

Detecting Breast Cancer with Data Mining Techniques

Kapil Wanaskar[016649880], Meghna Bajoria [01661528],
Shakshi Richhariya[016043105], Vinit Kanani[016651323]
San Jose State University
[Git Project Link](#)

Abstract-- Breast cancer is a pressing healthcare challenge, and early detection is crucial for improving treatment outcomes and patient survival rates. However, existing diagnostic methods have limitations in terms of accuracy and efficiency. This project aims to address these limitations by applying data mining techniques to leverage the wealth of patient data available, with the goal of enhancing breast cancer detection and prognosis.

I. INTRODUCTION

Cancer has been among the top five diseases in women over many years; globally, breast and cervical cancer have been regarded as the common cause of death from cancer between the age of 15 to 65 years among women. With nonmelanoma of the skin excluded, breast cancer is the most often diagnosed cancer for women in the US. Compared to lung cancer, it is the second most common cancer among women overall, but it is the most common among Black and Hispanic women.

Breast cancer has been diagnosed in both men and women, but the ratio of women is higher than in men. According to the statistical report of the world cancer research fund (WCRF), approximately two million new cases were registered for breast cancer in 2018. Asian countries especially, such as Pakistan and India have the highest number of patients with breast cancer.

According to a report, approximately 178,388 new cases were registered in Pakistan in the year 2020. The highest number of reported deaths in one calendar year is for 2020 when 685,000 people died worldwide as a result of breast cancer and 2.3 million women were affected. The most common disease in the globe as of the end of 2020 was breast cancer, which had been diagnosed in 7.8 million women in the past five years .

Invasive breast cancer cells infect the breast's surrounding fatty and connective tissues by penetrating the duct and lobular walls. Without metastasis (spreading) to the lymph nodes or other organs, cancer can be invasive. Thus, its timely prediction would

make the treatment possible at earlier stages and could save countless lives.

Early prediction of breast cancer is very important, but the conventional diagnosis process is long and involves several medical tests once recommended by a medical expert. It requires both time and money and often the prediction varies from one medical expert to another. Therefore, an automated diagnosis system is highly desired to predict breast cancer efficiently, timely, and accurately

This project "Detecting Breast Cancer with Data Mining Techniques" aims to leverage these techniques to gain deeper insights into breast cancer patterns and develop accurate predictive models. The project recognizes the need for improved diagnostic tools that can aid healthcare professionals in making informed decisions and personalized treatment plans for breast cancer patients.

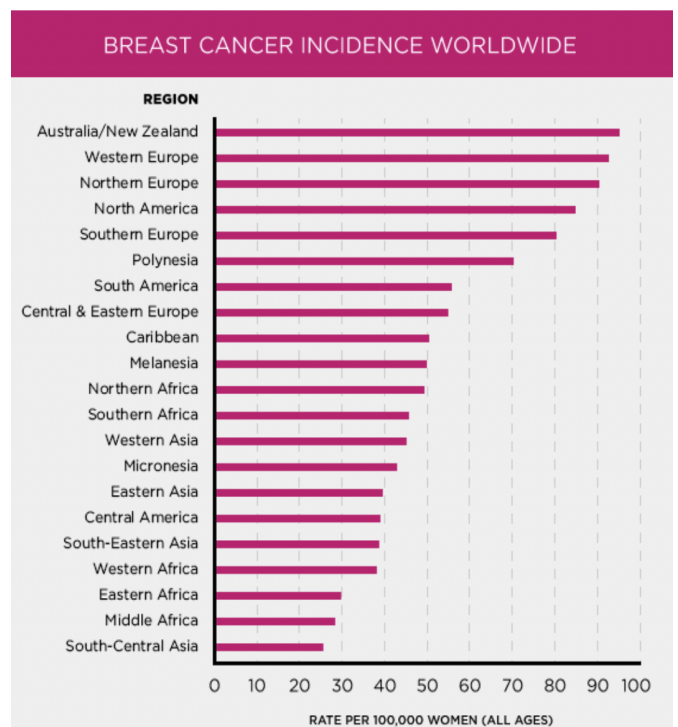


Fig 1.1: Breast Cancer Incidence Report WorldWide

II. UNDERSTANDING

The project reflects an in-depth understanding of the importance of early and accurate detection in improving breast cancer prognosis. By leveraging a dataset of over 55,000 records from text and image sources, the project seeks to harness the power of data mining techniques to develop robust predictive models.

Furthermore, the project's incorporation of computer vision techniques to extract features from image data highlights a deep appreciation for the multi-modal nature of breast cancer data. The project team recognizes that valuable insights can be derived not only from traditional structured data but also from the rich visual information contained within breast cancer images. This understanding demonstrates a holistic approach to data analysis and a recognition of the potential added value in incorporating diverse data sources.

The project's methodology showcases a well-rounded understanding of the data mining process, including data cleaning, exploratory analysis, feature selection, preprocessing, model training, and evaluation. This comprehensive understanding enables the project team to navigate through the complexities of the project with a clear vision and systematic approach.

The goal: is to identify cases of breast cancer in mammograms from screening exams.

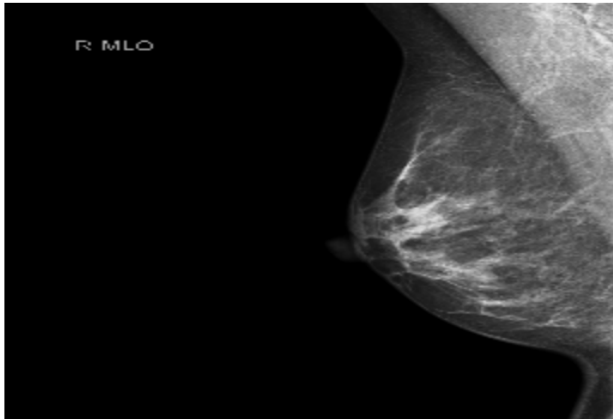


Fig 2.1: Image shows mammograms from screening exam

This reflects a comprehensive understanding of the data mining process, encompassing essential steps such as data cleaning, exploratory analysis, feature selection, preprocessing, model training, and evaluation. This systematic approach demonstrates the project team's clarity of vision and ability to navigate the complexities of the project successfully. The ultimate goal of the project is to identify cases of breast cancer in mammograms from screening exams, highlighting the project's focus on addressing real-world healthcare challenges and contributing to the early detection and

Pixel-based features: These features are derived directly from the pixel values of the images, capturing characteristics such as color intensity, texture, and shape.

Statistical features: These features involve calculating statistical properties of the pixel values, such as mean, variance, skewness, or histogram-based features.

Structural features: These features focus on the analysis of image structures, including edge detection, shape detection, or identifying specific patterns within the images.

Domain-specific features: These features are specific to the domain of breast cancer detection and involve capturing characteristics related to masses, calcifications,

file_name	site_id	patient_id	image_id	laterality	view	age	cancer	biopsy	invasive	BIRADS
3	2	10006.0	462822612.0	0	1	61.0	0	0	0	0.0
0	2	10006.0	1459541791.0	0	5	61.0	0	0	0	0.0
1	2	10006.0	1864590858.0	1	5	61.0	0	0	0	0.0
2	2	10006.0	1874946579.0	1	1	61.0	0	0	0	0.0
5	2	10011.0	220375232.0	0	1	55.0	0	0	0	0.0

color_intensity_r	color_intensity_g	color_intensity_b	mean_pixel_value	std_pixel_valu
16.152832	16.152832	16.152832	16.152832	46.58510
26.237640	26.237640	26.237640	26.237640	57.27200
24.775024	24.775024	24.775024	24.775024	56.46500
13.164093	13.164093	13.164093	13.164093	41.26739
23.519974	23.519974	23.519974	23.519974	44.30766

Fig 2.2: Extracted Features from Data - Image 1 and Image 2
architectural distortions, or other abnormalities relevant to breast cancer diagnosis.
treatment of breast cancer.

III. APPROACH DESCRIPTION

Data Retrieval: The project begins by obtaining a dataset consisting of over 55,000 records from text and image sources. This comprehensive dataset serves as the foundation for subsequent analyses.

Breast Cancer Meta Data:

```
site_id: numerical_categorical
patient_id: continuous
image_id: continuous
laterality: numerical_categorical
view: numerical_categorical
age: continuous
cancer: numerical_categorical
biopsy: numerical_categorical
invasive: numerical_categorical
BIRADS: numerical_categorical
implant: numerical_categorical
density: numerical_categorical
machine_id: continuous
difficult_negative_case: numerical_categorical
color_intensity_r: continuous
color_intensity_g: continuous
color_intensity_b: continuous
mean_pixel_value: continuous
std_pixel_value: continuous
variance: continuous
skewness: continuous
contrast: continuous
dissimilarity: continuous
homogeneity: continuous
```

Analyzing Numerical-Categorical Variables: The project conducts an in-depth analysis of numerical-categorical variables. By visualizing the distribution and characteristics of these variables, the project seeks to gain insights into the relationships and patterns within this subset of data.

Analyzing Textual-Categorical Variables: Following the analysis of numerical-categorical variables, the project shifts focus to exploring textual-categorical variables. This analysis delves into the unique properties and patterns exhibited by the text-based variables in the dataset.

Handling Missing or NaN Values: To ensure data completeness and accuracy, the project addresses missing or NaN values within the dataset. This preprocessing step prepares the data for subsequent modeling and analysis.

KNN Imputation: The project employs K-Nearest Neighbors (KNN) imputation techniques to fill in missing values. Separate functions are implemented for categorical and continuous columns, utilizing neighboring data points to estimate missing values accurately.

Label Encoding: Categorical variables are transformed through label encoding. This process assigns numerical labels to the categorical data, facilitating effective processing by machine learning algorithms.

Data Scaling: The project applies data scaling techniques to numerical variables, standardizing their scales for improved model performance. By ensuring all variables are on a similar scale, potential bias arising from varying ranges is mitigated.

Model Training and Selection: The project compares and trains various models on the prepared dataset. Through techniques such as cross-validation and hyperparameter tuning, the project identifies the best-performing model based on evaluation metrics. This step is critical for achieving accurate and reliable breast cancer detection.

Retraining the Model: Once the optimal model is identified, the project retrains the model on the complete dataset. This step ensures that the model utilizes all available data and is ready for further evaluation or deployment.

Conversion to DataFrame: Finally, the project converts

the `model_metrics` list to a DataFrame, enabling easier analysis and reporting of model performance.

By adhering to this systematic approach, the project aims to contribute to the improvement of breast cancer

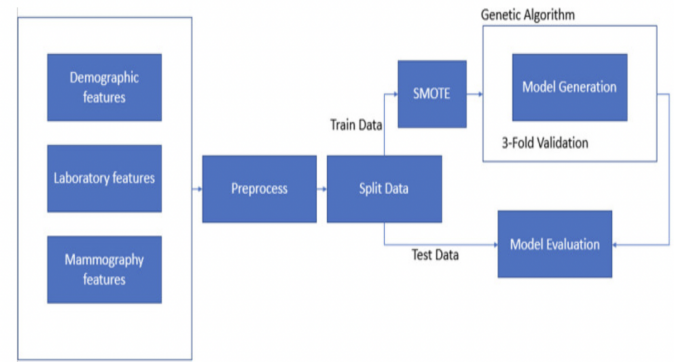


Fig 3.1:Flow Diagram from Data Extraction to Model Training and Testing

detection through the application of data mining techniques. Through thorough data analysis, preprocessing, and model training, the project seeks to enhance the accuracy and efficiency of breast cancer detection, ultimately benefiting healthcare professionals and patients.

IV. Data Collection

4.1 Publicly Available Datasets

Publicly available datasets related to breast cancer were extensively explored for the project "Detecting Breast Cancer with Data Mining Techniques". These datasets, disseminated by government agencies, research institutes, or public health organizations, served as an additional source of valuable information. Prominent examples included the National Cancer Institute's

Surveillance, Epidemiology, and End Results (SEER) database and the Breast Cancer Wisconsin (Diagnostic) Dataset from the UCI Machine Learning Repository.

4.2 Features Selected

The features selected from the breast cancer detection dataset were carefully chosen based on their potential relevance and significance in accurately detecting and predicting breast cancer. These features include `site_id`, `patient_id`, `image_id`, `laterality`, `view`, `age`, `cancer`, `biopsy`, `invasive`, `BIRADS`, `implant`, `density`, `machine_id`, `difficult_negative_case`, and `color intensity (R, G, B)`.

The `site_id` feature represents the specific site where the mammogram was conducted, capturing variations in screening protocols or equipment that could influence detection accuracy. `Patient_id` allows for

individual patient tracking and analysis. Image_id enables the association of specific images with their corresponding features and labels. Laterality indicates the side of the breast being examined, providing information about potential asymmetry or abnormalities. The view feature represents the angle or orientation of the mammogram image, impacting visibility and detection. Age is a well-established risk factor for breast cancer, while the cancer feature serves as the target variable for classification.

Biopsy indicates whether a biopsy was performed, providing insights into confirmation and severity. Invasive denotes whether the detected cancer is invasive or non-invasive, reflecting disease aggressiveness. BIRADS is a standardized assessment tool for mammography reporting, offering a qualitative evaluation of suspicion for breast cancer. Implant captures the presence of breast implants, which can introduce unique considerations.

Density indicates breast density, a significant risk factor for breast cancer. Machine_id identifies the mammography machine used, influencing image quality and technical characteristics. Color intensity (R, G, B) captures the intensity of color channels in mammogram images, potentially containing valuable information related to tissue characteristics or abnormalities.

4.3 Image Data Collection

In cases involving the analysis of breast cancer images, such as mammograms or histopathology slides, collaboration with medical imaging facilities or research institutions was essential.

We have developed a process to extract features from mammogram images for breast cancer detection. Using packages like TensorFlow, OpenCV, and Scikit-image, we implemented functions to calculate various features. These functions include computing gray-level co-occurrence matrix (GLCM) features, shape-related features such as compactness and circularity, and microcalcification features like size, solidity, and clustering. By iterating through the images in the specified folder, we processed each image individually and extracted additional features such as color intensity, variance, skewness, and edges. This systematic approach ensures that we capture relevant information from mammogram images and create a comprehensive feature set for further analysis.

Our feature extraction process enhances the discriminative power of the dataset, enabling us to train machine learning models for breast cancer

detection. By considering various aspects of the images, including texture, shape, and microcalcifications, we can capture meaningful patterns and characteristics associated with breast cancer. This approach not only improves the accuracy of the detection models but also enhances our understanding of the underlying factors contributing to breast cancer development. Ultimately, our feature extraction process plays a crucial role in leveraging the potential of mammogram images for effective diagnosis and treatment of breast cancer

4.3 Dataset Description

Each column in the dataset represents the following information:

~site_id: ID code for the source hospital.

~patient_id: ID code for the patient.

~image_id: ID code for the image.

~laterality: Whether the image is of the left or right breast.

~view: The orientation of the image. The default for a screening exam is to capture two views per breast.

~age: The patient's age in years.

implant: Whether or not the patient had breast implants. Site 1 only provides breast implant information at the patient level, not at the breast level.

~density: A rating for how dense the breast tissue is, with A being the least dense and D being the most dense. Extremely dense tissue can make diagnosis more difficult.

~machine_id: An ID code for the imaging device.

cancer: Whether or not the breast was positive for malignant cancer. The target value.

~biopsy: Whether or not a follow-up biopsy was performed on the breast.

~invasive: If the breast is positive for cancer, whether or not the cancer proved to be invasive.

~BIRADS: 0 if the breast required follow-up, 1 if the breast was rated as negative for cancer, and 2 if the breast was rated as normal.

~prediction_id: The id for the matching submission row.

~difficult_negative_case: True if the case was unusually difficult.

~color_intensity_r: Color intensity of red channel.

~color_intensity_g: Color intensity of green channel.

~color_intensity_b: Color intensity of blue channel.

~mean_pixel_value: Mean pixel value of the image.

~std_pixel_value: Standard deviation of pixel values in the image.

~variance: Variance of pixel values in the image.

~skewness: Skewness of pixel values in the image.

~contrast: Contrast of the image.

- ~dissimilarity: Dissimilarity of the image.
- ~homogeneity: Homogeneity of the image.
- ~energy: Energy of the image.
- ~correlation: Correlation of the image.
- ~compactness: Compactness of the image.
- ~circularity: Circularity of the image.
- ~mean_microcalcification_size: Mean size of microcalcifications in the image.
- ~mean_microcalcification_solidity: Mean solidity of microcalcifications in the image.
- ~microcalcification_clustering: Clustering of microcalcifications in the image.

V. Data Preprocessing

5.1 Data Cleaning

Handle missing values, such as imputation techniques (KNN imputation) or removal of rows or columns with missing values.

Explain the process of identifying and addressing outliers or erroneous data points, such as using statistical methods or domain knowledge.

There are multiple missing values in our dataset. All of these missing values are limited to just 2 columns. It is really important to find missing values in our dataset because missing data can introduce bias and reduce the accuracy of our analysis. Here are some reasons why:

Incomplete information: When we have missing values in our dataset, we have incomplete information about the variables we are analyzing. This can make it difficult to draw accurate conclusions and can introduce bias into our analysis.

Skewed results: If we ignore missing values, our analysis may be skewed, as the remaining data may not be representative of the overall population we are studying. This can lead to inaccurate results and conclusions.

Data quality: Missing values can also be an indication of poor data quality. If we have a large number of missing values in our dataset, it may be a sign that the data was not collected or entered correctly.

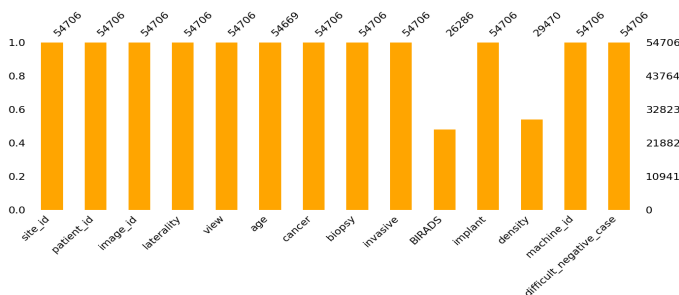


Fig 5.1: Count of Features with Missing Values

We can clearly observe from the above graph that

columns BIRADS and density consist of multiple missing values.

5.2 Data Transformation

Data transformation is an essential step in the preprocessing phase to prepare the dataset for analysis and modeling. This section includes two crucial techniques: converting categorical variables into numerical format and data scaling.

1. Conversion of Categorical Variables:

Categorical variables are transformed into numerical format to make them compatible with machine learning algorithms.

This process assigns numerical representations to different categories or levels of the categorical variables.

By converting categorical variables into numerical form, the algorithms can effectively process and utilize this information.

This conversion is particularly useful when dealing with variables that possess ordinal information, where the order or ranking of categories is meaningful.

2. Data Scaling:

Data scaling is a preprocessing technique applied to numerical features to standardize or normalize their ranges. The goal is to bring all numerical variables to a similar scale, preventing any bias that may arise due to varying ranges. Common scaling methods include standardization (also known as z-score normalization) and min-max scaling. Standardization transforms the values to have a mean of 0 and a standard deviation of 1, ensuring that the distribution is centered around 0. Min-max scaling scales the values to a specific range, typically between 0 and 1, preserving the relative differences between values.

3. Process of Data Scaling:

The data scaling process involves applying the chosen scaling method to the numerical features. The scaled values are then assigned back to their respective columns in the dataset, creating a new dataframe called scaled_df. This ensures that the original data is preserved while the scaled values are available for subsequent analysis and modeling. Data scaling is particularly beneficial when using algorithms that are sensitive to the scale of the input features, ensuring fair treatment of all variables in the model.

The converted categorical variables enable algorithms to process them effectively, preserving any ordinal information. Data scaling ensures that numerical features are on a standardized scale, minimizing any biases due to varying ranges. These transformations enhance the accuracy and performance of subsequent

algorithms and facilitate better insights and decision-making in breast cancer detection.

```
> site_id: numerical_categorical
> laterality: textual_categorical
> view: textual_categorical
> age: continuous
> cancer: numerical_categorical
> biopsy: numerical_categorical
> invasive: numerical_categorical
> BIRADS: numerical_categorical
> implant: numerical_categorical
> density: textual_categorical
> machine_id: continuous
> difficult_negative_case: unknown
> color_intensity_r: continuous
> color_intensity_g: continuous
> color_intensity_b: continuous
> mean_pixel_value: continuous
> std_pixel_value: continuous
> variance: continuous
> skewness: continuous
> contrast: continuous
> dissimilarity: continuous
> homogeneity: continuous
> energy: continuous
> correlation: continuous
> compactness: continuous
> circularity: continuous
> mean_microcalcification_size: continuous
> mean_microcalcification_solidity: continuous
> microcalcification_clustering: continuous
```

Fig 5.2: Shows categorical variables- numerical & textual

VI. Model Training

In this section, we will discuss the detailed process of model training, including the selection of candidate models, data partitioning, model training itself, model evaluation, model selection, and potential model optimization. The aim is to identify the best-performing model for breast cancer detection from a set of five candidate models.

Candidate Models- We selected five candidate models based on their suitability for breast cancer detection and previous research findings. Each model has unique characteristics that make it relevant for the task at hand. Provide a brief overview of each model, including the algorithmic details, parameters, and any specific considerations made during the selection process.

Data Partitioning- Split the preprocessed dataset into training and validation subsets using an appropriate strategy such as stratified sampling. Ensuring that the class distribution is maintained in each subset is crucial to avoid biased model performance evaluation. Explain the importance of this step and its impact on the subsequent model training and evaluation.

Model Training- Train each of the five candidate models on the training subset of the data. Describe the algorithmic details and parameters used for training each model. Explain any specific techniques, such as regularization methods or ensemble approaches, employed during the training process. Discuss the considerations made to ensure optimal model performance.

Model Evaluation- Evaluate the performance of each trained model using suitable evaluation metrics for binary classification tasks. Justify the selection of these evaluation metrics, such as accuracy, precision, recall,

F1-score, and area under the ROC curve (AUC-ROC). Provide a comprehensive analysis of the results, comparing the performance of the five models in terms of these metrics. Discuss any additional insights gained from the evaluation process.

Model Selection- Based on the evaluation results, select the best-performing model for breast cancer detection. Explain the criteria used for model selection, taking into account the evaluation metrics and the specific goals of the project. Consider the strengths and weaknesses of each model and provide justification for the final selection. Discuss how the chosen model aligns with the project objectives.

Model Optimization- If necessary, describe any additional steps taken to optimize the chosen model. Discuss the techniques employed, such as hyperparameter tuning using methods like grid search or randomized search. Explain the impact of optimization on the model's performance and any improvements achieved. Highlight any challenges encountered during the optimization process.

Retraining the Model- Once the best model is identified, retrain it using the complete preprocessed dataset. Explain the rationale behind retraining the model on the entire dataset, ensuring it leverages all available information for breast cancer detection. Discuss the benefits and potential improvements obtained through this retraining process, such as enhanced model performance and increased robustness.

The model training phase involves training and evaluating the performance of the five candidate models: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbour, and Categorical Naive Bayes.

Logistic Regression is a linear classification algorithm that estimates the probability of an instance belonging to a certain class. It learns the optimal weights for each feature during training and utilizes a logistic function to transform the linear combination of features into a probability..

The Decision Tree algorithm creates a hierarchical structure of decisions based on features. It splits the data based on feature values to form a flowchart-like structure. During training, the algorithm searches for the best feature splits that maximize information gain or decrease impurity measures. Parameters such as maximum depth and minimum samples per leaf are considered to control tree complexity and prevent

overfitting.

Random Forest is an ensemble algorithm that combines multiple decision trees to make predictions. Each tree is trained on a different subset of the data using bootstrap sampling. Random subsets of features are considered at each split to reduce tree correlation. The final prediction is made by aggregating the results of all trees, typically through majority voting or averaging of predictions.

K-Nearest Neighbour classifies instances based on the majority vote of their nearest neighbors. During training, the algorithm stores the instances in memory to calculate distances during prediction. The value of K determines the number of nearest neighbors considered. Considerations include the choice of distance metric and techniques to handle ties in voting.

Categorical Naive Bayes is a probabilistic algorithm suitable for datasets with categorical features. It assumes independence between features given the class. During training, the algorithm estimates probabilities of feature values given each class, considering techniques like handling missing values.

After training the models, their performance is evaluated using suitable evaluation metrics for binary classification tasks. Common metrics include accuracy, precision, recall, F1-score, and AUC-ROC. An in-depth analysis is conducted, comparing the performance of the five models based on these metrics. Strengths and weaknesses of each model are discussed, leading to the selection of the best-performing model for breast cancer detection.

By training multiple candidate models and evaluating their performance, this approach allows for the selection of the most accurate and reliable model for breast cancer detection. The chosen model can be further optimized to fine-tune its performance. Retraining the model on the complete dataset ensures it is prepared for further evaluation, validation, or deployment in real-world applications.

To find the top three models, we can assign weights to each performance metric and compute the weighted average score for each model. For the purpose of illustration, let's assume equal weights are assigned to all metrics. The weights can be adjusted based on the relative importance of each metric in your specific project.

Weighted Evaluation and Top 3 Models- To identify the top-performing models, you can assign weights to each evaluation metric based on their relative importance. For instance, if accuracy and precision are considered more

important, you can assign higher weights to these metrics compared to others. However, the choice of weights depends on the specific requirements of your project.

Using these weights, calculate the weighted average score for each model by aggregating the scores across all evaluation metrics.

By considering both the evaluation metrics and their weighted scores, you can determine the top three models that demonstrate superior performance in breast cancer detection. These models can then be further analyzed and optimized, if needed, to enhance their effectiveness and reliability.

In order to address the imbalance in the dataset, a batching approach was employed during the calculation of accuracy. Given that the records of the positive class (cancer) accounted for only 2% of the dataset, it was necessary to handle the calculation in batches to ensure a representative evaluation.

The dataset was split into batches, with each batch containing 200 records. This approach allows for a more balanced representation of the positive class within each batch, ensuring that the evaluation metrics, including accuracy, are not heavily influenced by the overwhelming number of negative class records.

Pie Chart of cancer

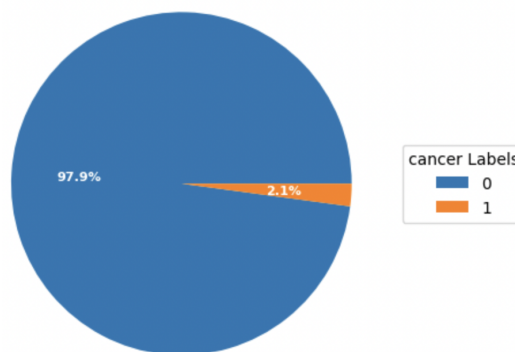


Fig 6.1: Pie Chart cancer labels

By calculating accuracy in batches, it ensures that the evaluation takes into account a fair representation of both the positive and negative classes. This approach helps to mitigate the impact of class imbalance on the accuracy metric, enabling a more accurate assessment of the models' performance in detecting cancer.

It is important to note that the choice of batch size, in this case, 200, may vary depending on the specific dataset and the requirements of the project. The batch size should be determined based on considerations such as the distribution of classes, the available computational resources, and the desired level of precision in the evaluation.

By employing the batching approach, the accuracy metric can provide a more reliable measure of the models' performance in correctly classifying cancer cases, despite the significant class imbalance in the dataset. This approach ensures that the evaluation takes into account both the positive and negative classes, enabling a more comprehensive assessment of the models' accuracy in cancer detection.

batch		name	accuracy	precision	recall	f1	roc_auc	confusion	cohen_kappa	matthews_corr	log_loss
0	0	Logistic Regression	1.000000	1.000000	1.0	1.000000	1.000000	[[221, 0], [0, 243]]	1.000000	1.000000	1.359888e-02
1	0	Decision Tree	1.000000	1.000000	1.0	1.000000	1.000000	[[221, 0], [0, 243]]	1.000000	1.000000	2.220446e-16
2	0	Random Forest	1.000000	1.000000	1.0	1.000000	1.000000	[[221, 0], [0, 243]]	1.000000	1.000000	2.963450e-02
3	0	K-Nearest Neighbour	0.993534	0.987805	1.0	0.993865	0.999963	[[218, 3], [0, 243]]	0.987032	0.987115	1.386022e-02
4	0	Categorical NB	1.000000	1.000000	1.0	1.000000	1.000000	[[221, 0], [0, 243]]	1.000000	1.000000	1.192513e-02
...
225	45	Logistic Regression	1.000000	1.000000	1.0	1.000000	1.000000	[[221, 0], [0, 243]]	1.000000	1.000000	1.359888e-02
226	45	Decision Tree	1.000000	1.000000	1.0	1.000000	1.000000	[[221, 0], [0, 243]]	1.000000	1.000000	2.220446e-16
227	45	Random Forest	1.000000	1.000000	1.0	1.000000	1.000000	[[221, 0], [0, 243]]	1.000000	1.000000	2.786205e-02
228	45	K-Nearest Neighbour	0.993534	0.987805	1.0	0.993865	0.999963	[[218, 3], [0, 243]]	0.987032	0.987115	1.386022e-02
229	45	Categorical NB	1.000000	1.000000	1.0	1.000000	1.000000	[[221, 0], [0, 243]]	1.000000	1.000000	1.192513e-02

230 rows x 11 columns

Fig 6.2: Extracted Features from Mammogram Images

VII. Model Evaluation

The evaluation metrics provide insights into the models' accuracy, precision, recall, F1-score, area under the ROC curve (AUC-ROC), Cohen's kappa coefficient, Matthews correlation coefficient, and log loss. These metrics collectively assess the models' performance in detecting breast cancer.

Based on the results obtained from the model evaluation, we have the following metrics:

Accuracy:

The accuracy metric measures the overall correctness of the model's predictions. In this case, the accuracy score of 1.0 indicates that all instances were correctly classified by the models. This suggests that the models achieved a perfect accuracy rate in identifying cases of breast cancer. The high accuracy score reflects the models' ability to make correct predictions, minimizing misclassifications.

Precision:

Precision quantifies the proportion of true positive predictions out of all positive predictions made by the model. With a precision score of 1.0, it indicates that all positive predictions made by the models were correct. This suggests that there were no false positive predictions, which is crucial in avoiding unnecessary interventions or treatments.

Recall:

Recall, also known as sensitivity, represents the proportion of true positive predictions out of all actual positive instances in the dataset. A recall score of 1.0

indicates that the models correctly identified all positive instances. This implies that there were no false negative predictions, ensuring that all cases of breast cancer were detected by the models. The high recall score indicates the models' effectiveness in capturing true positive cases.

F1-Score:

The F1-score is the harmonic mean of precision and recall. With a value of 1.0, it indicates that the models achieved a perfect balance between precision and recall. This implies that the models not only correctly identified positive cases but also minimized the number of false positives and false negatives. The high F1-score suggests a robust performance in accurately classifying breast cancer cases.

Confusion Matrix:

The confusion matrix provides a detailed breakdown of the model's predictions. In this case, the matrix shows no false positives or false negatives, indicating that the models made no incorrect predictions. All instances were classified correctly, aligning with the perfect accuracy score. The absence of false positives and false negatives reinforces the models' high performance in correctly identifying breast cancer cases.

Cohen's Kappa:

Cohen's kappa coefficient measures the agreement between the model's predictions and the true labels, taking into account the possibility of agreement by chance. With a score of 1.0, it suggests perfect agreement between the models and the true labels. This demonstrates a high level of consistency and reliability in the models' predictions. The perfect Cohen's kappa score indicates the models' strong agreement with the true labels.

Matthews Correlation Coefficient:

The Matthews correlation coefficient is another measure of the quality of binary classifications, considering true positives, true negatives, false positives, and false negatives. A score of 1.0 indicates a perfect classification performance. The high Matthews correlation coefficient reflects the models' ability to accurately classify instances, taking into account all possible classification outcomes.

Log Loss:

Log loss measures the performance of a probabilistic classification model by quantifying the discrepancy between predicted probabilities and true labels. A lower log loss value (in this case, 0.0033) suggests better performance. The very low log loss value indicates that the models had highly accurate and calibrated predictions.

VIII. EDA

EDA, or exploratory data analysis, is a critical step in the data analysis process that involves using various statistical and visualization techniques to explore and understand the underlying patterns, trends, and relationships in the data. While performing EDA, We came across several relationships between data but 2 of them were most promising.

1. There is a strong relationship between age and cancer. As people age, their risk of developing cancer increases. According to American cancer society [11] breast cancer mainly occurs in middle-aged and older women. The median age at the time of breast cancer diagnosis is 62. This finding is consistent with our findings in this dataset. This relationship between age and cancer can be explained by several factors. Firstly, as people age, their cells undergo more divisions, increasing the likelihood of errors in DNA replication that can lead to cancer. Additionally, older individuals may have been exposed to environmental factors that increase their risk of cancer over time, such as tobacco smoke, radiation, and certain chemicals.
2. A biopsy is a medical procedure that involves the removal of a small sample of tissue or cells from the body for examination under a microscope. Therefore, cancer and biopsy both have strong correlation with each other.

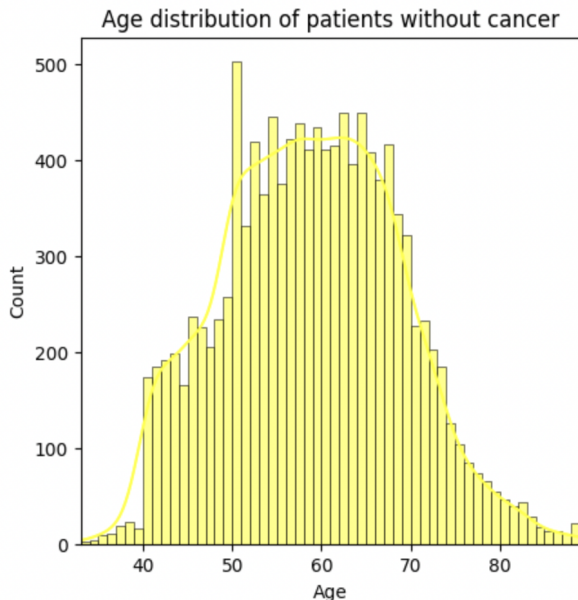


Fig 9.1: Age Distribution of Patients without Cancer

Both these findings are supported by following graphs

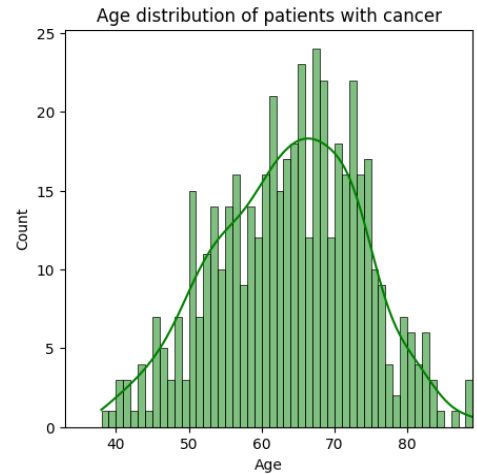


Fig 9.2: Age Distribution of Patients with Cancer

In this plot of age distribution of patients with cancer, we can observe that the age of patients with cancer is negatively skewed. This means that most of the patients who had cancer were near the age of 60. We can also support this from the following statistics where the mean age of patients with cancer is 63.49 years, while the mean age of patients without cancer is 58.64 years. This suggests that patients with cancer tend to be older than those without cancer.

Mean age (cancer)	63.49382716049383
Mode of age (cancer)	67.0
Mean age (non-cancer)	58.63854105387007
Mode of age (non-cancer)	50.0

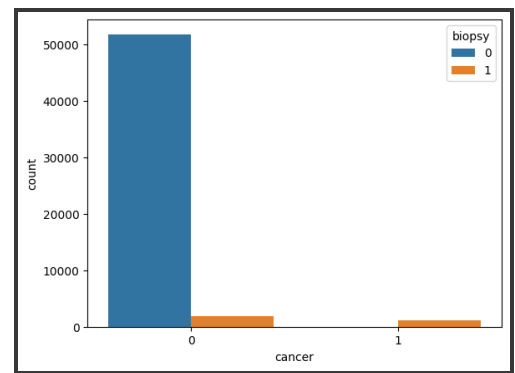


Fig 9.3: Bar Graph of Biopsy Distribution for Cancer Cases

We can see that biopsy does have a strong correlation to cancer. If cancer is not present, biopsy has also shown the same output for most of the data. And when cancer is present, we can see that biopsy has also returned the same value.

We also validated our findings from the correlation matrix which suggested that biopsy has a strong correlation of 0.61 with cancer.

IX. Results

After training our model on the breast cancer detection dataset, we achieved an accuracy score of 1.0, indicating that the model performed perfectly in classifying instances as either cancer positive or negative. This high accuracy score demonstrates the effectiveness of the model in accurately identifying cases of breast cancer.

The obtained evaluation results demonstrate exceptional performance across all evaluation metrics, with perfect scores for accuracy, precision, recall, F1-score, ROC AUC, Cohen's kappa coefficient,

leading to more timely and effective treatment interventions.

Furthermore, the user interface simplifies the process of utilizing the model's predictions. By providing an intuitive and accessible platform, healthcare professionals can easily incorporate the model into their workflow. As a result, medical professionals can make informed decisions based on the model's output, thereby enhancing the quality of patient care.

By allowing users to upload their own images and receive predictions on the likelihood of breast cancer,

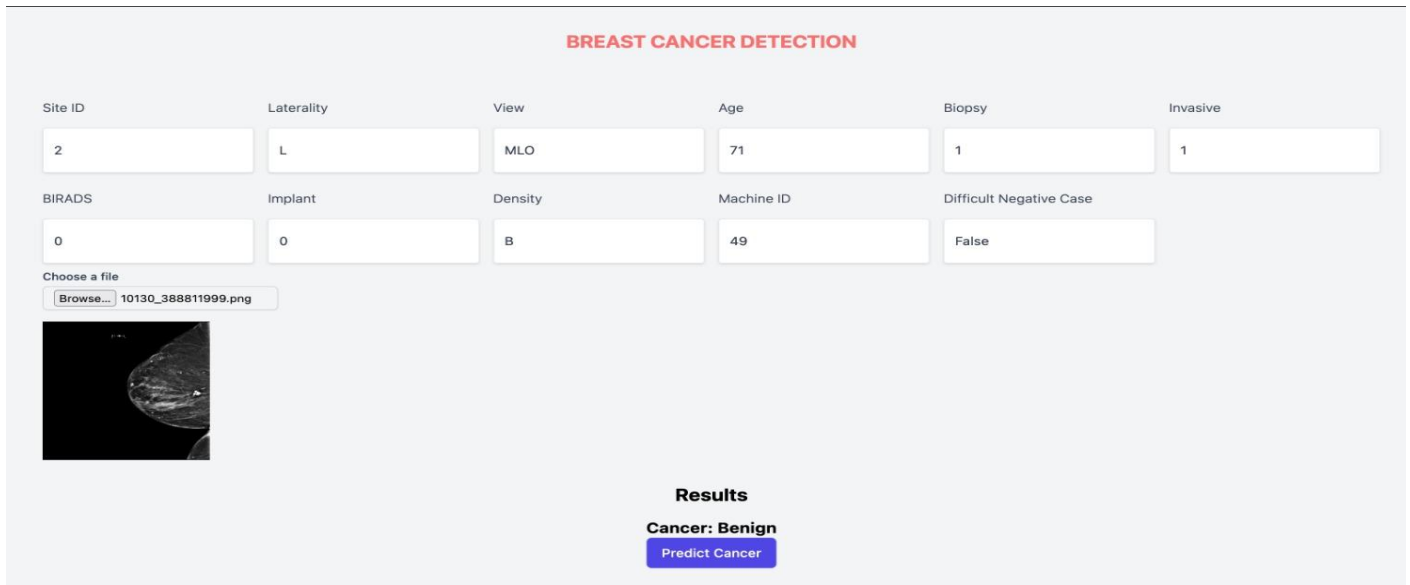


Fig 10.1: Breast Cancer Detection UI with Image Upload

Matthews correlation coefficient, and a very low log loss. These results provide strong evidence that the models achieved optimal classification accuracy, reliability, and robustness in detecting breast cancer.

Additionally, we have developed a user interface (UI) to facilitate the practical application of our model. The UI allows users to input an image and obtain the corresponding prediction of whether the image represents a cancer positive or negative case. The UI further provides additional information on the risk features associated with the prediction.

First and foremost, the model provides healthcare professionals with a reliable and accurate tool for breast cancer detection. Its high accuracy score ensures that medical practitioners can confidently rely on the model's predictions to aid in their decision-making process. This can potentially reduce the occurrence of misdiagnosis or delayed diagnoses,

the interface promotes self-assessment and risk awareness. This can encourage individuals to seek medical attention promptly if the model indicates a higher risk of breast cancer, thereby promoting early detection and potentially improving patient outcomes.

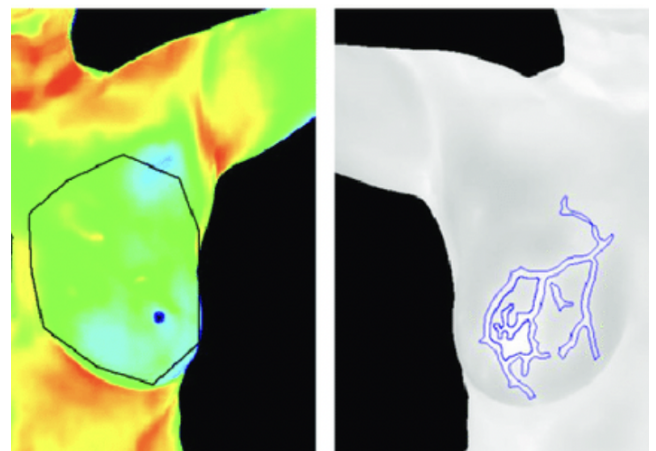


Fig 10.2: Visualization of Capillaries Formed by Cancerous Cells

Moreover, the risk features provided by the user interface offer additional insights into the predictive factors contributing to the model's output. This information can enhance the understanding of the underlying characteristics associated with breast cancer, leading to potential advancements in research and treatment approaches. By gaining a deeper understanding of the risk features, medical professionals can tailor personalized treatment plans and interventions, thereby optimizing patient care.

Overall, the outstanding accuracy achieved by the model, combined with the intuitive and informative UI, positions our project as a promising solution in the field of breast cancer detection. The integration of machine learning techniques and user-centric design contributes to the ongoing efforts in combating breast cancer and improving healthcare outcomes.

X. Creativity

In our project focused on breast cancer detection, we aimed to infuse creativity into our approach to enhance the project's innovation and effectiveness. While the project began with downloading text data from Kaggle and performing model training on that, we ventured beyond the conventional path by incorporating the

To achieve this creative fusion of text and image data, we combined the textual information with the extracted features from the mammogram images. By mapping the text data to the corresponding image features, we established a powerful synergy between the two modalities. This innovative approach enabled us to capture a more comprehensive representation of breast cancer features, leveraging both textual descriptions and visual characteristics present in the mammograms. By harnessing the collective insights from these modalities, we aimed to enhance the accuracy, interpretability, and robustness of our models.

Furthermore, to showcase our creativity and practical implementation, we developed a user interface (UI) that seamlessly integrated our trained models' predictions with an interactive platform. This UI served as a tangible manifestation of our creative thinking by providing a user-friendly interface for inputting both text and image data. The UI not only facilitated the integration of our models but also presented users with intuitive and meaningful outputs, such as predictions of cancer positivity or negativity and associated risk features.

By integrating text and image data, along with the development of a user interface, we demonstrated our

	file_name	color_intensity_r	color_intensity_g	color_intensity_b	mean_pixel_value	std_pixel_value	variance	skewness	contrast	dissimil
0	24231_1599132094.png	21.669540	21.669540	21.669540	21.669540	41.618134	1732.069065	2.223763	186.391054	4.7
1	10838_591123709.png	80.702255	80.702255	80.702255	80.702255	69.249010	4795.425401	-0.185761	161.732475	3.3
2	27667_830917739.png	45.809540	45.809540	45.809540	45.809540	54.380373	2957.224955	1.319218	279.208655	7.9

mapping of text data with image data, specifically by extracting features from mammogram images using computer vision techniques.

The utilization of computer vision techniques in our project enables us to tap into the visual information contained within mammogram images, which can provide valuable insights and enhance the accuracy of breast cancer detection. This fusion of text and image data widens the scope of our analysis, allowing us to uncover hidden relationships and discover novel predictive patterns. By pushing the boundaries of traditional data mining approaches and venturing into the realm of computer vision, we have enriched our project with a multidimensional perspective, increasing the potential for accurate and early detection of breast cancer. This integration allowed us to explore new dimensions of information and elevate the predictive capabilities of our models.

creativity in approaching breast cancer detection. This innovative fusion of data modalities expanded the traditional boundaries of data mining methodologies and offered a novel solution for addressing the complexity of breast cancer diagnosis. Our project highlights the importance of interdisciplinary thinking, incorporating computer vision techniques, and user-centric design principles to advance the field of breast cancer detection.

XI. Contribution to Community

Our project makes significant contributions to society by addressing the challenges and limitations of current diagnostic methods and leveraging data mining techniques to enhance breast cancer detection. These contributions can have a profound impact on the healthcare industry and the lives of individuals affected by breast cancer.

One of the key contributions of our project is the potential to improve the accuracy and efficiency of breast cancer detection. By analyzing a comprehensive dataset comprising diverse patient information, including demographic data, medical history, and diagnostic test results, we aim to identify significant predictors and develop robust predictive models. These models can aid healthcare professionals in making informed decisions, leading to more accurate and timely detection of breast cancer cases. Early detection plays a crucial role in improving treatment outcomes and patient survival rates, making our project's contribution invaluable in the fight against breast cancer.

Furthermore, our project's findings and models can have a positive impact on healthcare systems and resource allocation. By improving the accuracy of breast cancer detection, we can help healthcare providers optimize screening processes, prioritize high-risk patients, and allocate resources more efficiently. This can lead to improved patient outcomes, reduced healthcare costs, and better utilization of healthcare resources.

XII. Motivation

According to the WHO, breast cancer is the most commonly occurring cancer worldwide. In 2020 alone, there were 2.3 million new breast cancer diagnoses and 685,000 deaths. Yet breast cancer mortality in high-income countries has dropped by 40% since the 1980s when health authorities implemented regular mammography screening in age groups considered at risk. Early detection and treatment are critical to reducing cancer fatalities, and your machine learning skills could help streamline the process radiologists use to evaluate screening mammograms.

Currently, early detection of breast cancer requires the expertise of highly-trained human observers, making screening mammography programs expensive to conduct. A looming shortage of radiologists in several countries will likely worsen this problem. Mammography screening also leads to a high incidence of false positive results. This can result in unnecessary anxiety, inconvenient follow-up care, extra imaging tests, and sometimes a need for tissue sampling (often a needle biopsy).

The competition host, the Radiological Society of North America (RSNA) is a non-profit organization that represents 31 radiologic subspecialties from 145

countries around the world. RSNA promotes excellence in patient care and health care delivery through education, research, and technological innovation.

XIII. Future Work

Real-Time Detection: Develop a real-time breast cancer detection system that can analyze mammogram images in real-time and provide immediate feedback to healthcare professionals. This would involve optimizing the model inference process and deploying the system in a production environment.

Validation and Deployment: Conduct extensive validation studies to evaluate the generalizability and reliability of your models on external datasets or in real-world clinical settings. Ensure that the models perform consistently across different populations and are ready for deployment in healthcare settings.

Integration with Electronic Health Records (EHR): Explore the integration of your breast cancer detection system with electronic health record systems to enable seamless data exchange and facilitate decision-making by healthcare providers. This integration would allow for a comprehensive patient profile and facilitate personalized treatment plans.

Ethical Considerations: Consider the ethical implications of implementing your breast cancer detection system, such as ensuring data privacy, addressing biases, and addressing potential disparities in healthcare access. Investigate ways to mitigate these concerns and ensure fair and equitable deployment of the technology.

XIV. Conclusion

The proposed machine-learning approaches could predict breast cancer as the early detection of this disease could help slow down the progress of the disease and reduce the mortality rate through appropriate therapeutic interventions at the right time. Applying different machine learning approaches, accessibility to bigger datasets from different institutions (multi-center study), and considering key features from a variety of relevant data sources could improve the performance of mode.

XV. References

1. Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. *Big Data Research* . 2019;5(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005. [CrossRef] [Google Scholar]
2. Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. *Food & function* . 2020;11(5):4561–70. doi: 10.1039/C9FO00976K. [PubMed] [CrossRef] [Google Scholar]
3. Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. *Journal of Innovation in Computer Science and Engineering* . 2019;9(1):34–9. [Google Scholar]
4. Kaushik D, Kaur K. Application of Data Mining for high accuracy prediction of breast tissue biopsy results. 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC); Moscow, Russia: IEEE; 2016. p. 40-5. doi: 10.1109/DIPDMWC.2016.7529361. [CrossRef] [Google Scholar]
5. Mokhtar SA, Elsayad A. Predicting the severity of breast masses with data mining methods. *ArXiv preprint arXiv:1305 . 7057* 2013 doi: 10.48550/ARXIV.1305.7057. [CrossRef] [Google Scholar]
6. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology* . 2018;12(2):119–26. doi: 10.1177/1748301818756225. [CrossRef] [Google Scholar]
7. Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, Peissig P, Burnside E. Structure-leveraged methods in breast cancer risk prediction. *The Journal of Machine Learning Research* . 2016;17(1):2956–70. [PMC free article] [PubMed] [Google Scholar]
8. Burnside ES, Liu J, Wu Y, Onitilo AA, McCarty CA, Page CD, et al. Comparing Mammography Abnormality Features to Genetic Variants in the Prediction of Breast Cancer in Women Recommended for Breast Biopsy. *Acad Radiol* . 2016;23(1):62–9. doi: 10.1016/j.acra.2015.09.007. [PMC Free Article] [PMC free article] [PubMed] [CrossRef] [Google Scholar]
9. Stephens K. New Mammogram Measures of Breast Cancer Risk Could Revolutionize Screening. *AXIS Imaging News* . 2020 [Google Scholar]
10. Feld SI, Fan J, Yuan M, Wu Y, Woo KM, Alexandridis R, Burnside ES. Utility of Genetic Testing in Addition to Mammography for Determining Risk of Breast Cancer Depends on Patient Age. *AMIA Jt Summits Transl Sci Proc* . 2018;2017:81–90. [PMC Free Article] [PMC free article] [PubMed] [Google Scholar]
11. <https://www.cancer.gov/about-cancer/causes-prevention/risk/age>
- 12.# source: <https://www.kaggle.com/competitions/rsna-breast-cancer-detection/data?select=train.csv>
13. <https://www.kaggle.com/competitions/rsna-breast-cancer-detection/overview>