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

Zulfikar Ali Ansari[a,b], Manish Madhava Tripathi[a], Rafeeq Ahmed[b,*]

[a]*Department of Computer Science and Engineering, Integral University, Kursi Road, Lucknow, 225606, Uttar Pradesh, India*
[b]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, Andhra Pradesh, India*

Abstract

Radiomics allows for measuring tumor 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. The characteristics include many qualities such as shape, texture, intensity, and spatial relationships within the images. Radiomics is crucial for extracting features by turning medical images into quantitative data that capture detailed aspects of tissue architecture and physiology. The identified traits could significantly transform clinical decision-making in oncology and other areas. This study aims to enhance existing breast cancer diagnostic techniques by utilizing radiomics to detect the disease at an early stage. Our study intends to enhance diagnostic accuracy by utilizing machine learning models and dimensionality reduction approaches on radiomics characteristics. We provide a new technique that integrates dimensionality reduction with machine learning algorithms to examine radiomics characteristics collected from breast cancer images, improving early breast cancer detection. The proposed method is comprehensively evaluated, showing significant enhancements in diagnostic accuracy for early-stage breast cancer when compared to conven-

*Corresponding author.
Email addresses: zulfi78692@gmail.com (Zulfikar Ali Ansari), mmt@iul.ac.in (Manish Madhava Tripathi), rafeeq.amu@gmail.com (Rafeeq Ahmed)

Preprint submitted to International Journal of Computing and Digital Systems February 27, 2024

tional methods.The proposed model has an accuracy of 88.72% as compared to recent works as mentioned in Table3. The results suggest that radiomics-based techniques could enhance breast cancer screening by identifying subtle imaging indicators.

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

1. INTRODUCTION

Breast cancer remains a significant global health challenge, demanding accurate and timely diagnostic approaches for effective treatment and improved patient outcomes (Barrios, 2022). The advent of radiomics, an innovative field leveraging quantitative analysis of medical images, has shown promising prospects in augmenting traditional diagnostic methodologies(Panayides et al., 2020). Radiomics empowers the extraction of intricate information from radiological images, enabling the exploration of subtle patterns and characteristics that might elude visual inspection alone (Upreti, 2023). In this context, our study delves into the realm of breast cancer diagnosis, focusing on the integration of radiomics and advanced computational techniques to refine 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(Zhang et al., 2022). 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. This research aims to investigate the efficacy of various dimensionality reduction methodologies in enhancing breast cancer diagnosis based on radiomics features. By condensing the feature space while preserving diagnostically relevant information, our endeavor seeks to optimize classification models, enabling more accurate and interpretable outcomes. The subsequent sections of this paper detail the methodology employed, encompassing the Radiomics feature extraction, the selection of dimensionality reduction techniques, and the evaluation of refined classification models. Furthermore,

comprehensive analyses and discussions of the findings offer insights into the potential impact of dimensionality reduction approaches on improving breast cancer diagnosis, ultimately contributing to the advancement of precision medicine in oncology.

1.1. 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. Through the utilization of ML models along with dimensionality reduction techniques on radiomics features, our goal is to provide a valuable contribution to the advancement of a more efficient and nuanced diagnostic approach, ultimately propelling the field of early diagnosis of breast cancer. Research Contribution Our research significantly enhances the field of early breast cancer diagnosis by offering a new method that combines dimensionality reduction techniques with machine learning (ML) algorithms to analyze radiomics features. The results of our study indicate a significant enhancement in the precision of diagnosing early-stage breast cancer as compared to conventional techniques. By employing dimensionality reduction approaches, our research identifies crucial radiomics properties that greatly boost diagnostic performance

The remaining part of the paper is elaborated as follows: Section 2 highlights the related work on Breast Cancer detection and also includes the comparative study of recent works. Section 3 describes the methodology and provides a brief about feature extraction and selection. Section 4 presents the experimental results. Section 5 describes the discussion and comparative analysis. Section 6 concludes the paper and future scope is also enlisted.

2. Related work

Breast cancer (BC) was the predominant form of cancer among women in all European countries in 2018, and it was the primary cause of death from cancer in women across Europe. Zielonke et al. (2020). Making use of the Wisconsin BC (original) dataset, the primary purpose of this work is to evaluate the efficiency of several different methods. The author has done a detailed review of the ML inVerma et al. (2020) to forecast cancer and describe and compare in-depth learning techniques. The SVM, K-NN, RF, ANN, and LR models were used in a comparative analysis of five machine learning methods that were used to predict BC. This analysis was

supplied by another author. The Author Amkrane et al. (2020) suggested a novel method to predict breast tumor response to treatment. Advancements have been achieved in characterizing breast cancer subtypes using radiological images. Specific qualitative and visual information obtained from breast magnetic resonance imaging (MRI), mammography, or ultrasound has been demonstrated to correlate with the molecular subtypes of breast cancer Ma et al. (2019a). The major goal of the paper Wu and Hicks (2021) 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 Tagliafico et al. (2020) just explored Radiomics as an overview through Machine learning & Deep Learning on Breast cancer mainly. Currently, physicians receive assistance in analyzing these images via computer-aided detection/diagnosis (CAD) systems. CAD is a software used in clinical medicine that utilizes clinical data and algorithms to propose diagnoses and treatments Massafra et al. (2021) Medical imaging is essential for the diagnosis, staging, therapy planning, post-operative surveillance, and evaluation of response in the routine management of cancer. Magnetic Resonance Imaging (MRI) is the most precise and sensitive imaging technique for diagnosing and identifying lesions, particularly in cases of breast cancer, which is the most prevalent type of cancer among women. The Authors Yu et al. (2021) to use the study utilized machine learning methods to create a very effective preoperative evaluation methodology for determining the status of axillary lymph nodes (ALN) using magnetic resonance imaging (MRI) radiomics. Additionally, the study investigated the relationship Regarding patients with early-stage invasive breast cancer, there exists a relationship between radiomics and the tumor microenvironment. The Author Laajili et al. (2021) This study demonstrated the utilization of diverse radiomics features to assist decision-making in clinical tasks and the efficacy of different machine learning classifiers, in conjunction with multiple feature selection techniques, in reliably predicting breast cancer nodules. The reviewed article is to provide a concise overview of the current advancements in breast cancer research that utilize radiomics in Conti et al. (2021). The proposed radiomics fusion algorithm is utilized to categorize the chosen characteristics into malignant and benign Mahmood et al. (2021)Ċomparative studies in breast cancer detection assess the efficacy of various imaging modalities, technologies, or procedures for early diagnosis, screening, and characterization of breast abnormalities. Table 1 compares the increased benefit of radiomics analysis in

Table 1: Various Comparative Studies of Breast Cancer Detection

Author	Classifier / Methods	Dataset	Radiomics	Accuracy
Mamatha Sai Yarabarla et al. Yarabarla et al. (2019), 2019	RF algorithm	WBCD	No	69
Nasser Binsaif Binsaif et al. (2022), 2022	DT, RF, K-NN, ANN, SVM & LR	BC Database of Coimbra (UCI)	No	64
Farouk A. K. Al-Fahaidy et al. Al-Fahaidy et al. (2022) 2022	SVM	MIAS dataset of mammogram images	No	87.1
Mahendran Botlagunta et al. Botlagunta et al. (2023) 2023	XGBoost, LR, KNN, DT, RF, SVM,	BIACH & RI as a semi-structured	No	83
Liliana Losurdo et al. Losurdo et al. (2019), 2019	SVM classifier	CESM Image	Yes	NA
Wenjuan Ma et al. Ma et al. (2019b), 2018	NB	331 Chinese women data	Yes	79.6
Chi-en Amy Tai et al. Tai et al. (2023), 2023	CNN	253 patients	No	87.7
Carmelo Militello et al. Militello et al. (2022), 2022	SVM	111 patients	Yes	91
Almir G.V. Bitencourt, et al. Bitencourt et al. (2020). 2020	HER2 expression	311 patients.	Yes	89.7
Lukas Lenga et al. Lenga et al. (2021), 2021	DECT iodine map-derived radiomic signatures	77 patients	Yes	92.6
Hongwei Yu et al. Yu et al. (2020), 2020	TIL levels	43 Patients	Yes	74.4

breast cancer detection to standard imaging approaches.

3. Material & Methods

This section will provide a comprehensive analysis of the dataset, including information on its details, preprocessing techniques, and feature extraction methods that will be employed. This section displays the flowchart illustrating the methodology for forecasting BC. The proposed classification model employs a variety of methods, such as LR, SVM, DT, RF, MLP, and XGboost. To examine the characteristics, we Will utilize PCA, NMF, and SVD as shown in figure 1. The authors have utilized a dataset comprising breast cancer imaging data obtained from the Kaggle, the dataset included a substantial number of cases, encompassing diverse types and stages of breast cancer along with healthy controls.

3.1. Dataset Description

Examples of breast ultrasounds taken from women ranging in age from 25 to 75 years old are included in the data acquired at the beginning of

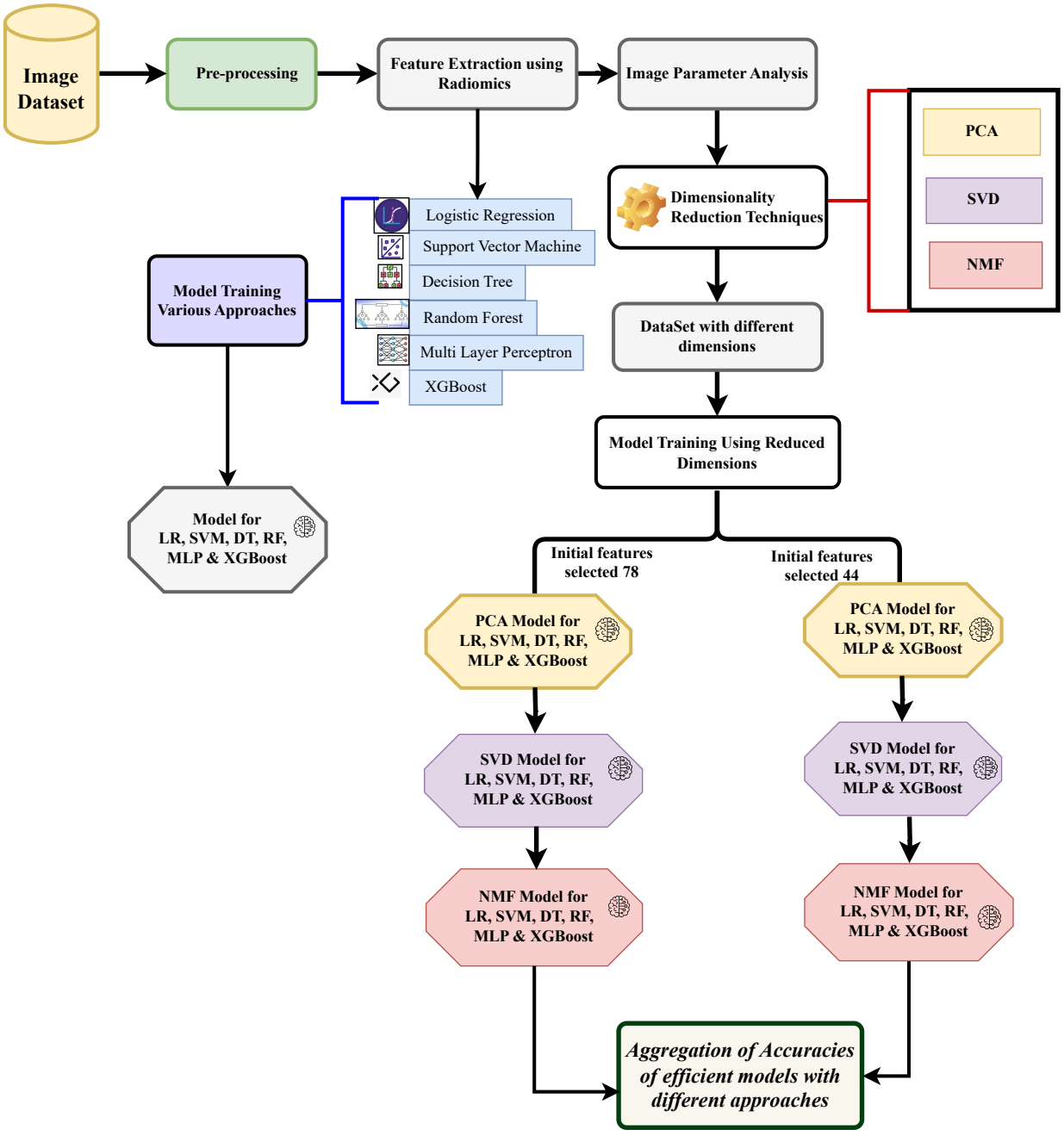

Figure 1: Flow diagram of the proposed approach

6

the study. It was in 2018 when these statistics were gathered. 600 female patients are included in the total number of patients. The collection of data includes 780 images (Benign 437, Malignant 210, Normal 133), each of which has an average size of 500 pixels by 500 pixels. PNG is the format that the images are in. Additionally, the original photos are exhibited alongside the ground truth images. These pictures are divided into three categories: normal, benign, and malignant. Normal photos are the most common. The dataset utilized in this work is a publicly accessible Kaggle website repository that hosts several datasets.

3.2. Preprocessing

Data preparation is a crucial step in eliminating noise inconsistencies, and redundant information to deliver high-quality data that boosts performanceGarcía et al. (2016). Within the context of the pre-processing of the data, the mean of the associated characteristic is utilized to fill in any absent values. The formatting of the dataset that was provided has been checked and found to be consistent. Following the removal of duplicates and missing values, we look for any missing values in this dataset. If we find any, we simply fill in the median of the value that was missing.

3.3. Feature Extraction

Image features are the basic qualities utilized to observe it. Unique properties are extracted. We must extract these features from an image dataset to sort photos by explicit features. There is no precise description of visual attributes, but size, form, etc. start them. Python algorithms and methods extract these features. Pictures contain a lot of information and pre-processing them helps. It aids in picture enhancement, retrieval, visualization, and identification. Python's Scikit-Image prepares images. The libraries or algorithms handle segmentation, color space manipulation, analysis, morphology, feature detection, etc. High-yielding computational devices can extract many quantitative features from tomographic pictures (CT, MR, PET, etc.). Radiomics converts medical pictures into high-dimensional data.

3.4. Radiomics

Radiomics is an evolving field within Medical imaging that encompasses the retrieval and examination of quantitative characteristics from radiographic images Kumar et al. (2012). It goes beyond traditional visual assessment by using advanced computational methods to capture a large amount of data

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 has the potential to deliver significant insights Tomaszewski and Gillies (2021). This methodology has been implemented in the field of oncology, where it is intended to enhance diagnostic accuracy, facilitate the assessment of prognostic factors, and provide assistance in the process of clinical decision-making. The radiomics method calculates the scalar values of the features from the predefined ROI (Region of interest) Mayerhoefer et al. (2020). Once the lesions are segmented, feature extraction is carried out using radiomics. In our model, Radiomics is used to extract manual features of 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:

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

Energy is a proportion of the size of voxel values in a picture. A bigger quality suggests a more prominent amount of the squares of these qualities.

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

A measure of the degree to which the distribution of attributes deviates from the mean value is referred to as skewness. This value can be either positive or negative, depending on the location where the tail is stretched and the mass of the dispersion is concentrated.

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

A proportion of the 'peakedness' of the conveying of attributes in the picture ROI is what is referred to as kurtosis. The value can be either positive or negative, depending on the position where the tail is stretched and the concentration of mass in the dispersion.

$$Sphericity = \frac{2\sqrt{\pi A}}{P} \tag{4}$$

Sphericity is the ratio of the perimeter of the tumor region to the diameter of a circle that has the same surface area as the tumor region. The metric quantifies the degree of sphericity of the tumor region about a circle. This metric is dimensionless and unaffected by both magnitude and orientation.

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

This element yields the biggest hub length of the ROI-encasing ellipsoid and is determined utilizing the biggest head part λ_{major}.

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

Elongation shows the connection between the two biggest head segments in the ROI shape.

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

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 calculates the instability of gray level intensity values in the image.

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

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.

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

SZNUN Calculates the instability of the size zone of the image.

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

DNUN Measures the analogy throughout the image, with a diminished value signifying homogeneity with dependencies in the image.

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

GLV measures the variance in the gray level.

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

DV measures the variance in dependence size in the image.

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

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.

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

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.

3.5. Feature Analysis

Within the machine learning process, feature analysis, feature engineering, and feature selection are all terms that describe the same process, and are very important components Behura (2021). The task involves the process of selecting, changing, or creating relevant features (input variables or attributes) from the raw data to improve the effectiveness of a machine-learning model. Effective feature analysis using a radiomics approach can lead to more accurate and efficient models, faster training times, and a better understanding of the underlying data. Radiomics analysis is a growing topic in medical imaging that entails extracting a large number of quantitative information from images. These properties can then be analyzed to provide important details about the basic biology of the imaged tissue. Radiomics has shown the potential to detect breast cancer by improving diagnosis, prognosis, and treatment planning.

3.5.1. DR

Dimensionality reduction (DR) is crucial in machine learning since it enables the conversion of high-dimensional data into a lower-dimensional space while preserving the essential properties of the data Bahri et al. (2021). It improves the effective utilization of data while also reducing the computing burden on automated processes. The use of this technique not only enhances the computational efficiency of data processing but also reduces the intricacy of the NN architecture and facilitates the effective construction of fuzzy inference rules Gupta and Janghel (2019).

PCA: Principal Component Analysis (PCA) is a widely employed method in the domains of Machine Learning (ML) and data analysis to reduce the dimensionality of a dataset. The primary objective is to decrease the number of attributes while maintaining a significant portion of the initial variability. The algorithm is specifically tailored to process a data matrix X with dimensions m * n, where m represents the number of observations and n represents the number of variables. Calculation of the covariance matrix in the context of breast cancer: Calculate the covariance matrix for the centered breast cancer data. The covariance between two features, represented as i and j, in the breast cancer dataset is calculated using the following formula in this specific scenario:

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

The result of this calculation is a covariance matrix with dimensions m x m, which holds great importance in the analysis of breast cancer and is symbolized as $\sum$. 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.

$$\sum v = \lambda v \tag{23}$$

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.

SVD: The Singular Value Decomposition (SVD) is a fundamental method used for matrix factorization, which finds extensive use in the domains of linear algebra and numerical linear algebra Kalman (1996). 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 of order n x d, we shall utilize three matrices: U, $\sum$, and V. Let C be a matrix in R_{nd}. U and V are orthogonal matrices of size n x n and d x d, respectively. The matrix $\sum$ is a nonnegative diagonal matrix belonging to the set of real matrices with dimensions n×d. Mathematically, the Singular Value Decomposition (SVD) factorizes a given matrix A in the following manner:

$$A = U\Sigma V^T \tag{24}$$

13

NMF: Non-negative matrix factorization (NMF) is a mathematical method employed in the fields of linear algebra and data analysis Fu et al. (2019). NMF, in contrast to SVD, decomposes a matrix into the product of two matrices that include only non-negative values. This property makes NMF highly advantageous for specific data analysis tasks such as feature extraction, image processing, and text mining. Mathematically, Non-negative Matrix Factorization (NMF) aims to find two non-negative 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. When considering factorization:

$$X \approx WH \tag{25}$$

Let X be the initial non-negative matrix that you wish to factorize. 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.

3.6. Methods Incorporated

In this section, we will discuss the whole classifier in a summarized form. Logistic Regression (LR) is a linear regression-based classification algorithm that estimates the probability of a binary or multiclass outcome by applying a logistic function to the input features, offering simplicity and interpretability Khan et al. (2023). 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 Azam et al. (2023). 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 Darveau (2023). Random Forest (RF), an ensemble of decision trees, mitigates overfitting by aggregating predictions from

multiple diverse trees, ideal for handling large datasets Baratchi. Multilayer Perceptron (MLP), a neural network, learns complex patterns through layers of nodes with non-linear activations, suitable for non-linear relationships Naskath et al. (2023). 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 Fatima et al. (2023).

3.7. 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

In this study, we employed a dimensionality reduction approach to radiomics features to enhance the diagnosis of breast cancer. 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. 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,

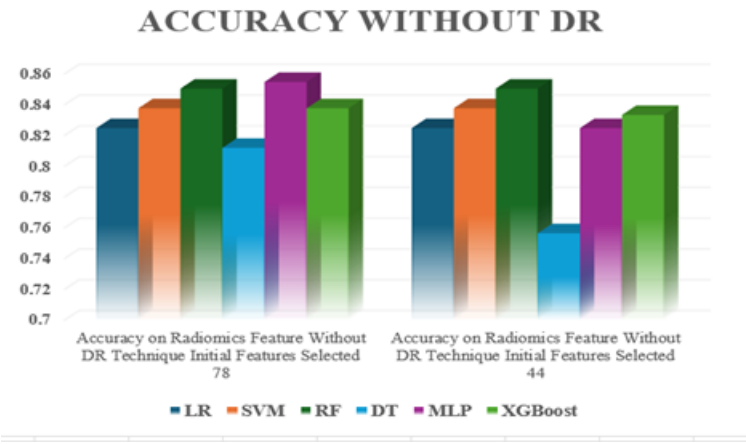

Figure 2: Comparison Accuracy all Models without DR

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.

16

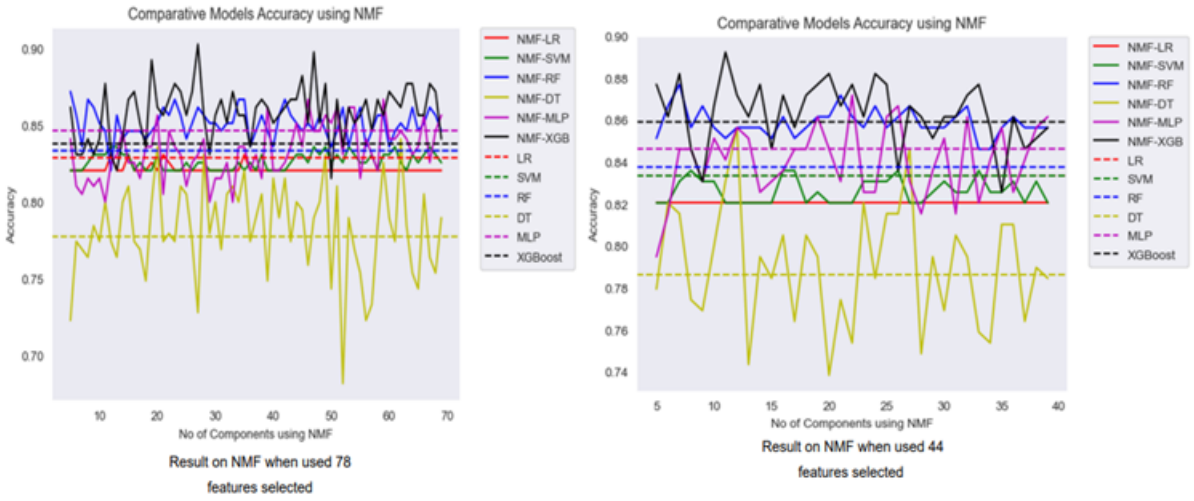

Figure 3: Result on NMF with two different features set

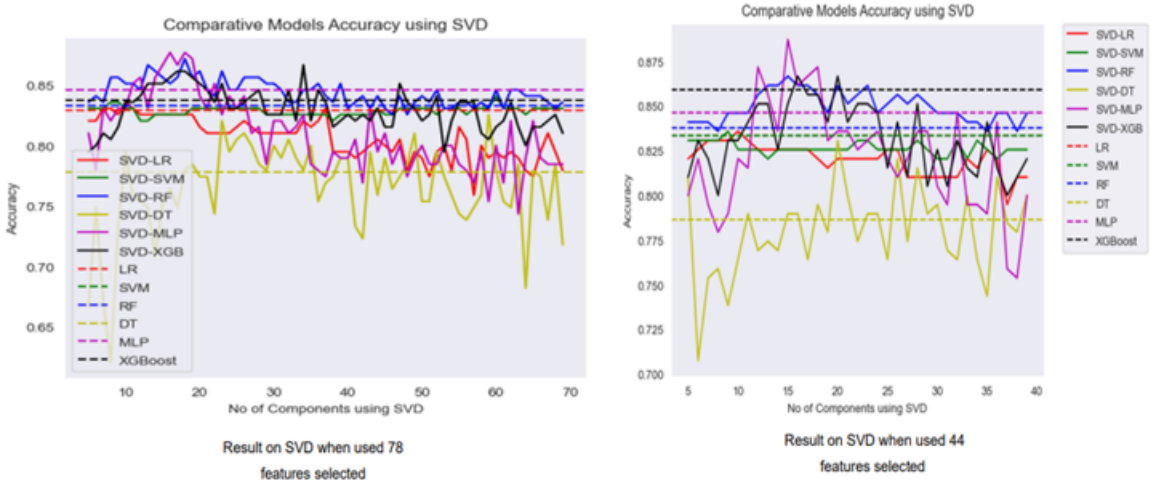

Figure 4: Result on SVD with two different features set

17

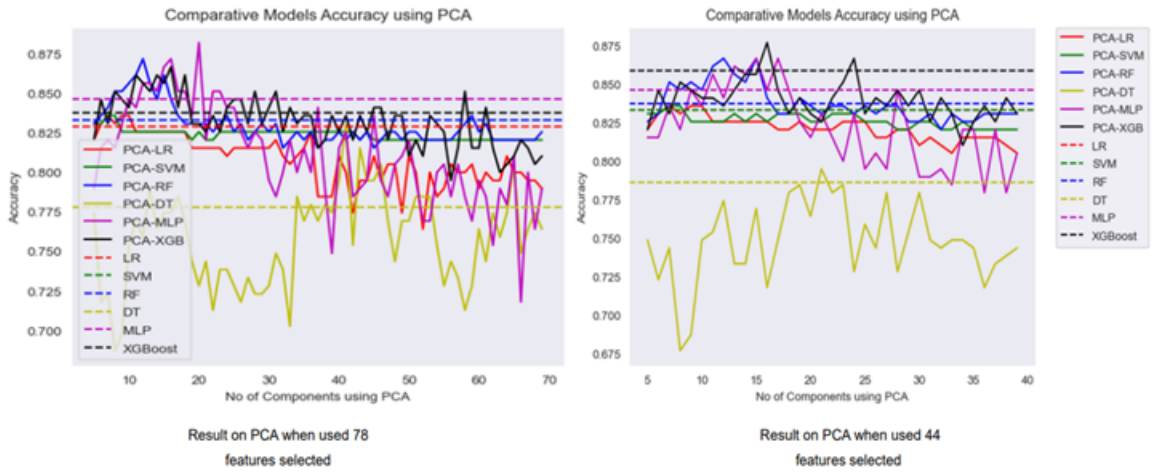

Figure 5: Result on PCA with two different features set

5. Discussion

Our study primarily aimed to evaluate the influence of dimensionality reduction on the diagnostic precision of breast cancer. Examination of the selected PCA, NMF, and SVD radiomics features that significantly contributed to the improved diagnostic accuracy. These discovered characteristics need additional examination as possible indicators of breast cancer. In table 2, we present a summary of our findings, which includes all the models evaluated and their corresponding top accuracy scores for each component. The results of the initial 78 features selected and the initial 44 features picked by PCA, SVD, and NMF on all models (LR, SVM, RF, DT, MLP, and XGboost) are summarised in table 2. In figure 6 we plot a graph of all models and finally, we got the best accuracy in all models through the DR technique on XGboost which is 88.72%.

The proposed methodology for breast cancer detection demonstrates significant improvements in accuracy compared to existing methods which is shown in table 3. Through a comprehensive comparative analysis, it is observed that the proposed approach integrates advanced radiomics features

18

Table 2: 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	0.841	17	0.8462
	SVM	10	0.8308	5	0.8308
	RF	18	0.8513	12	0.8462
	DT	41	0.8	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	0.8
	MLP	18	0.8564	17	0.8667
	XGBoost	12	0.8513	28	0.841
NMF	LR	12	0.841	10	0.841
	SVM	64	0.8462	16	0.841
	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 3: Comparison of Existing approaches with proposed approach

Author	**Classifier / Methods**	**Dataset**	**Radiomics**	**Accuracy**
Jing Zhou et al. 2020 Zhou et al. (2021)	SVM	306 patients	Yes	87
Isaac Daimiel Naranjo et al. 2021 Daimiel Naranjo et al. (2021)	multiparametric radiomics mode	93_Patients	Yes	85
Mohamed A. Hassanien et al. 2022 Hassanien et al. (2022)	ConvNeXt network, a deep convolutional neural network (CNN)	31 malignant and 28 benign / 3911 and 5245	Yes	87.17
JOONGYO LEE at al. 2023 Lee et al. (2023)	stacking model (SVM, RF,LR)	MRI between Jan'13 and Dec'17 were collected	Yes	78.4
Our Proposed	**LR, SVM,RF, DT, MLP & XGboost with DR**	**780 Images**	**Yes**	**88.72**

19

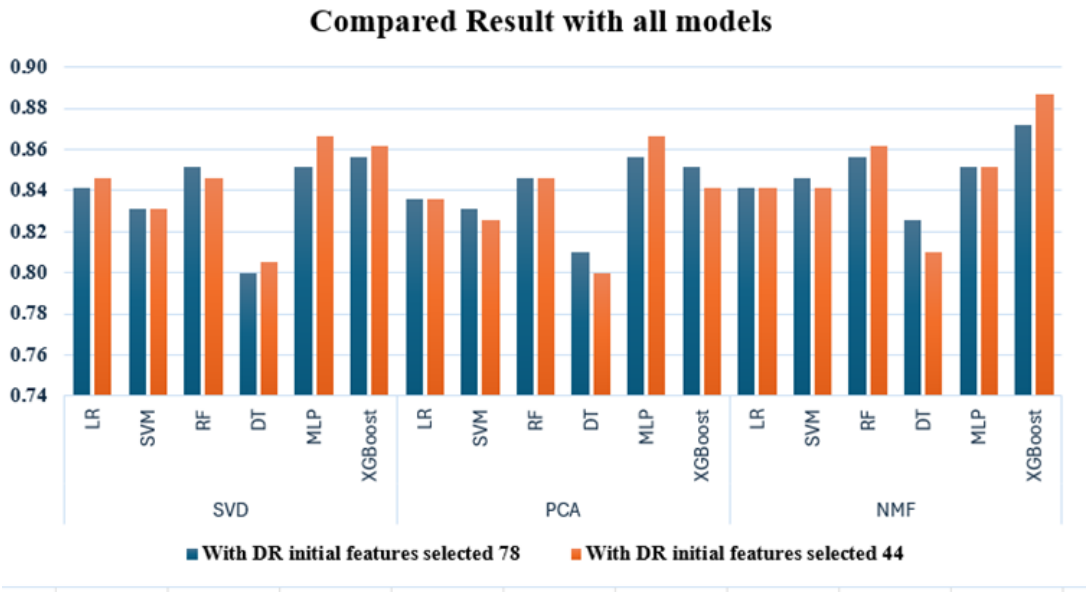

Figure 6: Accuracy of all models with two feature sets for NMF, SVD, and PCA

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, and frequency domain properties, contributes to a more comprehensive understanding of breast tissue properties. Additionally, the proposed approach embraces the power of artificial intelligence, 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.

20

6. Conclusion and Future scope

In this study, we explored the efficacy of employing dimensionality reduction techniques on radiomics features to enhance breast cancer diagnosis. Through rigorous experimentation and analysis, several key insights have been derived that contribute significantly to the field of breast cancer detection and diagnosis. our study underscores the significance of leveraging dimensionality reduction techniques in enhancing breast cancer diagnosis. The ability to distill intricate radiomics data into concise yet informative representations holds immense potential for improving diagnostic accuracy and aiding clinical decision-making in the realm of breast cancer detection. This study serves as a foundational step towards leveraging advanced data-driven methodologies to augment breast cancer diagnosis, paving the way for more effective, accurate, and personalized healthcare interventions in the field of oncology. These results open the door to more accurate and effective classification models, which will enable medical professionals to make better judgments and maybe enhance patient outcomes in the treatment of breast cancer. While this study presents promising outcomes, there are areas for further exploration. Future research could delve deeper into refining dimensionality reduction techniques specific to different imaging modalities. Moreover, exploring ensemble methods or hybrid approaches integrating multiple dimensionality reduction techniques may further amplify diagnostic accuracy.

Funding

The authors did not receive any funding.

Institutional Review Board Statement

Not applicable

Informed Consent Statement

Not applicable

Data Availability Statement

Available upon request

Conflicts of Interest

The authors declare no conflict of interest.

Ethics Approval and consent to participate

Not applicable

Consent for publication

Not applicable

References

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

Amkrane, Y., El Adoui, M., Benjelloun, M., 2020. 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. pp. 1–5.

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

Bahri, M., Bifet, A., Maniu, S., Gomes, H.M., 2021. Survey on feature transformation techniques for data streams, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 4796–4802.

Baratchi, M., . Supervisors: Nuno de mesquita césar de sá .

Barrios, C.H., 2022. Global challenges in breast cancer detection and treatment. The Breast 62, S3–S6.

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

Binsaif, N., et al., 2022. Application of machine learning models to the detection of breast cancer. Mobile Information Systems 2022.

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

Botlagunta, M., Botlagunta, M.D., Myneni, M.B., Lakshmi, D., Nayyar, A., Gullapalli, J.S., Shah, M.A., 2023. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Scientific Reports 13, 485.

Conti, A., Duggento, A., Indovina, I., Guerrisi, M., Toschi, N., 2021. Radiomics in breast cancer classification and prediction, in: Seminars in cancer biology, Elsevier. pp. 238–250.

Daimiel Naranjo, I., Gibbs, P., Reiner, J.S., Lo Gullo, R., Sooknanan, C., Thakur, S.B., Jochelson, M.S., Sevilimedu, V., Morris, E.A., Baltzer, P.A., et al., 2021. Radiomics and machine learning with multiparametric breast mri for improved diagnostic accuracy in breast cancer diagnosis. Diagnostics 11, 919.

Darveau, P., 2023. Support vector machines: Modeling the dual cognitive processes of an svm .

Fatima, S., Hussain, A., Amir, S.B., Ahmed, S.H., Aslam, S.M.H., 2023. Xgboost and random forest algorithms: An in depth analysis. Pakistan Journal of Scientific Research 3, 26–31.

Fu, X., Huang, K., Sidiropoulos, N.D., Ma, W.K., 2019. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. IEEE Signal Process. Mag. 36, 59–80.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M., Herrera, F., 2016. Big data preprocessing: methods and prospects. Big Data Analytics 1, 1–22.

Gupta, K., Janghel, R.R., 2019. Dimensionality reduction-based breast cancer classification using machine learning, in: Computational Intelligence: Theories, Applications and Future Directions-Volume I: ICCI-2017, Springer. pp. 133–146.

Hassanien, M.A., Singh, V.K., Puig, D., Abdel-Nasser, M., 2022. Predicting breast tumor malignancy using deep convnext radiomics and quality-based score pooling in ultrasound sequences. Diagnostics 12, 1053.

Kalman, D., 1996. A singularly valuable decomposition: the svd of a matrix. The college mathematics journal 27, 2–23.

Khan, R.A., Rashid, N., Shahzaib, M., Malik, U.F., Arif, A., Iqbal, J., Saleem, M., Khan, U.S., Tiwana, M., 2023. A novel framework for classification of two-class motor imagery eeg signals using logistic regression classification algorithm. Plos one 18, e0276133.

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S.A., Schabath, M.B., Forster, K., Aerts, H.J., Dekker, A., Fenstermacher, D., et al., 2012. Radiomics: the process and the challenges. Magnetic resonance imaging 30, 1234–1248.

Laajili, R., Said, M., Tagina, M., 2021. Application of radiomics features selection and classification algorithms for medical imaging decision: Mri radiomics breast cancer cases study. Informatics in Medicine Unlocked 27, 100801.

Lee, J., Yoo, S.K., Kim, K., Lee, B.M., Park, V.Y., Kim, J.S., Kim, Y.B., 2023. Machine learning-based radiomics models for prediction of locoregional recurrence in patients with breast cancer. Oncology Letters 26, 1–10.

Lenga, L., Bernatz, S., Martin, S.S., Booz, C., Solbach, C., Mulert-Ernst, R., Vogl, T.J., Leithner, D., 2021. Iodine map radiomics in breast cancer: prediction of metastatic status. Cancers 13, 2431.

Losurdo, L., Fanizzi, A., Basile, T.M.A., Bellotti, R., Bottigli, U., Dentamaro, R., Didonna, V., Lorusso, V., Massafra, R., Tamborra, P., et al., 2019. Radiomics analysis on contrast-enhanced spectral mammography images for breast cancer diagnosis: A pilot study. Entropy 21, 1110.

Ma, W., Zhao, Y., Ji, Y., Guo, X., Jian, X., Liu, P., Wu, S., 2019a. Breast cancer molecular subtype prediction by mammographic radiomic features. Academic radiology 26, 196–201.

Ma, W., Zhao, Y., Ji, Y., Guo, X., Jian, X., Liu, P., Wu, S., 2019b. Breast cancer molecular subtype prediction by mammographic radiomic features. Academic radiology 26, 196–201.

Mahmood, T., Li, J., Pei, Y., Akhtar, F., Imran, A., Yaqub, M., 2021. An automatic detection and localization of mammographic microcalcifications roi with multi-scale features using the radiomics analysis approach. Cancers 13, 5916.

Massafra, R., Bove, S., Lorusso, V., Biafora, A., Comes, M.C., Didonna, V., Diotaiuti, S., Fanizzi, A., Nardone, A., Nolasco, A., et al., 2021. Radiomic feature reduction approach to predict breast cancer by contrast-enhanced spectral mammography images. Diagnostics 11, 684.

Mayerhoefer, M.E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., Cook, G., 2020. Introduction to radiomics. Journal of Nuclear Medicine 61, 488–495.

Militello, C., Rundo, L., Dimarco, M., Orlando, A., Woitek, R., D'Angelo, I., Russo, G., Bartolotta, T.V., 2022. 3d dce-mri radiomic analysis for malignant lesion prediction in breast cancer patients. Academic Radiology 29, 830–840.

Naskath, J., Sivakamasundari, G., Begum, A.A.S., 2023. A study on different deep learning algorithms used in deep neural nets: Mlp som and dbn. Wireless Personal Communications 128, 2913–2936.

Panayides, A.S., Amini, A., Filipovic, N.D., Sharma, A., Tsaftaris, S.A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., et al., 2020. Ai in medical imaging informatics: current challenges and future directions. IEEE journal of biomedical and health informatics 24, 1837–1857.

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

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

Tomaszewski, M.R., Gillies, R.J., 2021. The biological meaning of radiomic features. Radiology 298, 505–516.

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

Verma, S.K., Arora, D., Bhardwaj, R., 2020. 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. pp. 169–171.

Wu, J., Hicks, C., 2021. Breast cancer type classification using machine learning. Journal of personalized medicine 11, 61.

Yarabarla, M.S., Ravi, L.K., Sivasangari, A., 2019. Breast cancer prediction via machine learning, in: 2019 3rd international conference on trends in electronics and informatics (ICOEI), IEEE. pp. 121–124.

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

Yu, Y., He, Z., Ouyang, J., Tan, Y., Chen, Y., Gu, Y., Mao, L., Ren, W., Wang, J., Lin, L., et al., 2021. Magnetic resonance imaging radiomics predicts preoperative axillary lymph node metastasis to support surgical decisions and is associated with tumor microenvironment in invasive breast cancer: A machine learning, multicenter study. EBioMedicine 69.

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

Zhou, J., Tan, H., Li, W., Liu, Z., Wu, Y., Bai, Y., Fu, F., Jia, X., Feng, A., Liu, H., et al., 2021. Radiomics signatures based on multiparametric mri

for the preoperative prediction of the her2 status of patients with breast cancer. Academic Radiology 28, 1352–1360.

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