Breast Cancer Detection using Machine Learning Approach

Conference Paper · July 2022

DOI: 10.1109/ICECET55527.2022.9872893

4 authors, including:

Abdus Sattar
Daffodil International University
69 PUBLICATIONS **236** CITATIONS

SEE PROFILE

Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET 2022)
20-22 July 2022, Prague-Czech Republic

Breast Cancer Detection using Machine Learning Approach

Rohena Begum Mim
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
rohena15-11388@diu.edu.bd

Afra Bente Islam
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
afra15-11487@diu.edu.bd

Sudipta Roy
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
sudipta15-11692@diu.edu.bd

Abdus Sattar
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
abdus.cse@diu.edu.bd

2022 International Conference on Electrical, Computer and Energy Technologies (ICECET) | 978-1-6654-7087-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECET55527.2022.9872893

***Abstract*—We have gathered the features of breast cancer and normal persons' cells both. To classify malignant and benign tumors, we used a supervised machine learning classifier algorithm. This paper shows the last update in this machine learning field on breast cancer in Bangladesh. We have used many classifiers of ML in this review. Most of cases it is difficult to identify the malignant tumors. For this, we hoped that with the help of math and the computational power of ML we can resolve this issue at a significant scale. Yet, there were a few difficulties with the process. Starting with featuring the dataset and creating a data frame we proceed to apply different types of machine learning classifiers. This paper presents an overview of the opinion examination challenges applicable to their methodologies and strategies.**

***Keywords*—** *Breast Cancer, Malignant tumor, Benign tumor, Feature Selection, Machine Learning.*

I. INTRODUCTION

The breast is made of a diffusion of tissues from very fatty tissues to very dense tissue. There is a community of lobes within these tissues. Each lobe is made of small tubular structures referred as lobules that include mammary glands. Small ducts join the glands, leaflets & leaves and carry the milk from the leaves to the nipples. The nipple is inside the center of the areola and is the dark region surrounding the nipple. Blood and lymphatic tubes run in the course of the chest. Blood is nourished cells. The lymphatic machine leaves the physical waste of the product. The lymphatic tube combines with lymph nodes, small and pointed organs that help to fight infectious illnesses. Groups of lymph nodes are in unique areas throughout the body, along with throat, bar and stomach. Local lymph node breasts are close to the chest like lymph nodes under the arm. Cancer changes healthy cells within the chest, develop from controls, and form mass or sheet of cells referred to as tumors. Tumors may be cancerous or benign. Cancer tumors are malignant, which means that that it is able to develop and unfold to the opposite frame element. Neighbor tumors mean that tumors can grow however aren't spreading. Breast cancer extends when most cancers grows into adjacent organs or other frame elements or at the time of the movement of cell to any other body parts through vessels or lymphatic is called metastasis. Breast cancer is malicious cell proliferation of the chest. If they remain untreated, the cancer is spreading to the region of the other body parts. Breast cancers is the maximum commonplace type of cancer in the US female and explain one among three most cancers diagnostics. In 2005, A predicted 211,240 new cases of invasive breast most cancers had been expected for girls in the United States [1]. In year 2005, approximately 1690 new instances of breast most cancers had been predicted for men [1]. The prevalence of breast most cancers increase after the age of forty. The maximum incidence (about 80% of invasive instances) happens in ladies over the age of fifty. By adding in attacking breast cancers, nearly 58,590 new cases of breast cancers are anticipated to rise in woman in 2005 [1][2]. About 88% of these cases are sectioned as ductal carcinoma in situ [3]. The detection of ductal carcinoma in situ cases are an immediate output of mammography screening reports. This technique additionally serves to detect invasive cancers at a slower stage than otherwise. In 2005, an estimated 40,870 deaths from breast cancer (410 women and 460 guys) have been anticipated. Breast cancer ranks 2nd among ladies' cancer related deaths [1]. According to the ultra-modern data, mortality declined significantly between 1992 and 1998, with the most important decline amongst young girls, both white and black [4]. This guide covers both non-invasive breast cancer and early and locally advanced breast most cancers. This consists of three stages, I, II and III. The stage of breast most cancers suggests how plenty the most cancers has grown, whether the most cancers have spread, and wherein it has unfolded.

II. LITERATURE REVIEW

Breast cancer is one of the leading cause death in todays' women. The lifetime odds that a woman will develop breast cancer is 10.7% and die cause of breast cancer is 4% [5]. It is expected that 19,200 new instances might be identified in 2000, with 5,500 girl death from the ailment within the same 12 months [6]. In Bangladesh the rate of breast cancer is 22.5 per 100000 female among 15 to 44 years old [7].

The reasons of breast most cancers are nevertheless not known sometimes genetic, but numerous factors are

978-1-6654-7087-2/22/$31.00 ©2022 IEEE

associated with a multiple risk of breast cancer. Age, personal history records of breast cancer, circle of relative records of breast cancers are the dangerous aspects. In spite of this popularity of those risk elements, about 50 percent of ladies who develop breast cancer have no specific or identifiable risk element apart from being a woman and getting old [2]. Risk factors for breast cancer any benign disorder of breast, post-menopausal hormone substitute such as estrogen, progestin, proliferative breast sickness without atypia, menarche 55 years, sedentary life style and lack of doing exercise or workout, nulliparous, first degree family with breast cancer, post-menopausal weight issues, too much socioeconomic magnificence, history of endometrial or ovarian cancers, significant chest radiation, atypical hyperplasia and first diploma circle of relatives [8][9][10][15][16]. The rate of breast cancer will increase dramatically with the age of those women who might have these risk factors in them. More than sixty five percent of breast cancer cases rise in woman over the age of sixty [12][13].

Breast cancer can be identified as abnormality in mammograms. However, this may show its form with the discharge of nipple, lesions of breast skin or breast ache. Suspicious palpable and mammographic lesions of the breast are examined through biopsy. Mostly breast loads, mainly in young women at pre-menopausal phase are benign. Here the majority is 75% to 85% of cancerous loads spread too quickly and harmfully at the final rate of 15 to 25 percent [1]. Carcinoma in situ is characterized through the proliferation of malignant cells within the ducts or lobules of the breast without invading stromal tissue. Ductal carcinoma in situ and lobular carcinoma in situ are the most important subtypes. Lobular carcinoma in situ is microscopic and lacks clinical and mammographic signal and symptoms unlike ductal carcinoma in situ. Lobular carcinoma in situ is more likely to have bacterial engagement. Here the cells are grouped into a small solid mass with small featured, uniform, spherical or oval shaped nuclei. The American Joint Committee on cancer states that TNM class is based on the premise-cancers with identical anatomical vicinity and histology proportion similar boom spread patterns. It is based on the scale of the primary tumor, nearby lymph nodes involvements and distant metastasis [17]. The collaboration of the classifications T, N and M shows the quantity of sickness [17].

Effective treating of breast cancer depends on the situation of the patient. If the tumor can be localized then the maximum general treatment is surgery. The maximum usually used procedure for surgical operation is lumpectomy with axillary node dissection and modification of mastectomy. A mastectomy absolutely eliminates the effected breast, the fascia of underlying breast and some of the axillary nodes. The usage of Radiation therapy has been increased for few years. This remedy is quite popular when the cancer is detected at an early stage. For most early stage breast cancer patients, radiation therapy is used with a lumpectomy and surgical examination of axillary lymph nodes. For broad space or not localized breast cancers, places such as the breast, armpit and chest wall can be irradiated after surgery. Various complications because of the spread of cancer to a distant part may be correctly treated with radiation. Hormonal or pharmacological treatment is also viable. Surgery and radiation therapy are highly effective eliminating cancerous tissues when the exact place of the most cancers is understood and when close by ordinary organs and tissues may be spared without damage. Chemotherapy, alternatively, spreads at some stage in the frame. It is capable of smashing cancer cells whenever they are applied. Chemotherapy is generally used as an adjuvant therapy. It is applied while the number one tumor has been removed by using surgical treatment or maybe radiation therapy with having another or second tumor is existed. It is utilized in a few conditions wherein the cancer is localized in one place. In many instances, breast cancer boom has been proven to depend on the hormonal conditions provided by the individual patients' body condition. Hormone therapy applies some other methods to control or suppress the increasing of sensitive hormone tumor. Sometimes the suppression of tumor growth can be gained by decreasing the tiers of right and useful hormones inside our body. It can be done by surgical elimination or by x-ray destruction of an organ that is producing that particular hormone including ovaries and adrenal glands. Now there are also tablets that counter acts the impacts of useful hormones. Suppression of tumor is from time to time accomplished by means of increasing ranges of some useful hormones by means of giving them a form of medicine. Some data summarized from the National Cancer Database states that both relative and discovered survival rates increases by using years after prognosis and gradually from level to stage away. Stage IV sicknesses are declining more dramatically A research followed with 407 patients with axillary nodes, meaning poor breast cancers patient went through surgical treatment between 1976 to 1987 shows that a 10 year recurrence price of nineteen percent. They found that the tumor size is 0.0006 [18] [19]. The businesses at highest threat of recurrence are patients having tumor smaller than two centimeters. And sufferers having tumors size 1.12 centimeter are poorly differentiated or anaplastic tumors [18]. Lymph node illnesses proved as the most vital unmarried variable as a predictor of relapse. An observation of total 416 patients found that once a year charge of breast most cancers recurrence regularly expanded at some stage in the primary four years [19]. The annual hazard of recurrence for sufferers with wonderful nodes in the first year became 5 percent these rose to 10% and 14% in year 3 and 4 respectively [17]. In evaluation, in the ones sufferers who having node-poor at diagnosis, the danger of recurrence became 1 percentage within the first year and rose to five percent in later years 3 and 4 [19].

It is typically familiar that most cancers ought to be detected as early as feasible. However, a delay in prognosis or starting the treatment on late can be happen for multiple reasons. Firstly, it is difficult to detect breast cancer at the early stage due to the lack of knowledge of common people about this section. Though people are getting to know about it day by day as the alarming rate has made them concern about this cancer type. So eventually it becomes a slow process for the people of Bangladesh. However, early symptoms and signs of breast cancer can be detected by the victims and the physicians. It can be done by via screening program or through a physician or maybe one can check oneself. As the symptoms are too small to detect at an early stage, one can get wrong result. This can cause delay. Hospital delays are of interest to the felony profession and claims for delays in breast cancer diagnosis are bursting out in range and value. Sometimes, the spreading system of the tumor is so high that the whole procedure gets too slow for

the patient. Youngers are found to be more serious about health at taking breast cancer seriously than elders. Younger sufferers are keener to take proper medication, treatment and cautiousness when it comes to health mostly a very small early stage symptoms. It has been counseled that older sufferers may also do that. It has been said that older sufferers may be less possibly to seeking for scientific interest for breast problems, mainly in the presence of other co-going on age-related diseases. Physician delay is typically described because the time between the primary visit and the time of the very last surgical treatment or biopsy if that changed into the most effective surgical operation. Physician delay reduced with affected person age. This can also be happen because of the reason that the dense frequency is higher and lumpy tissues of the breast on this age. A research report said that records approximately the primary symptom or any sign of breast cancer, and as a consequence of affected persons' delay, is no longer as correct as records about medical doctor postpone [8]. The time of first symptom can be tough to establish due to the fact patients generally do now not record a genuine date and consequently many patients may additionally misjudge the time. Another predicament relates to lag bias, the unfairness that occurs whilst monitoring groups do no longer start at similar degrees inside the original situation's history [7]. Because of this the measurement of survival rate is counted usually from the time a victim first notices the symptom not from the time of prognosis or screening.

A complete overview of the worldwide observant studies has been brought to light and cooperates in 1999 to analyze the impact of the delay that happens on the survival of the breast cancer patients [20]. 87 studies on 101,954 victims published from 1907 to 1996 with statistics including put off of affected persons [20]. They classified each study into three categories for analyzing. Category-1, includes real five year survival costs after analyzing with delays much less or more than three months, much less or more than six months, or both. Category-2, real five to twelve months of survival costs excluding survival rates. Category-3, covers the studies on those terms that didn't fall into first classes, along with some research with no facts however in which researchers commented on the connection among put off and survival. So, my study is going to control the time consumption and will speeded up the further required action by the patient or doctors as it is going to predict with the first mammogram reports to get any positive or negative answer.

A several studies tested that the impact behind the scheduled radiation on survival. The impact of delayed radiation after chemotherapy on most cancers sufferers with lymph node involvement discovered that the delay of greater than 120 days increases the threat of relapse [21]. Five year survival charges can be 82 percent and 87 percent for delayed and early institution respectively. For the 42 patient who took delayed irradiation, here was in total six neighborhood relapses for 5-12 months actuarial price of 14% [20] [21]. In examine of the impact of radiation postponed after chemotherapy. It is found that the local control says that the share of the patients having no neighborhood relapses are 98percent for the early radiotherapy institution, in comparison with 76percent for the 8-12 months not on time institution [17]. Overall survival became 80percent for the beginning institution as opposed to 52% for sufferers who underwent delayed irradiation. The 8 year actuarial disease does not fastened the survival charges for an early and behind scheduled sufferers are 71% and 48% respectively. Among patients who underwent conservative surgery [17] [20] [21].

Ache in breast is pronounced as just only a symptom in about 10 percent of breast cancer patients. In the Cimprich study it is shown that the overall pain rating is 1.3 out of five, with pre-menopausal women having 1.5 median and for post-menopausal the score is 1.2 out of five [22]. Some other discomforting issues stated through the SDS are intestinal disenchanted and nausea.

The psychological results of breast most cancers are studied more often, however the period among prognosis and surgical treatment has been little researched. Again, Cimprich's examine offers a few insight. On the SDS, sufferers determined that psychological issues are the most tough to handle while waiting for surgical treatment, along with insomnia and mood swing problems occupying the bad ratings [23]. A 2nd evaluation institution of healthful women was drawn from observe through Anderson, Anderson and deProsse. Cimprich evaluated fifty one patients, along with 39 menopausal and 12 pre-menopausal [23]. Cimprich stated that pre-menopausal victims had substantially better rankings than post-menopausal sufferers. Again two other studies Romsaas et al. and Stanton and Snider additionally assessed that after diagnosis earlier than surgical procedure and the ratings of these three businesses were averaged [24]. That shows that all patients who were expecting surgery at the time of assessment had moderately worst scores in significant areas such as depression, anxiety and despair than those who are in the group of wholesome patients [24]. However these symptoms are not a lot distinct. All ladies scheduled for surgical treatment had notably worse (lower) energy scores than evaluation businesses or postmenopausal women.

Other measures of the studies reviewed that measured pleasant of existence from the time of prognosis to surgery, no other items have been investigated that might suggest worsening of the affected person's condition over the route of this particular period and they resulted mainly from the information that a surgical intervention become in development [25].

III. PROPOSED METHODOLOGY

Throughout the process we used plenty of algorithms. But our proposed model, XG Boost is employed to detect tumor from the given data. By using this classifier we get 96.49% accuracy.

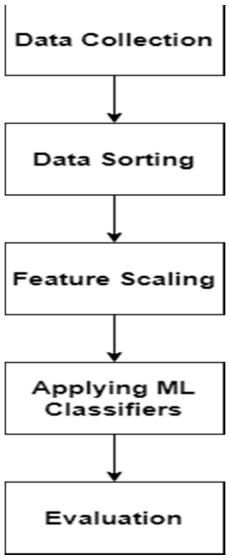

Fig. 1. Methodology Steps

A. Dataset

Initially we start with collecting data. We found some data from Kaggle. This work is very rare in the context of Bangladesh. As raw data are difficult to find, we gathered from online. First we collected the data then we categorized them according to the malignant classifications.

We have collected categorical data. These data are actual raw data of the malignant and benign cell containing people. Data Collection needs to be done seriously as these data will be used for different algorithms. We worked on feature selection through machine learning. Table I. shows the sample dataset that used for different algorithm.

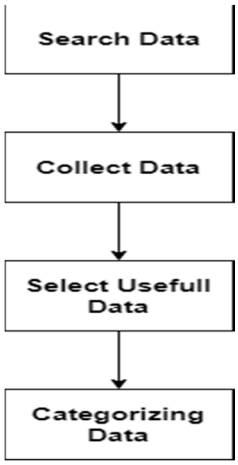

Fig. 2. Dataset Collection Procedure

There was two type of data. If the tumor is malignant then the input is 1. And cases having nonmalignant tumor is referred as 0 which suggests that significant patient doesn't have cancer in her body. We've got collected data of 570 patients. We've trained 570 dataset with 31 features and this model supported XG boost classifier. Here all feature data types are within the float data type.

We have data of two types of patient. If the tumor is malignant then it's 1 and if not then it is 0 which suggests that significant patient doesn't have cancer in her body.

B. Preprocessing

We have collected data from kaggle. And we have collected total 570 patients' reports to train our ML classifier. We have found mixed data. Image data and other non-usable data were there. So, we cut out all those data. We create a data frame to set our data. And then we did feature selection. We split our data in train and test. Then we used feature scaling to convert different units and magnitude to at least one unit. And in the end these pre-processing, we move forward to the ML classifier to search out the simplest one. We've got train and test our dataset with various ML algorithms and understand that the XG boost is that the most suited one that provides us the best accuracy rate.

TABLE I: Sample Dataset

C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
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	texture_v
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	17.3
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	23.4
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	25.5
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	26.5
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	16.6
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	23.7
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	27.6
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	28.1
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	30.7
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	40.6
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	33.8
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	27.2
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	29.9
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	27.6
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	32.0
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	37.1
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	30.8
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	31.4
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	30.8
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	19.2
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	20.4
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	15.6
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	19.0
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	35.5
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	31.5
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	21.4
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	33.2
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	27.2
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	36.7
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	19.5

C. *Model Architect*

Our motive was to applying machine learning classifiers to detect malignant cells which will lead us to get the result of breast cancers' positivity or negativity. We have applied multiple algorithms in our project. Our applied algorithms are,

XG Boost, Ada boost, KNN classifier, Logistic Regression classifier, Random Forest classifier, Decision tree classifier, Naïve Bayes classifier and Support Vector Machine classifier.

Machine learning is divided into two categories. First one is supervised learning classification and therefore the second is unsupervised learning classification. Betting on the used data with their availability, we are using some supervised learning classifiers here. Firstly we used this pair-plot to show the numeric distribution within the scatter so our data

is ready to visualize using these commands.

```
sns.pairplot(cancer_df, hue = 'target')
```

```
sns.pairplot(cancer_df, hue = 'target',
```

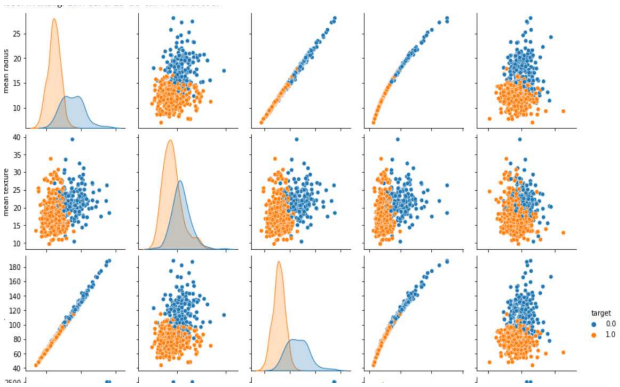

Fig. 3. Scatter plot

The malignant and tumor data are showed in two classes within this pair-plot. Our Scatter plot shows the total count of malignant and benign tumor patients using this command given below.

```
sns.countplot(cancer_df['target'])
```

Max samples mean radius is capable 1 in our case. These are shown within the given counterplot below.

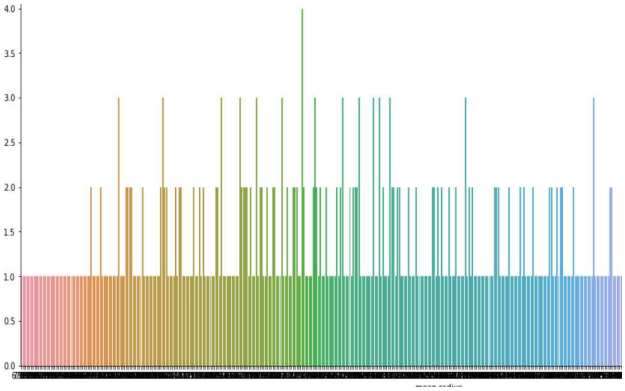

Fig. 4. Whole count of two cells.

We use Heatmap to determine the range of various features' value. We used heatmap with the matrix to find out the correlation between each feature and our target to visualize.

Fig. 5. Visualization through heatmap.

Our proposed model algorithm:

XG Boost Machine Learning classifier.

D. *Feature scaling*

After that we processed our data and split it in train and test. Then we used feature scaling to convert different units and magnitude to at least one unit. And in the end these pre-processing, we move forward to the ML classifier to search out the simplest one. We've got train and test our dataset with various ML algorithms and understand that the Xgboost is that the most suited one that provides us the best accuracy rate.

E. *Applying Machine Learning Classifier*

We have applied multiple algorithms in our project. Our applied algorithms are XG Boost, Ada boost, KNN classifier, Logistic Regression classifier, Random Forest classifier, Decision tree classifier, Naïve Bayes classifier and Support Vector Machine classifier.

Support Vector Machine or SVM is one of the mostly used Supervised Learning algorithms. This is often used for classification and regression issues both.

K-Nearest Neighbour is that the easiest of all Machine Learning algorithms which is predicated on Supervised Learning technique. KNN algorithms assumption is that the similarity between newer data and the available or older data and then put the newer data into one specific category which is mostly like the available categories. The basic Naïve Bayes theorem says that each and every feature makes an equal and independent contribution to the given result. The decision tree classifier is built a classification model. They are doing it creating a choice tree. Each node of the tree specifically applies a test on a specific attribute. Each section or branch are descended from that node corresponds to at least one of the foremost possible values for that specific attribute. A random forest may be called a meta-estimator. It fitted different types of decision tree classifiers on different sub-samples of our dataset. Then uses by creating an average to boost the prognostic accuracy rate and also controls over-fitting. An Ada boost started fitting with our very original dataset and then fitted its extra copies of the classifier on an identical dataset but weighted of incorrectly classified instances are adjusted. XG boost is highly flexible, portable

and efficient. In our model we decide XG boost that offers us the most effective accuracy rate among other classifiers.

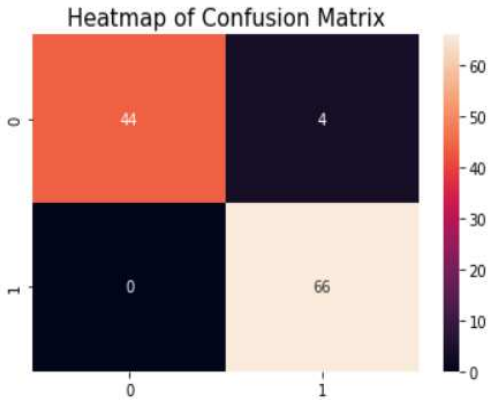

Fig. 6. Error showing model.

Confusion matrix is that the summary of predictions on a classification problem. It gives us actuality understanding about the errors that are made by our classifiers and also shows their type.

Here, our model shows type two error zero. Cross validation could be a procedure won't to calculate machine learning models on an awfully limited data sample. It is actually used in any applied machine learning model to find out the abilities of the unseen data.

F. Model Training

We've got collected data of 570 patients. We've used 570 dataset with 31 features and this model supported XG boost classifier. Our targeted column was one, having two binary variables 0 and 1. For training our data length was 455 and to test our data length was 114.

IV. MODEL PERFORMANCE

A. Training, Testing and Cross Validation

After training couple of ML algorithms, we saw that Ada boost, Naïve Bayes, XG Boost and Random forest classifiers are giving the highest accuracy rate then the other classifiers than others that we have tried. From them we choose XG Boost classifier as it gives us the highest accuracy rate than others. It gives us accuracy at the rate of 96.49%. Here to search out if our selected ML model is over-fitted, under fitted or perfectly fitted also referred as normalized we have done cross-validation for XG Boost. Our output is, that shows that it's a bit over-fitted.

```
Cross validation accuracy of XGBoost model = [1.        0.97826087 0.95652174 0.97826087
 1.        0.97777778 0.97777778 0.88888889]

Cross validation mean accuracy of XGBoost model =  0.9603864734299516
```

Fig. 7. Cross-validation report.

B. Model performance

After completing ML model, to deploy it we saved it first. We use the pickle to avoid wasting it. Thus we discover the most effective suited machine learning algorithm for our project which is XG Boost classifier. Confusion matrix of XG boost model:

```
Confusion matrix of XGBoost model:
 [[44  4]
 [ 0 66]]

Accuracy of XGBoost model =  0.9649122807017544
```

Fig. 9. Accuracy report

Our Confusion matrix showing in figure 7 is mainly used to justify our models' performance. For each and every type's dataset, inputs in the following matrix are True positive rate, false positive rate, True negative rate and false negative rate. And the accuracy rate is the division of the absolute numbers of the prediction and the predictions that are correct. This is the by which we found our confusion matrix.

C. Result compression

In recent years, a lot of research has been done to detect cancer cells through machine learning classifiers. We have tried to use some classifiers to detect the best results for breast cancer identification.

To predict breast cancer at early stage with the best accuracy this model is built for. Throughout our research our goal was to find the best classifier to predict breast cancer. And compare to other classifiers, XG Boost gave the best performance. So, we have selected it for our model.

V. CONCLUSION

The features are all non-null, inferring that there are no null values present in any of the features. The presence of null values throws off any algorithm, resulting in unforeseen outcomes. As a result, various imputing procedures must be used to replace the null values. The relationship between two variables is depicted in a pair plot. The lack of a barrier between two classes can be seen in all of the plots, especially in the pair plot. The Counterplot also reveals that the number of benign instances (goal '0') outnumbers the number of malignant ones (target '1'). This is referred to as an unbalanced dataset. The correlation matrix depicts the relationship between several characteristics or variables. The correlation between two variables might be anything between -1 and +1. Here, a correlation matrix expresses that -1 is a non-positive correlation, which means if one value increases then the other will decrease and vice versa. For +1, the opposite is true. When we have two strongly correlated independent variables, we should eliminate one of them to avoid the multi-collinearity problem. The coefficients for the two strongly correlated variables will be unreliable in those circumstances. None of the factors in our situation had a high correlation score. As a result, none of the variables had to be represented.

In comparison to other ML algorithms, XG Boost performs exceptionally well, as evidenced by the literature. When we train with default parameters, we acquire accuracy of 98.24%. Finding the most optimum values for the parameters improves accuracy rather frequently. It frequently utilizes far more resources than its competitors. If resource consumption is a significant concern, we may need to sacrifice final accuracy and use Naive Bayes instead. XG Boost, on the other hand, only fails to classify two benign

situations. It's possible that the two instances were anomalous. Outlier detection, as previously noted, can be utilized to improve the performance of all of the techniques.

REFERENCES

[1] American Cancer Society Cancer Facts and Figures. Atlanta, GA, American Cancer Society, 2005.

[2] Surveillance, Epidemiology, and End Results Program. Available at http://www.seer.cancer.gov. Accessed August 16, 2004.

[3] K. Kerlikowske. Epidemiology of ductal carcinoma in situ. J Natl Cancer Inst Monogr. 2010;2010(41):139-41.

[4] D.M. Parkin et al. Incidence in Five Continents. 1997, Lyon: IARC Press, VIII:

[5] N. Howlader et al. (eds). SEER Cancer Statistics Review, 1975–2017, National Cancer Institute. Bethesda, MD, based on November 2019 SEER data submission, posted to the SEER web site, April 2020.

[6] C.K. Anders CK et al., Breast cancer before age 40 years. *Semin Oncol*, 2009. 36(3): p. 237–49.

[7] S.A. Begum et al., Attitude and Practice of Bangladeshi Women towards Breast Cancer: A Cross Sectional Study. Mymensingh Med J. 2019 Jan;28(1):96-104. PMID: 30755557.

[8] S.J. London et al.. A prospective study of benign breast disease and the risk of breast cancer. JAMA. 1992;267:941–944. [PubMed] [Google Scholar]

[9] W.D. Dupont WD, Parl FF, Hartmann WH, Brinton LA, Winfield AC, Worrell JA, Schuyler PA, Plummer WD. Breast cancer risk associated with proliferative breast disease and atypical hyperplasia. Cancer. 1993;71:1258–1265. [PubMed] [Google Scholar]

[10] J. Wang, Constantino JP, Tan-Chiu E, Wickerham DL, Paik S, Wolmark N. Lower-category benign breast disease and the risk of invasive cancer. J Natl Cancer Inst. 2004;96:616–620. [PubMed] [Google Scholar]

[11] L.C. Hartmann, T. A. Sellers, M. H. Frost,W.L. Lingle, A.C. Degnim , Ghosh K, et al. Benign breast disease and the risk of breast cancer. New Engl J Med. 2005;353:229–237. [PubMed] [Google Scholar]

[12] L.C. Collins, H. J. Baer, R. M. Tamimi RM, J. L. Connolly, G.A. Colditz, S.J. Schnitt. Magnitude and laterality of breast cancer risk according to histologic type of atypical hyperplasia. Cancer. 2007;109:180–187. [PubMed] [Google Scholar]

[13] A.C. Degnim, D. W. Visscher, H. K. Berman, M. H. Frost, T.A. Sellars , R.A. Vierkant, et al. Stratification of breast cancer risk in women with atypia: a Mayo cohort study. J Clin Oncol. 2007;25:2671–2677. [PubMed] [Google Scholar]

[14] M.J. Worsham, U. Raju, M. Lu, A. Kapke, A. Bottrell, J. Cheng, et al. Risk factors for breast cancer from benign breast disease in a diverse population. Breast Cancer Res Treat. 2008 Oct 4;

[15] C. Byrne, J. L. Connolly, G.A. Colditz, S.J. Schnitt. Biopsy confirmed benign breast disease, postmenopausal use of exogenous female hormones, and breast carcinoma. Cancer. 2000;89:2046–2052.

[16] G.C. Kabat, J. G. Jones, N. Olson, A. Negassa, C. Duggan, M. Ginsberg , R.A. Kandel, A.G. Glass, T.E. Rohan. A multi-center prospective cohort study of benign breast disease and risk of subsequent breast cancer.

[17] American Joint Committee on Cancer. Breast. In: AJCC Cancer Staging Manual. 8th ed. New York, NY: Springer; 2017:589.

[18] G.N. Peters, M. Wolff, C.D. Haagensen. Tubular carcinoma of the breast. Clinical pathologic correlations based on 100 cases. Ann Surg 1981; 193:139–149.

[19] R.W. McDivitt, W. Boyce, D. Gersell. Tubular carcinoma of the breast. Clinical and pathological observations concerning 135 cases. Am J Surg Pathol 1982; 6:401–411.

[20] MA, Westcombe AM, Love SB, Littlejohns P, Ramirez AJ. Influence of delay on survival in patients with breast cancer: a systematic review. Lancet. 1999 Apr 3;353(9159):1119-26. doi: 10.1016/s0140 6736(99)02143-1. PMID: 10209974.

[21] Flores-Balcázar Christian et al.. Impact of Delayed Adjuvant Radiotherapy in the Survival of Women with Breast Cancer. Cureus. 2018 Jul 30;10(7):e3071. doi: 10.7759/cureus.3071. PMID: 30510860; PMCID: PMC6267615.

[22] Baseline health status and setting impacted minimal clinically important differences in COPD: an exploratory study; Harma Alma, Corina de Jong, Danijel Jelusic, Michael Wittmann, Michael Schuler.

[23] Nour Abuhadra et al., Early-stage Triple-negative Breast Cancer: Time to Optimize Personalized Strategies, The Oncologist, 10.1093/oncolo/oyab003, 27, 1, (30-39), (2022).

[24] J. Kamath, et al. Symptom Distress Associated with Biopsy in Women with Suspect Breast Lesions", *International Scholarly Research Notices*, vol. 2012, Article ID 898327, 9 pages, 2012.

[25] World Health Organization. Regional Office for Europe. (2020). Screening programmes: a short guide. Increase effectiveness, maximize benefits and minimize harm. World Health Organization. Regional Office for Europe.