International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 16, No.1 (Sep-24)

http://dx.doi.org/10.12785/ijcds/1601114

Quantifying Breast Cancer: Radiomics, Machine Learning, and Dimensionality Reduction for Enhanced Image-Based Diagnosis

Zulfikar Ali Ansari[1,2], Manish Madhava Tripathi[2] and Rafeeq Ahmed[3,*]

[1]*CSE Department, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India*
[2]*CSE Department, Integral University, Lucknow, Uttar Pradesh, India*
[3]*CSE Department, Government Engineering College West Champaran, Bettiah, Bihar, India*

Received 26 Feb. 2024, Revised 6 Jun. 2024, Accepted 28 Jun. 2024, Published 26 Sep. 2024

Abstract: Radiomics allows for measuring tumour heterogeneity, discovering prognostic biomarkers, early detection and diagnosis, and combining with machine learning to improve clinical decision-making. Radiomics is essential for obtaining quantitative characteristics. From medical pictures, such as those acquired from radiological scans such as MRI, CT, or PET scans, our study intends to enhance diagnostic accuracy by utilizing machine learning models such as Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, Multilayer Perceptron, and XGBoost and then applying dimensionality reduction approaches like PCA, SVD, and NMF to examine radiomics characteristics collected from breast cancer images, improving early breast cancer detection. These ML models are powerful classification and predictive analytics models, while dimensionality reduction techniques simplify complex datasets by reducing features, improving visualization, and minimizing noise. The proposed method comprehensively evaluates the accuracy using train-test data with 20% as test data, showing significant enhancements in diagnostic accuracy for early-stage breast cancer compared to conventional methods. The proposed model has an accuracy of 88.72% as compared to recent works. The high accuracy of our model shows that it could be used to find cancer early, which is important for getting treatment right away. Some limitations include that imaging methods aren't always the same, sample numbers are too small to be useful in real life, computing costs are high, and Clinical validation is needed. Future studies should focus on bigger studies to make them more reliable. These studies should combine genomic data with radiomics to get a fuller picture.

Keywords: Radiomics Features, Breast Cancer Detection, Digital Image Processing, Machine Learning

1. INTRODUCTION

Breast cancer is a serious worldwide health concern, requiring accurate and prompt diagnostic techniques for successful treatment and better patient outcomes [1]. The advent of radiomics, an innovative field leveraging quantitative analysis of medical images, has shown promising prospects for augmenting traditional diagnostic methodologies.[2]. Radiomics permits the acquisition of intricate details from radiological pictures, allowing the discovery of subtle patterns and traits that would otherwise evade eye scrutiny [3]. In this respect, our investigation goes into the domain of breast cancer detection, concentrating on the integration of radiomics and modern computational tools to enhance the classification process. The richness and complexity of radiomics features extracted from various imaging modalities offer a comprehensive representation of tissue characteristics, aiding in the characterization of breast lesions and tumor behavior[4]. However, the sheer volume and intricacy of these radiomics features pose challenges, often leading to high-dimensional datasets. This abundance of information can potentially introduce noise, redundancies, and computational inefficiencies, hindering the development and deployment of robust classification models. Hence, the application of dimensionality reduction techniques emerges as a pivotal strategy to distill crucial information while mitigating computational complexities [5]. There is a lack of research in effectively combining statistical analysis, clinical analysis, decision support systems, and factors that impact the classification of early-stage breast cancer diagnosis. Existing methods frequently lack a cohesive approach, resulting in potential inefficiencies in diagnosing and treating medical conditions. This is because deep learning solely focuses on image classification, which is insufficient in the field of medical science [6]. Medical professionals require information on the factors that cause cancer, making a decision support system vital for the patient's well-being. Therefore, by correcting this deficiency, it is possible to improve the accuracy of diagnosis and provide more precise instruc-

E-mail address: zulfi78692@gmail.com, mmt@iul.ac.in, rafeeq.amu@gmail.com

TABLE I. List of Abbreviation

S/N	Keyword	Full form
1	BC	Breast Cancer
2	ML	Machine Learning
3	PCA	Principal Component Analysis
4	SVD	Singular Value Decomposition
5	NMF	Non-negative matrix factorization
6	DR	Dimensionality Reduction
7	SVM	Support Vector Machine
8	WBCD	Wisconsin Breast Cancer Database
9	RF	Random Forest
10	DT	Decision Tree
11	MLP	Multilayer Perceptron
12	ANN	Artificial Neural Networks
13	KNN	K-Nearest Neighbor
14	CM	Confusion Matrix
15	XGBoost	Extreme Gradient Boosting

tions for preventive measures, ultimately leading to better patient outcomes. This integration would guarantee that medical professionals had a comprehensive comprehension and powerful instruments for the early identification, enabling prompt and efficient therapies. This research aims to investigate the efficacy of various dimensionality reduction methodologies in enhancing breast cancer diagnosis based on radiomics features. Dimensionality reduction techniques aid physicians in identifying subtle malignancy patterns in mammograms, ultrasounds, and MRIs by simplifying intricate, multi-dimensional medical images while retaining all relevant information. By condensing the feature space while preserving diagnostically relevant information, our endeavor seeks to optimize classification models, enabling more accurate and interpretable outcomes.

A. *Contribution:*

The research is motivated by the need to overcome the limits of current breast cancer diagnostic methods and utilize the promise of radiomics for early diagnosis.

- The research used 780 pictures from the UCI Machine Learning Repository to pull out 120 radiomics features. These included important features like RunVariance, RunEntropy, Energy, Elongation, and MinorAxisLength.
- When dimensionality reduction (DR) methods were not used, the Random Forest (RF) model got 85.04 percent accuracy with 44 features and the Multi-layer Perceptron (MLP) model got 85.47 percent accuracy with 78 features.
- When Non-negative Matrix Factorization (NMF) was used, the XGBoost (XGB) model got 87.18 percent accurate with 78 features and 88.72 percent accurate with 44 features. The MLP model got 85.64 percent with 78 features and 86.67 percent with 44 features when Principal Component Analysis (PCA) was used.
- The XGB model got 85.64 percent with 78 features, while the MLP model got 86.67 percent with 44 features. Overall, NMF combined with XGB gave the best results, showing that DR methods greatly improve model performance and diagnostic accuracy.

The remainder of the paper is explained in the following manner: In Section 2, we talk about Related Work, which compares and contrasts current studies in this area and also talks about related research on breast cancer detection. An explanation of the Method is given in Section 3 along with a quick rundown of Radiomics, Methods Incorporated, and DR Techniques. In Section 4, you can see what the experiment showed. Section 5 talks about the Discussion, the comparative analysis, the study's limitations, and what it all means. Section 6 wraps up the piece and talks about its future scope.

2. Related work

In 2018, breast cancer (BC) accounted for the majority of cancer related deaths among women in all of Europe and was the most common kind of cancer among women in all of Europe [7]. The author has conducted an extensive analysis of machine learning in [8] to predict cancer and to characterize compare and contrast deep learning methods. In the publication [9], the author proposed a unique approach to predict the therapeutic response for breast tumors. Advancements have been achieved in characterizing breast cancer subtypes using radiological images. Based on their molecular composition, several characteristics observed on breast imaging tests such as MRIs, mammograms, and ultrasounds can be associated with distinct forms of breast cancer [10]. The major goal of the paper [11] was to offer an effective approach for identifying cancers utilizing mammography pictures of breasts and an ML algorithm. Second, based on the proposed strategy in the first phase, this investigation aims to develop a CAD program for the detection of BC. The Author [12] just explored Radiomics as an overview through Machine learning & Deep Learning on Breast cancer mainly. Currently, physicians receive assistance in analyzing these images via CAD systems. CAD (Computer-Aided Diagnosis) refers to software applications used in healthcare to support medical professionals in analyzing medical imaging data and patient information [13]. These applications can -highlight potential abnormalities, suggest diagnoses, and provide additional information. Medical imaging is critical for the diagnosis, staging, therapy planning, postoperative monitoring, and response evaluation in the routine care of cancer [14]. Among the different breast cancer imaging modalities, Magnetic Resonance Imaging (MRI) has higher sensitivity in lesion identification, possibly due to its multi-parametric character, which includes features such as T1-weighted and diffusion-weighted imaging [15]. However, it is critical to recognize the limits of individual modalities and use a multi-modal approach, which may include methods such as mammography and ultrasound, as well as clinical data and AI-powered analysis, for thorough diagnosis

and risk assessment [16]. The authors in the paper [17] used machine learning techniques to create a preoperative axillary lymph nodes (ALN) status assessment approach based on MRI radiomics characteristics. This technique sought to improve the preoperative assessment process for targeted therapies. The study sought to investigate the potential link between radiomics characteristics and the tumor microenvironment (TME) in individuals with early-stage invasive breast cancer. This objective could be computationally addressed by 1) extracting quantitative radiomics features from the acquired MRI images using image analysis software or libraries like PyRadiomics or Radiomics R package [18], 2) employing machine learning models such as correlation analysis, SVMs, or RFs to identify statistically significant relationships between the extracted features and TME characteristics [19], and 3) visualizing the identified relationships using dimensionality reduction. The author [17] demonstrated how radiomics features improve clinical decision-making and how several machine learning classifiers, together with numerous feature selection strategies, reliably predict breast cancer nodules. The author has provided a concise overview of the current advancements in breast cancer research that utilize the radiomics approach [20]. The proposed radiomics fusion algorithm is utilized to categorize the chosen characteristics into malignant and benign [21]. Comparative studies in breast cancer detection assess the efficacy of various imaging modalities, technologies, or procedures for early diagnosis, screening, and characterization of breast abnormalities[22]. They demonstrated that MRI machine learning radiomics can predict how long breast cancer patients will live without return after surgery and measure lncRNAs without surgery. This method can make good predictions, but it might need a big set of data to learn on [23], [24]. The strength is that the Stacking model performed better than the others. However, one weakness is that the AUCs for some individual models were lower than those for others [25]. These models are being used in mammography, ultrasonography, and MRI, among others, to diagnose and assess risk. However, there are problems that make it hard for AI to be widely used in clinical practice, such as the need for strict validation, interpretability, and technical considerations [26]. AI systems detect tiny abnormalities, decipher ambiguous images, and perform quantitative analysis, improving imaging accuracy, but they require substantial validation, data standardization, regulatory compliance, and ethical consideration. Table II compares the increased benefit of radiomics analysis in breast cancer detection to standard imaging approaches.

3. Methods

This section will thoroughly examine the dataset, including details on its structure, Preprocessing techniques, and feature extraction methodologies to be used. A flowchart of the methods for anticipating breast cancer (BC) will be shown. The suggested classification model uses a variety of approaches, including Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Multi-Layer Perceptron (MLP), and XGBoost. The dataset features will be analyzed using PCA, NMF, and SVD, as shown in Figure 1. The scientists used a Kaggle dataset containing breast cancer imaging data, which included a large number of cases representing various kinds and stages of breast cancer, as well as healthy controls.

A. Preprocessing

The data that was collected at the beginning of the study includes examples of breast ultra-sounds that were taken from women ranging in age from 25 to 75 years old. The total number of patients includes 600 female patients out of the total number of patients. Each of the 780 photographs that are included in the collection of data has an average size of 500 pixels by 500 pixels. Additionally, there are 437 benign images, 210 malignant images, and 133 normal images. PNG is the format that the images are in for the most part. It is essential to execute data pre-processing in order to ensure that data quality issues, such as noise, inconsistencies, and redundancy, are minimized, which ultimately results in improved performance of machine learning models [37]. The techniques of missing value imputation were utilized in this particular instance. While missing values for categorical features were imputed using the mode, missing values for numerical attributes were imputed using the median of their respective feature distributions. Mean was used to impute missing values for numerical attributes. *https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset [38]*

B. Feature Extraction

Image features are the essential attributes that are employed to analyze and perceive it. These characteristics can be derived to discern distinctive qualities within the image data. To classify images based on specific characteristics, it is necessary to extract these characteristics from the dataset of images. Providing an exact description of visual qualities is not feasible; however, size, form, and other characteristics serve as the first foundations of these properties. These traits encompass characteristics such as tumor texture, shape, and intensity. Currently, our algorithms play a crucial role in capturing these patterns, which are vital for the detection and diagnosis of breast cancer. Its aid enables the enhancement, retrieval, visualization, and recognition of images. We have processed images using Python's Scikit-Image. The libraries are responsible for performing tasks such as segmentation, color space modification, analysis, morphology, feature detection, and other skills. Advanced computational equipment can efficiently extract diverse quantitative data from tomographic pictures, including CT, MR, PET, and other types. Radiomics refers to the transformation of medical pictures into high-dimensional data.

C. Radiomics

Radiomics is an evolving field within Medical imaging that encompasses retrieving and examining quantitative characteristics from radiographic images [39]. It goes beyond traditional visual assessment by using advanced computational methods to capture a large amount of data

TABLE II. Various Comparative Studies of Breast Cancer Detection

Dataset detail	Classifier / Methods	Dataset	Radiomics	Accuracy
UCI (80% training and 20% testing), 2022 [27]	DT, RF, K-NN, ANN, SVM & LR	116 Samples	No	64
MIAS Dataset (242 training and 82 testing) 2022 [28]	SVM	324 samples	No	87.1
BIACH and RI (80% Training and 20% Testing) 2023 [29]	XGBoost, LR, KNN, DT, RF, SVM,	1449 samples	No	83
CESM Image (80% Training images and 20% Testing images) 2019 [30],	SVM classifier	51 Samples	Yes	NA
Real data 70% Training images and 30% Testing images) 2018 [31],	NB	331 Chinese women data	Yes	79.6
ACRIN (203 Training images and 50 Testing images) 2023 [32],	CNN	253 patients	No	87.7
Real Data (90 Training images and 21 Testing images)2022 [33],	SVM	111 patients	Yes	91
HER2 overexpressing breast cancer patients (249 Training images and 62 Testing images)2020[34].	HER2 expression	311 patients.	Yes	89.7
Real data Jan 2017 to Feb 2019 (80% Training images and 20% Testing images)2021[35],	DECT iodine map-derived radiomics signatures	77 patients	Yes	92.6
Real data Jan 2018 to Dec 2018 (80% Training images and 20% Testing images)2020 [36],	TIL levels	43 Patients	Yes	74.4

from medical images, such as CT scans, MRI, or PET scans. These data include shape, intensity, texture, and spatial relationships of pixels or voxels within the images. Radiomics contributes to a better knowledge of the intricate properties of tumors and can potentially deliver significant insights [40]. This technique has been used to the field of cancer with the goals of improving prognostic factor evaluation, improving diagnostic accuracy, and aiding in clinical decision-making. The radiomics method calculates the scalar values of the features from the predefined ROI (Region of interest) [41]. Once the lesions are segmented, feature extraction is carried out using radiomics. In the proposed models, Radiomics statistics have been used to extract different categories of features [42] from the ROI of Breast Cancer images. Therefore, a total of 78 features are extracted, and the radiomics features are normalized to the 0-1 range. Various features are calculated as:

$$Energy = \sum_{j=1}^{N_p} (Y(j) + a)^2 \tag{1}$$

In image processing, ”energy” refers to a statistical measure of the distribution of voxel values in an image. It measures the voxel values' magnitude, representing the overall intensity variance within the picture. Higher energy levels suggest more high-intensity voxel values are concentrated in the picture, which suggests more overall contrast and variety.

$$Skewness = \frac{\frac{1}{N_p}\sum_{j=1}^{N_p}(Y(j) - \bar{Y})^3}{(\sqrt{\frac{1}{N_p}\sum_{j=1}^{N_p}(Y(j) - \bar{Y})^2})^3} \tag{2}$$

Skewness is used to quantify how asymmetrical attribute values are distributed concerning their mean. It evaluates how far the distribution deviates from symmetry concerning the mean value. Positive skewness means the distribution has a longer or fatter tail toward higher attribute values, with most values left of the mean. Negative skewness indicates a longer or fatter tail toward lower attribute values, with most values to the right of the mean. A fully symmetric distribution with mirror-image tails has a skewness of zero. Skewness can be positive (right-skewed) or negative (left-skewed), with 0 skewness indicating perfect symmetry.

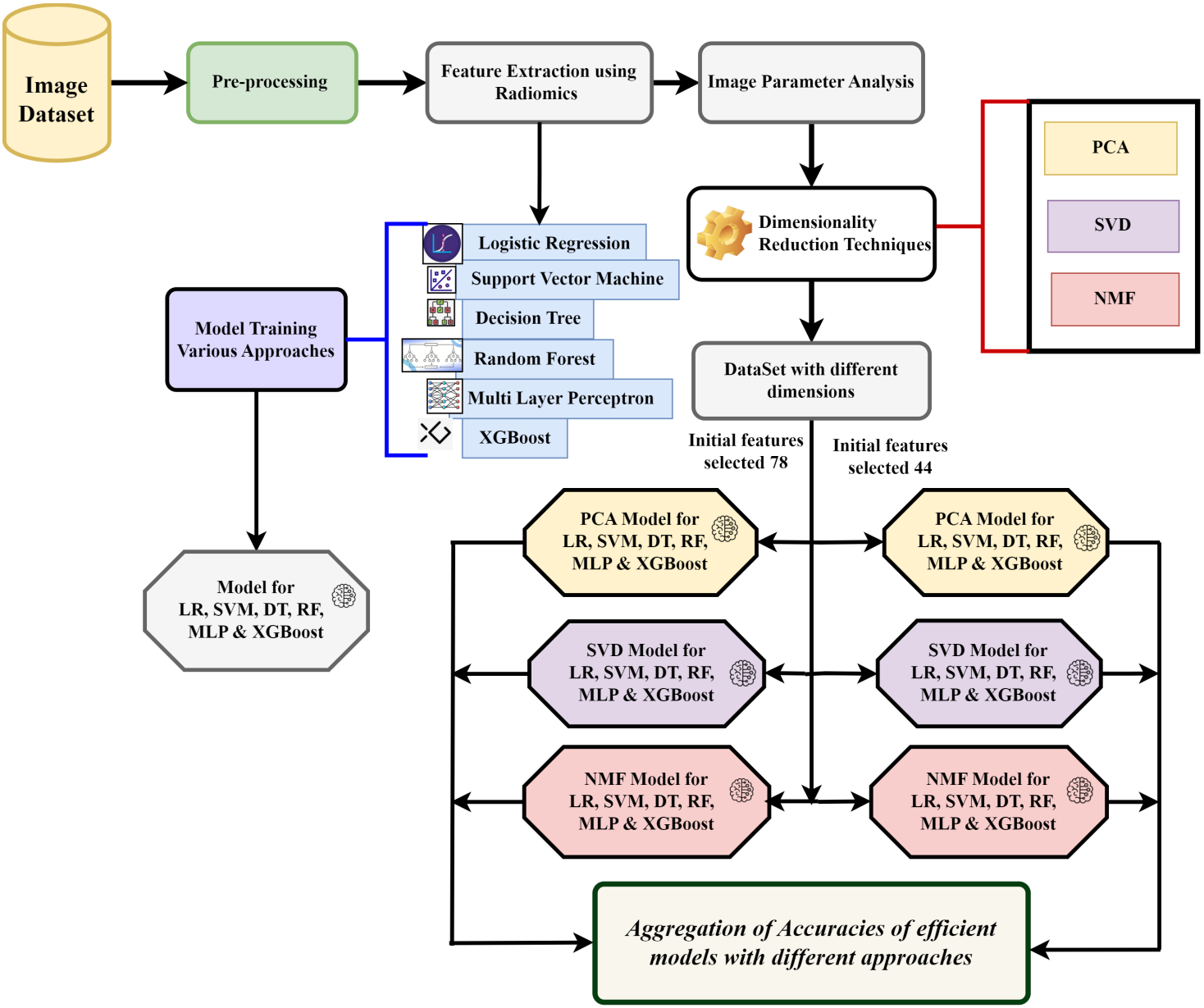

Figure 1. Flow diagram of the proposed approach

$$Kurtosis = \frac{\frac{1}{N_p}\sum_{j=1}^{N_p}(Y(j)-\bar{Y})^4}{(\sqrt{\frac{1}{N_p}\sum_{j=1}^{N_p}(Y(j)-\bar{Y})^2})^2} \quad (3)$$

Kurtosis measures a probability distribution's peakedness and tail extremity compared to a normal distribution. Calculated using the fourth standardized moment, it shows the distribution's extreme outlier propensity. Mésokurtic (normal kurtosis, value around 3), leptokurtic (high kurtosis, value greater than 3, indicating fatter tails and a sharper peak), and platykurtic (low kurtosis, value less than 3). Less than 3 excess kurtosis is used to compare to the normal distribution, with positive values indicating leptokurtic and negative values platykurtic distributions. In finance and quality control, this measure helps estimate risk and spot abnormalities.

$$Sphericity = \frac{2\sqrt{\pi A}}{P} \quad (4)$$

The ratio of the tumor region's perimeter to the diameter of a circle with a surface area equal to the tumor region's is known as sphericity. This metric evaluates how much the

tumor area resembles a spherical. Interestingly, sphericity is a dimensionless measure that is unaffected by changes in orientation or size. It is said to be spherical when the differences between all pairs of groups (or levels) have the same variation.

$$Major \quad axis \quad length = \sqrt[4]{\lambda_{minor}} \tag{5}$$

This element, which is calculated using the largest head portion, produces the largest hub length of the ROI-encasing ellipsoid λ_{major}.

$$Elongation = \sqrt[4]{\frac{\lambda_{minor}}{\lambda_{major}}} \tag{6}$$

The ROI shape's elongation illustrates the relationship between its two largest head segments.

$$Difference \quad Entropy = \sum_{c=0}^{N_g-1} cp_{m-n}(c)(p_x(a)p_y(b) + \epsilon) \tag{7}$$

Difference Entropy is a proportion of the irregularity/fluctuation in neighborhood power esteem contrasts.

$$Contrast = \sum_{a=1}^{N_g-1}\sum_{a=1}^{N_g-1}(a-b)^4 p(a,b) \tag{8}$$

Contrast is a proportion of the nearby power variety, preferring values from the inclining (a=b). A bigger worth connects with a more noteworthy dissimilarity in force esteems among adjoining voxels.

$$CP = \sum_{a=1}^{N_g-1}\sum_{a=1}^{N_g-1}(a+b-\mu_m-\mu_n)^4 p(a,b) \tag{9}$$

CP is defined by the evaluation of the asymmetry and skewness of the GLCM.

$$GLNU = \frac{\sum_{a=1}^{N_g}(\sum_{b=1}^{N_s}(a-b)^4 p(a,b))^2}{N_z} \tag{10}$$

GLNU, short for Gray-Level Non-Uniformity, is a radiomics characteristic obtained from the examination of medical imaging. The measure quantifies the range of gray levels in an image, indicating the diversity of textures within a specific area of interest. Elevated GLNU levels imply increased heterogeneity, which may be linked to intricate tissue architectures or pathological conditions such as malignancies.

$$LGLZE = \frac{\sum_{a=1}^{N_g}(\sum_{b=1}^{N_s}\frac{P(a,b)}{a^2}}{N_z} \tag{11}$$

LGLZE estimates the circulation of lower Gray level size zones, with a higher worth showing a more noteworthy extent of lower dim level qualities and size zones in the picture. Low Gray-Level Zone Emphasis (LGLZE) is a radiomics feature utilized in the examination of medical pictures to measure the dispersion of low gray-level zones inside a specific area of interest. LGLZE quantifies the ratio of low-intensity regions to the overall number of zones in an image.

$$SZNUN = \frac{\sum_{a=1}^{N_g}(\sum_{b=1}^{N_s}(a-b)^4 p(a,b))^2}{N_z^2} \tag{12}$$

Short-Zone Non-Uniformity Normalized (SZNUN) is a radiomics feature that is calculated using the Gray-Level Size Zone Matrix (GLSZM). The statement describes the process of measuring the proportion of small, uniform areas within a picture compared to the overall volume of the image while taking into account the total number of these areas. This characteristic quantifies the degree of variation in small areas with a particular intensity level, offering valuable information on the consistency of texture within a specific region of interest.

$$DNUN = \frac{\sum_{b=1}^{N_d}(\sum_{a=1}^{N_g}(a-b)^4 p(a,b))^2}{N_z^2} \tag{13}$$

DNUN Measures the analogy throughout the image, with a diminished value signifying homogeneity with dependencies in the image. The Difference of Normalized Uniformity (DNUN) is a radiomics characteristic that measures the level of uniformity within a given region of interest in a medical image, taking into account a specific method for normalization. It quantifies the differences in gray-level patterns, providing valuable information on the texture and diversity of the tissue under examination.

$$GLV = \sum_{a=1}^{N_g}\sum_{b=1}^{N_d} P(a,b)(a-\mu)^2 \tag{14}$$

Gray-level variance (GLV) is a radiomics characteristic that measures the amount of variation in the intensity of gray-level values inside a specific area of interest in medical

pictures. This metric quantifies the level of diversity or variety in texture within the tissue under examination.

$$DV = \sum_{a=1}^{N_g} \sum_{b=1}^{N_d} P(a,b)(b-\mu)^2 \tag{15}$$

DV measures the variance in dependence size in the image. In the field of radiomics and radiation therapy, the abbreviation "DV" commonly refers to "Dose Volume," specifically in relation to Dose-Volume Histogram (DVH). The Dose Volume Histogram is an essential tool used in the planning and evaluation of radiation therapy.

$$Coarseness = \frac{1}{\sum_{a=1}^{N_g} x_a y_a} \tag{16}$$

The coarseness of an individual voxel indicates the rate at which it is changing within its neighborhood. Greater values indicate lower spatial change rates and a local texture that is more uniform.

$$Busyness = \frac{\sum_{a=1}^{N_g} x_a y_a}{\sum_{a=1}^{N_g} (\sum_{b=1}^{N_g} |a x_a - b x_b|)} \tag{17}$$

An indication of how a pixel differs from its neighbor. Busyness is a measure of the rapid pixel and neighborhood intensity adjustments in an image. High values indicate a busy image.

$$Strength = \frac{\sum_{a=1}^{N_g} \sum_{b=1}^{N_g} (x_a + y_a)(a-b)^2}{\sum_{a=1}^{N_g} y_a} \tag{18}$$

An image's strength refers to its primitives. The intensity of the primitive is high when it is easily distinguished and observable, e.g., a still image with many coarse variations in gray levels but slowly changing intensity.

$$RV = \sum_{a=1}^{N_g} \sum_{b=1}^{N_r} P(a,b|\theta)(b-\mu)^2 \tag{19}$$

The variance of runs for run lengths is defined as RV. RV stands for the three-dimensional area of interest (ROI) from which radiomics features are derived. The volume in question refers to a specific structure within the body, such as a tumor, organ, or other object of interest. It is examined using imaging techniques like CT, MRI, or PET scans.

$$RP = \frac{N_r(\theta)}{N_p} \tag{20}$$

In RP, the ratio between the number of runs and the number of voxels in the ROI is used to quantify the coarseness of the texture.

$$SRE = \frac{\sum_{a=1}^{N_g} \sum_{b=1}^{N_r} \frac{P(a,b|\theta)}{b^2}}{N_r(\theta)} \tag{21}$$

A greater value indicates a shorter run length or finer texture. SRE measures the distribution of short-run length.

D. Feature Analysis

Feature analysis, feature engineering, and feature selection are all words that refer to the same procedure inside the machine learning process, and they are all vital components [43]. The work involves choosing, modifying, or creating relevant features (also known as input variables or characteristics) from unprocessed data to improve a machine learning model's performance. A radiomics approach for statistical analysis can result in quicker training times, more accurate and efficient models, and a deeper comprehension of the underlying data. Radiomics analysis, which involves deriving a great deal of quantitative data from pictures, is a rapidly developing field in medical imaging. Subsequent analysis of these attributes can yield significant insights into the fundamental biology of the tissue under observation. A growing area in medical imaging called "radiomics analysis" deals with taking a wealth of quantitative data out of pictures. These features that were retrieved capture different aspects of the tissue that were scanned and provide important information about its basic biological properties. By using extensive imaging data, radiomics has shown to have significant potential in improving diagnosis, prognosis, and treatment planning for breast cancer.

1) DR

Dimensionality reduction (DR) algorithms are important in machine learning because they allow for the translation of high-dimensional data into lower-dimensional spaces while retaining crucial information [44]. This technique has various advantages, including reduced data complexity, enhanced computational efficiency, simplified model architecture, and effective rule construction.
Non-negative Matrix Factorization (NMF) is a technique for data analysis that is distinct from SVD [45]. Unlike SVD, NMF splits a data matrix into two matrices that only include non-negative values. This characteristic makes NMF especially helpful for tasks like image processing and text mining, where negative values may not be significant.

PCA is a widely utilized dimensionality reduction technique in machine learning (ML) and data analysis. Its primary objective is to project a high-dimensional dataset

(X) onto a lower-dimensional subspace while capturing the most significant variance in the original data.

Centering the Data: Before computing the covariance matrix, the data needs to be centered by subtracting the mean of each feature from its corresponding values in each sample. This removes the influence of the mean and ensures values are centered around zero, improving the accuracy of the covariance calculation.

Covariance matrix calculation: The covariance matrix ($\sum$) is a square matrix of size n x n, where each element $\sum_{ij}$ represents the covariance between the i-th and j-th features. The formula for calculating the covariance between features i and j is:

$$cov(Feature_i, Feature_j) = \frac{1}{n-1}\sum_{k=1}^{n}(Feature_{ki} - \mu_i)(Feature_{kj} - \mu_j) \quad (22)$$

This computation yields a covariance matrix of size m x m, which is denoted by $\sum$ and is very significant in the study of breast cancer. In the context of breast cancer, an eigenvalue decomposition was performed on the covariance matrix $\sum$, which is unique to the dataset of breast cancer. This process entails identifying the eigenvectors and eigenvalues. Within this particular framework, the eigenvectors serve as significant orientations, commonly known as principal components, while the related eigenvalues show the extent of variability along these orientations.

$$\sum v = \lambda v \quad (23)$$

Here, v signifies an eigenvector and λ represents the eigenvalue. Principal component selection in the context of breast cancer: The eigenvalues were arranged in decreasing order, and the matching eigenvectors indicate the major components that are specific to the breast cancer data. The determination of the number of primary components to keep is influenced by multiple criteria, including those about the explained variance, which hold particular significance in the context of breast cancer study.

Singular Value Decomposition (SVD): One of the DR effective methods known as SVD divides a matrix into three smaller matrices. This makes the data's hidden structure visible, which facilitates comprehension and analysis. Tasks like dimensionality reduction, data compression, and finding odd patterns in the data can all benefit from the usage of SVD. Three matrices are involved in the decomposition: a diagonal matrix with singular values on the left, a right singular matrix, and a left singular matrix [46]. The technique has significant importance in many domains, including but not limited to data compression, DR, signal processing, and ML. SVD decomposes a matrix into three simpler matrices, revealing the underlying structure and important characteristics of the original matrix. To decompose any matrix C_{nxd}, we employ three matrices which are U_{nxn}, $\sum_{nxd}$, and V_{dxd}. U and V are orthogonal matrices. The matrix $\sum$ is a non-negative diagonal matrix belonging to the set of real matrices. Mathematically, the Singular Value Decomposition (SVD) factorizes a given matrix C in the following manner:

$$C = U\Sigma V^T \quad (24)$$

Non-negative Matrix Factorization (NMF): Non-negative Matrix Factorization (NMF) is a technique for data analysis that is distinct from SVD [45]. Unlike SVD, NMF splits a data matrix into two matrices that only include non-negative values. This characteristic makes NMF especially helpful for tasks like image processing and text mining, where negative values may not be significant.

This is how it works mathematically. Given a non-negative data matrix X, NMF seeks to identify two smaller matrices, W (m x r) and H (r x n), where r is typically smaller than m and n, given a non-negative matrix X of size m x n. W and H each containing only non-negative values. This factorization aids in the extraction of hidden patterns and characteristics from the data. When considering factorization:

$$X \approxeq WH \quad (25)$$

Here X is the initial non-negative matrix that needs to be factorized. Matrix W is a non-negative matrix with dimensions m x r. Each column in matrix W represents a fundamental vector, and these fundamental vectors are used to approximate the data in matrix X. H is a matrix of size r x n, where r and n are non-negative values. The columns of matrix H correspond to the coefficients of the basis vectors in matrix W that are utilized to rebuild the columns of matrix X. The objective of NMF is to choose optimal values for matrices W and H, such that their multiplication yields an approximation of the original matrix X that is as near as possible while guaranteeing that all members in W and H are non-negative.

E. Methods Incorporated

This section provides a high-level overview of the classifier. Logistic Regression (LR) is a popular choice for its simplicity and interpretability. It leverages linear regression principles to estimate the probability of an outcome, making it suitable for both binary and multiclass classification tasks [47]. Decision Trees (DT) are non-linear models that recursively split the feature space based on thresholds, creating hierarchical structures for decision-making, and are easily interpretable, and capable of handling diverse data types [48]. Support Vector Machines (SVM) create optimal hyperplanes to separate classes in high-dimensional

space, excelling in complex classification tasks through maximizing the margin between different classes [49]. Random Forest (RF) is an ensemble learning approach that integrates predictions from many decision trees. This method minimizes overfitting by utilizing different trees trained on random subsets of data and characteristics. Because of its parallelizability and ability to capture intricate correlations within data, RF is highly suited for dealing with massive datasets [50]. Multilayer Perceptron (MLP), a neural network, learns complex patterns through layers of nodes with non-linear activations, suitable for non-linear relationships [51]. XGBoost, an extreme gradient boosting technique, sequentially builds an ensemble of weak learners to correct previous models' errors, offering high predictive accuracy and robustness to missing values [52].

F. Performance Analysis

Classification utilizes assessment metrics such as Accuracy, Precision, Recall, Specificity, and F-measure. The components of CM, which furnish information regarding anticipated and realized results, are employed to formulate these metrics. The equations provided represent the performance metrics in real-world scenarios. True Positive is denoted as TM (True Malignant), while True Negative is denoted as TB (True Benign). Equations:

$$Accuracy = \frac{TM + TB}{TM + TB + FM + FB} \tag{26}$$

$$Recall = \frac{TM}{TM + FB} \tag{27}$$

$$Precision = \frac{TM + TB}{TM + TB + FM + FB} \tag{28}$$

$$F1_{Score} = \frac{\mathbf{2} \times Precision \times Recall}{Precision + Recall} \tag{29}$$

$$Specificity = \frac{TN}{FP + TN} \tag{30}$$

4. Experiment and Result Analysis

The dataset, comprising 780 images taken from the UCI Machine Learning Repository, underwent a comprehensive analysis to extract a multitude of radiomics features from medical images. In this study, we investigated the influence of feature extraction techniques extracted breast cancer images on the accuracy of breast cancer detection. Since radiomics features are 120 features grouped into different categories, we applied DR techniques to further improve the proposed models. Some of the prominent radiomics features by their entropies are RunVariance, RunEntropy, Energy, Elongation, and MinorAxisLength. The subsequent reduction of feature dimensionality aimed to improve the efficiency and interpretability of the diagnostic process. Finally, we trained firstly ML models which are LR, SVM, RF, DT, MLP, and XGboost without DR techniques on 78 features selected only, and we got maximum accuracy on MLP which is 85.47% as shown in figure 2. Again, we have done the same procedure on Reduced features which is only 44 Features, and we can see in figure 2 that the RF model achieved the best accuracy as compared to other models which is 85.04%. Now we used DR techniques on the same models, and we got the best accuracy on XGB with NMF which is 87.18% as compared to other models, only using 78 features selected which is shown in figure 3. Again, the same procedure followed only used 44 features (Some features reduced) then we analyzed here the Accuracy improved as compared to 78 features on the same model which is XGboost, the accuracy is 88.72% but some other models are also varied, and some models gave the same accuracy as showed in figure-3. Again, we incorporate another DR technique which is PCA on the same models and the Initial 78 features selected and after reducing the feature, we observed that the best accuracy provided by MLP as compared to other models which are 85.64% and 86.67% respectively on the initial 78 features selected and after reducing feature as shown in figure 5. Finally, we applied the 3rd DR Technique on all models which is SVD, and we got here again best accuracy on other models as on the initial 78 features selected XGboos provided a better result which is 85.64% as shown in figure 4 and again when we reduced features just selected 44 features we got best accuracy on MLP again which is 86.67% as shown in figure 4.

A. Software Tools and Technique

We have employed Jupyter Notebook and Numpy, which support large multi-dimensional arrays and matrices and contain mathematical functions for array manipulation were used. For Data Framing and Series for easy and efficient structured data management used Pandas. Matplotlib can create line plots, scatter plots, bar charts, and histograms, whereas Sea-born has a sophisticated interface for creating attractive and instructive statistical graphics. PyRadiomics uses SimpleITK for image processing, Scipy for mathematical functions, and PyWavelets for wavelet transformations to efficiently extract quantitative radiomics features from medical pictures in Python. The software efficiently analyses medical images and masks specifying regions of interest (ROIs) to extract form, first-order statistics, and texture metrics like GLCM and GLRLM. Scikit-Learn supports regression, classification, clustering, and dimensionality reduction for machine learning models. It is designed to work smoothly with NumPy and Pandas, making it a flexible and easy-to-use tool for building and testing machine-learning models.

5. Discussion

Our study's primary goal was to assess the impact of feature extraction approaches on breast cancer detection accuracy. Examining the selected radiomics characteristics, such as RunVariance, RunEntropy, Energy, Elongation, and MinorAxisLength, found that they all contribute signifi-

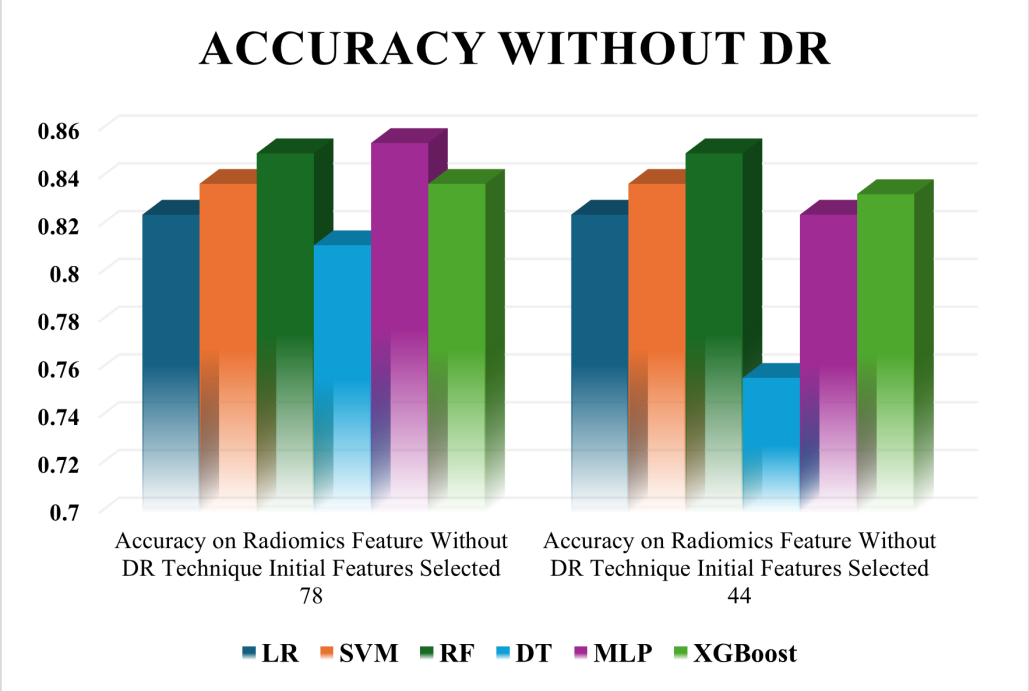

Figure 2. Comparison Accuracy all Models without DR

cantly to increased diagnostic accuracy. These traits require further investigation as possible indications of breast cancer.

Table III summarises our findings, including all of the models assessed and the highest accuracy ratings for each characteristic. The findings include the performance of models employing the initial set of 78 features and a subset of 44 features retrieved using several classifier algorithms (LR, SVM, RF, DT, MLP, and XGBoost).

Figure 6 shows a graphical depiction of the performance of all models. We plotted a graph of all models and finally, the best accuracy in all models through the DR technique is XGboost with 88.72%.

The proposed methodology for breast cancer detection demonstrates significant improvements in accuracy compared to existing methods which is shown in table IV. Through a comprehensive comparative analysis, it is observed that the proposed approach integrates advanced radiomics features and takes advantage of a more diverse and accurate set of quantitative metrics extracted from medical images. In contrast to traditional methods, the proposed methodology includes machine learning algorithms that efficiently analyze complex patterns and relationships within imaging data, enhancing the accuracy of lesion detection and classification. The inclusion of innovative features, such as texture analysis, shape descriptors, frequency domain properties, etc., contributes to a more comprehensive understanding of breast tissue properties. Additionally, the proposed approach embraces the power of artificial intelli-

TABLE III. Summarized results with all models.

DR Techniques	Model	With DR, Initial Features =78		With DR, Initial Features =44	
		Component	Max Accuracy	Component	Max Accuracy
SVD	LR	12	0.8410	17	0.8462
	SVM	10	0.8308	5	0.8308
	RF	18	0.8513	12	0.8462
	DT	41	0.8000	28	0.8051
	MLP	16	0.8513	16	0.8667
	XGBoost	20	0.8564	15	0.8615
PCA	LR	11	0.8359	18	0.8359
	SVM	13	0.8308	8	0.8256
	RF	17	0.8462	11	0.8462
	DT	19	0.8103	14	0.8000
	MLP	18	0.8564	17	0.8667
	XGBoost	12	0.8513	28	0.841
NMF	LR	12	0.8410	10	0.8410
	SVM	64	0.8462	16	0.8410
	RF	5	0.8564	34	0.8615
	DT	5	0.8256	17	0.8103
	MLP	48	0.8513	27	0.8513
	XGBoost	32	0.8718	13	0.8872

TABLE IV. Comparison of Existing approaches with proposed approach

Author	Classifier / Methods	Dataset	Radiomics	Accuracy
Jing Zhou et al. 2021 [53]	SVM	306 patients	Yes	87
Isaac Daimiel Naranjo et al. 2021 [54]	multiparametric radiomics mode	93_Patients	Yes	85
Mohamed A. Hassanien et al. 2022 [55]	ConvNeXt network, a deep convolutional neural network (CNN)	31 malignant and 28 benign / 3911 and 5245	Yes	87.17
JOONGYO LEE at al. 2023 [56]	stacking model (SVM, RF,LR)	MRI between Jan'13 and Dec'17 were collected	Yes	78.4
Yingyu Lin at al. 2024 [57]	Six Robust ML models	268 Breast cancer patients	Yes	82.5
Yimiao Yu at al. 2024 [58]	LR, SVM, RF and XGB	329 images	Yes	87.7
Proposed Approach	**LR, SVM,RF, DT, MLP & XGboost with DR**	**780 Images**	**Yes**	**88.72**

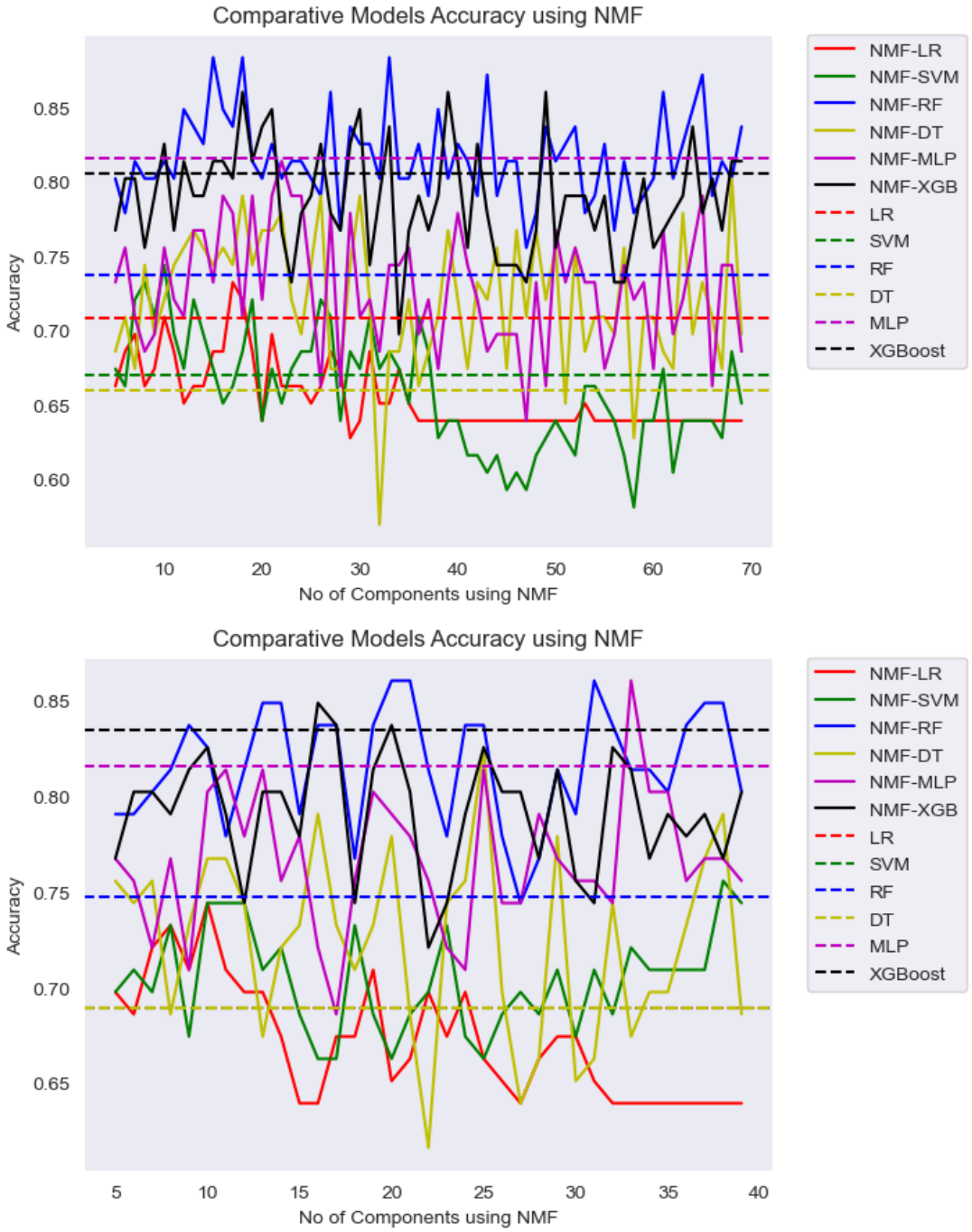

Figure 3. Result on NMF with two different features set

gence, enabling dynamic adaptation to evolving datasets and improving its predictive capabilities over time. Comparative studies highlight the superior performance of the proposed methodology and showcase its ability to significantly raise the accuracy of breast cancer detection, ultimately contributing to more reliable and timely diagnosis for improved patient outcomes.

Our work assesses the diagnostic accuracy of ML models, particularly XGBoost, and dimensionality reduction methods for early-stage breast cancer diagnosis. We use radiomics characteristics and machine learning methods to achieve 88.72% diagnostic accuracy on train-test splits with 20% testing. Figure 7 shows the Receiver Operating Characteristic (ROC) curve analysis for the XGBoost

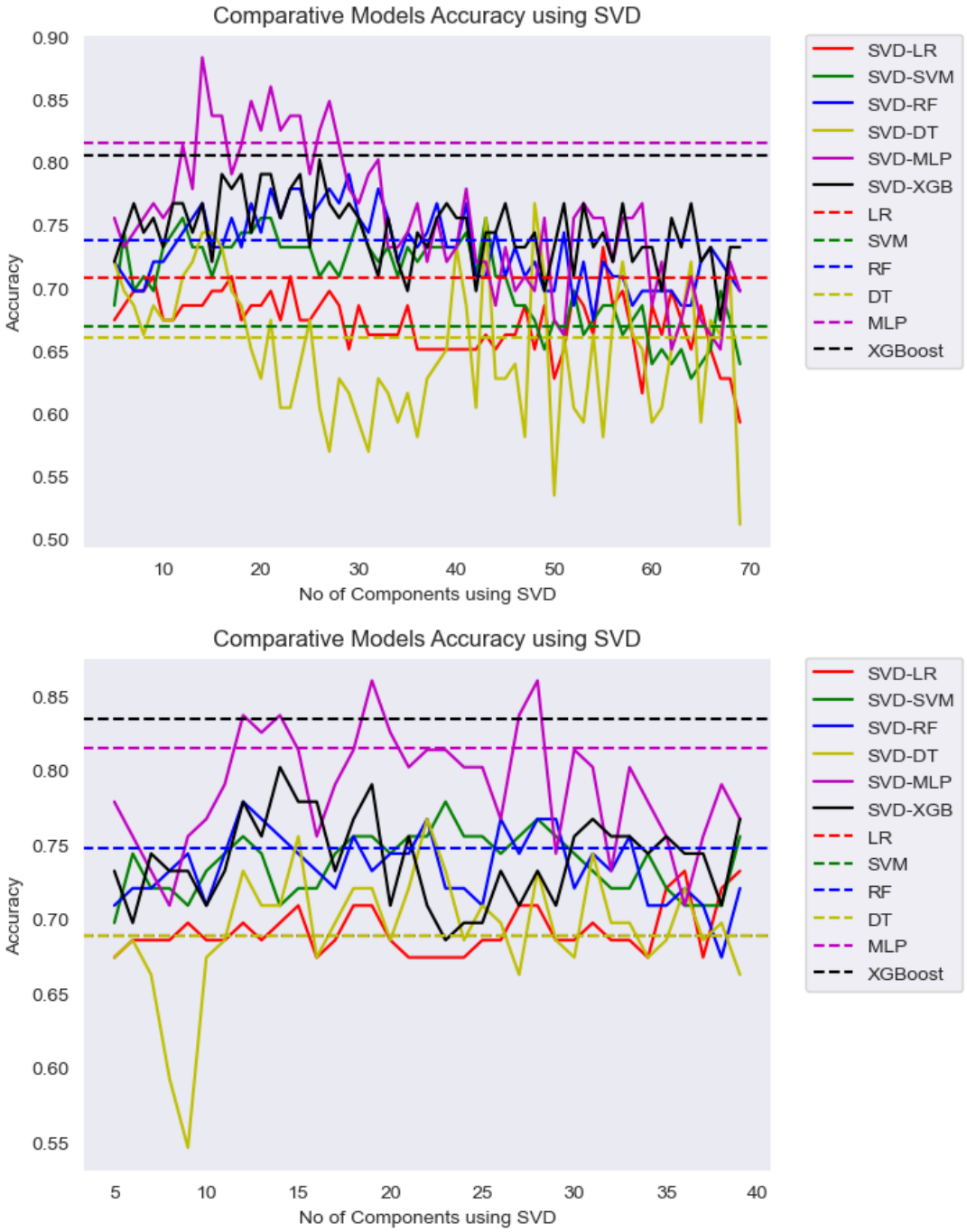

Figure 4. Result on SVD with two different features set

model, proving its robust classification between benign and malignant cases.

A. *Ethical Implication*

The combination of radiomics and machine learning has significant potential for improving breast cancer diagnosis, but it also presents various ethical concerns. A significant issue is the preservation of patient privacy, given the utilization of extensive datasets that contain sensitive information. This requires the implementation of rigorous data protection mechanisms to prevent the identification of individuals and illegal access to the data. Moreover, algorithmic decision-making biases might result in discrepancies in the accuracy of diagnoses and recommendations for treatment among various demographic groups. To reduce these biases, it

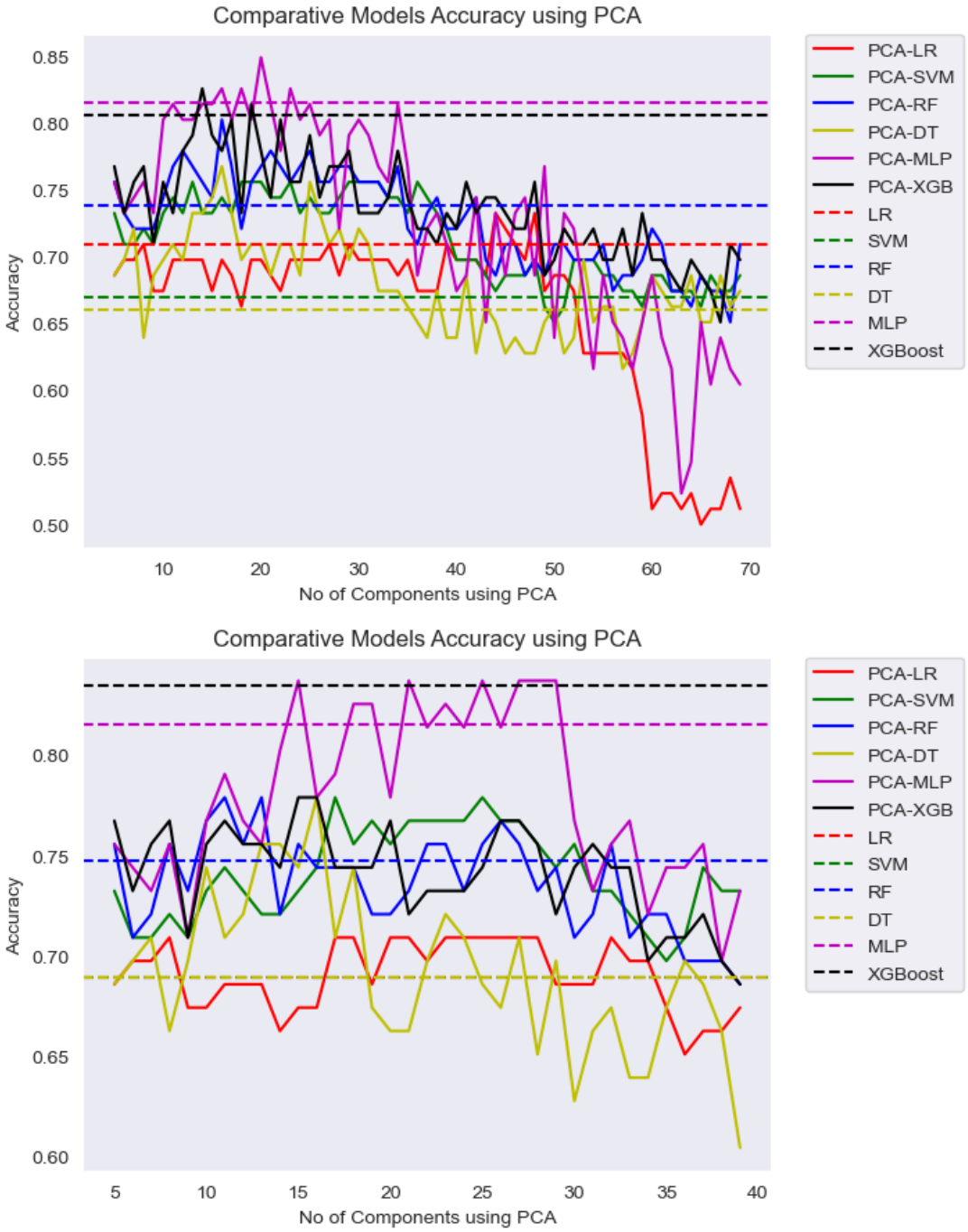

Figure 5. Result on PCA with two different features set

is crucial to ensure that the training data is diverse and representative and to continuously validate the model.

B. Significance of this study

The findings of our study improve the precision of diagnostic procedures beyond current methodologies, establishing it as a highly promising instrument for the early diagnosis of breast cancer. Our model demonstrates a potential to enhance patient outcomes by detecting and treating conditions early, with an accuracy range of 80-90%, which aligns with previous study findings. The exceptional precision of our machine learning model highlights its potential incorporation into diagnostic protocols to aid radiologists in detecting subtle abnormalities, hence improving breast cancer detection and diagnosis. The effectiveness of advanced

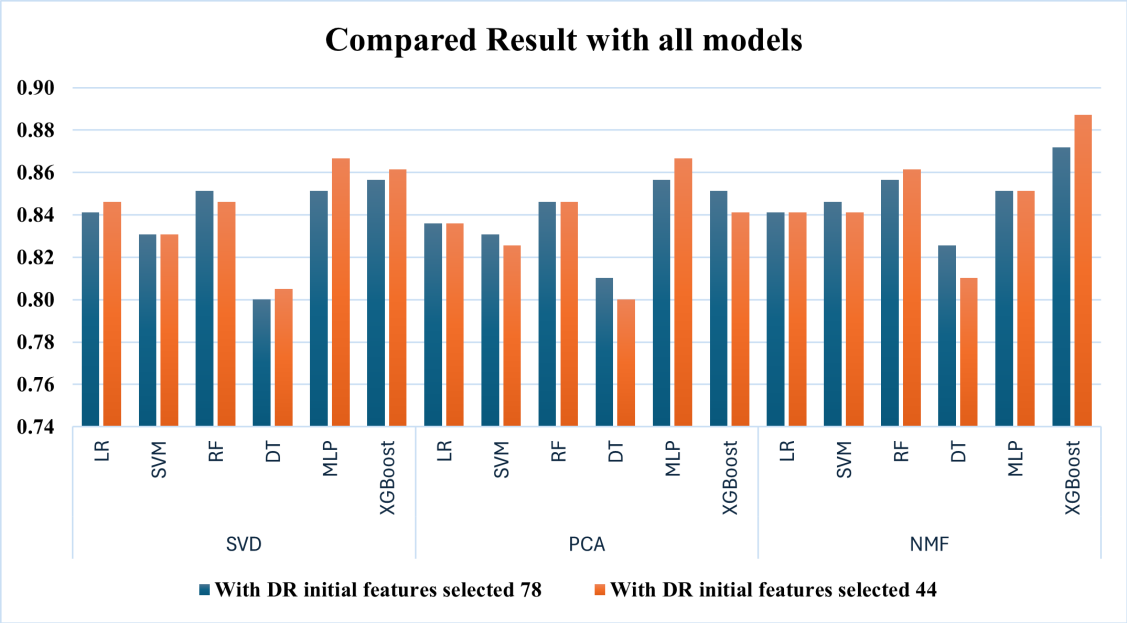

Figure 6. Accuracy of all models with two feature sets for NMF, SVD, and PCA

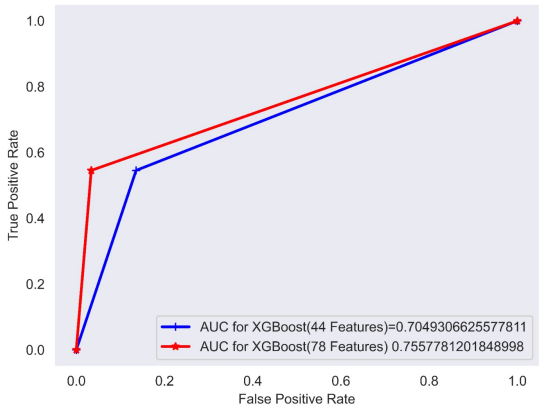

Figure 7. ROC Curve

feature extraction approaches confirms the usefulness of radiomics characteristics in offering precise quantitative data, hence enhancing the accuracy of diagnosis. Our paper demonstrates notable progress in the field of radiomics by showing how machine learning approaches and dimensionality reduction techniques can improve the performance of models and decrease the computing expenses for real-world applications that involve complex imaging data.

C. Limitation

The findings of our study demonstrate that the utilization of radiomics, machine learning, and dimensionality reduction techniques has the potential to enhance the accuracy of breast cancer diagnosis based on medical imaging. Nevertheless, there exist certain technical complications that require resolution. Varying institutions may employ distinct imaging techniques, apparatus, and patient cohorts, hence influencing the efficacy and applicability of the model. If the training datasets contain biases, these models may not perform well when applied to a diverse community. There may be instances in clinical settings where the necessary computational capacity to utilize machine learning models is not accessible. Moreover, the inherent complexity and lack of comprehensibility of these models can impede doctors from embracing and having faith in them.

6. Conclusion and Future scope

In this work, we explore the effectiveness of extracting attributes using radiomics statistics from breast cancer images to improve cancer detection using machine learning

approaches. Several crucial findings have evolved from rigorous testing and research, considerably advancing breast cancer detection and diagnosis. Our findings highlight the relevance of feature extraction from radiomics data in enhancing breast cancer detection. The capacity to extract pertinent features such as RunVariance, RunEntropy, Energy, Elongation, MinorAxisLength, etc. from complicated radiomics data has enormous promise for improving diagnostic accuracy and assisting clinical decision-making in breast cancer diagnosis. This study is a critical step towards using sophisticated data-driven approaches to improve breast cancer detection, resulting in more effective, accurate, and personalized cancer treatment strategies. These findings open the door for creating more accurate and effective categorization algorithms, allowing medical practitioners to make more informed decisions, and perhaps improving patient outcomes in breast cancer therapy.
In the future, researchers should ensure that our model works on bigger, more varied datasets from multiple institutions. This will make it more generalizable and reliable. AI-driven diagnostics will also be easier for clinicians to understand and accept if interpretability methods like SHAP values and decision trees are developed and used together. Advanced dimensionality reduction methods, like those based on deep learning, can help keep useful data while lowering the amount of data that needs to be stored.

Acknowledgment

This work is acknowledged under Integral University manuscript No IU/R&D/2024-MCN0002772

Funding

No Funding

Institutional Review Board Statement

Not applicable

Informed Consent Statement

Not applicable

Data Availability Statement

Not Applicable

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] C. H. Barrios, "Global challenges in breast cancer detection and treatment," *The Breast*, vol. 62, pp. S3–S6, 2022.

[2] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc *et al.*, "Ai in medical imaging informatics: current challenges and future directions," *IEEE journal of biomedical and health informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.

[3] G. Upreti, "Advancements in skull base surgery: Navigating complex challenges with artificial intelligence," *Indian Journal of Otolaryngology and Head & Neck Surgery*, pp. 1–7, 2023.

[4] X. Zhang, Y. Zhang, G. Zhang, X. Qiu, W. Tan, X. Yin, and L. Liao, "Deep learning with radiomics for disease diagnosis and treatment: challenges and potential," *Frontiers in oncology*, vol. 12, p. 773840, 2022.

[5] C. Yu, X. Bi, and Y. Fan, "Deep learning for fluid velocity field estimation: A review," *Ocean Engineering*, vol. 271, p. 113693, 2023.

[6] A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit, "A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, p. 5930, 2023.

[7] N. Zielonke, A. Gini, E. E. Jansen, A. Anttila, N. Segnan, A. Ponti, P. Veerus, H. J. de Koning, N. T. van Ravesteyn, E. A. Heijnsdijk *et al.*, "Evidence for reducing cancer-specific mortality due to screening for breast cancer in europe: A systematic review," *European journal of cancer*, vol. 127, pp. 191–206, 2020.

[8] S. K. Verma, D. Arora, and R. Bhardwaj, "Breast cancer survival rate prediction in mammograms using machine learning," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, 2020, pp. 169–171.

[9] Y. Amkrane, M. El Adoui, and M. Benjelloun, "Towards breast cancer response prediction using artificial intelligence and radiomics," in *2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*. IEEE, 2020, pp. 1–5.

[10] W. Ma, Y. Zhao, Y. Ji, X. Guo, X. Jian, P. Liu, and S. Wu, "Breast cancer molecular subtype prediction by mammographic radiomic features," *Academic radiology*, vol. 26, no. 2, pp. 196–201, 2019.

[11] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *Journal of personalized medicine*, vol. 11, no. 2, p. 61, 2021.

[12] A. S. Tagliafico, M. Piana, D. Schenone, R. Lai, A. M. Massone, and N. Houssami, "Overview of radiomics in breast cancer diagnosis and prognostication," *The Breast*, vol. 49, pp. 74–80, 2020.

[13] R. Massafra, S. Bove, V. Lorusso, A. Biafora, M. C. Comes, V. Didonna, S. Diotaiuti, A. Fanizzi, A. Nardone, A. Nolasco *et al.*, "Radiomic feature reduction approach to predict breast cancer by contrast-enhanced spectral mammography images," *Diagnostics*, vol. 11, no. 4, p. 684, 2021.

[14] L. A. Daamen, I. Q. Molenaar, and V. P. Groot, "Recent advances and future challenges in pancreatic cancer care: Early detection, liquid biopsies, precision medicine and artificial intelligence," *Journal of Clinical Medicine*, vol. 12, no. 23, p. 7485, 2023.

[15] C. Cong, X. Li, C. Zhang, J. Zhang, K. Sun, L. Liu, B. Ambale-Venkatesh, X. Chen, and Y. Wang, "Mri-based breast cancer classification and localization by multiparametric feature extraction and combination using deep learning," *Journal of Magnetic Resonance Imaging*, vol. 59, no. 1, pp. 148–161, 2024.

[16] N. Luo, X. Zhong, L. Su, Z. Cheng, W. Ma, and P. Hao, "Artificial intelligence-assisted dermatology diagnosis: from unimodal to multimodal," *Computers in Biology and Medicine*, p. 107413, 2023.

[17] R. Laajili, M. Said, and M. Tagina, "Application of radiomics features selection and classification algorithms for medical imaging

decision: Mri radiomics breast cancer cases study," *Informatics in Medicine Unlocked*, vol. 27, p. 100801, 2021.

[18] E.-N. Cheong, J. E. Park, S. Y. Park, S. C. Jung, and H. S. Kim, "Achieving imaging and computational reproducibility on multiparametric mri radiomics features in brain tumor diagnosis: Phantom and clinical validation," *European Radiology*, pp. 1–16, 2023.

[19] M. Hosseinzadeh, A. Gorji, A. Fathi Jouzdani, S. M. Rezaeijo, A. Rahmim, and M. R. Salmanpour, "Prediction of cognitive decline in parkinson's disease using clinical and dat spect imaging features, and hybrid machine learning systems," *Diagnostics*, vol. 13, no. 10, p. 1691, 2023.

[20] A. Conti, A. Duggento, I. Indovina, M. Guerrisi, and N. Toschi, "Radiomics in breast cancer classification and prediction," in *Seminars in cancer biology*, vol. 72. Elsevier, 2021, pp. 238–250.

[21] T. Mahmood, J. Li, Y. Pei, F. Akhtar, A. Imran, and M. Yaqub, "An automatic detection and localization of mammographic microcalcifications roi with multi-scale features using the radiomics analysis approach," *Cancers*, vol. 13, no. 23, p. 5916, 2021.

[22] M. I. Tsarouchi, A. Hoxhaj, and R. M. Mann, "New approaches and recommendations for risk-adapted breast cancer screening," *Journal of Magnetic Resonance Imaging*, vol. 58, no. 4, pp. 987–1010, 2023.

[23] Y. Yu, W. Ren, Z. He, Y. Chen, Y. Tan, L. Mao, W. Ouyang, N. Lu, J. Ouyang, K. Chen *et al.*, "Machine learning radiomics of magnetic resonance imaging predicts recurrence-free survival after surgery and correlation of lncrnas in patients with breast cancer: a multicenter cohort study," *Breast Cancer Research*, vol. 25, no. 1, p. 132, 2023.

[24] J. Lee, S. K. Yoo, K. Kim, B. M. Lee, V. Y. Park, J. S. Kim, and Y. B. Kim, "Machine learning-based radiomics models for prediction of locoregional recurrence in patients with breast cancer," *Oncology Letters*, vol. 26, no. 4, pp. 1–10, 2023.

[25] L. Luo, X. Wang, Y. Lin, X. Ma, A. Tan, R. Chan, V. Vardhanabhuti, W. C. Chu, K.-T. Cheng, and H. Chen, "Deep learning in breast cancer imaging: A decade of progress and future directions," *IEEE Reviews in Biomedical Engineering*, 2024.

[26] A. Sahu, P. K. Das, and S. Meher, "Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms," *Physica Medica*, vol. 114, p. 103138, 2023.

[27] N. Binsaif *et al.*, "Application of machine learning models to the detection of breast cancer," *Mobile Information Systems*, vol. 2022, 2022.

[28] F. A. Al-Fahaidy, B. Al-Fuhaidi, I. AL-Darouby, F. AL-Abady, M. AL-Qadry, and A. AL-Gamal, "A diagnostic model of breast cancer based on digital mammogram images using machine learning techniques." *Applied Computational Intelligence & Soft Computing*, 2022.

[29] M. Botlagunta, M. D. Botlagunta, M. B. Myneni, D. Lakshmi, A. Nayyar, J. S. Gullapalli, and M. A. Shah, "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Scientific Reports*, vol. 13, no. 1, p. 485, 2023.

[30] L. Losurdo, A. Fanizzi, T. M. A. Basile, R. Bellotti, U. Bottigli, R. Dentamaro, V. Didonna, V. Lorusso, R. Massafra, P. Tamborra *et al.*, "Radiomics analysis on contrast-enhanced spectral mammography images for breast cancer diagnosis: A pilot study," *Entropy*, vol. 21, no. 11, p. 1110, 2019.

[31] W. Ma, Y. Zhao, Y. Ji, X. Guo, X. Jian, P. Liu, and S. Wu, "Breast cancer molecular subtype prediction by mammographic radiomic features," *Academic radiology*, vol. 26, no. 2, pp. 196–201, 2019.

[32] C.-e. A. Tai, H. Gunraj, N. Hodzic, N. Flanagan, A. Sabri, and A. Wong, "Enhancing clinical support for breast cancer with deep learning models using synthetic correlated diffusion imaging," in *International Workshop on Applications of Medical AI*. Springer, 2023, pp. 83–93.

[33] C. Militello, L. Rundo, M. Dimarco, A. Orlando, R. Woitek, I. D'Angelo, G. Russo, and T. V. Bartolotta, "3d dce-mri radiomic analysis for malignant lesion prediction in breast cancer patients," *Academic Radiology*, vol. 29, no. 6, pp. 830–840, 2022.

[34] A. G. Bitencourt, P. Gibbs, C. R. Saccarelli, I. Daimiel, R. L. Gullo, M. J. Fox, S. Thakur, K. Pinker, E. A. Morris, M. Morrow *et al.*, "Mri-based machine learning radiomics can predict her2 expression level and pathologic response after neoadjuvant therapy in her2 overexpressing breast cancer," *EBioMedicine*, vol. 61, 2020.

[35] L. Lenga, S. Bernatz, S. S. Martin, C. Booz, C. Solbach, R. Mulert-Ernst, T. J. Vogl, and D. Leithner, "Iodine map radiomics in breast cancer: prediction of metastatic status," *Cancers*, vol. 13, no. 10, p. 2431, 2021.

[36] H. Yu, X. Meng, H. Chen, X. Han, J. Fan, W. Gao, L. Du, Y. Chen, Y. Wang, X. Liu *et al.*, "Correlation between mammographic radiomics features and the level of tumor-infiltrating lymphocytes in patients with triple-negative breast cancer," *Frontiers in Oncology*, vol. 10, p. 412, 2020.

[37] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics*, vol. 1, no. 1, pp. 1–22, 2016.

[38] "Breast ultrasound images dataset," https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset, accessed: 05/18/2024.

[39] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher *et al.*, "Radiomics: the process and the challenges," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[40] M. R. Tomaszewski and R. J. Gillies, "The biological meaning of radiomic features," *Radiology*, vol. 298, no. 3, pp. 505–516, 2021.

[41] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, and G. Cook, "Introduction to radiomics," *Journal of Nuclear Medicine*, vol. 61, no. 4, pp. 488–495, 2020.

[42] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[43] A. Behura, "The cluster analysis and feature selection: Perspective of machine learning and image processing," *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 249–280, 2021.

[44] M. Bahri, A. Bifet, S. Maniu, and H. M. Gomes, "Survey on

feature transformation techniques for data streams," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 4796–4802.

[45] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications." *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, 2019.

[46] D. Kalman, "A singularly valuable decomposition: the svd of a matrix," *The college mathematics journal*, vol. 27, no. 1, pp. 2–23, 1996.

[47] R. A. Khan, N. Rashid, M. Shahzaib, U. F. Malik, A. Arif, J. Iqbal, M. Saleem, U. S. Khan, and M. Tiwana, "A novel framework for classification of two-class motor imagery eeg signals using logistic regression classification algorithm," *Plos one*, vol. 18, no. 9, p. e0276133, 2023.

[48] Z. Azam, M. M. Islam, and M. N. Huda, "Comparative analysis of intrusion detection systems and machine learning based model analysis through decision tree," *IEEE Access*, 2023.

[49] Z. Jin, T. G. Hei, and K. Jin, "Application of support vector machines for categorizing biological and medical data," vol. 13105, pp. 1099–1107, 2024.

[50] T. Mahesh, V. Vinoth Kumar, V. Vivek, K. Karthick Raghunath, and G. Sindhu Madhuri, "Early predictive model for breast cancer classification using blended ensemble learning," *International Journal of System Assurance Engineering and Management*, vol. 15, no. 1, pp. 188–197, 2024.

[51] J. Naskath, G. Sivakamasundari, and A. A. S. Begum, "A study on different deep learning algorithms used in deep neural nets: Mlp som and dbn," *Wireless Personal Communications*, vol. 128, no. 4, pp. 2913–2936, 2023.

[52] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam, "Xgboost and random forest algorithms: An in depth analysis," *Pakistan Journal of Scientific Research*, vol. 3, no. 1, pp. 26–31, 2023.

[53] J. Zhou, H. Tan, W. Li, Z. Liu, Y. Wu, Y. Bai, F. Fu, X. Jia, A. Feng, H. Liu *et al.*, "Radiomics signatures based on multiparametric mri for the preoperative prediction of the her2 status of patients with breast cancer," *Academic Radiology*, vol. 28, no. 10, pp. 1352–1360, 2021.

[54] I. Daimiel Naranjo, P. Gibbs, J. S. Reiner, R. Lo Gullo, C. Sooknanan, S. B. Thakur, M. S. Jochelson, V. Sevilimedu, E. A. Morris, P. A. Baltzer *et al.*, "Radiomics and machine learning with multiparametric breast mri for improved diagnostic accuracy in breast cancer diagnosis," *Diagnostics*, vol. 11, no. 6, p. 919, 2021.

[55] M. A. Hassanien, V. K. Singh, D. Puig, and M. Abdel-Nasser, "Predicting breast tumor malignancy using deep convnext radiomics and quality-based score pooling in ultrasound sequences," *Diagnostics*, vol. 12, no. 5, p. 1053, 2022.

[56] J. Lee, S. K. Yoo, K. Kim, B. M. Lee, V. Y. Park, J. S. Kim, and Y. B. Kim, "Machine learning-based radiomics models for prediction of locoregional recurrence in patients with breast cancer," *Oncology Letters*, vol. 26, no. 4, pp. 1–10, 2023.

[57] Y. Lin, J. Wang, M. Li, C. Zhou, Y. Hu, M. Wang, and X. Zhang, "Prediction of breast cancer and axillary positive-node response to neoadjuvant chemotherapy based on multi-parametric magnetic resonance imaging radiomics models," *The Breast*, p. 103737, 2024.

[58] Y. Yu, Z. Wang, Q. Wang, X. Su, Z. Li, R. Wang, T. Guo, W. Gao, H. Wang, and B. Zhang, "Radiomic model based on magnetic resonance imaging for predicting pathological complete response after neoadjuvant chemotherapy in breast cancer patients," *Frontiers in Oncology*, vol. 13, p. 1249339, 2024.

Zulfikar Ali Ansari holds a Bachelor of Technology (B.Tech.) in Computer Science & Engineering from Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India, obtained in 2012. He subsequently earned a Master of Technology (M.Tech.) in the same field from Integral University, Lucknow, India, in 2019. Currently pursuing his Ph.D. in Computer Science & Engineering at Integral University Lucknow, India, Mr. Ansari's research interests focus on Artificial Intelligence, Machine Learning, Deep Learning, and Explainable AI.

Dr. Manish Madhava Tripathi has a distinguished figure in computer science and engineering education and boasts a prolific research portfolio spanning over two decades. With expertise in medical image processing, big data analytics, and machine learning, Prof. Tripathi has authored 60 papers indexed in Google Scholar, boasting an impressive H-index of 8 and numerous citations. His groundbreaking research has significantly contributed to advancements in medical image watermarking, paving the way for enhanced diagnostic accuracy in healthcare. He has mentored numerous Ph.D. and M.Tech scholars, nurturing the next generation of researchers. Prof. Tripathi's commitment to academic excellence extends to his active participation in organizing workshops and conferences on emerging technologies like Python, machine learning, and cybersecurity. His leadership in research projects and role as a reviewer for esteemed conferences underscore his esteemed status in the scientific community, exemplifying his unwavering dedication to advancing the frontiers of computer science research.

Dr. Rafeeq Ahmed works as an Assistant Professor in the CSE Department, at Government Engineering College West Champaran, India. He has a Ph.D. from Jamia Millia Islamia, an M.Tech (Software Engineering), and a B.Tech (Computer Engineering) from Aligarh Muslim University. He has been given a gold medal in M.Tech. UGC has also awarded him the Maulana Azad National Fellowship (MANF). He has teaching experience of more than 10 years. He has worked on the organizing committee for the international conferences ICACSE 2019 and ICACSE 2021. He has published 20+ International Journals, patents, and Conference papers in SCI/SCOPUS-indexed journals in text mining, Big Data, Recommendation systems, IoT, and many others. He has also received the best paper award at the International Conference SIGMA-2018 held at NSIT, New Delhi. He is a reviewer of many reputed Q1/Q2/Q3 Journals.