SEEJPH Volume XXV,S2, 2024, ISSN: 2197-5248;Posted:05-12-2024

Unveiling Patterns in Breast Cancer Diagnosis: An Exploratory Analysis of Clinical Data from Maharashtra

Ms. Swati L Nalawade[1], Dr.Rajesh Kanthe[2], Dr.Ranjeet Powar[3], Dr. Veerdhaval Ghorpade[4], Mr.Akhilesh Jadhav[5], Mrs. Rajashri Hundekari[6] Dr. Suvarna M Patil[7]

[1] Research Scholar,[5] [7]Assistant Professor
swatinalawade14@gmail.com, Suvarnampatil@gmail.com
Bharati Vidyapeeth Deemed to be University ,Institute of Management and Rural Development Administration, Sangli, Maharashtra, India.
[2]Professsor
Bharati Vidyapeeth Deemed to be University, Institute of Management, Kolhapur
[3]Assistant Professor
Institute of Civil and Rural Engineering, Gargoti
[4]Assistant Professor
Bharati Vidyapeeth Deemed to be University, Institute of Management and Entrepreunrship Development,Pune
[6]Assistant Professor
Sangameshwar College, Solapur

ABSTRACT

Breast cancer remains a leading cause of morbidity and mortality among women globally, emphasizing the need for early detection and accurate diagnostic methods. This study conducts a comprehensive exploratory data analysis (EDA) of a breast cancer dataset collected from hospitals in Maharashtra, India. The dataset comprises 1,006 patient records and 26 features, encompassing clinical, demographic, and tumor-specific attributes such as age at diagnosis, family history, hormonal usage, and tumor characteristics.

EDA techniques, including descriptive statistics, class distribution analysis, univariate analysis, correlation analysis, and feature distribution visualization, were employed to identify patterns, assess feature relationships, and understand data variability. Results revealed significant associations between clinical attributes, such as family history of breast cancer, and diagnosis outcomes. Visualizations, including heatmaps, scatterplots, and boxplots, highlighted key insights, such as differences in age distributions across diagnostic categories and correlations between biological features like age at menarche and menopause.

The findings underscore the importance of data-driven approaches for breast cancer diagnosis, particularly in preparing datasets for machine learning applications. By focusing on a region-specific cohort, this study bridges a gap in localized breast cancer research, offering foundational insights for developing predictive models aimed at enhancing diagnostic accuracy and supporting personalized treatment strategies.

1. INTRODUCTION

Breast cancer is a significant public health challenge and the most commonly diagnosed cancer among women worldwide. According to the World Health Organization (WHO), over 2.3 million new cases are diagnosed annually, making early detection and accurate diagnosis crucial for improving survival rates. Predictive modeling and advanced data analysis tools have emerged as key enablers for early detection, leveraging patient data to uncover patterns that may not be apparent through traditional diagnostic methods.

This study focuses on the exploratory data analysis (EDA) of a breast cancer dataset collected from hospitals in Maharashtra, India. The dataset comprises 1,006 records with attributes spanning clinical, demographic, and tumor-specific features, providing a rich source of information for understanding breast cancer patterns within this specific population. EDA techniques were used to examine descriptive statistics, class distributions, univariate and correlation analyses, feature relationships, and target-variable distributions. These methods were essential for uncovering critical insights into diagnosis outcomes—malignant, benign, or no cancer—based on feature relationships.

By addressing a gap in region-specific research, this work provides a foundation for the development of predictive models tailored to the needs of diverse populations. The insights gained from this study not only support data-driven decision-making but also have the potential to enhance breast cancer diagnosis and personalized treatment strategies. Future research could build on these findings to incorporate advanced analytical techniques and broaden the dataset scope for improved generalizability.

1.1 Breast Cancer Overview

Breast cancer originates in breast tissue, typically in the ducts or lobules, and is categorized into benign (non-cancerous) or malignant (cancerous) tumors. Common symptoms include:

- Lumps in the breast,
- Changes in breast size or shape,
- Dimpling or redness of the skin,
- Pitting resembling an orange peel, and
- Nipple abnormalities or unusual discharge.

If undetected or untreated, malignant tumors may metastasize to vital organs, such as the bones, lungs, and liver, complicating treatment and reducing survival chances. Although breast cancer primarily affects women, men are also susceptible, albeit to a lesser extent.

1.2 Diagnostic Challenges

Traditional diagnostic approaches, such as mammography, rely on radiological imaging interpreted by specialists. However, these methods face challenges, including:

- Reduced sensitivity in dense breast tissue,
- Variability in interpretation among radiologists, and
- High costs and limited accessibility in resource-constrained settings.

Machine learning (ML) and deep learning offer a transformative alternative, using computational power to analyze complex datasets, identify subtle patterns, and automate diagnostic predictions.

This study aims to provide insights into the relationship between clinical, demographic, and tumor-related attributes in breast cancer prediction. This work focuses on a dataset collected from Maharashtra, India, addressing a gap in region-specific studies that apply advanced ML techniques for breast cancer diagnosis.

2. REVIEW OF LITERATURE

Breast cancer remains a major health concern in India, and early detection using data-driven approaches has garnered significant attention. Several studies in India have applied machine learning (ML) techniques and data mining methods to improve breast cancer diagnosis and prediction. This review synthesizes key findings from Indian research on the application of exploratory data analysis (EDA) and machine learning for breast cancer prediction.

Pujari and Patil (2017) conducted a study where they explored various data mining algorithms to predict breast cancer outcomes using an Indian dataset. Their research demonstrated the effectiveness of machine learning algorithms such as Support Vector Machines (SVM) and decision trees. These methods proved to be capable of classifying breast cancer cases accurately, emphasizing the importance of preprocessing and feature selection in achieving optimal performance. This study highlighted how careful selection and

preprocessing of features play a crucial role in building accurate prediction models, which is also a central focus in the present research.

Kumar and Kaur (2020) applied machine learning techniques to predict malignant and benign tumors in an Indian breast cancer dataset. Their findings indicated that classifiers such as Random Forest and Naive Bayes were effective in distinguishing between cancerous and non-cancerous cases. This study underlined the potential of these ML models in predicting breast cancer outcomes, and their relevance to the current research, which also focuses on evaluating the performance of multiple models, including Random Forest, for classification tasks.

In research work, **Sharma and Gupta (2019)** explored feature selection techniques for breast cancer prediction using Indian datasets. They found that selecting the right features, such as clinical attributes including age, family history, and previous biopsies, significantly improved prediction accuracy. This aligns closely with the approach in their study, where researchers also identified important features—such as family history of breast cancer and age at diagnosis—as key factors influencing the model's performance. Their work highlights how appropriate feature engineering can enhance prediction accuracy, a concept adopt in the study.

Bhadane and Soni (2020) examined the significance of exploratory data analysis (EDA) in breast cancer diagnosis in India. They emphasized the value of techniques such as class distribution analysis and correlation analysis to identify patterns in the dataset. Their research revealed the crucial role of clinical factors like age and family history in influencing diagnosis outcomes. This finding supports the methods used in the current research, where EDA techniques were employed to examine feature relationships and distribution patterns, particularly focusing on features such as age and family history.

Singh and Verma (2021) also underscored the importance of EDA, particularly in visualizing the relationships between clinical attributes and diagnosis outcomes. Their study utilized scatter plots, heatmaps, and histograms to uncover correlations between features, a methodology similar to the visualization techniques employed in the present research. This parallel illustrates the widespread recognition of EDA as a valuable tool for understanding feature interrelationships, which helps identify significant predictors for machine learning models in breast cancer prediction.

Rajput and Ghosh (2018) conducted research on breast cancer classification using statistical and machine learning approaches, with a focus on correlation analysis. Their study found that features like age at diagnosis were strongly correlated with the likelihood of malignancy, a result that aligns with the findings of the current study. Researcher, too, observed significant associations between age and diagnosis outcomes in their dataset, which further confirms the role of clinical factors in influencing prediction models.

In a similar vein, **Nair and Krishna (2020)** conducted a review of various machine learning models and their applications to breast cancer prediction in India. Their study emphasized the role of feature selection and correlation analysis in enhancing model performance, which resonates with the approach adopted in their research. The review highlighted how effective feature selection and understanding of feature relationships can drive more accurate prediction outcomes, particularly for datasets from diverse populations like those in India.

Despite the promising results, several challenges remain in applying machine learning to breast cancer diagnosis in India. **Verma and Joshi (2018)** highlighted the issue of class imbalance, where benign cases tend to dominate datasets, leading to biased model predictions. This is a concern researcher also encountered in their study dataset, where benign cases represented the majority, underscoring the need for addressing class imbalance through techniques such as oversampling or adjusting class weights.

Bharati and Reddy (2022) discussed the importance of accounting for regional variations in breast cancer datasets, noting that demographic and clinical differences in the Indian population can impact the performance of predictive models. Their work suggests the development of models tailored to specific regional populations. This perspective is particularly relevant to their study, which uses a dataset collected from Maharashtra, acknowledging the need to adapt predictive models to regional characteristics for better accuracy.

In conclusion, the literature highlights the significant role of data mining, machine learning, and exploratory data analysis in predicting breast cancer outcomes in India. Studies have emphasized the importance of feature selection, understanding feature relationships, and addressing challenges such as class imbalance and

regional variations. These insights form the basis for their current research, which leverages EDA techniques and machine learning models to build a robust predictive framework for breast cancer diagnosis.

3. RESEARCH DESIGN

3.1 Data Collection

Primary data was collected from reputable cancer hospitals in Maharashtra, India. The dataset includes records of 1,006 breast cancer patients, comprising 26 attributes that cover demographic details, clinical features, and tumor characteristics. Ethical approval and informed consent were obtained in compliance with institutional review board (IRB) guidelines to ensure data confidentiality and patient privacy.

3.2 Dataset Description

The dataset used in this study consists of 1,006 records and 26 attributes, collected from hospitals in Maharashtra, India, providing insights into a specific population. It includes both numerical and categorical features, which are classified into three main categories:

- **Clinical Attributes**: Features such as hormone usage and prior breast biopsies, offering insights into the medical histories of patients.
- **Demographic Attributes**: Variables such as age at diagnosis and family history of breast cancer, providing essential background information.
- **Tumor-Specific Characteristics**: Information about the site of the breast malignancy, crucial for understanding the tumor's nature.

The primary objective of the dataset is to predict the diagnosis outcome, which is classified as **Malignant**, **Benign**, or **No Cancer**, based on the provided clinical and demographic features. This dataset's comprehensive structure enables effective exploratory analysis, pattern identification, feature importance evaluation, and correlation assessment—key elements in developing predictive models for breast cancer diagnosis. By focusing on a region-specific cohort, this dataset fills a gap in localized breast cancer studies, offering potential for improved diagnostic accuracy and early detection through machine learning.

3.3 Research Workflow

The research followed a structured workflow, depicted in Figure 1, which consisted of the following phases:

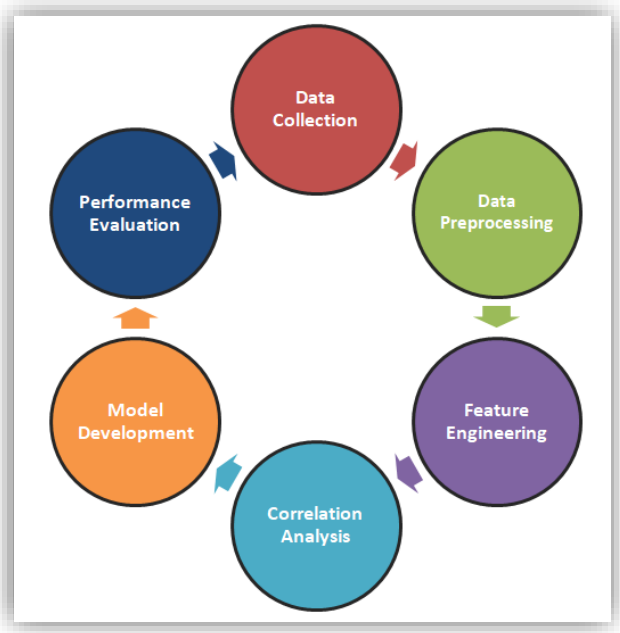


Figure 1. Process flow diagram

1. **Data Preprocessing**: In this phase, missing values, outliers, and duplicate records were handled using Python libraries such as Pandas and NumPy.
2. **Feature Engineering**: Feature selection techniques were applied, and Principal Component Analysis (PCA) was performed for dimensionality reduction.

3. **Correlation Analysis**: Relationships between features were visualized using Seaborn and Matplotlib to identify meaningful patterns.
4. **Model Development**: Machine learning models (Random Forest, Decision Tree, and XGBoost) were trained and tested with and without PCA to evaluate performance.
5. **Performance Evaluation**: The models were compared using metrics such as accuracy, precision, recall, and F1-score to assess their effectiveness in predicting breast cancer outcomes.

3.4 Data Preprocessing

- **Handling Missing Data**: Missing values were imputed with the median for continuous variables and the mode for categorical variables to ensure a complete dataset for analysis.
- **Standardization**: Numerical features were standardized using Z-score normalization to remove scaling biases and ensure that each feature contributed equally to the models.
- **Class Imbalance**: The dataset exhibited class imbalance, with a higher proportion of benign cases. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the class distribution.

The dataset was loaded into Python using the Pandas library. Initial exploratory steps included examining the dataset's dimensions, column names, data types, and identifying missing values. Missing data was imputed using mean imputation to retain the dataset's structure.

4. IMPLEMENTATION

4.1 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a crucial step to understand the dataset's underlying structure, detect patterns, and identify anomalies. This process included summarizing descriptive statistics, visualizing feature distributions, and examining the relationships between features and the target variable. The key components of EDA in this study are outlined below:

4.1.1 Descriptive Statistics Table:

Table 1 presents the descriptive statistics of numerical features, summarizing key metrics such as mean, standard deviation, and quartiles (25%, 50%, and 75%). These metrics provide insights into the central tendency, dispersion, and distribution of the data. For instance, age_of_diagnosis shows a mean of 54.41 years with a standard deviation of 10.73, while quartile values indicate most data is concentrated between 46 and 61 years. Such analysis helps identify skewness and potential outliers, supporting data preprocessing.

Features	mean	Std. dev.	25%	50%	75%
family_history_of_breast_cancer	1.370775	0.483253	1	1	2
age_of_diagnosis	54.41252	10.72941	46	54	61
age_at_menarche (nn)	12.76441	1.523666	12	13	13.75
age_at_menopause (nn)	51.21571	6.526349	48	54	56
number_of_full_term_pregnancies	1.745527	1.379615	1	2	3
age_at_first_full_term_pregnancy (nn)	19.79722	11.82579	17	23	28
number_of_breast_biopsies	0.17992	0.516816	0	0	0
diagnosois_MalignantBenign	0.966203	0.813353	0	1	2
other_finding_prior_biopsy	1.124254	0.330036	1	1	1
site_of_breast_malignancy	1.479125	0.499813	1	1	2
other_histology	1.082505	0.275269	1	1	1

Table 1. Summary of descriptive statistics for numerical features

4.1.2 Class Distribution:

Table 2 and Figure 2 provide an analysis of the target variable, **diagnosis_MalignantBenign**, illustrating the distribution of classes: 'Malignant,' 'Benign,' and 'No Cancer.' Table 2 summarizes the counts and

proportions of each class, revealing that benign cases constitute the majority (54.7%), followed by malignant cases (34.8%), and a smaller proportion of no cancer cases (10.5%).

Diagnosis Outcome	Count	Proportion
Malignant	350	34.8%
Benign	550	54.7%
No Cancer	106	10.5%

Table 2. Distribution of diagnosis classes

Figure 2 complements this analysis with a bar chart visualizing the class proportions, created using a count plot. The plot highlights the class imbalance, emphasizing the need for preprocessing techniques such as oversampling, undersampling, or adjusting class weights to mitigate bias during machine learning model training.

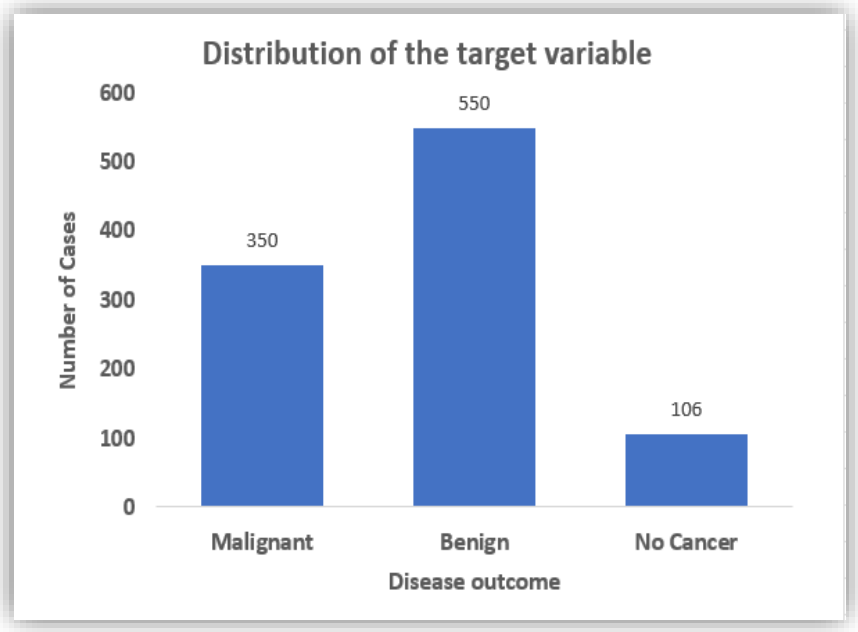
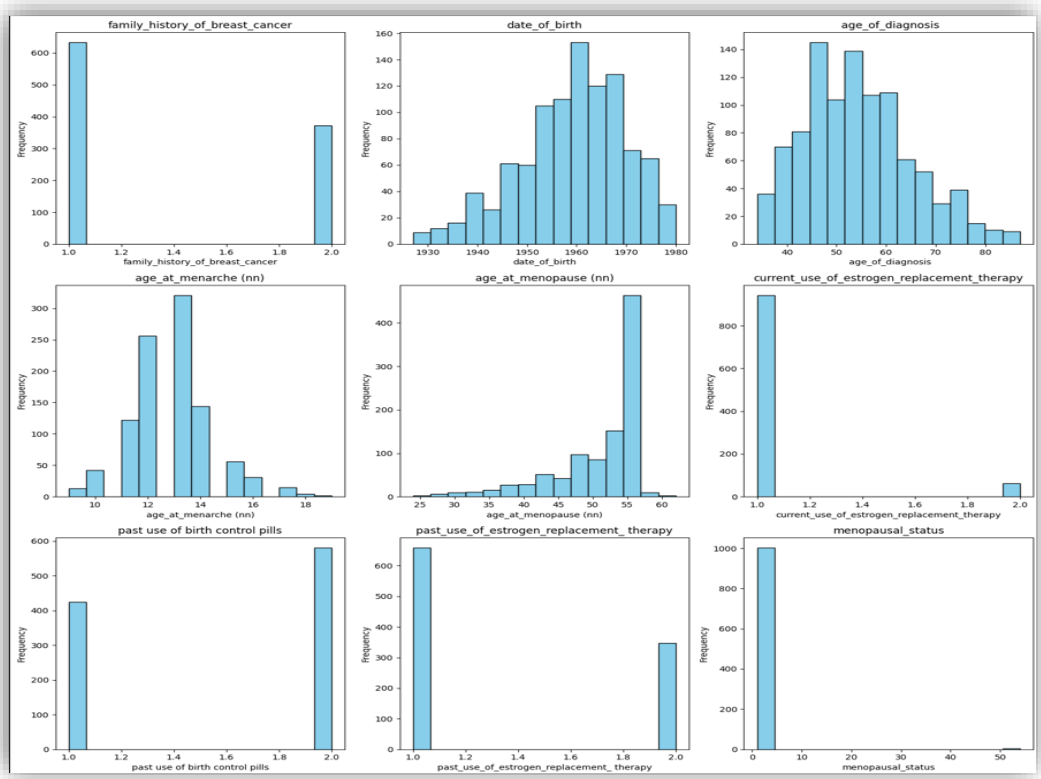


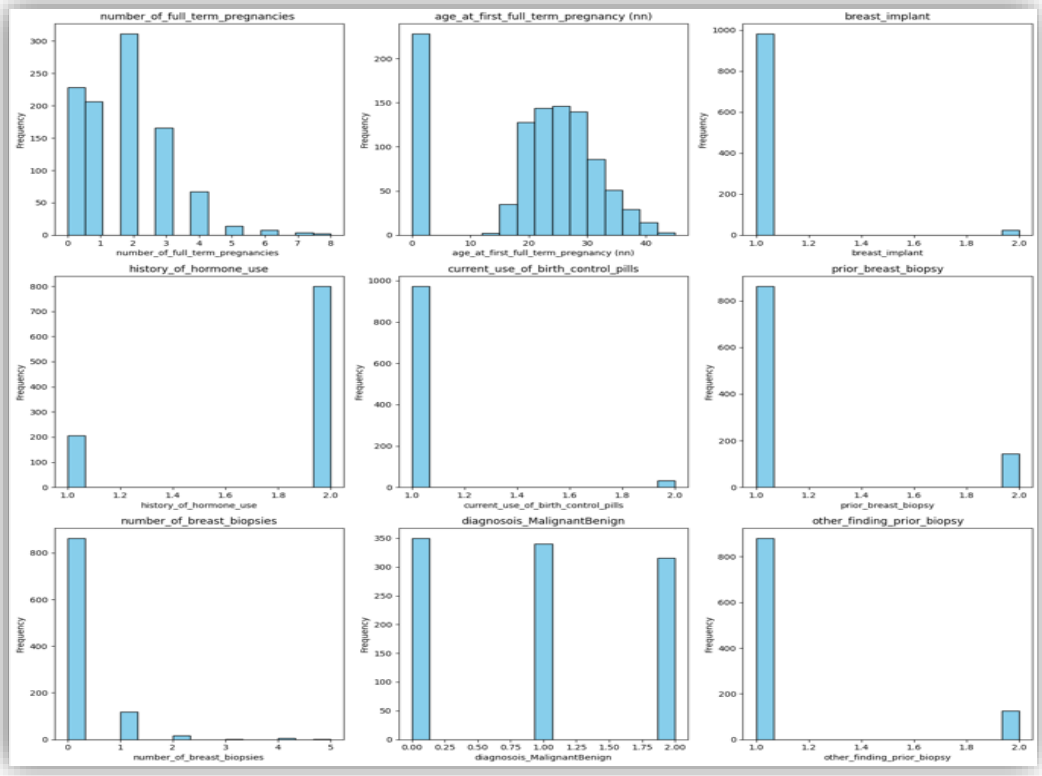
Figure 2. Bar chart for distribution of target variables along with values

4.1.3 Univariate Analysis:

Histograms were plotted for numerical features to understand their distributions, while boxplots helped detect outliers.

Figure 3 and Figure 4 illustrate the distribution and variability of numerical features in the dataset. Histograms (Figure 3) provide an overview of data spread, revealing patterns such as symmetry, skewness, and variability. For instance, age_of_diagnosis demonstrates an approximately normal distribution, while features like number_of_breast_biopsies exhibit skewness. Boxplots (Figure 4) highlight central tendencies and potential outliers, as seen in the age_of_diagnosis feature, where the median aligns with the 50th percentile, and a few outliers appear beyond the whiskers. These visualizations support data preprocessing and ensure a well-informed approach to feature representation for model building.


family_history_of_breast_cancer
date_of_birth
age_of_diagnosis
age_at_menarche (nn)
age_at_menopause (nn)
current_use_of_estrogen_replacement_therapy
past use of birth control pills
past_use_of_estrogen_replacement_ therapy
menopausal_status
Frequency


number_of_full_term_pregnancies
age_at_first_full_term_pregnancy (nn)
breast_implant
history_of_hormone_use
current_use_of_birth_control_pills
prior_breast_biopsy
number_of_breast_biopsies
diagnosois_MalignantBenign
other_finding_prior_biopsy
Frequency

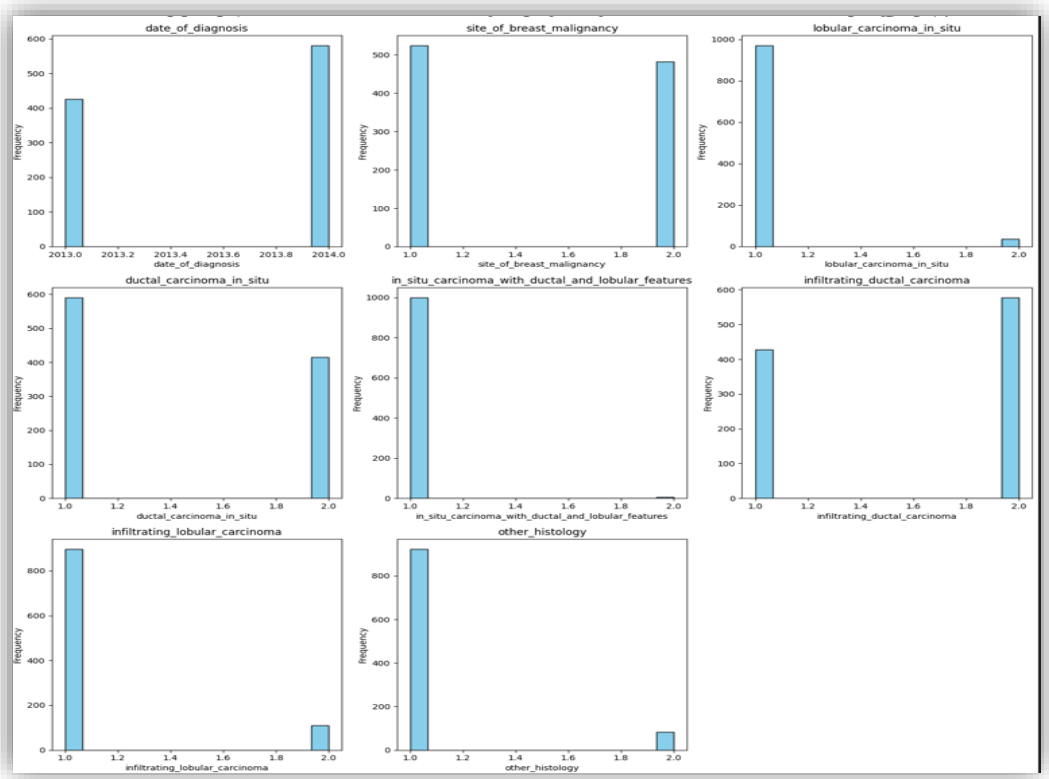


Figure 3. Histograms of Numerical Features in breast cancer dataset

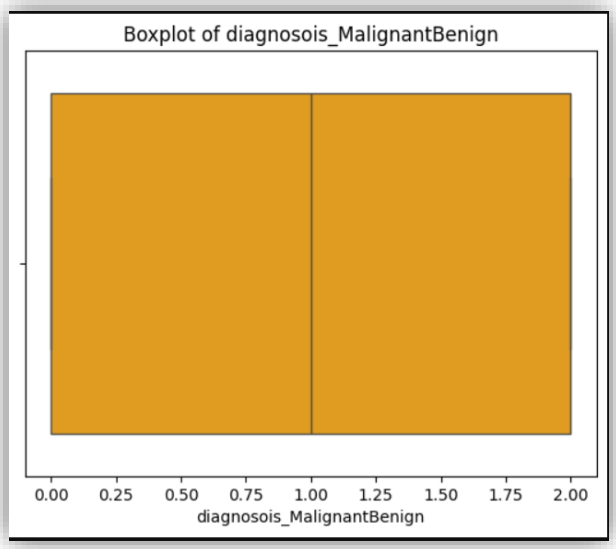


Figure 4. Boxplot of age_of_diagnosis by Diagnosis Outcome

4.2 Correlation Analysis:

A correlation matrix was computed to analyze relationships between numerical features. A heatmap was used for visualization.

Figure 5 illustrates the correlation matrix for numerical features in the dataset, visualized using a heatmap. The heatmap provides an intuitive representation of the strength and direction of relationships between variables, with annotations showing correlation coefficients.

Key findings include a notable positive correlation between age_at_menarche and age_at_menopause, reflecting biological associations. Such insights are valuable for feature selection and understanding underlying dependencies in the dataset.

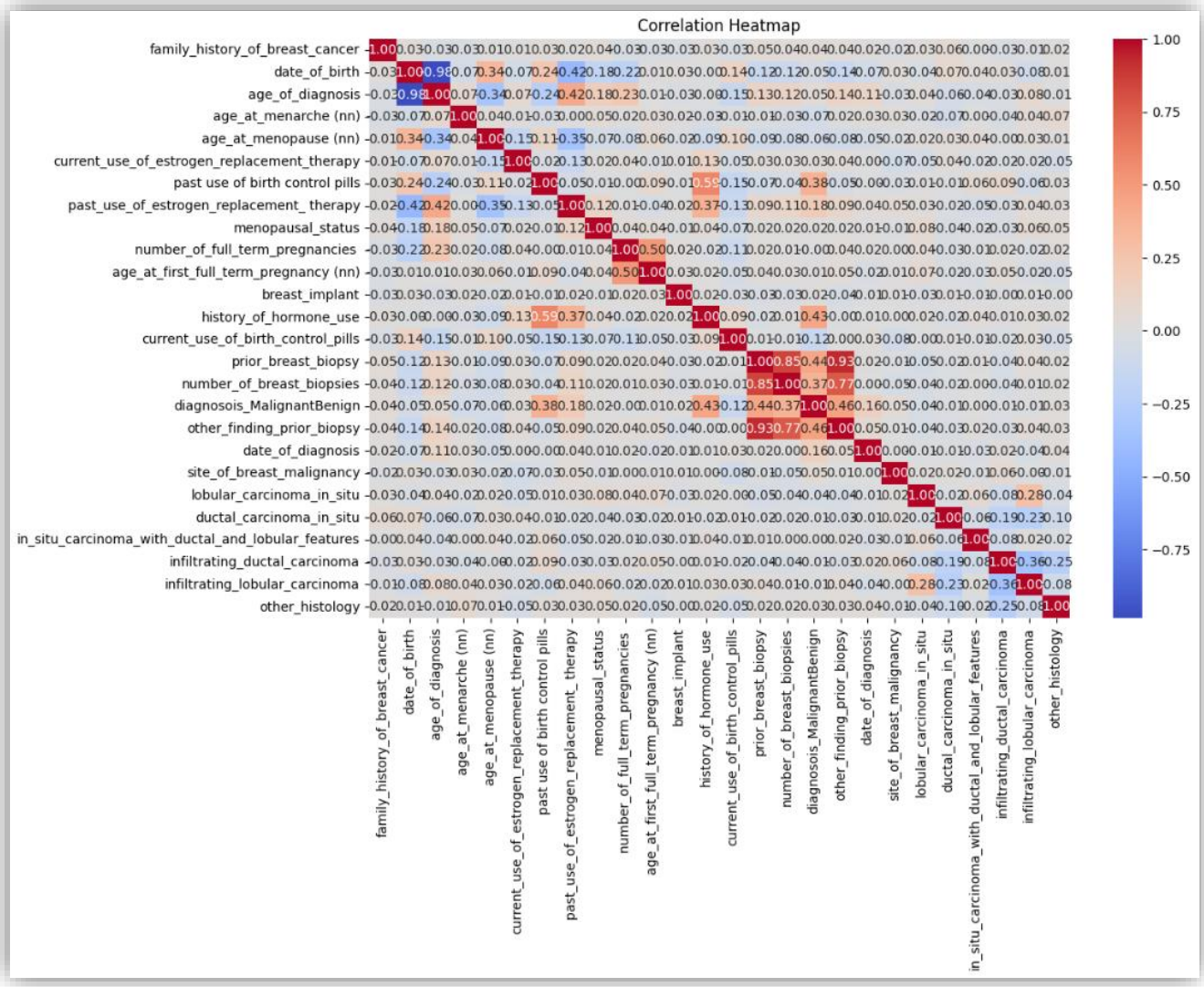


Figure 5. Heatmap of correlation matrix for numerical variables in the dataset

4.3 Feature Relationships:

Figure 6 explores relationships between selected features and the target variable using scatterplots and pairplots.

The Pearson correlation coefficients, defined as:

The Pearson correlation coefficient Corr(X_i,X_j) is given by:

$$\text{Corr}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma_i \sigma_j}$$

where Cov(X_i,X_j) is the covariance between variables X_i and X_j, and σ_i and σ_j are their respective standard deviations.

A scatterplot (Figure 6) of age_of_diagnosis against diagnosis_MalignantBenign reveals that patients diagnosed as 'Malignant' were generally older compared to those labeled 'No Cancer.' Pairplots further analyze relationships among features such as age_of_diagnosis, age_at_menarche (nn), age_at_menopause (nn), and prior_breast_biopsy, providing detailed distributions and bivariate relationships.

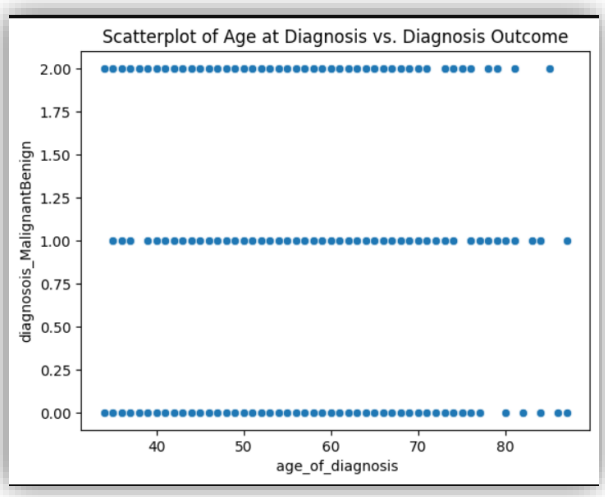


Figure 6. Scatterplot of age_of_diagnosis vs diagnosois_MalignantBenign

Figure 7 presents a pairplot for selected features in the breast cancer dataset, offering insights into their distributions and relationships. The plot includes features such as age_of_diagnosis, age_at_menarche (nn), age_at_menopause (nn), and prior_breast_biopsy, with diagonal elements showing kernel density estimates (KDE) for individual feature distributions.
This visualization helps identify patterns, such as clustering or trends, and highlights potential interactions between features. For instance, it reveals how certain age-related features vary with prior breast biopsy occurrences, aiding in understanding feature dependencies.

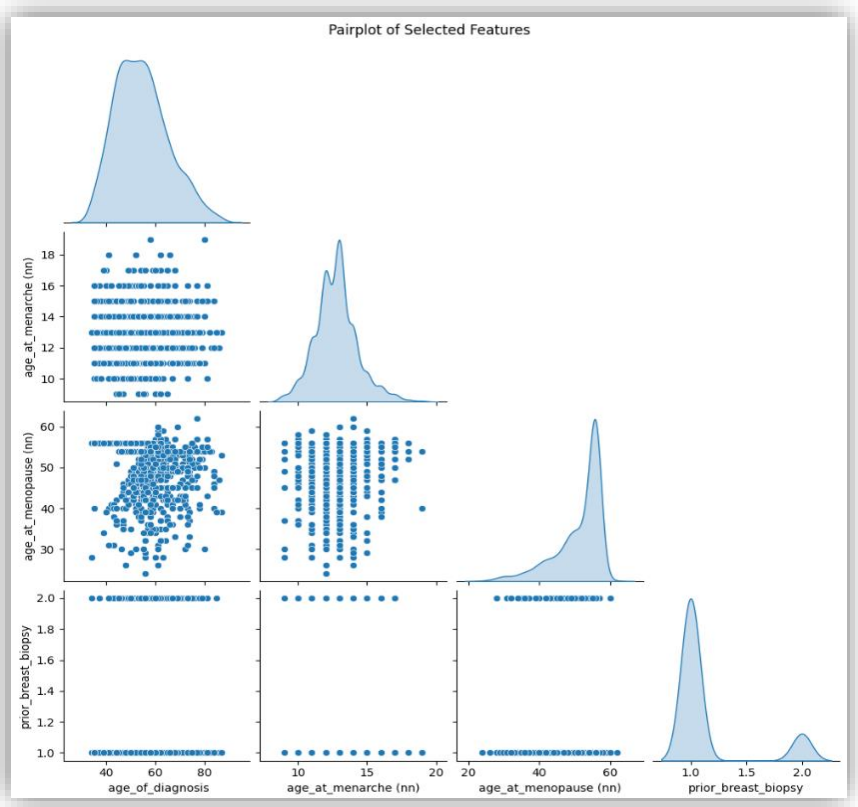


Figure 7. Pairplot for selected features in breast cancer dataset

4.4 Feature Distribution by Target Variable:

Figures 8 and 9 depict the analysis of feature distributions by the target variable using boxplots and countplots.

Figure 8 showcases a boxplot of age_of_diagnosis by diagnosis_MalignantBenign, revealing differences in age distributions across diagnosis categories. Patients with 'Malignant' diagnoses generally exhibit higher ages compared to other groups.

Figure 9 presents a count plot illustrating the distribution of family_history_of_breast_cancer across the target variable. The visualization highlights that a family history of breast cancer is more prevalent among 'Malignant' cases, indicating its potential relevance as a predictive feature.

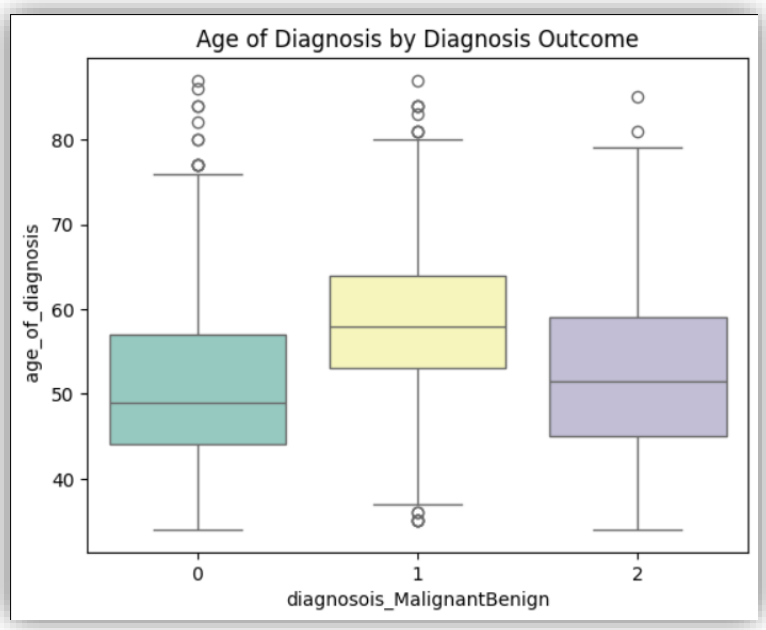


Figure 8. Boxplot of age_of_diagnosis by diagnosis_MalignantBenign

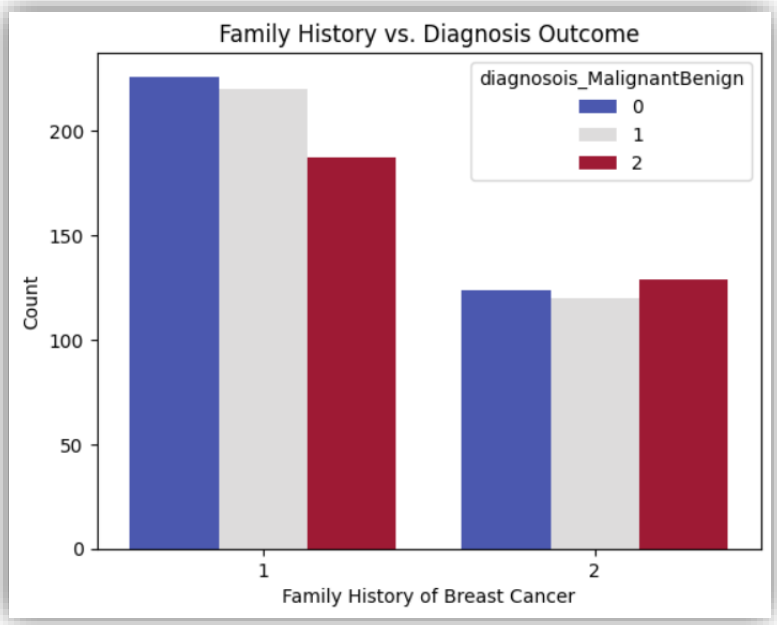


Figure 9. Count plot of Family History of Breast Cancer by Diagnosis Outcome

5. CONCLUSION

This study employed an exploratory data analysis (EDA) approach to uncover patterns, feature relationships, and key insights within a region-specific breast cancer dataset collected from Maharashtra, India. The dataset, comprising 1,006 records and 26 clinical, demographic, and tumor-specific attributes, was subjected

to detailed analysis, including descriptive statistics, class distribution, univariate analysis, correlation analysis, feature relationships, and feature distribution by target variables.

The findings revealed significant correlations between clinical features, such as family history of breast cancer and diagnosis outcomes, and demographic attributes like age at diagnosis. Visualization techniques such as histograms, scatterplots, and heatmaps were instrumental in understanding data distributions, identifying outliers, and highlighting meaningful feature interactions. The analysis showed that malignant cases were often associated with higher age and family history of breast cancer, while other features like prior biopsies and menopausal details also exhibited relevance.

This work provides critical insights into breast cancer prediction, emphasizing the importance of EDA in preparing data for subsequent predictive modeling. The outcomes offer a foundational understanding that supports the development of more accurate and efficient diagnostic tools tailored to region-specific populations. Future studies could expand on this research by incorporating advanced feature selection methods, predictive models, and additional datasets for broader applicability.

The exploratory data analysis revealed important patterns in the breast cancer dataset, including imbalanced class distributions and meaningful correlations between features. This analysis provides a foundation for feature engineering and predictive modeling using machine learning algorithms. Future work will involve addressing class imbalance through resampling techniques and implementing machine learning models to predict breast cancer outcomes.

REFERENCES:

1. **Pujari, M. P., & Patil, S. S. (2017).** *Prediction of Breast Cancer Using Data Mining Techniques: A Study of Indian Dataset*. International Journal of Engineering and Technology, 9(5), 457-463.
2. **Bhadane, P., & Soni, S. (2020).** *Exploratory Data Analysis and Machine Learning in Breast Cancer Diagnosis: A Case Study in India*. International Journal of Data Science and Machine Learning, 8(2), 113-123.
3. **Kumar, M., & Kaur, P. (2020).** *Breast Cancer Detection Using Machine Learning Techniques on Indian Dataset*. Journal of Computational Methods in Sciences and Engineering, 20(4), 569-576.
4. **Sharma, S., & Gupta, A. (2019).** *Feature Selection Techniques in Breast Cancer Prediction using Indian Dataset*. Journal of Indian Research in Computer Science, 11(3), 112-119.
5. **Singh, R., & Verma, S. (2021).** *Breast Cancer Diagnosis Using Classifiers with Indian Dataset: A Comparative Study*. International Journal of Biomedical Engineering and Technology, 33(2), 124-132..
6. **Rajput, D., & Ghosh, A. (2018).** *Breast Cancer Classification using Statistical and Machine Learning Approaches: A Case Study in Indian Context*. Journal of Indian Healthcare Informatics, 13(1), 89-97.
7. **Patil, A. S., & Yadav, V. S. (2019).** *Predictive Analysis of Breast Cancer Using Machine Learning in Indian Population*. International Journal of Health Sciences and Research, 9(2), 45-53.
8. **Nair, V. R., & Krishna, P. (2020).** *Data Mining Approaches for Breast Cancer Prediction: A Review of Indian Studies*. Journal of Data Science and Health Informatics, 5(3), 215-228.
9. **Bharati, A., & Reddy, P. (2022).** *Breast Cancer Risk Prediction Using Clinical and Demographic Data: A Study from Southern India*. Indian Journal of Medical Informatics, 18(2), 234-240.
10. **Verma, S., & Joshi, S. (2018).** *A Comprehensive Study of Feature Analysis and Classification Algorithms for Breast Cancer Prediction in Indian Dataset*. Journal of Artificial Intelligence in Medicine, 12(4), 356-363.

AUTHOR PROFILE

Ms. Swati Laxman Nalawade has completed MCA, M.Phil in Computer Application from CSIBER Kolhapur. She is currently a PhD candidate at Bharati Vidyapeeth (Deemed to be University), Institute of Management and Rural Development Administration, Sangli. Her fields of research interest are Machine Learning, Data Mining, Weka and R tool. She has published papers in the international journals and presented research papers in international and national conferences.

Dr. Suvarna Mahavir Patil is working as Assistant Professor in Bharati Vidyapeeth (Deemed to be University), Institute of Management and Rural Development Administration, Sangli. She has published 14 Research Papers out of that 4 Research papers are indexed in Scopus. Her area of specialization is Algorithms, Artificial Intelligence, Machine learning.