International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

Detection of Breast Cancer using Machine Learning Techniques

Ragul T N[1], Karthikeyan I[1], Vaidyanathan K[1],and Dr.G R Hemalakshmi[2]
[1]Department of Computer Science and Engineering, National Engineering College, Kovilpatti,Tamil Nadu.
[2]Assistant Professor (Senior Grade), Department of Computer Science, National Engineering College, Kovilpatti, Tamil Nadu, India.

***Abstract*: Abnormal growth of breast cells causes breast cancer. These cells divide faster than healthy cells and continue to accumulate, forming a lump or mass. The cells can spread through the breast to the lymph nodes or other body parts. But fortunately, it is also curable cancer in its early stages. Breast cancer is among the 20 leading causes of death worldwide, affecting approximately 10% of the world's female population. As the number of people with breast cancer increases, effective predictive measures for the early diagnosis of breast cancer improve the prognosis and survival of patients. This study helps experts research preventive measures against breast cancer through early diagnosis using machine learning techniques. In this project, supervised Learning is used to analyze all features to determine whether a patient is affected by a benign or malignant tumour. The evaluation is performed on several patient datasets that contain features such as radius, texture, perimeter, area, and smoothness. Supervised Learning is a method in which a machine is trained on data in which inputs and outputs are labelled. A model can learn training data and process future data to predict outcomes. Therefore, machine learning techniques are of great importance in the early detection of breast cancer. These techniques support professionals and doctors in the early detection of breast cancer to prevent the development of the disease.**

Keywords: Unsupervised Learning, predictive analysis

1. INTRODUCTION

Breast cancer is a type of cancer that occurs in the breast tissue. It occurs when cells in tissue grow uncontrollably and form a lump or tumour. Breast cancer is the most common and the second most common cancer in women worldwide. Although breast cancer can occur in both men and women, it is more common in women. It is estimated that 1 in 8 women in the United States will develop breast cancer in their lifetime.

There are many types of breast cancer, and treatment varies by type, stage, and other factors. The most common types of breast cancer include ductal carcinoma in situ, invasive ductal carcinoma, and invasive lobular carcinoma. Breast cancer can be detected early with screening tests such as mammography, which can help improve treatment options. Early diagnosis is essential because it improves prognosis and increases the chances of survival.

Breast cancer is caused by the abnormal growth of breast cells, which divide more rapidly than healthy cells and accumulate to form a lump or mass. The cancer cells may spread to other parts of the body or lymph nodes. Although breast cancer is among the top 20 causes of death worldwide and affects around 10% of women globally, it can be cured in its early stages.

To improve early diagnosis and increase patient survival rates, experts are exploring preventive measures for breast cancer using machine learning techniques. In this project, supervised Learning is used to analyze features such as radius, texture, perimeter, area, and smoothness and determine whether a patient has a benign or malignant tumour. Supervised Learning is a method where the machine is trained on labelled input and output data. The model learns from the training data to predict outcomes for future data. Therefore, machine learning techniques are critical for early breast cancer detection and support doctors and experts in preventing the disease.

2. RELATED WORK

Tianyu Shen et al. l. proposed a hierarchical fused model based on deep Learning and fuzzy Learning for breast cancer diagnosis based on lesion segmentation and disease grading. The critical point is to alleviate the drawbacks of deep Learning in terms of interpretability, generalization ability, and few-shot Learning. The proposed model consists of a pixel-wise segmentation unit based on ResU-seg Net, a feature extraction unit based on domain knowledge, and a severe grading classification unit based on the IT2PFCM-fused feedforward neural network. The feature representation and rule-based Learning integrated with domain knowledge ensured the interpretability of the system. Both segmentation and disease grading performance in a few-shot learning manner are improved. Cross-dataset research proved the improvement of generalization ability.

Ravi K. Samala et al. l. works demonstrate that multi-stage transfer learning can utilize the knowledge gained through source tasks from unrelated and related domains. And show that the limited data availability in a target domain can be alleviated with pre-training the Convolutional Neural Network using data from similar auxiliary domains. And also show that the gain in

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

Convolutional Neural Network performance from the additional stage of fine-tuning with the auxiliary data depends on the relative sizes of the available training samples in the target and the auxiliary domains and the proper selection of the transfer learning strategy. Furthermore, when the training sample size is small, the variance in the performance of the trained Convolutional Neural Network is significant. Reporting the best performance through exhaustive searches using a "test" set can be overly optimistic. It is, therefore, essential to validate the generalizability of the trained Convolutional Neural Network with unknown independent cases.

Yongjin Zhou et al. l. proposed that the most significant addition of this work is a Convolutional Neural Network-based radionics technique on shear wave elastography for breast cancer diagnosis. It is the first attempt to use radionics based on Convolutional Neural Network to automatically extract high-throughput features from shear-wave elastography to classify malignant and benign breast tumours. And another significant contribution is that segmenting the tumour is unnecessary; it can reduce a lot of work and improve the classification model performance. Because this method doesn't need segmentation in advance and manual feature extraction, it has excellent potential to be applied to Computer-Aided Diagnosis Systems. The classification model is extendable and flexible that can be trained again to generalize for the new dataset.

Jun Xu et al. l. proposed a Stacked Sparse Auto-encoder framework for automated nuclei detection of breast cancer histopathology. The Stacked Sparse Autoencoder model can capture high-level feature representations of pixel intensity in an unsupervised manner. These high-level features enable to decrease the discrepancy between input and reconstruction as much as possible by learning encoder and decoder networks which yields a set of weights and biases.

Michiel Kallenberg et al., l. proposed a technique that builds a feature hierarchy from raw data. Breast density segmentation and grading of mammographic texture are two distinct tasks that can be addressed when the learnt features are applied as the input to a straightforward classifier. The suggested model picks up features at various scales. A unique sparsity that considers both lifetime and population sparsity is introduced to manage the model's capacity. Additionally examined, the approach using three various clinical datasets. Additionally, cutting-edge research demonstrates that the learnt breast density scores correlate with manual ones and that the learned texture scores diagnose breast cancer. The model is straightforward to use and generalizes to a wide range of different segmentation and scoring issues.

Bolei Xu et al. l. present a brand-new deep hybrid attention network to classify breast cancer histological images. The network's intricate attention technique can automatically identify the valuable parts from the photos in the Break His dataset; as a result, the network does not need to resize the image to prevent information loss. Comparing our framework selection mechanism to the previous partially observable Markov decision process-based approach, training time can be cut in half and test our methodology on a publicly available dataset, where it achieves about 98% accuracy at four different magnifications.

Jingxin Liu et al. l., Proposed a method that employs a single fully convolutional network to extract every nucleus region, mimicking the decision-making process of pathologists (tumour and nontumor). This multi-column convolutional neural network uses the outputs of the first two fully convolutional networks. The image describing the staining intensity as an input serves as the high-level decision-making mechanism to directly output the H Score of the input Tissue microarray image. A second fully convolutional network to extract the tumour nuclei region. This first end-to-end system uses a Tissue microarray image as the input and directly produces a clinical score. It will discuss experimental findings that show the H-Scores predicted by the model have a robust and statistically significant correlation with the scores of seasoned pathologists and that the discrepancy between the H-Scores of the algorithm and the pathologists is comparable to the inter-subject discrepancy between the pathologists.

Mandeep Rana et al. l. proposed that the performance of each algorithm varies based on the dataset and parameter choices. K-Nearest Neighbor methodology has produced the best outcomes overall. Naive Bayes and logistic regression have also shown promising results in diagnosing breast cancer. As previously stated, Support Vector Machine is a powerful technique for predictive analysis. In light of the previous finding, we conclude that Support Vector Machine with a Gaussian kernel is the best technique for predicting whether breast cancer will cure.

Vikas Chaurasia et al. l. applied three breast cancer survival prediction models to two parameters: patients with benign and malignant cancer. Employed the Naive Bayes, Reverse Path Forwarding Network, and J48 data mining techniques here. The University of California Irvine Machine Learning repository provided a dataset, which was obtained. To create the prediction models, and used data selection, preprocessing, and transformation. In this study, survivability was represented by a binary categorical survival variable computed from the variables in the raw dataset, where benign is represented by a value of and malignant is represented by a value. It employed a 10-fold cross-validation process to assess the three methods' unbiased prediction performance. Divided the dataset into 0 mutually exclusive partitions using a stratified sampling technique to do this. For each of the three prediction models, repeat this procedure. This gave a less biased way to compare the three models' prediction performance. The acquired findings showed that the J48 came in third with a classification accuracy of 93.41%, followed by Reverse Path Forwarding Network in second place with a classification accuracy of 96.77%. The Naive Bayes did the best, scoring 97.36%. To better understand the relative contributions of the independent variables to predicting survivability, also performed sensitivity analysis and specificity analysis on Naive Bayes, Reverse Path Forwarding Network, and J48 in addition to the prediction model. According to the sensitive data, the prognosis factor Class is the most significant predictor.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

Vanlalhmangaihsanga et al. l., Proposed that, in contrast to other models, the K-Nearest Neighbor and Logistic Regression have insultingly poor accuracy during the training process. Compared to Decision Tree and Random Forest classifiers, which have an accuracy of 75%, the Support Vector Machine also performed significantly better in classification errors, including correctly and erroneously categorized instances. Deploying the model on the testing data set after training it and assessing the algorithms, effectiveness on the training dataset is known. Surprisingly, the accuracy of the top-performing classifiers was 97% for the Logistic Regression and Random Forest, compared to 57% for the Support Vector MachineLogistic Regression and the Random Forest classifier are on equal footing when looking at accuracy alone. But when efficiency, sensitivity, and specificity are considered, random forest classifiers perform significantly better.

3. METHODOLOGY

The data consists of 569 patients and 32 characteristics. These characteristics formed 32 columns in the dataset. Features are Id, Diagnosis, Radius_mean, Texture_mean, Perimeter_mean, Area_mean, Smoothnes_mean, Concavity_mean, Compactness_mean, Concavepoints_mean, Symmetry_mean, fractal_dimension_mean, Radius_se, Texture_se, Perimeter_se, Area_se, Smoothnes_se, Concavity_se, Compactness_se, Concavepoints_se, Symmetry_se, fractal_dimension_se, Radius_worst, Texture_worst, Perimeter_worst, Area_worst, Smoothnes_worst, Concavity_worst, Compactness_worst, Concavepoints_worst, Symmetry_worst, fractal_dimension_worst.

The target is the classification which is either benign breast cancer or malignant breast cancer. The data needs cleaning because it has Null, and the numeric features must be forced to float. We were instructed to get rid of all rows with Null. At first, the dataset is fetched using the Panda's library, and then we save the data inside a Panda's data frame. Initially, it counted the rows and columns in the dataset; there were 569 and 32 columns. This dataset consists of many null values; it counts the columns with null values then the columns with null values drop because the model cannot process the null values.

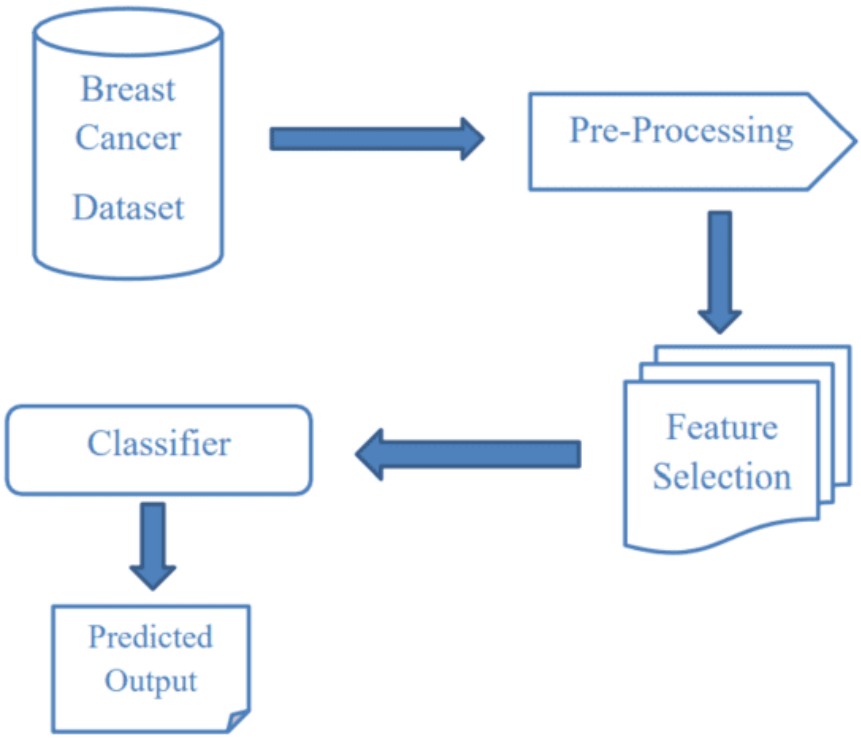
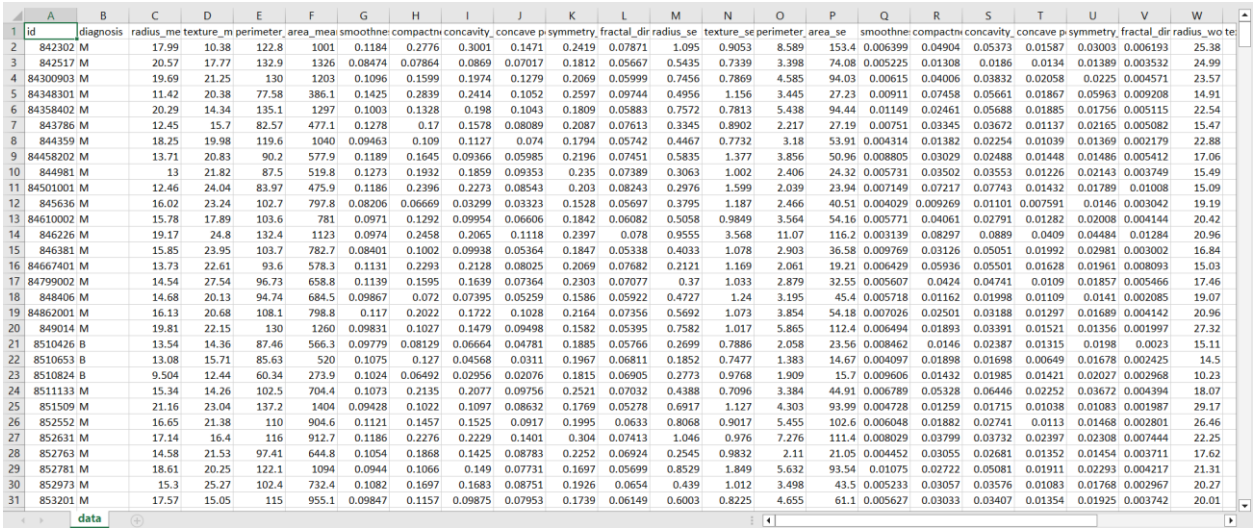


FIG 3.1: WORKFLOW DIAGRAM

The Data are retrieved from the input dataset by using Panda's library. Pandas provide a unique method to retrieve rows from a Data frame. Data frame. Loc [] method is a method that takes only index labels and returns a row or data frame if the index label exists in the caller data frame. The Data are retrieved from the input dataset by using Panda's library.
Pandas provide a unique method to retrieve rows from a Data frame. Data frame. Loc [] method is a method that takes only index labels and returns a row or data frame if the index label exists in the caller data frame. Figure.3.2 Data set of the patient

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	id	diagnosis	radius_me	texture_m	perimeter_	area_mea	smoothne	compactn	concavity_	concave p	symmetry_	fractal_dir	radius_se	texture_se	perimeter_	area_se	smoothne	compactn	concavity_	concave p	symmetry_	fractal_dir	radius_wo
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532	24.99
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571	23.57
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208	14.91
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005082	15.47
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003749	15.49
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003042	19.19
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004144	20.42
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04484	0.01284	20.96
15	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003002	16.84
16	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008093	15.03
17	84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005466	17.46
18	848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002085	19.07
19	84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01689	0.004142	20.96
20	849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01356	0.001997	27.32
21	8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387	0.01315	0.0198	0.0023	15.11
22	8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698	0.00649	0.01678	0.002425	14.5
23	8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606	0.01432	0.01985	0.01421	0.02027	0.002968	10.23
24	8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.006789	0.05328	0.06446	0.02252	0.03672	0.004394	18.07
25	851509	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.004728	0.01259	0.01715	0.01038	0.01083	0.001987	29.17
26	852552	M	16.65	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633	0.8068	0.9017	5.455	102.6	0.006048	0.01882	0.02741	0.0113	0.01468	0.002801	26.46
27	852631	M	17.14	16.4	116	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413	1.046	0.976	7.276	111.4	0.008029	0.03799	0.03732	0.02397	0.02308	0.007444	22.25
28	852763	M	14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924	0.2545	0.9832	2.11	21.05	0.004452	0.03055	0.02681	0.01352	0.01454	0.003711	17.62
29	852781	M	18.61	20.25	122.1	1094	0.0944	0.1066	0.149	0.07731	0.1697	0.05699	0.8529	1.849	5.632	93.54	0.01075	0.02722	0.05081	0.01911	0.02293	0.004217	21.31
30	852973	M	15.3	25.27	102.4	732.4	0.1082	0.1697	0.1683	0.08751	0.1926	0.0654	0.439	1.012	3.498	43.5	0.005233	0.03057	0.03576	0.01083	0.01768	0.002967	20.27
31	853201	M	17.57	15.05	115	955.1	0.09847	0.1157	0.09875	0.07953	0.1739	0.06149	0.6003	0.8225	4.655	61.1	0.005627	0.03033	0.03407	0.01354	0.01925	0.003742	20.01

data

Figure.3.2 Data set of the patient

Figure 3.2 contains the data and attributes taken into inconsideration for the diction of Breast Cancer Detection. Data visualization is the discipline of understanding data by placing it into visual form to interactively and efficiently convey insights so that the patterns, trends and correlations of the data that might not otherwise be detected can be visualized in large data sets. It removes the noise from the data and highlights valuable information. As visualization makes it easier to detect patterns, trends and outliers and provides precise, better and reliable results, it is implemented in this paper by creating a count plot, pair plot and heat map. In this work, data visualization is done with the help of the seaborn library. Data visualization is the discipline of understanding data by placing it into visual form to interactively and efficiently convey insights so that the patterns, trends and correlations of the data that might not otherwise be detected can be visualized in large data sets. It removes the noise from the data and highlights valuable information. As visualization makes it easier to detect patterns, trends and outliers and provides precise, better and reliable results, it is implemented in this paper by creating a count plot, pair plot and heat map. In this work, data visualization is done with the help of the seaborn library.

3.1. Support Vector Machine (SVM)

Support vector machine is the modern, high-speed machine learning algorithm for solving multiclass classification problems for large datasets based on a simple iterative approach. The SVM model is created in the dataset's linear CPU time. SVM can be used for the high dimensional dataset in the sparse and dense format. A support Vector Machine is a supervised classifier algorithm. It is used kernel trick for solving the classification problem. Based on these transformations, the ideal edge is found between the possible outputs. SVM is used for the nonlinear kernel, such as RBF. For the linear kernel, SVM is an appropriate choice. SVM classifier is sufficient for all linear problems. This algorithm gave an accuracy of about 95.12%.

3.2.Recurrent Neural Network (RNN)

A recurrent neural network (RNN) is an artificial neural network that uses sequential or time series data. Well-known programmes like Siri, voice search, and Google Translate include these deep learning algorithms. They are frequently employed for ordinal or temporal issues in language translation, natural language processing (NLP), speech recognition, and picture captioning. Recurrent neural networks (RNNs) use training data to learn as feedforward and convolutional neural networks (CNNs) do. They stand out because of their "memory," which allows them to affect the current input and output using data from previous inputs. Recurrent neural networks' output relies on the sequence's fundamental components, but typical deep neural networks presume that inputs and outputs are independent. Unidirectional recurrent neural networks cannot consider future events in their forecasts, although they would also contribute to the output of a particular sequence.

3.3.Convolutional Neural Networks (CNNs)

CNN is an artificial neural network widely used for image/object recognition and classification. Deep Learning thus recognizes objects in an image by using a CNN. CNNs play a significant role in diverse tasks/functions like image processing problems, computer vision tasks like localization and segmentation, video analysis, recognizing obstacles in self-driving cars, and speech recognition in natural language processing. As CNNs play a significant role in these fast-growing and emerging areas, they are trendy in Deep Learning.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

CNN is a different neural network that can find important information in time series and visual data. It is valuable for image-related tasks like image identification, object categorization, and pattern recognition. A CNN uses concepts from linear algebra, such as matrix multiplication, to find patterns in a picture. CNNs may also categorize signal and audio data. The structure of a CNN is comparable to the connection structure of the human brain. Like the brain has billions of neurons, CNNs also have neurons but are structured differently. The neurons in CNN are designed to resemble the brain's frontal lobe, which processes visual inputs.

This configuration ensures that the entire visual field is covered, eliminating the piecemeal picture processing of standard neural networks. Compared to the older networks, a CNN performs better with image and speech or audio signal inputs. A deep-learningA convolutional layer, a pooling layer, and a fully connected (FC) layer are the three layers that makeup CNN. The first layer is the convolutional layer, while the final layer is the FC layer. The complexity of the CNN grows from the convolutional layer to the FC layer. Due to the image's growing complexity, CNN can recognize more essential details and intricate aspects until it locates the item. This increasing complexity allows the CNN to successfully identify more significant portions and complex features of an image until it finally identifies the object.

3.4. Naïve Bayes (NB)

The Bayes theorem is the foundation of the supervised learning technique, the Naive Bayes algorithm, employed to resolve distributional classification issues. It is mainly utilized in high-dimensional training datasets for text categorization. One of the easiest and most effective classifiers is the Naive Bayes model. Classification algorithms aid in the development of quick machine-learning models with rapid prediction capabilities. It is a probabilistic classifier, which implies that it bases its assumptions on the likelihood that an item exists. Spam filtration, Sentimental analysis, and material classification are a few well-known applications of the Naive Bayes algorithm. Naive Bayes (NB) algorithm for breast cancer detection and demonstrated the certainty results as 93%

4. RESULT AND DISCUSSION

As expected, the model predicts whether the patient has a benign or malignant level of tumors.

Accuracy Check

Figure 4.1 Accuracy Check of Algorithm

The above Figure 4.1 shows the model's accuracy analyzed with the actual values of a breast cancer diagnosis. The accuracy check of an algorithm that detects breast cancer is crucial in ensuring the reliability and effectiveness of the system. The algorithm's accuracy is measured by comparing its results with the actual diagnosis of a set of patients. This process is commonly known as validation or testing. The accuracy check of a breast cancer detection algorithm involves the evaluation of the algorithm's sensitivity and specificity. Sensitivity refers to the algorithm's ability to correctly identify patients with breast cancer, while specificity refers to correctly identifying patients without breast cancer. An algorithm with high sensitivity and specificity indicates a more reliable system to detect breast cancer accurately. Therefore, an accurate algorithm with high sensitivity and specificity is essential for early diagnosis and improved patient outcomes.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

Data Encoding

```
df.iloc[:,1]= labelencoder_Y.fit_transform(df.iloc[:,1].values)
print(labelencoder_Y.fit_transform(df.iloc[:,1].values))#Encoding categorical data values (
from sklearn.preprocessing import LabelEncoder
labelencoder_Y = LabelEncoder()
df.iloc[:,1]= labelencoder_Y.fit_transform(df.iloc[:,1].values)
print(labelencoder_Y.fit_transform(df.iloc[:,1].values))
```

```
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 0 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 1 0 1 1 0 0 0 0 1 0 1 1
 0 1 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 1 0 0 0 1 1 0 0 0 0 1 0 0 1 0 0
 0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 0 0 1 0 0 1 0 0 0 0 1 0
 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 0 0 0 1 0 0 1 1 1 1 0 1
 0 1 0 0 0 1 0 0 1 1 1 0 1 1 1 1 0 1 1 1 0 1 0 1 0 0 1 0 1 1 1 1 0 0 1 1 0 0
 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 1 1 1 0 0
 0 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1
 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0
 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 1 0 0 0 0 0 1 0 0
 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 1 0 1 0 1 1
 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 0]
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 0 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 1 0 1 1 0 0 0 0 1 0 1 1
 0 1 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 1 0 0 0 1 1 0 0 0 0 1 0 0 1 0 0
 0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 0 0 1 0 0 1 0 0 0 0 1 0
 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 0 0 0 1 0 0 1 1 1 1 0 1
 0 1 0 0 0 1 0 0 1 1 1 0 1 1 1 1 0 1 1 1 0 1 0 1 0 0 1 0 1 1 1 1 0 0 1 1 0 0
 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 1 1 1 0 0
 0 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1
 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0
 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 1 0 0 0 0 0 1 0 0
 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 1 0 1 0 1 1
 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 0]
```

Figure 4.2 Encoding of data

Figure 4.2 shows that the categorical data in column 'diagnosis' is encoded/ transformed from M and B to 1 and 0 using Label Encoder from sklearn. Preprocessing.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

Data Classification:

```
In [8]: #Visualize this count
        sns.countplot(df['diagnosis'],label="Count")

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x233b7e1f848>
```

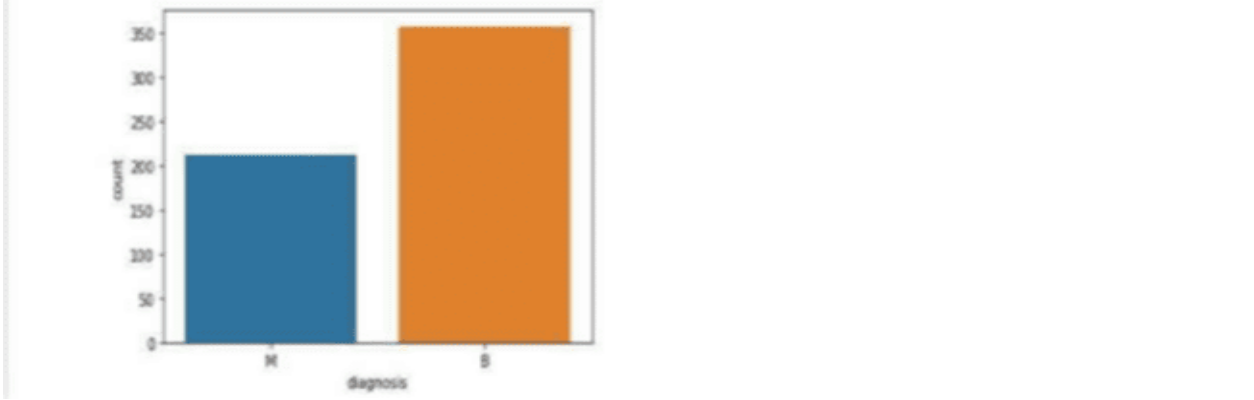


Figure 4.3 Count Plot of Benign and Malignant Tumor

Above, Figure 4.3 shows the plot representing the class distribution of diagnosed malignant and benign patients. Two hundred twelve malignant diagnosed patients, i.e., around 38% of the data and 357, i.e., 62% of patients diagnosed with a benign tumor.

Pair Plot:

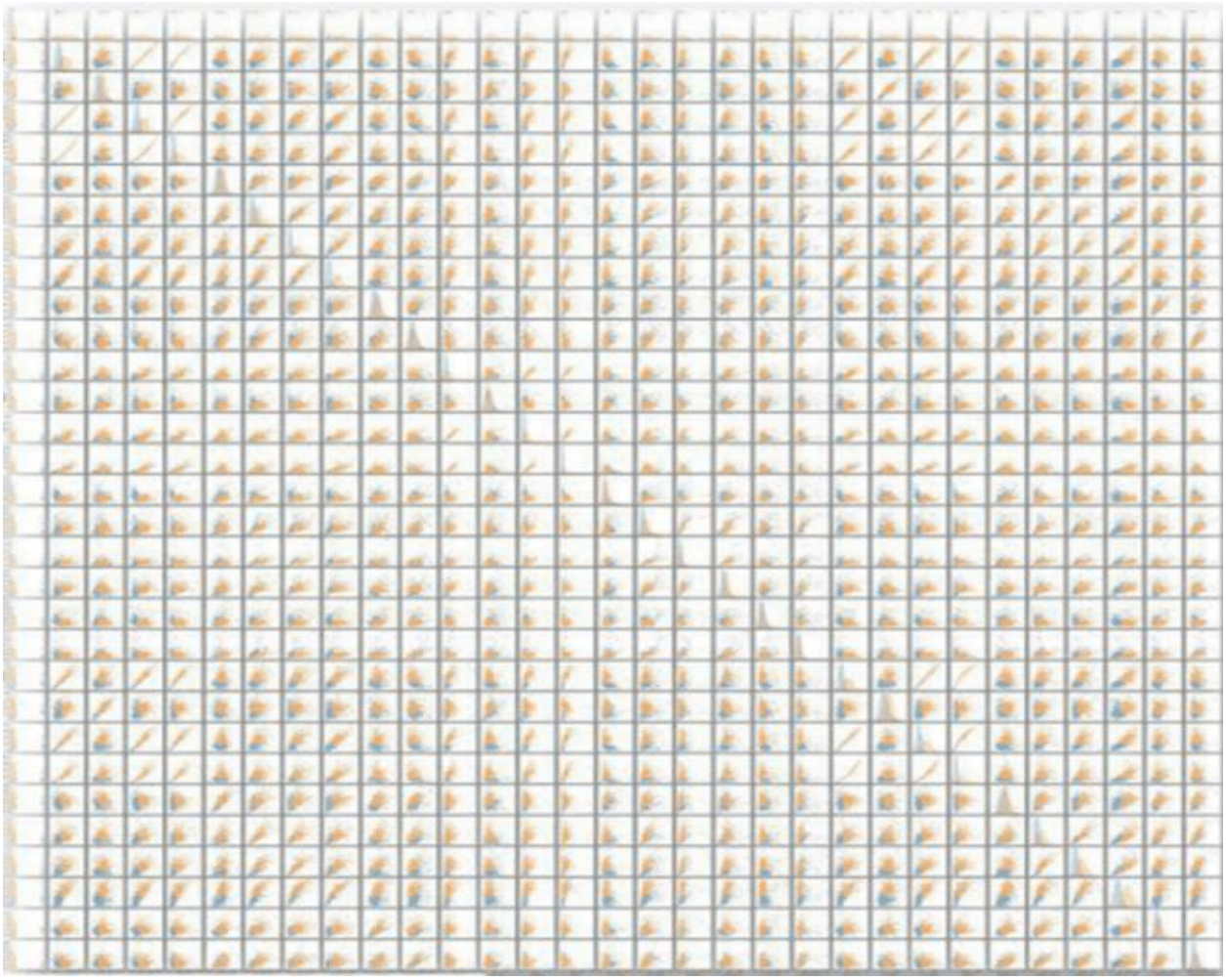


Figure 4.4 Creating Pair Plot

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

The above Figure.4.4 represents the pair plot of all the columns highlighting the diagnosis points. The orange points are for one, and the blue points are for 0. The pair is used to show the numeric distribution in the scatter plot.

Heat Map

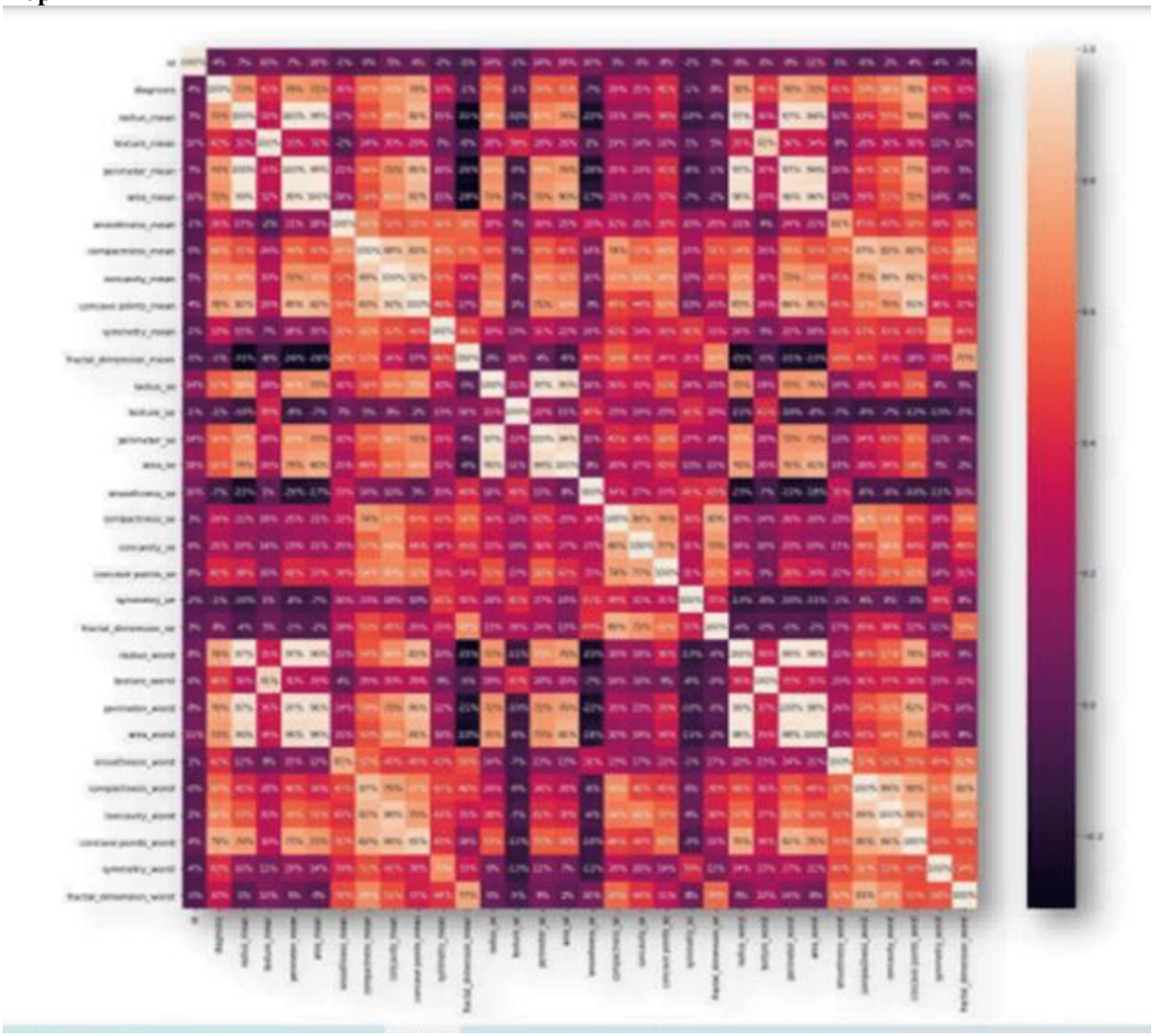


Figure 4.5 Function of Heat Map

The above Figure 4.5 shows the focus is on the light and the dark areas. It shows the strength of correlation.

4. CONCLUSION

This study attempts to analyze various supervised machine-learning algorithms and select the most accurate model for breast cancer detection. The work focused on advancing predictive models with the help of Python to achieve better accuracy in predicting correct outcomes. The analysis of the result signifies that integration of data, feature scaling, and different classification methods and analysis provide markedly successful tools in prediction. It has also been observed that the model misdiagnosed a few patients with cancer when they were not having cancer and vice versa. Although the model is accurate when dealing with people's lives, further research in building the most accurate and precise model must be carried out for the better performance of classification techniques and to get the accuracy as close to 100% as possible. Thus, the tuning of each of the models is necessary for the building of a more reliable model.

5. REFERENCES

[1] Paweł Filipczuk, Thomas Stevens, Adam Krzyżak and Roman Monczak "Hierarchical Fused Model With Deep Learning and Type-2 Fuzzy Learning for Breast Cancer Diagnosis" IEEE Transactions on fuzzy systems,

[2] Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Caleb D. Richter, and Kenny H. Cha "Breast Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on Multi-Stage Transfer Learning Using Deep Neural Nets " IEEE transactions on medical imaging, vol. 38, no. 3, march 2019.

[3] Yongjin Zhou, Jingxu Xu, Qiegen Liu, Cheng Li, Zaiyi Liu, Meiyun Wang, Hairong Zheng, and Shanshan Wang "A Radiomics Approach With CNN for Shear-Wave Elastography Breast Tumor Classification" IEEE Transactions on biomedical engineering, vol. 65, no. 9, September 2018

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

[4] Jun Xu*, Member, IEEE, Lei Xiang, Qingshan Liu, Senior Member, IEEE, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi "Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images " IEEE transactions on medical imaging, vol. 35, no. 1, January 2016.
[5] Michiel Kallenberg*, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm "Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring " IEEE transactions on medical imaging, vol. 35, no. 5, May 2016.
[6] Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, Jon Garibaldi IEEE, Ian O. Ellis, Andy Green, Linlin Shen, and Guoping Qiu "A Deep Selective Attention Approach to Breast Cancer Classification" IEEE Transactions on medical imaging, vol. 39, no. 6, June 2020.
[7] Jingxin Liu, Bolei Xu, Chi Zheng, Yuanhao Gong, Jon Garibaldi, Daniele Soria, Andrew Green, Ian O. Ellis, Wenbin Zou, and Guoping Qiu "An End-to-End Deep Learning Histochemical Scoring System for Breast Cancer TMA " IEEE transactions on medical imaging, vol. 38, no. 2, February 2019.
[8] Mandeep Rana, Pooja Chandorkar and Alishiba Dsouza "Breast cancer diagnosis and recurrence prediction using machine learning techniques" International Journal of Research in Engineering and Technology Volume: 04 Issue: 04 | Apr-2015.
[9] Vikas Chaurasia, BB Tiwari and Saurabh Pal “Prediction of benign and malignant breast cancer using data mining techniques” Journal of Algorithms and Computational Technology Vol. 12(2),2018.
[10] D. Dubey, S.Kharya and S.Soni "Breast cancer detection using machine learning." International Journal of Computer Science and Information Technologies vol. 8, no. 6, June 2021