previous order history. This is true across all new-age marketing segments like Book sites, Movie services, Hospitality sites, etc. and it is done by implementing personalized marketing. This uses Artificial Neural Networks to identify the customer's likes, dislikes, previous shopping history, etc., and then tailor the marketing campaigns accordingly.
- Healthcare: Artificial Neural Networks are used in Oncology to train algorithms that can identify cancerous tissue at the microscopic level at the same accuracy as trained physicians. Various rare diseases may manifest in physical characteristics and can be identified in their premature stages by using Facial Analysis on the patient photos. So, the full-scale implementation of Artificial Neural Networks in the healthcare environment can only enhance the diagnostic abilities of medical experts and ultimately lead to the overall improvement in the quality of medical care all over the world.

- Personal Assistants: I am sure you all have heard of Siri, Alexa, Cortana, etc., and also heard them based on the phones you have!!! These are personal assistants and an example of speech recognition that uses Natural Language Processing to interact with the users and formulate a response accordingly. Natural Language Processing uses artificial neural networks that are made to handle many tasks of these personal assistants such as managing the language syntax, semantics, correct speech, the conversation that is going on, etc.

Advantages of Deep Learning:

- Automatic feature learning: Deep learning algorithms can automatically learn features from the data, which means that they don’t require the features to be hand-engineered. This is particularly useful for tasks where the features are difficult to define, such as image recognition.
- Handling large and complex data: Deep learning algorithms can handle large and complex datasets that would be difficult for traditional machine learning algorithms to process. This makes it a useful tool for extracting insights from big data.
- Improved performance: Deep learning algorithms have been shown to achieve state-of-the-art performance on a wide range of problems, including image and speech recognition, natural language processing, and computer vision.
- Handling non-linear relationships: Deep learning can uncover non-linear relationships in data that would be difficult to detect through traditional methods.
- Handling structured and unstructured data: Deep learning algorithms can handle both structured and unstructured data such as images, text, and audio.
- Predictive modeling: Deep learning can be used to make predictions about future events or trends, which can help organizations plan for the future and make strategic decisions.
- Handling missing data: Deep learning algorithms can handle missing data and still make predictions, which is useful in real-world applications where data is often incomplete.
- Handling sequential data: Deep learning algorithms such as Recurrent Neural Networks (RNNs) and Long Short-term Memory (LSTM) networks are particularly suited to handling sequential data such as time series, speech, and text. These algorithms have the ability to maintain context and memory over time, which allows them to make predictions or decisions based on past inputs.
- Scalability: Deep learning models can be easily scaled to handle an increasing amount of data and can be deployed on cloud platforms and edge devices.

- Generalization: Deep learning models can generalize well to new situations or contexts, as they are able to learn abstract and hierarchical representations of the data.

Disadvantages of Deep Learning:

- High computational cost: Training deep learning models requires significant computational resources, including powerful GPUs and large amounts of memory. This can be costly and time-consuming.
- Overfitting: Overfitting occurs when a model is trained too well on the training data and performs poorly on new, unseen data. This is a common problem in deep learning, especially with large neural networks, and can be caused by a lack of data, a complex model, or a lack of regularization.
- Lack of interpretability: Deep learning models, especially those with many layers, can be complex and difficult to interpret. This can make it difficult to understand how the model is making predictions and to identify any errors or biases in the model.
- Dependence on data quality: Deep learning algorithms rely on the quality of the data they are trained on. If the data is noisy, incomplete, or biased, the model’s performance will be negatively affected.
- Data privacy and security concerns: As deep learning models often rely on large amounts of data, there are concerns about data privacy and security. Misuse of data by malicious actors can lead to serious consequences like identity theft, financial loss and invasion of privacy.
- Lack of domain expertise: Deep learning requires a good understanding of the domain and the problem you are trying to solve. If the domain expertise is lacking, it can be difficult to formulate the problem and select the appropriate algorithm.
- Unforeseen consequences: Deep learning models can lead to unintended consequences, for example, a biased model can discriminate against certain groups of people, leading to ethical concerns.
- Limited to the data it’s trained on: Deep learning models can only make predictions based on the data it has been trained on. They may not be able to generalize to new situations or contexts that were not represented in the training data.
- Black box models: some deep learning models are considered as “black box” models, as it is difficult to understand how the model is making predictions and identifying the factors that influence the predictions.

Types of neural networks-

Deep Learning models can automatically learn features from the data, which makes them well-suited for tasks such as image recognition, speech recognition, and natural language processing. The most widely used architectures in deep learning are feedforward neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs).

- Feedforward neural networks (FNNs) are the simplest type of ANN, with a linear flow of information through the network. FNNs have been widely used for tasks such as image classification, speech recognition, and natural language processing.
- Convolutional Neural Networks (CNNs) are specifically for image and video recognition tasks. CNNs are able to automatically learn features from the images, which makes them well-suited for tasks such as image classification, object detection, and image segmentation.
- Recurrent Neural Networks (RNNs) are a type of neural network that is able to process sequential data, such as time series and natural language. RNNs are able to maintain an internal state that captures information about the previous inputs, which makes them well-suited for tasks such as speech recognition, natural language processing, and language translation.

Deep learning hardware requirements-Deep learning requires a tremendous amount of computing power. High performance graphical processing units (GPUs) are ideal because they can handle a large volume of calculations in multiple cores with copious memory available. However, managing multiple GPUs on-premises can create a large demand on internal resources and be incredibly costly to scale.

Artificial intelligence vs Machine learning vs Deep learning

Artificial intelligence	Machine learning	Deep learning
AI refers to the broad field of computer science that focuses on creating intelligent machines that can perform tasks that would normally require human intelligence, such as reasoning, perception, and decision-making.	ML is a subset of AI that focuses on developing algorithms that can learn from data and improve their performance over time without being explicitly programmed	DL is a subset of ML that focuses on developing deep neural networks that can automatically learn and extract features from data**.**
AI is a computer algorithm which exhibits intelligence through decision making.	ML is an AI algorithm which allows system to learn from data.	DL is a ML algorithm that uses deep (more than one layer) neural networks to analyze data and provide output accordingly.
The aim is to basically increase chances of success and not accuracy.	The aim is to increase accuracy not caring much about the success ratio.	It attains the highest rank in terms of accuracy when it is trained with large amount of data.
Three broad categories/types Of AI are: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI)	Three broad categories/types Of ML are: Supervised Learning, Unsupervised Learning and Reinforcement Learning	DL can be considered as neural networks with a large number of parameters layers lying in one of the four fundamental network architectures: Unsupervised Pre-trained Networks, Convolutional Neural Networks, Recurrent Neural Networks and Recursive Neural Networks
The efficiency Of AI is basically the efficiency provided by ML and DL respectively.	Less efficient than DL as it can't work for longer dimensions or higher amount of data.	More powerful than ML as it can easily work for larger sets of data.

Examples of AI applications include Google's AI-Powered Predictions, Ridesharing Apps Like Uber and Lyft, Commercial Flights Use an AI Autopilot, etc.	Examples of ML applications include Virtual Personal Assistants: Siri, Alexa, Google, etc., Email Spam and Malware Filtering.	Examples of DL applications include Sentiment based news aggregation, Image analysis and caption generation, etc.

Table :4- Difference Between AI, ML &DL

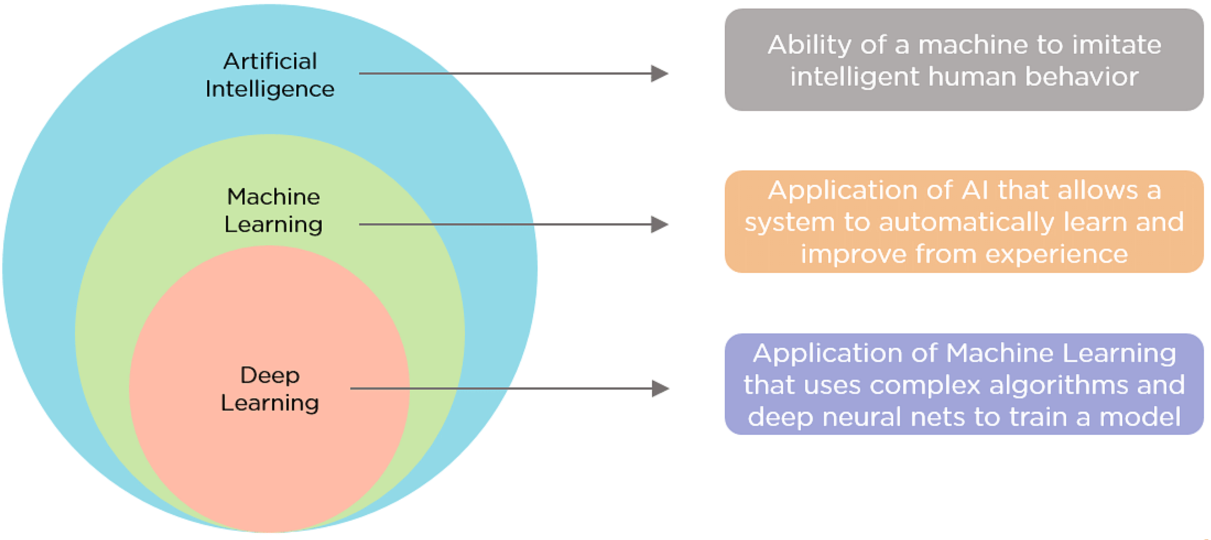

Artificial Intelligence
Machine Learning
Deep Learning
Ability of a machine to imitate intelligent human behavior
Application of AI that allows a system to automatically learn and improve from experience
Application of Machine Learning that uses complex algorithms and deep neural nets to train a model

Fig 10: - Venn Diagram of AI, ML & DL [GeeksforGeeks]

1.4 Explainable AI(XAI)

Explainable AI is a set of tools and frameworks to help you understand and interpret predictions made by your machine learning models, natively integrated with a number of Google's products and services. With it, you can debug and improve model performance, and help others understand your models' behaviour. You can also generate feature attributions for model predictions in AutoML Tables, Big Query ML and Vertex AI, and visually investigate model behaviour using the What-If Tool.

It Is a Set of Tools That Helps Us to Understand The Predictions Made By Our Machine Learning & Deep Learning Model. AI Algorithms Are Often Known As "Black Box" Means

We Don’t Know What Happening Inside We Just Give Input Get Output, So Explainable AI Helps Us to Understand How It Is Generating the Output.

Ex-

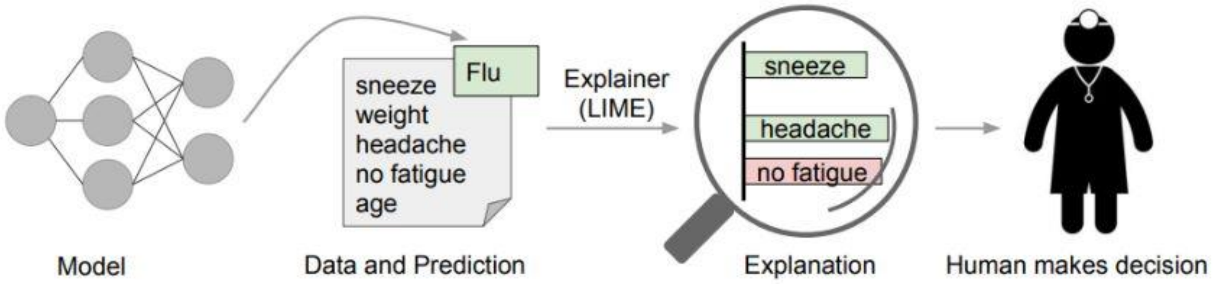


Fig 11: - Flu Prediction [Toward Data Science]

2. LITERATURE SURVEY

[1] Riddhi R. Gujar et al. “Breast Cancer Prediction Using Machine Learning” [ResearchGate 2023]

This paper reviews the use of convolutional neural networks (CNNs) in medical scans to predict breast cancer using mammograms and ultrasounds. It discusses CNN architecture, training data, methods, and performance evaluation criteria. The paper explores the benefits and drawbacks of CNNs compared to conventional techniques. While CNNs can lower false-positive outcomes and increase accuracy rates, further testing and study are needed to ensure their reliability. Moral concerns like data privacy and bias must be considered.

[2] Rashika Pandita et al. “Analysis of Breast Cancer Prediction Using Machine Learning Techniques: Review Paper” [ResearchGate 2023]

Machine learning plays a crucial role in detecting and treating various human diseases, such as breast cancer. Deep learning, a child of AI, focuses on images, providing accurate and efficient predictions. This technology has been gaining popularity in various sectors, including cancer detection. This review paper aims to analyze thirteen studies on breast cancer prediction and classification, examining various machine learning algorithms and methodologies for diagnosis. The paper presents both theoretical and non-theoretical papers executed in this field.

[3] Rahul Karmakar et al. “BCPUML: Breast Cancer Prediction Using Machine Learning Approach—A Performanc” [ResearchGate 2023]

Breast cancer is a prevalent global malignancy, affecting women worldwide. Early detection and diagnosis are crucial for survival. Machine learning algorithms are being used to diagnose breast cancer and perform performance analysis. The study uses five classifiers, including K-Nearest Neighbors, Random Forest, Decision Trees, Logistic Regression, and Support Vector Machines, on the WISCONSIN (Diagnostic) data set. Cross-validation approaches are applied to achieve exact accuracies.

Performance analysis is conducted based on each observation. This study highlights the importance of machine learning in breast cancer prediction and performance analysis.

[4] Muhammad Waqas Arshad et al. “PREDICTION AND DIAGNOSIS OF BREAST CANCER USING MACHINE LEARNING AND ENSEMBLE CLASSIFIERS” [ResearchGate 2023]

This paper investigates breast cancer fatalities annually, the most common cancer and the main cause of death in women worldwide. An accurate cancer prognosis is crucial for a healthy life.

Machine learning approaches, such as Random Forest, Logistic Regression, Xtreme Gradient, and AdaBoost Classifier, are used to assess and compare their effectiveness in early detection and prediction of breast cancer. The main objective is to identify the most effective ensemble and machine learning classifiers for accurate breast cancer detection and diagnosis.

[5] Amin Mohamed Ahsan et al. "Breast Cancer-Risk Factors and Prediction Using Machine-Learning Algorithms and Data Source: A Review of Literature" [ResearchGate 2023]

This review explores breast cancer concerns, focusing on risk factors, data sources, and machine learning algorithms for prediction. It discusses age, family history, lifestyle choices, and environmental factors, and reviews clinical, genomic, and lifestyle data sources. The paper reviews machine learning algorithms, including supervised and unsupervised learning, and assesses their performance in predicting BC risk using different data sources. The findings will be useful for healthcare professionals and researchers in the field of breast cancer.

[6] Xinkang Li et al. "Prediction of ADMET Properties of Anti-Breast Cancer Compounds Using Three Machine Learning Algorithms" [ResearchGate 2023]

This paper uses machine learning algorithms, including PLS-DA, AdaBoost, and LGBM, to predict the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of anti-breast cancer compounds. LGBM achieved the highest accuracy, precision, recall, and F1-score, making it a valuable tool for virtual screening and drug design researchers. The results suggest that LGBM can establish reliable models for predicting molecular ADMET properties, making it a valuable tool in the field.

[7] D. Shanthi et al. "A study of deep learning techniques for predicting breast cancer types'' [ResearchGate 2023]

Breast cancer is a global health issue causing high mortality rates. Traditional machine learning techniques struggle with feature extraction, while novel deep learning techniques improve diagnosis using imaging modalities like mammograms, magnetic resonance imaging, and ultrasounds. This survey analyzes the challenges of classical models and compares traditional and deep learning models.

[8] Reza Rabiei et al. "Prediction of Breast Cancer using Machine Learning Approaches" [ResearchGate 2022]

This paper aims to predict breast cancer using machine-learning approaches, using data from the Motamed Cancer Institute (ACECR) database. The study used random forest, neural network, GBT, and genetic algorithms. Models were trained with demographic and laboratory features, and then with all demographics, laboratory, and mammographic features. Combining multiple risk factors in modeling can aid early diagnosis and care plans. Data collection, storage, and management are crucial for effective disease management.

[9] Farzane Tajdini et al. "Breast Cancer Diagnosis and Prediction Using Machine Learning" [ResearchGate 2022]

This paper analyzes recent developments in cancer diagnosis using machine learning and deep learning (ML/DL) for six different types of cancer. It covers 40 papers from 2016 to 2021, focusing on breast cancer diagnosis. The study examines methodology, feature extraction methods, data modalities, and accuracy considering current difficulties. Breast cancer is the second leading cause of cancer-related death in women, with 4.5 to 5% more new cases yearly. Early detection and proper diagnosis can increase patient survival rates.

Mammography is used for breast cancer diagnosis, but waiting for reports can be time-consuming. Machine learning can help identify breast cancer quickly and accurately, improving the accuracy and efficiency of cancer prediction. The primary goals of the evaluated publications are reliable diagnosis and classification.

[10] Shiekhah A. al Binali et al. "Breast Cancer Subtypes Prediction Using Omics Data and Machine Learning Models" [ResearchGate 2022]

This paper explores the use of omics data to understand breast cancer subtyping, focusing on mRNA, DNA methylation, and copy number variation. The cancer genome atlas portal was used to acquire the data, which is highly imbalanced and challenging for machine learning methods. The objective is to identify the best predictive approach and relevant omics data for BC subtype prediction. Three issues are addressed: reducing dataset dimensionality, balancing datasets, and determining the best machine learning model. Multiple dimensionality reduction methods and classifiers were considered, and a comprehensive experimental study was conducted to identify the best approach. The results showed competitive and even better results compared to state-of-the-art approaches, with a significant drop in feature count.

[11] Ramya Challa et al. "Breast Cancer Prediction Using Machine Learning" [ResearchGate 2022]

Breast cancer is a global disease-causing numerous death annually. Early detection and diagnosis are challenging, but machine learning and data mining techniques can help predict chronic diseases like cancer. This study aims to determine the accuracy of classification algorithms like Support Vector Machine, J48, Naïve Bayes, and Random Forest and suggest the best algorithm.

[12] M. Thangavel et al. "Enhancing the Prediction of Breast Cancer Using Machine Learning and Deep Learning Techniques" [ResearchGate 2022]

This paper aims to develop a model to predict breast cancer classification using machine learning classifications and supervised learning algorithms. The model can analyze thousands of biopsies in seconds, saving lives by accurately predicting benign or malignant cancer. The model utilizes deep learning and convolutional neural networks to analyze biopsy images, ensuring accurate predictions and saving lives.

[13] Jonathan M. Ji et al. "A Novel Machine Learning Systematic Framework and Web Tool for Breast Cancer Survival Rate Assessment" [ResearchGate 2022]

Cancer research, particularly in breast cancer, relies on molecular profiling and genomic technology. However, translating vast patient data into clinically meaningful results remains a challenge. Traditional statistical methods are inadequate for tackling unstructured data. Machine learning has the potential to supersede current capabilities in understanding the correlations between gene set alterations, drug responses, and prognosis of breast cancer patients. This information would benefit scientists and physicians in developing personalized treatment strategies.

This machine learning project uses multiple approaches, including a deep learning algorithm, to build models for detecting and visualizing prognostic indicators of breast cancer patient survival rates. This project used clinical and genomic data from 1,980 primary breast cancer samples from the METABRIC database. Eight classical machine learning models and a deep learning Convolutional Neural Network (CNN) model were trained, with the deep learning model outperforming all other classifiers and achieving the highest accuracy (AUC = 0.900). The project was built in Google Colab using Python, data visualization, TensorFlow, and Keras. The CNN model demonstrated its potential for real-time prediction by end-users. A web application was developed to facilitate easy interaction with the model and obtain quick and accurate predictions.

[14] Xia Jiang et al. “Deep Learning and Machine Learning with Grid Search to Predict Later Occurrence of Breast Cancer Metastasis Using Clinical Data” [ResearchGate 2022]

This paper aims to predict breast cancer metastasis occurrence in patients using non-image clinical data. Deep neural network (DNN) learning has gained popularity for image detection and prediction, but its performance in non-image clinical data remains unanswered. The research uses grid search to improve prediction performance, but its effect on other machine learning methods is understudied. The study developed prediction models using DFNN, naïve Bayes, logistic regression, SVM, LASSO, decision tree, KNN, random forest, AdaBoost, and XGBoost, and used grid search to tune hyperparameters for all methods. Weather compared feedforward deep learning models to nine other machine learning methods, revealing that DFNN ranks 6th, 4th, and 3rd in predicting 5-year, 10-year, and 15-year BCM, respectively.

The top performing methods in 5-year BCM were XGB, RF, and KNN, while XGB, RF, and NB were the top performers in 10-year BCM. SVM, LR, LASSO, and DFNN were the top performers in 15-year BCM. RF and XGB outperform other methods when data is less balanced, while SVM, LR, LASSO, and DFNN outperform other methods when data is more balanced. The results indicate that deep learning with grid search performs similarly to other machine learning methods when using non-image clinical data. However, DFNN's computation time is significantly longer than the other nine methods.

[15] Santhosh Voruganti et al. “Breast Cancer Prediction using CNN and Machine Learning Algorithms with Comparative Analysis” [ResearchGate 2021]

Breast cancer is the second leading cause of cancer death in women, following lung cancer. It occurs when abnormal breast cells rise, and early treatment can prevent it. This paper focuses on diagnosing breast cancer using classification algorithms and data mining tools. It uses machine learning algorithms to predict breast cancer using datasets like clump thickness, uniformity, and tissue images. The accuracy, precision, recall, f-score, and ROC of each algorithm are calculated. Data mining of intelligence from previously diagnosed patients opens up new medical advancements.

[16] Yuhong Huang et al. "Prediction of Tumour Shrinkage Pattern to Neoadjuvant Chemotherapy Using a Multiparametric MRI-Based Machine Learning Model in Patients with Breast Cancer" [ResearchGate 2021]

This research uses a Kaggle breast cancer dataset and pre-processes for missing values. The model predicts breast cancer disease with accuracy between 52.63% and 98.24%, with Random Forest showing better accuracy (98.24%) compared to other machine learning algorithms. The study aims to provide awareness and diagnosis for individuals with breast cancer.

[17] Muktevi Sri Venkatesh et al. "Prediction of Breast Cancer Disease using Machine Learning Algorithms" [ResearchGate 2020]

This research uses a Kaggle breast cancer dataset and pre-processes for missing values. The model predicts breast cancer disease with accuracy between 52.63% and 98.24%, with Random Forest showing better accuracy (98.24%) compared to other machine learning algorithms. The study aims to provide awareness and diagnosis for individuals with breast cancer.

[18] M. S. Dawngliani et al. "Prediction of Breast Cancer Recurrence Using Ensemble Machine Learning Classifiers" [ResearchGate 2020]

This study proposes new criteria for predicting breast cancer patient survival using ensemble machine learning techniques. It analyzes a breast cancer dataset with 23 attributes and 575 samples from the Mizoram State Cancer Institute of Aizawl, India. The study compares ensemble machine learning classifiers and uses 10-fold cross-validation and ROC curves to evaluate their performance. Attributes are ranked based on their contribution to the prediction.

[19] Vinoth S. M. E et al. "Accurate Breast Cancer Prediction using Machine Learning Techniques." [ResearchGate 2020]

Machine learning (ML) is increasingly used in disease prediction, including breast cancer prediction. This work implements breast cancer images, preprocesses them, uses 2D median filters, and employs contrast-limited adaptive histogram equalization. Segmentation and GLCM feature extraction are employed, and an artificial neural network (ANN) is used for accurate classification.

[20] Elham Yousef Kalafi et al. “Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data” [ResearchGate 2019]

This study uses machine learning and deep learning methods to predict breast cancer survival in 4,902 patient records. Results show MLP, RF, and DT classifiers achieve 88.2%, 83.3, and 82.5% accuracy, with tumor size being the most important feature.

3.Dataset Preparation

Data Collection

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
The dataset includes ten features from each one of the cells in the sample. The Used Dataset includes 699 rows and 11 columns in which the last column represents classes of cancer cells and the rest ten columns are features of cancer cells. Nine real-valued features are computed for each cell nucleus.

Attribute Information:

a) Sample code number.

b) Clump Thickness.

c) Uniformity of Cell Size.

d) Uniformity of Cell Shape.

e) Marginal Adhesion.

f) Epithelial Cell Size.

g) Bare Nuclei.

h) Bland Chromatin.

i) Bare Nuclei.

Data Visualization

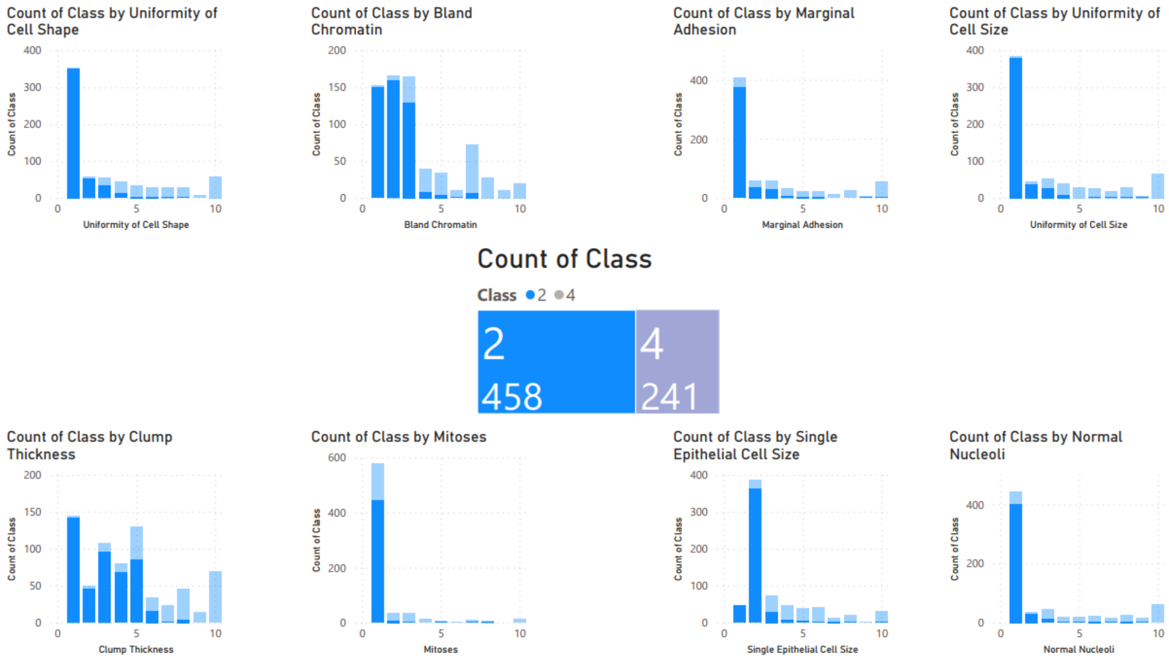


Fig 12: - Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

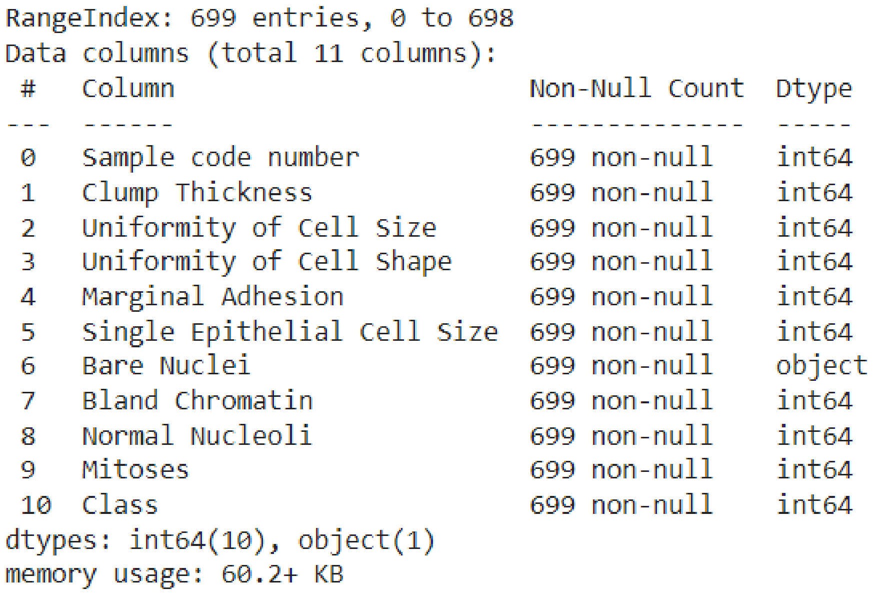
Dataset Preparation

Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.

Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

Use the SciKit-Learn library to Split the dataset into training and test data using the train_test_split method. Divide dataset into train-dataset and test-dataset in ratio 80% by 20%. Divide train and test dataset into Input as X_train , X_test and Output as Y_train, Y_test. Now the input train-dataset contains 559 rows and 10 columns and the input test-dataset contains 132 rows and 10 columns. Whereas the output train-dataset contains 559 rows and 1 columns and output test dataset contains 132 rows and 1 columns.

Data Exploration

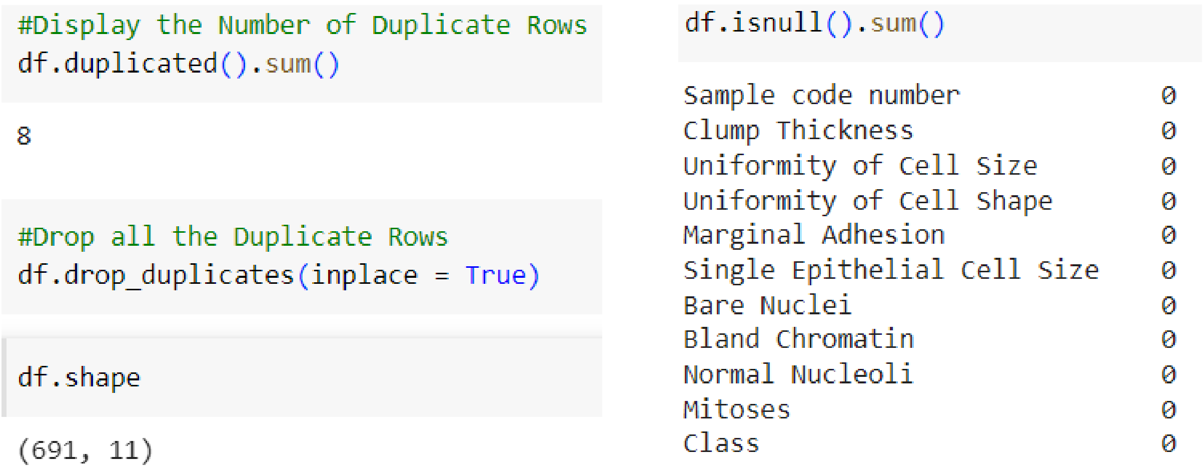


```
RangeIndex: 699 entries, 0 to 698
Data columns (total 11 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   Sample code number           699 non-null    int64 
 1   Clump Thickness              699 non-null    int64 
 2   Uniformity of Cell Size      699 non-null    int64 
 3   Uniformity of Cell Shape     699 non-null    int64 
 4   Marginal Adhesion            699 non-null    int64 
 5   Single Epithelial Cell Size  699 non-null    int64 
 6   Bare Nuclei                  699 non-null    object
 7   Bland Chromatin              699 non-null    int64 
 8   Normal Nucleoli              699 non-null    int64 
 9   Mitoses                      699 non-null    int64 
 10  Class                        699 non-null    int64 
dtypes: int64(10), object(1)
memory usage: 60.2+ KB
```

Fig 13: - Data Exploration

Dataset includes 699 rows and 11 columns in which the first column represents Sample code number, last column represents classes of cancer cells, and the rest nine columns are features of cancer cells. Datasets consist of 8 duplicates entries and 0 null values in each column. Dataset doesn't contain any outliers.

Data Cleaning and Validation



```
#Display the Number of Duplicate Rows
df.duplicated().sum()

8

#Drop all the Duplicate Rows
df.drop_duplicates(inplace = True)

df.shape

(691, 11)
```

```
df.isnull().sum()

Sample code number             0
Clump Thickness                0
Uniformity of Cell Size        0
Uniformity of Cell Shape       0
Marginal Adhesion              0
Single Epithelial Cell Size    0
Bare Nuclei                    0
Bland Chromatin                0
Normal Nucleoli                0
Mitoses                        0
Class                          0
```

Fig 14: - Data Cleaning & Validation

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.
Dataset contains 8 duplicate entries. Clean the datasets by removing all duplicates entries. After removing duplicate entries, the dataset now contains 691 rows and 11 columns. Dataset doesn't contain outliers and null values.
Data validation is the practice of checking the integrity, accuracy and structure of data before it is used for a business operation. Data validation operation results can provide data used for data analytics, business intelligence or training a machine learning model. For data validation, have to assume that data is collected from authenticated sources therefore data is accurate.

- **Data formatting**

  Each data format represents how the input data is represented in memory. This is important as each machine learning application performs well for a particular data format and worse for others. All the features have datatype int64 format except Bare Nuclei which has object datatype format. Since Bare Nuclei also contains Integer values, so, have to change the datatype of Bare Nuclei from Object to int64 datatype format.

- **Feature Engineering**
  Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling. All 9 features are relevant with respect to this Breast Cancer dataset therefore no requirement of feature engineering in this dataset.

- **Feature Extraction**
  Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data. All 9 features are in numerical features in this Breast Cancer dataset therefore no requirement of feature extraction in this dataset.

- **Feature Selection**
  Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. Out of 11 attributes in the dataset Sample code number feature is not required in classification purpose so, remove it from the dataset. From rest 10 attributes Class is the dependent feature and rest 9 are independent features with equal importance that can be verified using correlation between features.

3. Methodology

A training set, a validation set, and optionally a test set are often separated into two or three subsets of the dataset. The validation set is used to fine-tune hyperparameters and compare several models, while the test set is set aside for the final assessment of the chosen model. The training set is used to train the ML model.Then, we used a variety of machine learning (ML) methods, such as decision trees, random forests, support vector machines (SVM), logistic regression, Gaussian Naive Bayes, and artificial neural networks (ANNs), to predict breast cancer. The right algorithm should be chosen based on the size of the dataset, the complexity of the features, the need for interpretability, and the computational resources that are available. Then, we used a variety of the ML and DL algorithms listed below:

Logistic Regression

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

```
from sklearn.linear_model import LogisticRegression
log=LogisticRegression(random_state=0)
log.fit(X_train,Y_train)
log_pred=log.predict(X_test)
```

Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. This algorithm can be used for regression and classification problems — yet, is mostly used for classification problems. A decision tree follows a set of if-else conditions to visualise the data and classify it according to the conditions. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

```
from sklearn.tree import DecisionTreeClassifier
tree=DecisionTreeClassifier(random_state=0,criterion="entropy")
tree.fit(X_train,Y_train)
tree_pred=tree.predict(X_test)
```

Random Forest

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

```
from sklearn.ensemble import RandomForestClassifier
forest=RandomForestClassifier(random_state=0,criterion="entropy",n_estimators=10)
forest.fit(X_train,Y_train)
forest_pred=forest.predict(X_test)
```

Gaussian Naive Bayes

Naive Bayes is a generative model. (Gaussian) Naive Bayes assumes that each class follows a Gaussian distribution. The difference between QDA and (Gaussian) Naive Bayes is that Naive Bayes assumes independence of the features, which means the covariance matrices are diagonal matrices. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

```
from sklearn.naive_bayes import GaussianNB
gnb=GaussianNB()
gnb.fit(X_train,Y_train)
gnb_pred=gnb.predict(X_test)
```

Support Vector Machine (SVM)

Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

```
from sklearn.svm import SVC
svc = SVC(kernel='linear')
svc.fit(X_train, Y_train)
```

```
svc_pred=svc.predict(X_test)
```

Convolutional Neural Network (CNN)

Convolutional neural networks are a sort of deep learning algorithm frequently used for computer vision applications including object identification and picture categorization. Due to its capacity to automatically learn hierarchical representations, CNNs have demonstrated to be quite efficient in analyzing visual data. Convolutional, Activation, Pooling, Fully Connected, and Output layers make up the CNN algorithm. Fit the model into X_train and Y_train, and then make predictions for X_test.

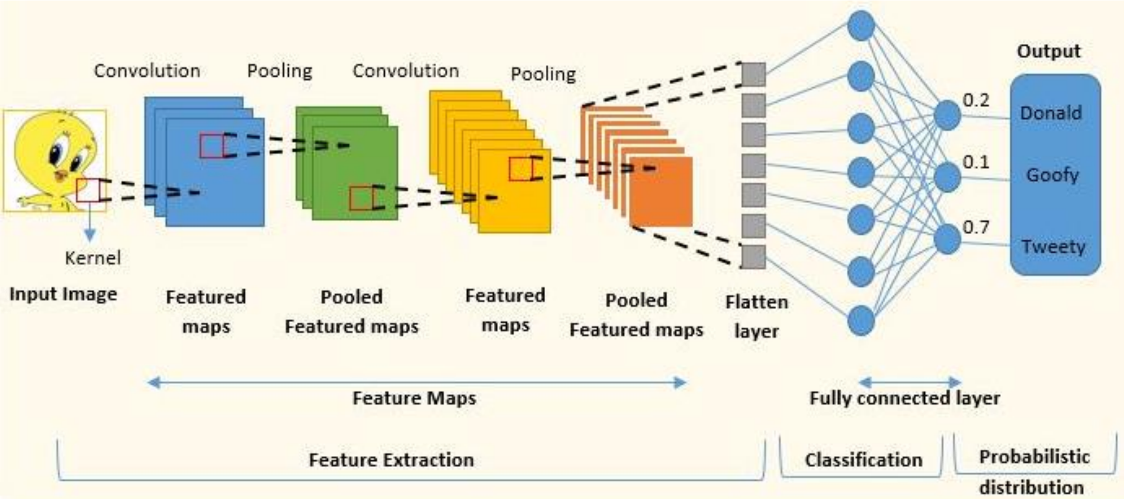


Fig 15: - Architecture of CNN [Analytics Vidhya]

Feed Forward Neural Network (FFNN)

Feed Forward Neural Network is a type of neural network which is used to a sort of artificial neural network in which data only travels in one way, from the input layer to the output layer, without looping around or repeating itself. It is made up of neurons and functions like as activation, forward propagation, learning, loss, and inference.

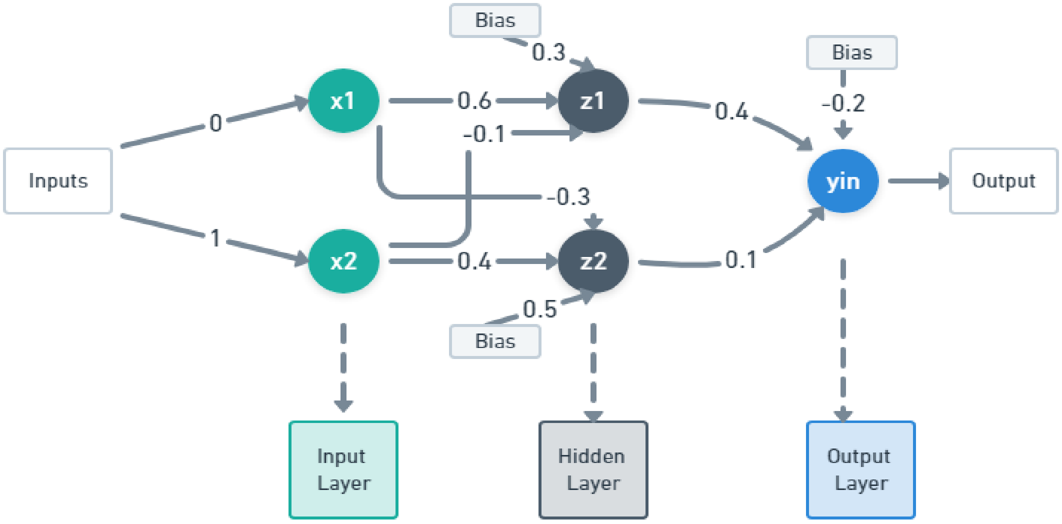


Fig 16: - Working of FFNN [GeeksforGeeks]

Recurrent Neural Network (RNN)

Recurrent Neural Network is a kind of artificial neural network made to process time series, speech, text, and video, among other sequential data types.

RNNs include loops in their internal structure that enable them to process input sequences in a recursive fashion, in contrast to typical neural networks, which only process input data in a single direction. This implies that RNNs can draw on their recollection of previous inputs to guide how they process the inputs they receive today.

LSTM- The recurrent neural network (RNN) variant known as LSTM, or "Long Short-Term Memory," was created to address the issue of disappearing gradients in conventional RNNs. Compared to conventional RNNs, LSTMs have a more intricate internal structure made up of numerous gates that control the information flow across the network.

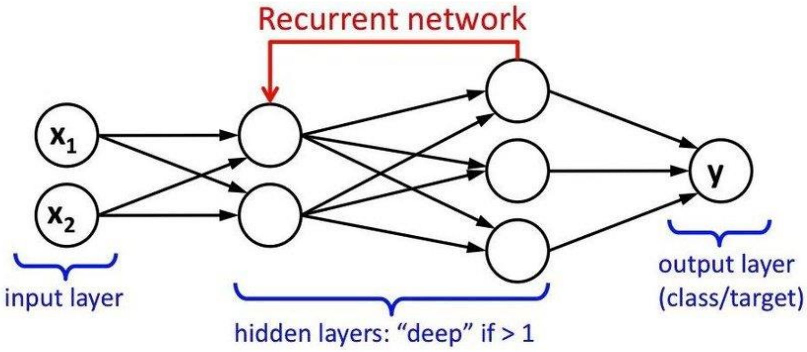


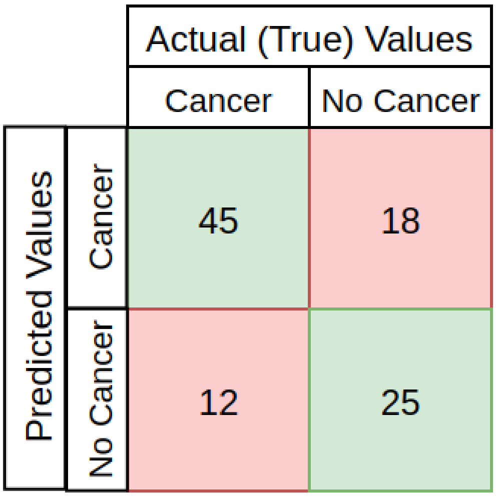
Fig 17: - Basic Architecture of RNN [Towards Data Science]

An input gate, an output gate, and a forget gate are some of these gates. The output gate regulates how much of the current state is used as output, the forget gate regulates how much of the prior state is forgotten, and the input gate regulates how much new input is permitted into the network.

5. Experiment and Results

Confusion Matrix

A confusion matrix is sometimes used to illustrate classifier performance based on the above four values (TP, FP, TN, FN). These are plotted against each other to show a confusion matrix:



		Actual (True) Values	
		Cancer	No Cancer
Predicted Values	Cancer	45	18
	No Cancer	12	25

Fig: - Confusion Matrix [Towards Data Science]

- TP: 45 positive cases correctly predicted.
- TN: 25 negative cases correctly predicted.
- FP: 18 negative cases are misclassified (wrong positive predictions).
- FN: 12 positive cases are misclassified (wrong negative predictions).

Precision

Precision is a measure of how many of the positive predictions made are correct (true positive).

Recall

Recall is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.

F1 Score

F1-Score is a measure combining both precision and recall. It is generally described as the harmonic mean of the two.

Accuracy

The base metric used for model evaluation is often Accuracy, describing the number of correct predictions over all predictions:

5.1 ML Approach

Algorithm	Precision	Recall	F1 score
Logistic Regression	98.18	94.73	96.42
Decision Tree	89.09	92.45	90.74
Random Forest	**98.18**	**96.42**	**97.29**
Gaussian Naïve Bayes	98.18	91.52	94.73
SVM	98.18	94.73	96.42

This table illustrates the values for the accuracy, recall, and F1 score of all the used machine learning methods. The table demonstrates that, in comparison to other models, Gaussian Naive Bayes has more accuracy, Decision Tree greater recall, and Random Forest greater F1 score.

Accuracy Table:

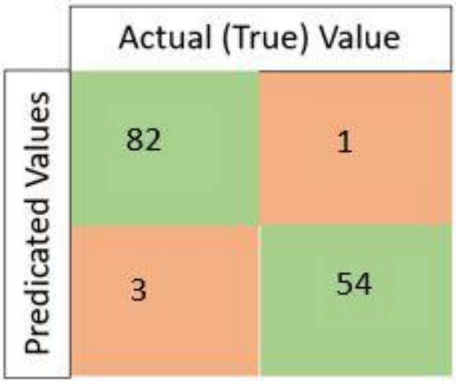
Algorithm	Accuracy
Logistic Regression	97.14
Decision Tree	92.85
Random Forest	**97.85**
Gaussian Naive Bayes	95.71
SVM	97.14

The accuracy of several used algorithms is displayed in this table. According to the table, the Random Forest has the highest accuracy, at 97.85 percent.

Algorithm	Bagging Score	Boosting Score
Logistic Regression	96.42	96.4
Decision Tree	96.42	96.4
Random Forest	**97.31**	**97.1**
Gaussian Naïve Bayes	96.24	97.1
SVM	96.42	95.5

This table emphasizes the strengths of these algorithms by displaying the Bagging and Boosting scores of all used machine learning algorithms. According to the table, Random Forest and Gaussian Naive Bayes both have higher Bagging and Boosting scores than the competition.

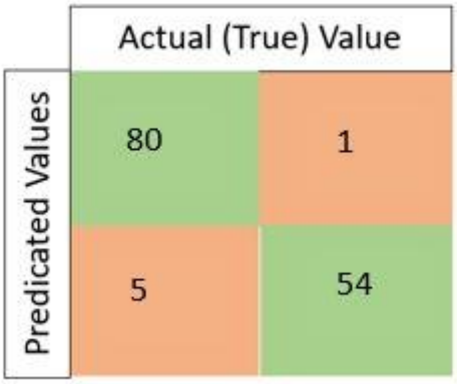
Confusion Matrix:



	Actual (True) Value	
Predicated Values	82	1
	3	54

Fig 18: -Confusion Matrix of Logistic Regression

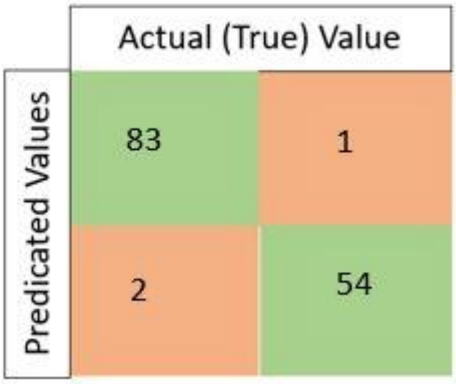
After using Logistic Regression, this confusion matrix shows the true positive, false positive, true negative, and false negative values. False negative values are 1, false positive values are 3, and genuine negative values are 85.



	Actual (True) Value	
Predicated Values	80	1
	5	54

Fig 19: -Confusion Matrix of Gaussian Naïve Bayes

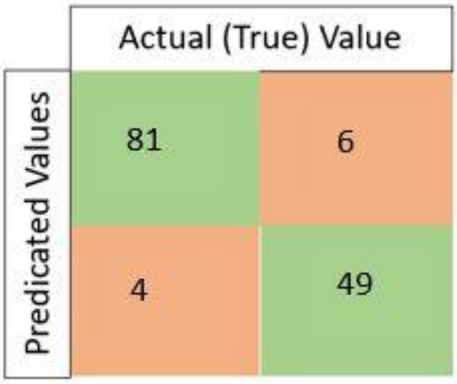
Following the use of Gaussian Naive Bayes, the true positive, false positive, true negative, and false negative values are represented by this confusion matrix. False negative values are 1, false positive values are 5, and genuine positive values are 80, 54, and 1, respectively.



	Actual (True) Value	
Predicated Values	83	1
	2	54

Fig 20: -Confusion Matrix of Random Forest

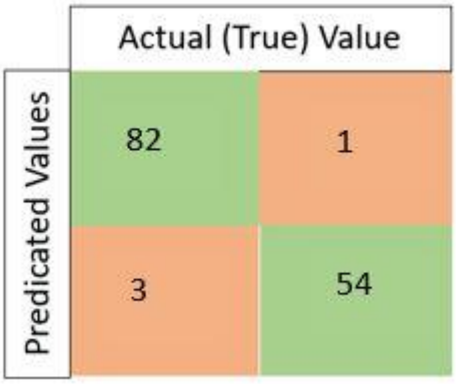
After using Logistic Regression, this confusion matrix shows the true positive, false positive, true negative, and false negative values. False negative values are 1, false positive values are 2, and genuine negative values are 83, 54, and 1, respectively.



	Actual (True) Value	
Predicated Values	81	6
	4	49

Fig 21: -Confusion Matrix of Decision Tree

After using Logistic Regression, this confusion matrix shows the true positive, false positive, true negative, and false negative values. Fake negative values are 6, fake positive values are 4, and genuine negative values are 49. The true positive value is 81.



	Actual (True) Value	
Predicated Values	82	1
	3	54

Fig 22: -Confusion Matrix of SVM

After using Logistic Regression, this confusion matrix shows the true positive, false positive, true negative, and false negative values. False negative values are 1, false positive values are 3, and genuine positive values are 82, 54, and 1, respectively.

5.2 DL Approach

Algorithm	**precision**	**Recall**	**F1 score**
CNN	75.0	92.75	96.24
FFNN	**98.65**	**92.75**	**99.29**
RNN	82.0	89.26	93.38

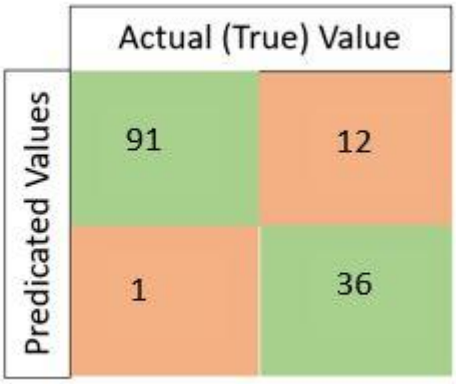
This table emphasizes the strengths of all used deep learning algorithms by displaying their accuracy, recall, and F1 scores. The table demonstrates that FFNN outperforms competitors in terms of accuracy, recall, and F1 score.

Accuracy Table:

Algorithm	**Accuracy**
CNN	90.71
FFNN	**97.61**
RNN	95.67

This table displays the accuracy of several deep learning methods that have been used. According to the table, the FFNN has the highest accuracy among the rest, at 97.61.

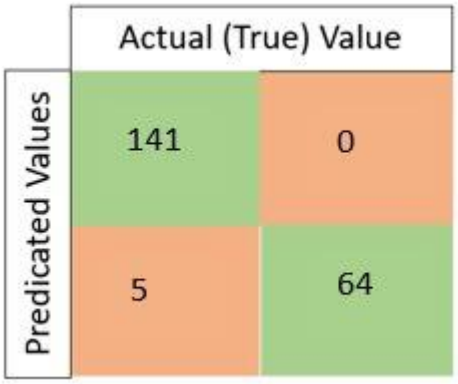
Confusion Matrix:



91	12
1	36

Fig 23: -Confusion Matrix of CNN

After using Logistic Regression, this confusion matrix shows the true positive, false positive, true negative, and false negative values. False negative values are 12, true negative values are 36, false positive values are 1, true positive values are 91.



141	0
5	64

Fig 24: -Confusion Matrix of FFNN

After using Logistic Regression, this confusion matrix shows the true positive, false positive, true negative, and false negative values. False negative values are 0, false positive values are 5, and genuine negative values are 64. The true positive value is 1.

6.Explainability of the Model

EXPLAINABLE AI(XAI)

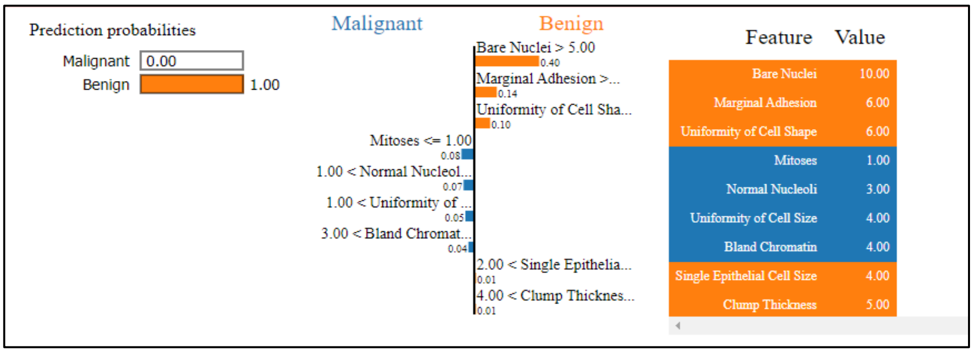
Explainable AI refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts. It contrasts with the concept of the “black box” in machine learning where even their designers cannot explain why the AI arrived at a specific decision. XAI is an implementation of the social right to explanation.

Using LIME-LIME stands for Local Interpretable Model Agnostic Explanation. The simplicity and use of LIME are its greatest strengths. Despite being extensive, LIME's main concept is quite clear-cut and straightforward. Let's explore what the name itself means first: Model agnosticism is a trait of LIME that allows it to provide justifications for any specific supervised learning model by considering it as a stand-alone "black box." LIME can therefore support practically any model that is currently in use. Local explanations refer to LIME providing justifications that are accurate in the immediate neighborhood of the observation or sample being explained.

Lime using ML Model	Decision Tree	Random Forest	Logistic Regression	Gaussian NB	SVC
Prediction Probabilities for Malignant	1.00	0.98	0.98	1.00	0.97
Bare Nuclei	0.41	0.32	0.32	0.28	0.33
Marginal Adhesion	0.13	0.26	0.27	0.24	0.26
Uniformity of Cell Shape	0.09	0.21	0.20	0.23	0.14
Mitoses	-0.06	-0.22	-0.22	-0.19	-0.31
Normal Nucleoli	-0.05	0.00	0.02	-0.05	0.02
Uniformity of cell Size	-0.04	-0.00	0.01	-0.04	-0.01
Single Epithelial cell Size	0.03	0.01	-0.01	-0.03	-0.00
Clump Thickness	0.03	0.04	0.03	-0.00	0.04
Bland Chromatin	0.00	0.01	0.02	-0.02	0.04

Table:7- Explaining the algorithms using LIME

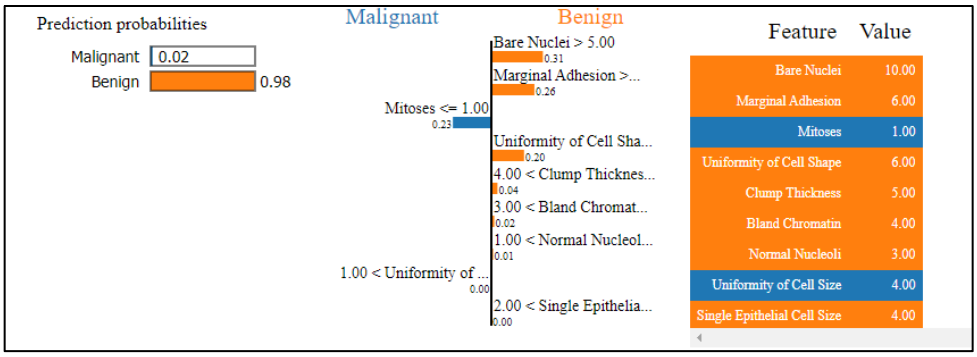
Using LIME, this table displays each characteristic of the dataset. After using machine learning and deep learning methods, LIME is used to explain each feature value in the dataset.



Feature	Value
Bare Nuclei	10.00
Marginal Adhesion	6.00
Uniformity of Cell Shape	6.00
Mitoses	1.00
Normal Nucleoli	3.00
Uniformity of Cell Size	4.00
Bland Chromatin	4.00
Single Epithelial Cell Size	4.00
Clump Thickness	5.00

Fig 25: - Explaining Decision Tree Using LIME

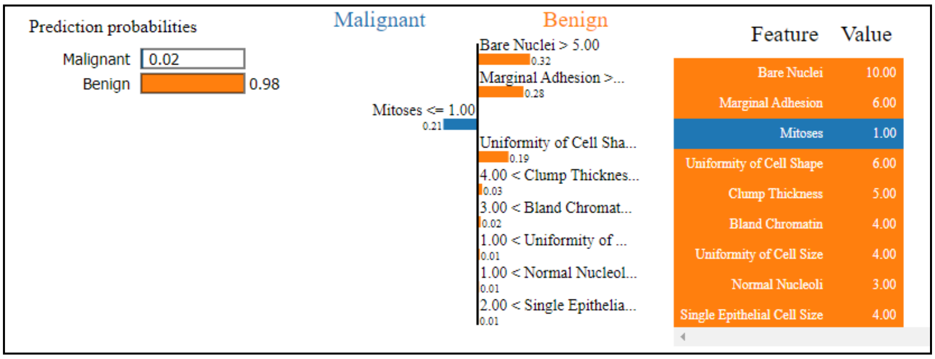
The explanation is broken down into three parts: Prediction probabilities are displayed in the leftmost area. Nine key aspects are returned in the middle portion. It would be in the two hues orange and blue for the binary classification challenge. Orange attributes support class 1, whereas blue attributes support class 0. Class 1 is supported by Benign, and Class 0 is supported by Malignant. The relative relevance of these qualities is represented by float point numbers on the horizontal bars. Each part has the same color coding. It includes the top 9 variables' actual values. In the decision tree algorithm, the features with the highest feature relevance include bare nuclei, marginal adhesion, uniformity of cell size, single epithelial cell size, and clump thickness.



Feature	Value
Bare Nuclei	10.00
Marginal Adhesion	6.00
Mitoses	1.00
Uniformity of Cell Shape	6.00
Clump Thickness	5.00
Bland Chromatin	4.00
Normal Nucleoli	3.00
Uniformity of Cell Size	4.00
Single Epithelial Cell Size	4.00

Fig26: - Explaining Random Forest Using LIME

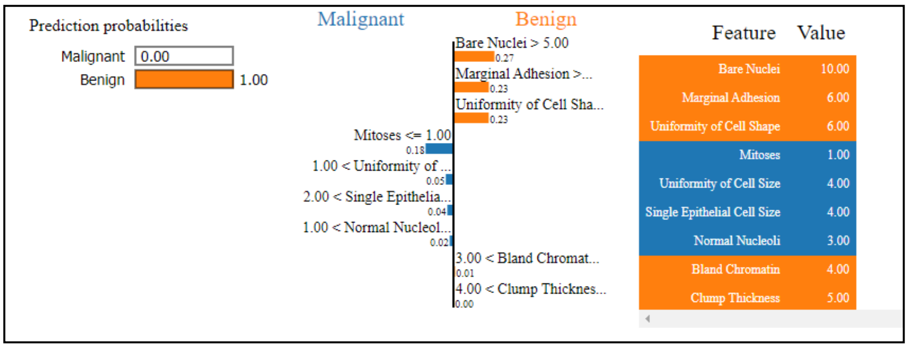
Mitosis and uniformity of cell size are low-priority features in the Random Forest method, whereas the other features are of utmost relevance.



Feature	Value
Bare Nuclei	10.00
Marginal Adhesion	6.00
Mitoses	1.00
Uniformity of Cell Shape	6.00
Clump Thickness	5.00
Bland Chromatin	4.00
Uniformity of Cell Size	4.00
Normal Nucleoli	3.00
Single Epithelial Cell Size	4.00

Fig 27: - Logistic Regression Using LIME

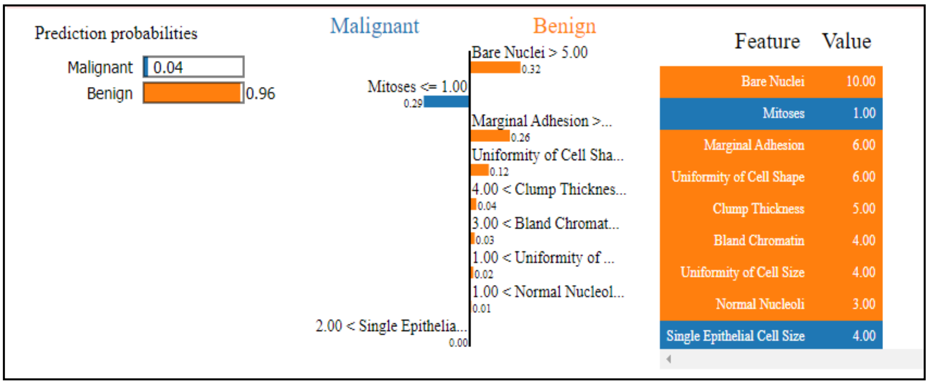
Mitosis is a low-priority feature in the Logistic Regression method, while the remaining features are of the highest relevance.



Feature	Value
Bare Nuclei	10.00
Marginal Adhesion	6.00
Uniformity of Cell Shape	6.00
Mitoses	1.00
Uniformity of Cell Size	4.00
Single Epithelial Cell Size	4.00
Normal Nucleoli	3.00
Bland Chromatin	4.00
Clump Thickness	5.00

Fig 28: - Gaussian Naive Bayes Using LIME

Mitosis and uniform cell size, single epithelial cell size, and normal nucleoli are low-priority features in the Gaussian Naive Bayes method, whereas the other features are of maximum relevance.



Feature	Value
Bare Nuclei	10.00
Mitoses	1.00
Marginal Adhesion	6.00
Uniformity of Cell Shape	6.00
Clump Thickness	5.00
Bland Chromatin	4.00
Uniformity of Cell Size	4.00
Normal Nucleoli	3.00
Single Epithelial Cell Size	4.00

Fig29: - SVM Tree Using LIME

Mitosis and Epithelial Cell Size are two features in the Support Vector Machine method that are of lower value than the others.

Using SHAP- SHAP is a mathematical method to explain the predictions of machine learning models. It is based on the concepts of game theory and can be used to explain the predictions of any machine learning model by calculating the contribution of each feature to the prediction.

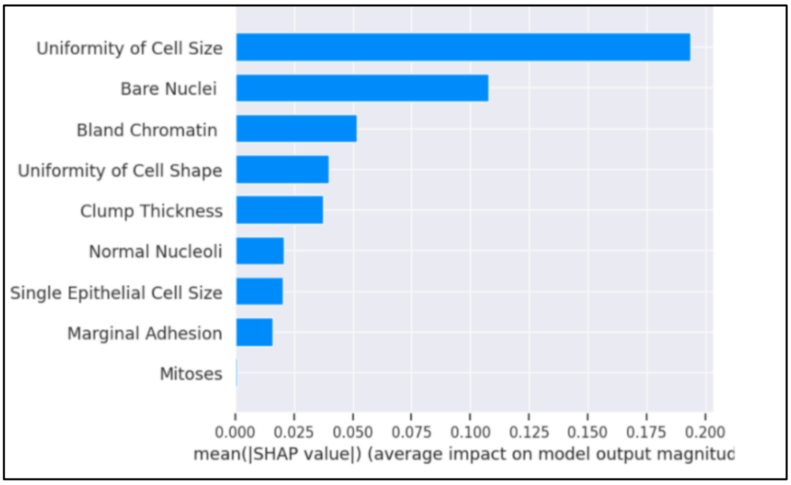


Fig30: - Summary plot using SHAP.

We can determine which characteristics have the most impact on your model's predictions by looking at the summary graphic. This information can give you insights into the underlying links between data and predictions as well as the decision-making process used by the model.

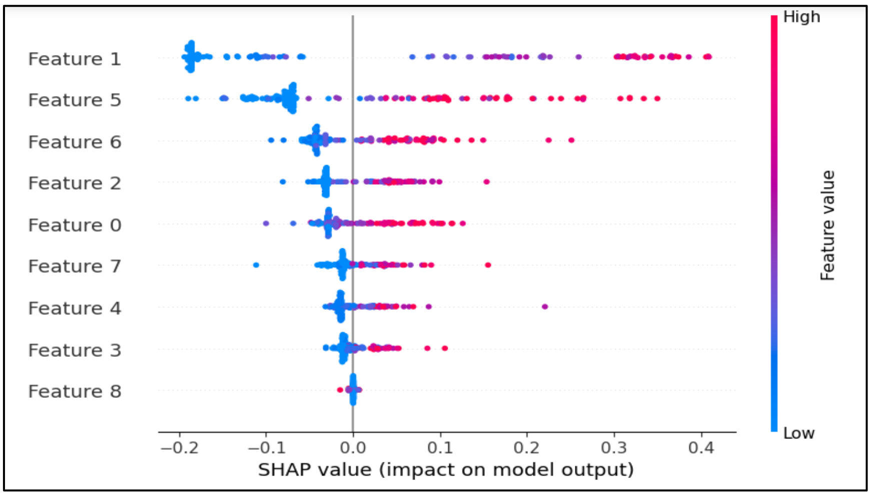


Fig31: - Beeswarm plot using SHAP.

The dispersion of data points along a continuous axis is shown using the visualization technique known as a Beeswarm plot. When applied to a specific instance or collection of instances, it may be used to express the SHAP values of specific characteristics. This figure demonstrates that Feature 1 has the largest distribution, followed by Feature 5, and that Feature 8 contributes very little to this prediction.

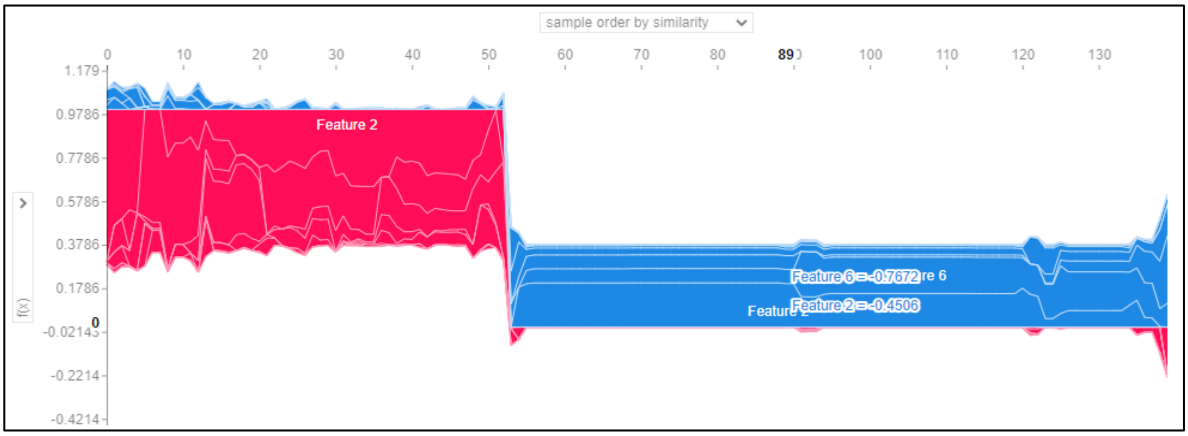


Fig32: - force plot using SHAP for Decision Tree Classifier

A force plot interprets machine learning model predictions, showing the contribution of each attribute to a prediction. Positive contributions increase predictions, while negative contributions decrease them. Feature 2 contributes the most, followed by feature 6 and feature 2.

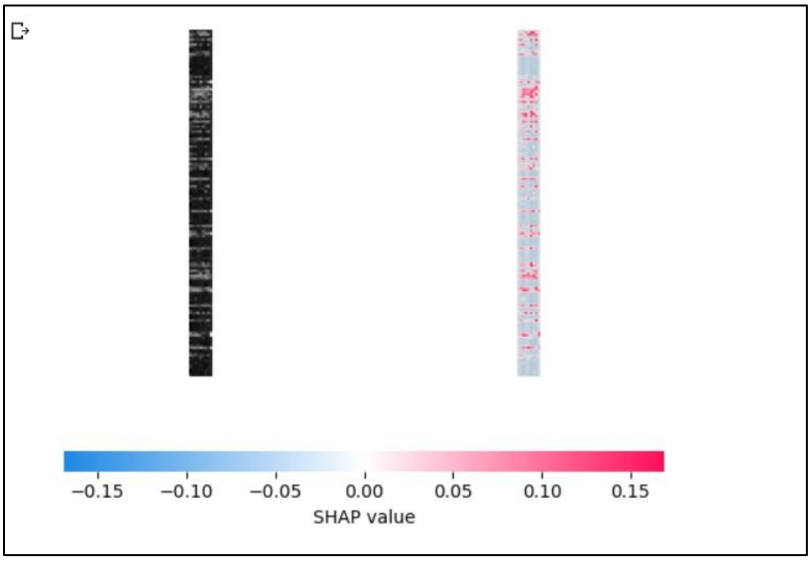


Fig33: - Image plot of CNN

A useful method for evaluating CNN behavior and comprehending how they take in, process, and interpret visual data is the use of image plots.

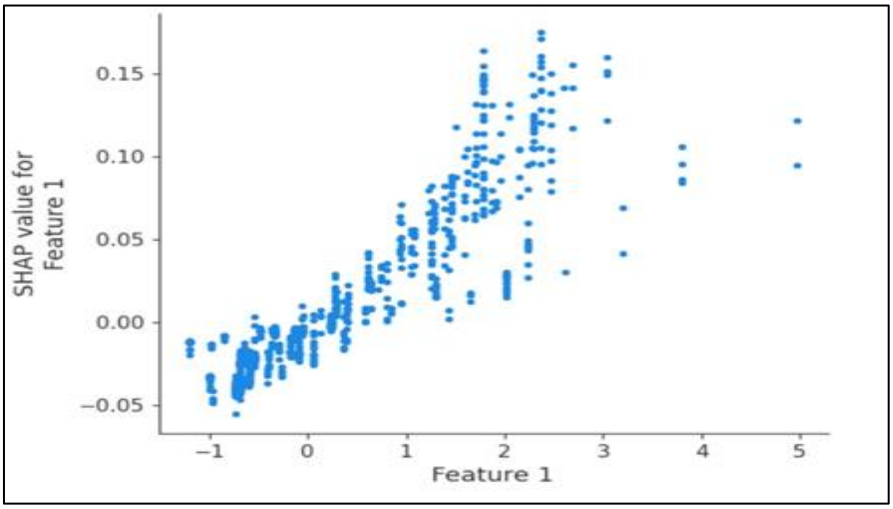


Fig34: - Dependence plot of CNN

By applying convolutional filters to the input pictures and learning hierarchical features that capture spatial dependencies, CNNs create dependencies. Later layers use these dependencies for activities like categorization or object detection.

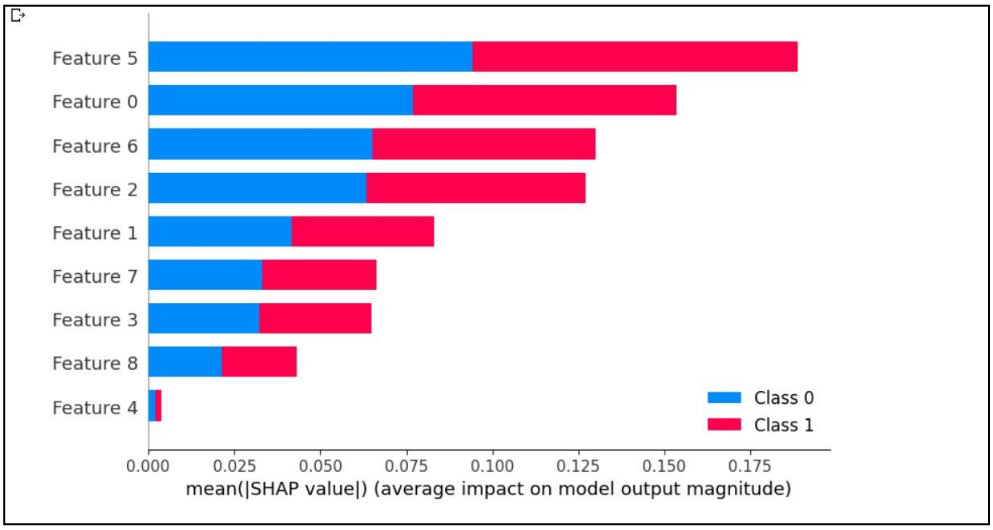


Fig35: - Summary plot of FFNN

This graph demonstrates that feature 5 has the biggest contribution to this prediction, followed by feature 0 and feature 4, which has a minimal influence.

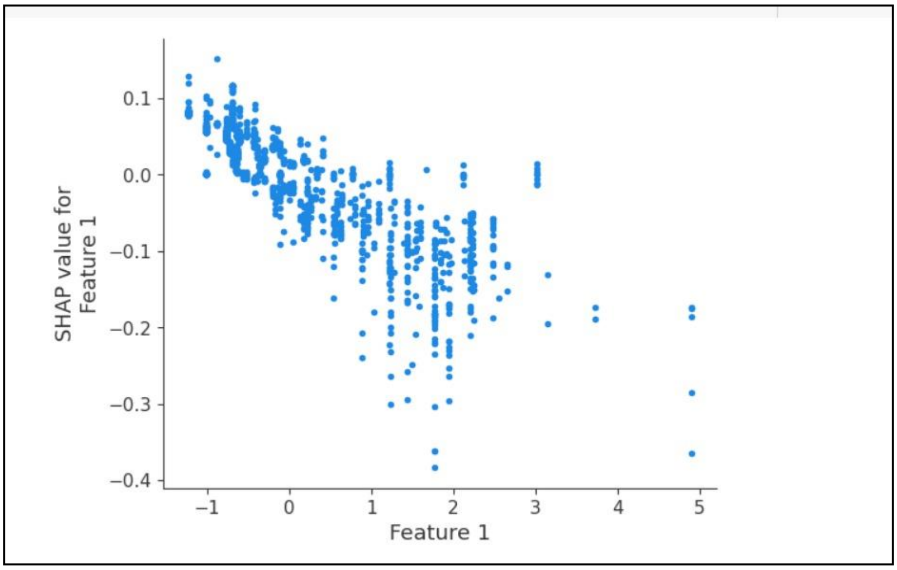


Fig36: - Dependence plot of FFNN

A dependency plot is a visualization approach that aids comprehension of the association between a certain input characteristic and the expected output of the model. We can see how the FENN's predictions vary when the selected input feature fluctuates while maintaining the fixed values of other characteristics by evaluating the Dependence Plot. It gives information on the connection between the feature and the model's output and may be used to locate any complicated or non-linear dependencies that the FFNN has discovered.

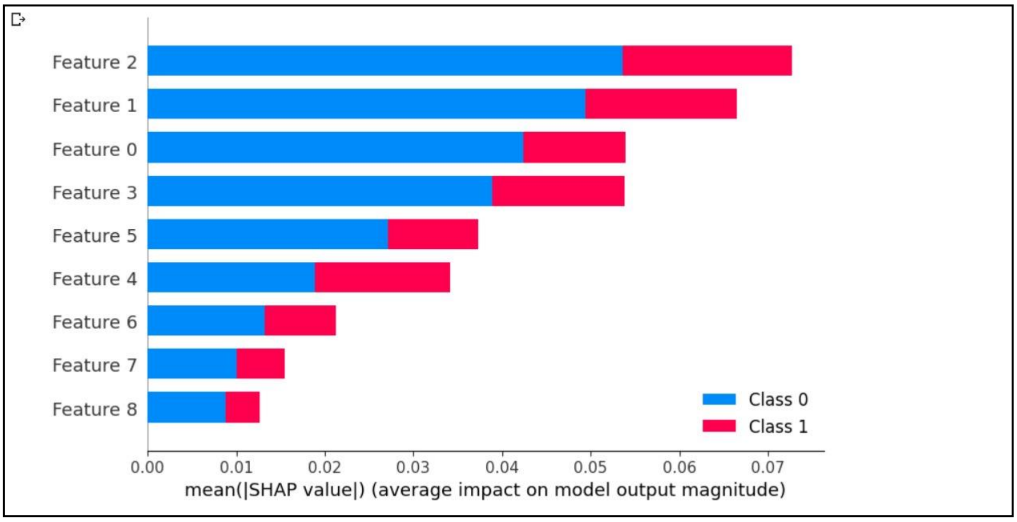


Fig37: - Summary plot of RNN

This graph demonstrates that feature 2 has the biggest contribution to this prediction, followed by feature 1, while feature 8 has a very little influence.

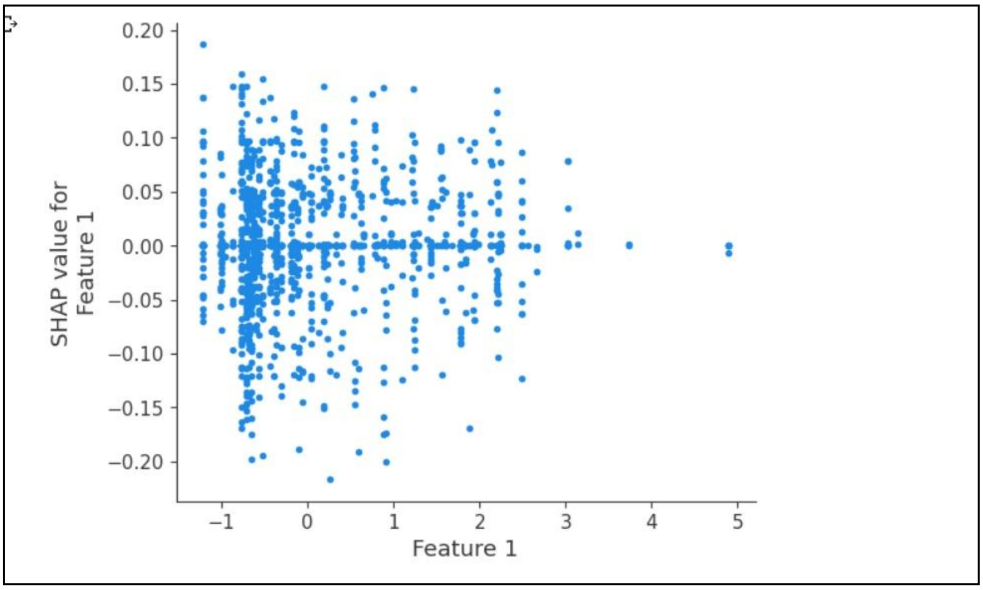


Fig38: - Dependence plot of RNN

Recurrent connections, which link the hidden state at each time step to the input and the hidden state from the previous time step, are how RNNs create dependencies. As a result, the network may discover relationships and patterns in sequential data and generate predictions using the context that earlier inputs have supplied.

7. CONCLUSION:

In this paper we have reviewed different types of machine learning and deep learning algorithms for predicting breast cancer. And we have also explained the algorithms using the Explainable AI. In explainable AI we have used SHAP and LIME. Our goal is to find out the most suitable algorithm that can predict the occurrences of breast cancer more effectively.

Using LIME explainable AI analyse the predicted probablity for different ML models like Decision Tree, Random Forest, Logistic Regression and SVC.Since Decision Tree based on logical classification its predicted probablity is 1. Analyzing from given table for different features of dataset found that for 50th datapoint Bare Nuclei is the most important features in prdecting it as Malignant for all different models. For example in Decision tree, Bare Nuclei has 41% positive contribution towards predicting class as Malignant. Similarly for SVC, Bare Nuclei has 33% positive contribution towards predicting class as Malignant.

Similarly, Mitoses is the least important features in prdecting it as Malignant for all different models. For example in Random forest, Mitoses has 22% negative contribution towards predicting class as Malignant. Similarly for Gaussian NB, Mitoses has 33% positive contribution towards predicting class as Malignant. Rest of other features are in between these two features. For different aspects, can also check for different datapoint other than 50th which has predicted class as Benign.

The results analysis shows that the combination of multidimensional data with various feature selection, classification, and dimensionality reduction approaches might offer advantageous tools for inference in this field. It is necessary to conduct additional study in this area to improve the effectiveness of classification algorithms and enable them to make predictions on a wider range of factors. In order to attain high accuracy, we want to parametrize our categorization systems. We are investigating a variety of datasets and the potential applications of machine learning techniques to further characterise breast cancer. We aim to maximise accuracy while lowering mistake rates.

REFERENCES

[1] Riddhi R. Gujar et al. “Breast Cancer Prediction Using Machine Learning” [ResearchGate 2023].

[2] Rashika Pandita et al. “Analysis of Breast Cancer Prediction Using Machine Learning Techniques: Review Paper” [ResearchGate 2023].

[3] Rahul Karmakar et al. “BCPUML: Breast Cancer Prediction Using Machine Learning Approach—A Performanc” [ResearchGate 2023].

[4] Muhammad Waqas Arshad et al. “PREDICTION AND DIAGNOSIS OF BREAST CANCER USING MACHINE LEARNING AND ENSEMBLE CLASSIFIERS” [ResearchGate 2023].

[5] Amin Mohamed Ahsan et al. “Breast Cancer-Risk Factors and Prediction Using Machine-Learning Algorithms and Data Source: A Review of Literature” [ResearchGate 2023].

[6] Xinkang Li et al. “Prediction of ADMET Properties of Anti-Breast Cancer Compounds Using Three Machine Learning Algorithms” [ResearchGate 2023].

[7] D. Shanthi et al. “A study of deep learning techniques for predicting breast cancer types" [ResearchGate 2023].

[8] Mainak Sanyal et al. “Deep Learning Techniques For Breast Cancer Diagnosis -A Short Review” [ResearchGate 2023].

[9] Haitham Elwahsh et al. “A New Approach for Cancer Prediction based on Deep Neural Learning” [ResearchGate 2023].

[10] Reza Rabiei et al. “Prediction of Breast Cancer using Machine Learning Approaches” [ResearchGate 2022].

[11] Farzane Tajdini et al. “Breast Cancer Diagnosis and Prediction Using Machine Learning” [ResearchGate 2022].

[12] Shiekhah A. al Binali et al. “Breast Cancer Subtypes Prediction Using Omics Data and Machine Learning Models” [ResearchGate 2022].

[13] Ramya Challa et al. “Breast Cancer Prediction Using Machine Learning” [ResearchGate 2022].

[14] M. Thangavel et al. “Enhancing the Prediction of Breast Cancer Using Machine Learning and Deep Learning Techniques” [ResearchGate 2022].

[15] Jonathan M. Ji et al. “A Novel Machine Learning Systematic Framework and Web Tool for Breast Cancer Survival Rate Assessment” [ResearchGate 2022].

[16] Xia Jiang et al. “Deep Learning and Machine Learning with Grid Search to Predict Later Occurrence of Breast Cancer Metastasis Using Clinical Data” [ResearchGate 2022].

[17] Santhosh Voruganti et al. “Breast Cancer Prediction using CNN and Machine Learning Algorithms with Comparative Analysis” [ResearchGate 2021].

[18] Yuhong Huang et al. “Prediction of Tumour Shrinkage Pattern to Neoadjuvant Chemotherapy Using a Multiparametric MRI-Based Machine Learning Model in Patients with Breast Cancer” [ResearchGate 2021].

[19] Muktevi Sri Venkatesh et al. “Prediction of Breast Cancer Disease using Machine Learning Algorithms” [ResearchGate 2020].

[20] M. S. Dawngliani et al. “Prediction of Breast Cancer Recurrence Using Ensemble Machine Learning Classifiers” [ResearchGate 2020].

[21] Vinoth S. M. E et al. “Accurate Breast Cancer Prediction using Machine Learning Techniques.”[ResearchGate 2020].

[22] Elham Yousef Kalafi et al. “Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data” [ResearchGate 2019].

[23] Ramachandiran et al. “SYSTEMATIC ANALYSIS OF BREAST CANCER PREDICTION USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS” [ResearchGate 2019].

[24] Saria Eltalhi et al. “Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review” [ResearchGate 2019].

[25] Elham Yousef Kalafi et al. “Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data” [ResearchGate 2019].

[26] Siyabend Turgut et al., “Microarray Breast Cancer Data Classification Using Machine Learning Methods” [IEEE 2018].

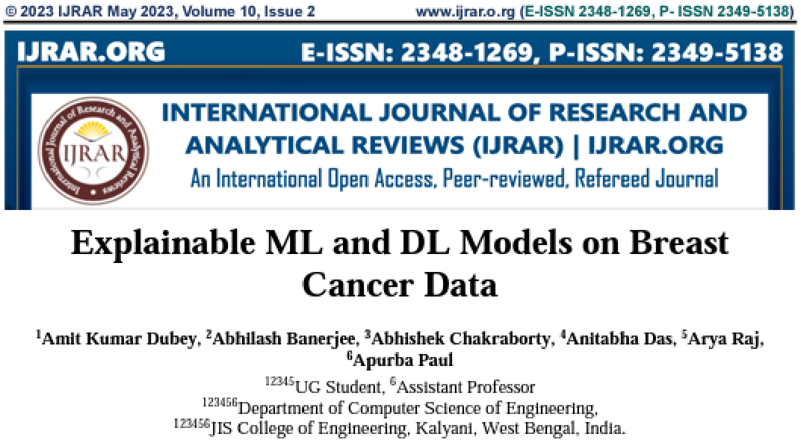
PUBLICATION FROM THE WORK

We have submitted our research paper to the UGC Approved index journal.

Journal Details:

Journal Name: - International Journal of Research and Analytical Review (IJRAR)

Status: - Research Paper Accepted



© 2023 IJRAR May 2023, Volume 10, Issue 2 www.ijrar.o.rg (E-ISSN 2348-1269, P- ISSN 2349-5138)

IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

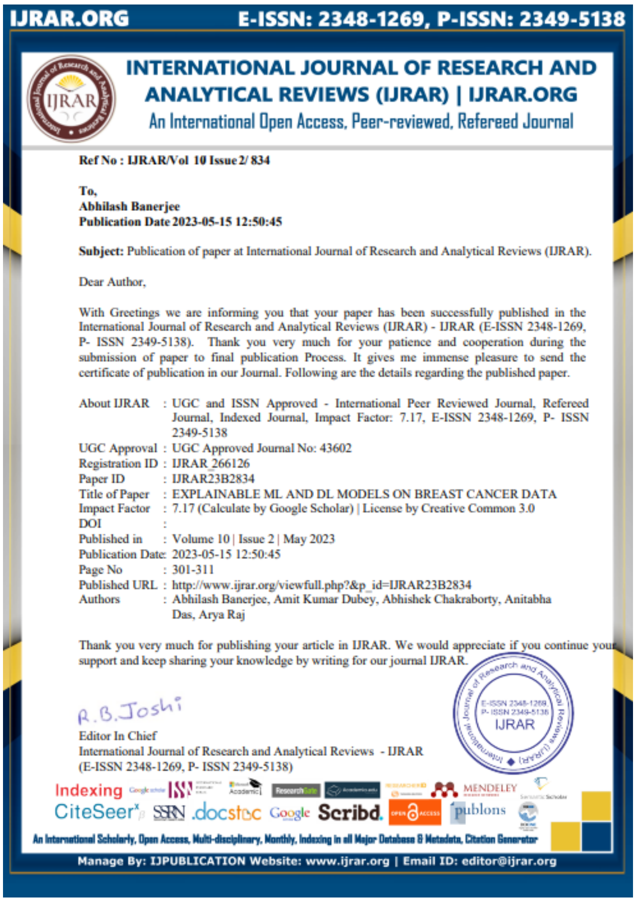
Explainable ML and DL Models on Breast Cancer Data

[1]Amit Kumar Dubey, [2]Abhilash Banerjee, [3]Abhishek Chakraborty, [4]Anitabha Das, [5]Arya Raj, [6]Apurba Paul

[12345]UG Student, [6]Assistant Professor

[123456]Department of Computer Science of Engineering,

[123456]JIS College of Engineering, Kalyani, West Bengal, India.



IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Ref No : IJRAR/Vol 10 Issue 2/ 834

To,
Abhilash Banerjee
Publication Date 2023-05-15 12:50:45

Subject: Publication of paper at International Journal of Research and Analytical Reviews (IJRAR).

Dear Author,

With Greetings we are informing you that your paper has been successfully published in the International Journal of Research and Analytical Reviews (IJRAR) - IJRAR (E-ISSN 2348-1269, P- ISSN 2349-5138). Thank you very much for your patience and cooperation during the submission of paper to final publication Process. It gives me immense pleasure to send the certificate of publication in our Journal. Following are the details regarding the published paper.

About IJRAR	: UGC and ISSN Approved - International Peer Reviewed Journal, Refereed Journal, Indexed Journal, Impact Factor: 7.17, E-ISSN 2348-1269, P- ISSN 2349-5138
UGC Approval	: UGC Approved Journal No: 43602
Registration ID	: IJRAR_266126
Paper ID	: IJRAR23B2834
Title of Paper	: EXPLAINABLE ML AND DL MODELS ON BREAST CANCER DATA
Impact Factor	: 7.17 (Calculate by Google Scholar) | License by Creative Common 3.0
DOI	:
Published in	: Volume 10 | Issue 2 | May 2023
Publication Date:	2023-05-15 12:50:45
Page No	: 301-311
Published URL	: http://www.ijrar.org/viewfull.php?&p_id=IJRAR23B2834
Authors	: Abhilash Banerjee, Amit Kumar Dubey, Abhishek Chakraborty, Anitabha Das, Arya Raj

Thank you very much for publishing your article in IJRAR. We would appreciate if you continue your support and keep sharing your knowledge by writing for our journal IJRAR.

International Journal of Research and Analytical Reviews (IJRAR) E-ISSN 2348-1269 P-ISSN 2349-5138 IJRAR

R.B.Joshi
Editor In Chief
International Journal of Research and Analytical Reviews - IJRAR
(E-ISSN 2348-1269, P- ISSN 2349-5138)