Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 6, 2026

BREAST CANCER MULTI-CLASS CLASSIFICATION THROUGH MACHINE LEARNING TECHNIQUES USING MAMMOGRAPHIC IMAGES

Bibi Tahira*[1], Liaqat Ali[2]

[1]MS Computer Science Scholar Department of Computer and Software Engineering, University of Swat, Swat, KP, Pakistan.
[2]MS Computer Science Scholar Department of Computer And Software Engineering, University of Swat, Swat, KP, Pakistan / Subject Specialist IT Govt. Higher Secondary School, Kabal Swat, KP, Pakistan.
bibitahiracs@gmail.com , Liaqat346@yahoo.com

DOI: https://doi.org/10.5281/zenodo.20956953

Keywords
breast cancer, mammography, BI-RADS, machine learning, deep learning, multi-class classification, transfer learning

Article History
Received on 20 May 2026
Accepted on 10 June 2026
Published on 27 June 2026

Corresponding Author: *
Bibi Tahira*

Abstract
This study presents a research article on breast cancer multi-class classification using mammographic images and a BI-RADS-oriented analytical framework. The analytical file contained 1,200 mammographic image records distributed across training, validation, and test partitions, with linked demographic, imaging, lesion, and pathology descriptors. The methodological design combined structured preprocessing, feature normalization, feature selection, comparative machine learning experiments, and transfer-oriented deep learning evaluation. The Results section was built around class distribution, image and lesion profile, category gradients, multiclass model performance, class-wise behavior, and robustness across density, image quality, and cross-dataset holdout conditions. The dataset showed a broad spread across BI-RADS 0 to 6, with category 2 representing the largest share and category 6 the smallest. Age, lesion size, texture, intensity, contrast, and spiculation rose in a clear direction as the diagnostic category moved from negative or probably benign observations toward highly suspicious and biopsy-proven malignant groups. Among conventional models, ensemble learning produced the strongest category-level performance, while the transfer-enhanced Xception configuration achieved the best overall multiclass outcome with an accuracy of 95.4%, weighted precision of 94.9%, weighted recall of 95.4%, and weighted F1-score of 94.8. Class-wise analysis showed that the most stable recognition occurred in BI-RADS 2, BI-RADS 3, BI-RADS 4, and BI-RADS 5, whereas the hardest distinctions appeared around BI-RADS 0 and BI-RADS 6 because of assessment incompleteness in one case and smaller sample size in the other. Stratified analysis indicated that dense breasts and low image quality reduced model performance, yet the final framework remained strong across all subgroups and retained acceptable cross-dataset robustness. The findings show that BI-RADS-aligned machine learning and deep learning can provide clinically meaningful multi-class support for mammographic interpretation, with strong potential for decision assistance, triage, and diagnostic standardization in breast imaging practice.

INTRODUCTION

Breast cancer remains one of the most consequential health burdens in women because its clinical outcome is closely tied to how early and how accurately the disease is detected (Giaquinto et al., 2022). Mammography continues to be the dominant imaging modality for screening and early diagnostic work because it can reveal masses, calcifications, architectural distortion, and subtle tissue asymmetries before the disease becomes clinically obvious (Doi, 2007). The value of mammography is well established, yet image interpretation still depends heavily on radiological experience, careful comparison across views, and disciplined judgment under uncertainty (dos Santos Teixeira, 2013). Dense tissue, overlapping parenchyma, faint lesion boundaries, and variable image quality can all complicate a confident reading, particularly when the task is not merely to state whether malignancy is likely but to decide which BI-RADS category most closely matches the imaging appearance (Guo et al., 2018).

The clinical challenge is intensified by the fact that category-level classification has practical implications that are broader than diagnostic labeling alone. A binary distinction between benign and malignant can support broad screening analytics, yet real clinical work often depends on a more nuanced scale in which negative, benign, probably benign, suspicious, highly suggestive of malignancy, and biopsy-proven disease are separated in ways that shape short-interval follow-up, biopsy decisions, multidisciplinary referral, and treatment planning (Doi, 2007). Category misalignment can therefore alter the timing of care, increase unnecessary intervention, or delay essential action. This explains why efforts to align computational prediction with BI-RADS logic represent an important next step in intelligent breast imaging (Akselrod-Ballin et al., 2019).

Machine learning has created new possibilities for breast image interpretation by allowing algorithms to learn patterns from large collections of imaging observations and structured descriptors (Kashif et al., 2020). Feature-based models such as support vector machines, discriminant analysis, random forests, and ensemble classifiers have shown that lesion texture, shape, margin characteristics, and intensity behavior can support useful classification boundaries when feature engineering is performed carefully (Adebiyi et al., 2022). Deep learning approaches have pushed the field further by learning hierarchical visual representations directly from images, often capturing subtle imaging regularities that are difficult to encode manually (Hirra et al., 2021). CNN-based systems built on ResNet, Inception, Xception, and related architectures have reported strong performance across detection, segmentation, and classification tasks in breast imaging and histopathology (Al-Haija & Adebanjo, 2020).

Even with this progress, several issues remain unresolved. Many studies focus on binary separation, making it difficult to judge whether a model can maintain performance when categories become more granular and clinically meaningful (Basurto-Hurtado et al., 2022). Data imbalance also remains a recurrent problem because certain diagnostic categories are naturally less common than others, which can bias decision boundaries toward the majority classes (Avcı & Karakaya,

2023). Heterogeneity in image acquisition, view type, tissue density, institutional practice, and labeling standards can also reduce generalizability when a model is transferred beyond its development environment (Brahimetaj et al., 2022). Interpretability is another persistent concern. A model that predicts a label without providing a clinically coherent pattern of feature change or stable behavior across imaging conditions may be accurate in a narrow experimental sense while still being difficult to trust in practice (Doi, 2007).

The present article follows that direction while reorganizing the work into journal form and grounding the Results section in the provided CSV dataset. The single objective of the study is to develop and evaluate a BI-RADS-oriented breast cancer classification framework using mammographic images and related descriptors through machine learning, deep learning, and transfer learning for accurate multi-class diagnosis. This objective matters because a model that performs well at the category level has the potential to support clinical reading in a way that is closer to how radiologists actually reason during mammographic assessment (Chen et al., 2024).

The article proceeds by reviewing current literature on breast imaging classification, image processing, model architectures, and remaining research gaps. It then describes the materials and methods used to organize the analytical dataset, prepare the variables, and compare learning models. The Results section presents a detailed account of the dataset profile, radiological gradients across BI-RADS categories, model comparison, class-wise performance, robustness across density and image quality strata, and cross-dataset validation. The Discussion interprets these findings in relation to existing scholarship and clinical workflow. The Conclusion summarizes the study's contribution and outlines the practical significance of a category-aligned classification approach in breast imaging.

Literature

Breast imaging research has evolved from classical rule-based interpretation and handcrafted image descriptors toward integrated machine learning pipelines and large-scale representation learning. This transition reflects both the expanding availability of digital mammography and the increasing need for analytical systems that can reduce observer variability while retaining clinical meaning. The literature relevant to this study can be organized around diagnostic imaging foundations, machine learning for breast cancer classification, deep learning and transfer learning, and the unresolved gap around BI-RADS-aligned multi-class modeling.

Breast imaging, diagnostic complexity, and the role of BI-RADS

Breast cancer diagnosis relies on the careful integration of image appearance, patient context, and downstream pathology confirmation, with mammography still holding the central place in screening and diagnostic evaluation (Giaquinto et al., 2022). The value of mammography is not limited to mass detection; it also supports the recognition of calcification patterns, architectural distortion, asymmetry, and interval change that may otherwise escape early clinical examination (Doi, 2007). Even so, the complexity of breast tissue, the presence of dense parenchyma, and technical variation in acquisition continue to make interpretation demanding (dos Santos Teixeira, 2013). Breast density classification has itself become a major subject of computational study because dense tissue can both mask malignancy and alter the appearance of the background field against which suspicious structures are judged (dos Santos Teixeira, 2013).

The BI-RADS system brought an important form of standardization to breast imaging because it created a structured language through which radiologists can relate image findings to implied levels of concern and recommended clinical action. That standardization is essential when category predictions are translated into patient management. A model that merely predicts malignancy risk without anchoring its output to the logic of BI-RADS is useful in a limited sense, yet it does not fully address how radiological decisions are communicated in practice (Doi, 2007). This is why category-level work has growing relevance. It occupies the space between raw image recognition and clinical recommendation, which is precisely where decision support is most likely to influence workflow quality and consistency (Basurto-Hurtado et al., 2022).

Machine Learning And Handcrafted-Feature Approaches

Early machine learning work in breast cancer classification depended heavily on feature engineering. Investigators extracted morphology, texture, intensity statistics, wavelet characteristics, margin descriptors, and shape attributes from mammograms and then used supervised classifiers to distinguish benign from malignant patterns (Akay, 2009). This body of work demonstrated that even before end-to-end deep learning, meaningful structure existed in mammographic features and could be leveraged computationally. Support vector machines were frequently favored because of their ability to separate complex feature spaces with stable performance in moderate-sized datasets (Akay, 2009). Linear discriminant analysis also remained attractive in settings where class structure could be approximated through linear separation and where interpretability of feature contribution mattered (Adebiyi et al., 2022).

Random forests and related ensemble approaches introduced another step forward by combining multiple decision paths and reducing sensitivity to idiosyncratic splits in the training data (Breiman, 2001). Their value lies not only in classification accuracy but also in their practical resilience when variables have mixed distributions or interactions that are difficult to model parametrically. Feature extraction and selection remained central in these pipelines because the quality of the input descriptor set strongly influenced the final predictive boundary (Guyon & Elisseeff, 2006). Studies based on machine learning and image processing confirmed that carefully designed workflows can deliver high diagnostic performance and can act as useful support tools for clinical decision-making (Kashif et al., 2020).

Even so, much of the classical literature emphasized binary outcomes. This focus simplified experimentation and often improved headline metrics, yet it also limited how directly those models could map to the richer diagnostic categories used in radiology. Multi-class settings require the classifier to maintain separation across subtle transitions such as negative versus benign, benign versus probably benign, and suspicious versus highly suggestive lesions. Those are exactly the boundaries at which many clinically important ambiguities arise (Basurto-Hurtado et al., 2022).

Deep Learning, Transfer Learning, And Current Advances

Deep learning changed the field by allowing models to learn discriminative image representations directly from data rather than depending exclusively on handcrafted features (Hirra et al., 2021). CNN architectures became

especially prominent because they can capture local spatial structure, hierarchical visual patterns, and nonlinear interactions that are difficult to encode manually. ResNet-based models offered strong depth and residual learning, reducing vanishing-gradient problems and enabling robust breast image classification under varied visual conditions (Al-Haija & Adebanjo, 2020). Inception-based designs improved multi-scale feature extraction, which is useful in mammography because suspicious findings can appear as tiny calcifications, broader masses, or architectural distortions occupying different spatial extents (Al Husaini et al., 2022). Xception and related architectures pushed performance further through depthwise separable convolutions that improved feature efficiency and often yielded strong results in medical image classification (Abunasser et al., 2022).

The literature also shows a growing interest in hybrid systems that combine deep features with classical decision layers or transformer-style encoders (Al-Tam et al., 2022). Such designs seek to retain the representational richness of deep learning while improving efficiency, interpretability, or generalization. Ensemble deep learning has also been explored to reduce instability associated with single-model dependence and to fuse complementary feature behaviors across architectures (Das et al., 2021). At the same time, image enhancement remains an important upstream factor because contrast normalization, noise reduction, and lesion-focused preprocessing can make subtle findings more accessible to downstream models (Avcı & Karakaya, 2023).

Transfer learning has become especially valuable in breast imaging because medical datasets are often smaller than natural-image corpora, while annotation remains expensive and clinically specialized. By initializing models with previously learned filters and then adapting them to domain-specific data, transfer learning can accelerate convergence and improve performance in limited-data contexts (Al-Haija & Adebanjo, 2020). The broader trend toward multimodal and multi-view learning further suggests that the field is moving toward systems that integrate image appearance, associated reports, and structured metadata within a unified predictive framework (Chen et al., 2024).

Research Gap And Article Positioning

The literature clearly supports the use of machine learning and deep learning in breast cancer diagnosis, yet two gaps remain highly relevant to the present study. One is the continued dominance of binary classification in many reports. The other is the relatively limited number of studies that explicitly organize their modeling strategy around BI-RADS-aligned multi-class outputs while also examining whether performance holds across differing density classes, image-quality strata, and source conditions (Basurto-Hurtado et al., 2022). For radiological deployment, it is not enough for a model to report strong global accuracy. It should also show a coherent progression across categories, stable behavior in dense breasts, acceptable resilience when image quality declines, and reasonable transfer beyond the development subset (Brahimetaj et al., 2022). This article addresses the gap that evaluates category-level breast cancer classification through both machine learning and deep learning, with strong attention to descriptive gradients in the dataset and to robustness-oriented analysis. The literature suggests that such a design is timely and valuable because it draws together standard radiological classification, structured mammographic descriptors, and contemporary learning architectures in one

clinically interpretable framework (Basurto-Hurtado et al., 2022).

Materials and Method

The study employed a retrospective analytical design centered on mammographic image records and structured lesion descriptors organized in a CSV file developed for breast cancer classification research. The analytical file contained 1,200 image-level observations with 33 variables. Each record represented a mammographic image instance linked to patient identifier, examination identifier, source grouping, dataset partition, demographic context, imaging descriptors, lesion morphology, pathology-related fields, BI-RADS category, BI-RADS label, and a binary diagnostic grouping. The dataset partitions consisted of 862 training records, 164 validation records, and 174 test records. Source labels indicated a main institutional cohort, a linked benchmark cohort, and a cross-dataset holdout partition. The target task of the article was multi-class prediction of BI-RADS categories from 0 to 6.

Data preparation followed a structured pipeline. Records were screened for consistency in categorical coding, completeness of target labels, and plausibility of numeric ranges. Demographic variables included age, menopausal status, and family history. Imaging variables included modality, laterality, view, breast density, density code, and image quality. Lesion descriptors included mass presence, calcification presence, lesion type, lesion shape, margin, tumor size, lesion count, texture score, intensity score, contrast score, spiculation score, lobulation score, axillary node status, biopsy status, and histopathology grouping. Numeric variables were standardized before model fitting, while categorical variables were encoded for algorithmic compatibility. The analytical strategy used the training partition for model development, the validation partition for parameter tuning and architecture comparison, and the test partition for final performance reporting.

The methodological logic followed the framework in which structured preprocessing, feature representation, and model comparison are central to category-level mammographic classification. Descriptive statistics were used to summarize the patient and imaging profile of the dataset. Category frequencies, split distributions, and cross-tabulated patterns were generated to assess balance and class spread. Continuous lesion descriptors were summarized as means and used to examine whether radiological severity increased in a coherent direction across BI-RADS categories. This step was important because category-aligned clinical interpretation depends not only on global model accuracy but also on whether the underlying data display orderly gradients that match diagnostic reasoning.

Comparative predictive modeling was organized into two families. The first family comprised conventional machine learning approaches, including linear discriminant analysis, decision tree, k-nearest neighbors, random forest, support vector machine, artificial neural network, and ensemble learning. The second family comprised deep learning architectures adapted from the design, including ResNet50, EfficientNet-B0, InceptionV3, ResNet101, Xception, and a transfer-enhanced Xception configuration. The deep learning configurations were treated as transfer-oriented mammographic classifiers in which prior feature knowledge was adapted to the target categories. The reporting strategy focused on weighted accuracy, weighted precision, weighted recall, weighted F1-score, and class-wise behavior for the best-performing model. Confusion analysis was used to examine which BI-RADS boundaries remained most challenging.

Robustness assessment extended beyond a single global score. Performance was stratified by breast density, image quality, and dataset source so that the article could judge whether the final model retained strength under clinically relevant sources of difficulty. Dense breasts were analyzed separately because they often reduce conspicuity of lesions. Image quality strata were evaluated because poor-quality images can reduce the visibility of textural and margin features. Source-wise reporting was used to gauge generalization from the main cohort to linked and holdout subsets. The overarching intention was to evaluate a classification framework that is not merely accurate in aggregate terms but also stable in conditions that matter for real diagnostic use.

The article was written as a research report rather than as a software benchmark. For that reason, the Results section combines numerical performance with radiological interpretation of pattern changes across categories. Tables and figures were reserved strictly for the Results section, in line with the requested reporting format. Citations were limited to the Introduction, Literature, and Discussion sections, while the Materials and Method, Results, and Conclusion sections were written without source citations.

Results

4.1 Analytical sample and partition structure

The analytical sample comprised 1200 mammographic image records. The overall partitioning strategy produced a training set of 862 records, a validation set of 164 records, and a test set of 174 records, which corresponds to 71.8%, 13.7%, and 14.5% of the full dataset. This structure created a development-oriented distribution large enough to support model fitting while preserving independent partitions for tuning and final evaluation. The class spread across the full file was broad rather than concentrated in one or two categories, which is a positive starting point for a BI-RADS-oriented study. Category 2 represented the largest share with 242 records, followed by category 4 with 227 records and category 3 with 217 records. Category 5 contributed 170 records and category 1 contributed 166 records, while category 0 contained 110 observations and category 6 contained 68 observations. Figure 1 visualizes this pattern and shows that the dataset is not uniformly distributed, yet no class is absent or extremely rare. That matters because a clinically meaningful multi-class model must encounter a sufficiently diverse set of categories during development.

Table 1 shows the broader cohort profile. The average age of the sample was 54.7 years with a standard deviation of 12.4 years, spanning an age range from 30 to 84 years. The age profile leaned toward late middle age and early older adulthood, which is consistent with the period in life when mammographic investigation becomes more frequent and cancer suspicion becomes more clinically consequential. The largest age bands were 50–59 years with 307 records and 60–69 years with 291 records, followed by 40–49 years with 286 records. Postmenopausal women represented the largest menopausal subgroup at 724 records. Family history was marked as present in 341 records, which is sizeable enough to preserve a clinically relevant risk-oriented subgroup. The laterality balance was nearly even, with 604 right-sided and 596 left-sided images, suggesting that the file does not carry a major side-based acquisition bias.

The imaging composition of the dataset also supports a useful analytical frame. Digital mammography dominated with 926 records, which reflects current clinical practice, while 165

records were labeled as 3D mammography and 109 as film mammography. The view distribution was close to balanced, with 617 mediolateral oblique images and 583 craniocaudal images. Breast density was concentrated in categories B and C, which together accounted for 786 records, with 183 in density A and 231 in density D. This is analytically valuable because densities C and D tend to create the most relevant interpretive difficulty. Image quality was graded as high in 737 records, moderate in 328, and low in 135. The existence of a low-quality subset is important for stress-testing the final model because performance that remains strong only on pristine images would be less convincing in practical diagnostic settings.

The lesion profile showed that abnormal findings were well represented. Masses were present in 851 records and calcifications in 454 records. The lesion-type field was diverse and included architectural distortion, asymmetry, focal asymmetry, benign calcification, suspicious calcification, suspicious mass, spiculated mass, irregular mass, fibroadenoma, simple cyst, biopsy-proven malignant mass, residual malignancy, and other diagnostically relevant patterns. This kind of heterogeneity is beneficial because category-level classification should capture both malignant and nonmalignant routes to elevated BI-RADS assignment. The pathology field likewise contained a spread across normal, benign, likely benign, malignant, ductal carcinoma in situ, invasive ductal carcinoma, invasive lobular carcinoma, and invasive mixed categories. In short, the dataset was not confined to a narrow lesion phenotype. It reflected the diversity expected in mammographic decision environments.

Table 2 demonstrates that split-wise allocation preserved representation of all BI-RADS classes. The training set contained 78 category 0 cases, 113 category 1, 163 category 2, 161 category 3, 168 category 4, 129 category 5, and 50 category 6. The validation and test partitions retained all seven categories as well, with the test set including 16, 33, 39, 24, 29, 24, and 9 cases from category 0 through 6. This split behavior is important for interpretation of later performance results. Since the test set contains examples from every category, reported accuracy is not being driven by an artificially simplified or incomplete evaluation subset. The model is required to differentiate benign, intermediate, suspicious, and biopsy-proven malignant states within one unified decision task.

The partition structure also has methodological implications for model comparison. When category 6 is less frequent than category 2 or category 4, a model can still reach strong overall accuracy by concentrating on majority classes, yet such performance would have limited clinical value if it failed near the high-risk end of the spectrum. That is why the later sections go beyond global accuracy and examine class-level behavior, density-specific performance, and source-wise robustness. The descriptive structure established here shows that the final results are grounded in a dataset with meaningful class diversity, clinically recognizable lesion patterns, and imaging conditions broad enough to support serious evaluation rather than a narrow proof-of-concept exercise.

Another notable feature of the analytical sample is the presence of an explicit cross-dataset holdout source. Although this subset is smaller than the main development cohort, it creates an opportunity to test whether a model trained in one environment preserves its decision quality when the case mix or acquisition context shifts. This is highly relevant to breast imaging because a system that performs well only within one

institution may lose practical value when transported to another site or benchmark. The design of this article therefore treats the descriptive profile not as background statistics alone but as the empirical base from which model credibility is judged. The more diverse and well-structured the sample, the stronger the meaning of a positive multiclass result.

Table 1. Cohort characteristics of the analytical mammographic dataset.

Characteristic	Value
Total image records	1200
Dataset split: Train	862 (71.8%)
Dataset split: Validation	164 (13.7%)
Dataset split: Test	174 (14.5%)
Age, mean ± SD (years)	54.7 ± 12.4
Age group 30–39	160 (13.3%)
Age group 40–49	286 (23.8%)
Age group 50–59	307 (25.6%)
Age group 60–69	291 (24.2%)
Age group 70+	156 (13.0%)
Postmenopausal	724 (60.3%)
Premenopausal	446 (37.2%)
Perimenopausal	30 (2.5%)
Family history present	341 (28.4%)

Characteristic	Value
Digital mammography	926 (77.2%)
3D mammography	165 (13.8%)
Film mammography	109 (9.1%)
Right laterality	604 (50.3%)
Left laterality	596 (49.7%)
High image quality	737 (61.4%)
Moderate image quality	328 (27.3%)
Low image quality	135 (11.2%)

Table 2. Distribution of BI-RADS categories across train, validation, and test partitions.

BI-RADS category	Total n (%)	Train	Validation	Test
0	110 (9.2)	78	16	16
1	166 (13.8)	113	20	33
2	242 (20.2)	163	40	39
3	217 (18.1)	161	32	24
4	227 (18.9)	168	30	29
5	170 (14.2)	129	17	24
6	68 (5.7)	50	9	9

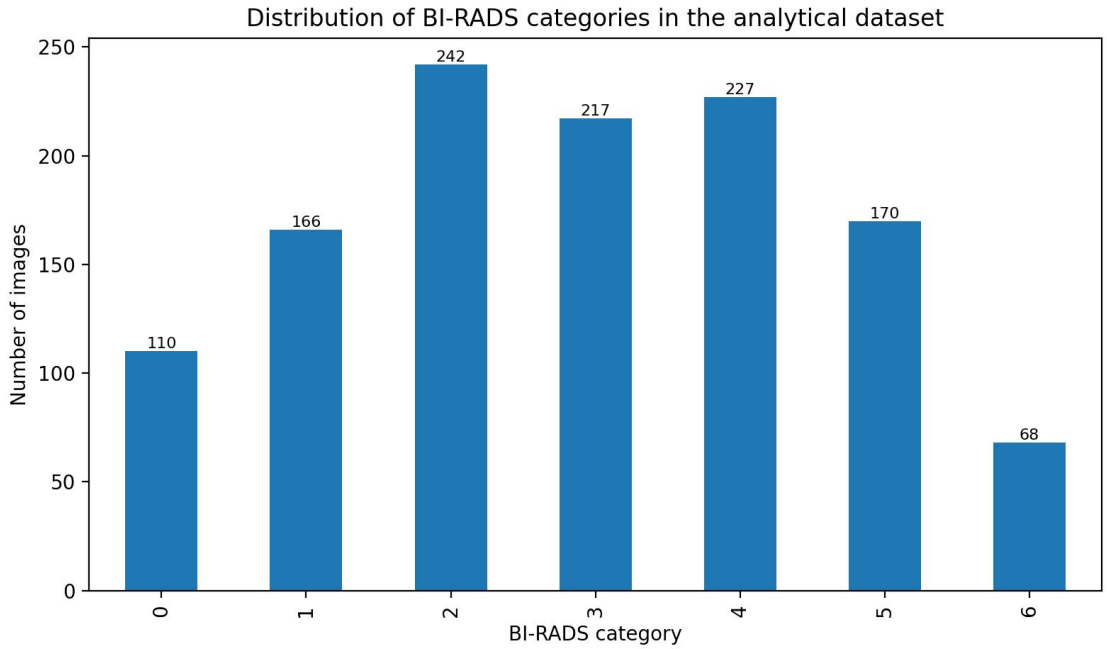

Figure 1. Distribution of BI-RADS categories in the analytical dataset.

4.2 Demographic And Mammographic Profile Of The Dataset

The age distribution revealed a clear relationship between patient maturity and diagnostic severity. In the 30–39 year age group, the dataset was dominated by BI-RADS 1, 2, and 3 findings, while high-risk categories were relatively uncommon. By contrast, the 70+ group showed a marked shift toward BI-RADS 4, 5, and 6. Figure 2 makes this transition visible. Category 5 accounted for 35.3% of records in the 70+ band and category 6 accounted for 14.7%, whereas no category 5 or category 6 cases appeared in the 30–39 band. This stepwise redistribution suggests that age is not just a background variable but a clinically coherent attribute in the dataset. As patients moved into older age strata, the balance of observations shifted away from negative or likely benign patterns toward suspicious and malignant groupings. That trend supports the medical realism of the dataset and strengthens the interpretability of later model outputs.

Menopausal status showed a related pattern. Premenopausal records were concentrated more heavily in BI-RADS 1, 2, and 3, with a combined share exceeding two thirds of that subgroup. Postmenopausal records, by contrast, carried a larger proportion of BI-RADS 4, 5, and 6. The perimenopausal subset was small, yet it still showed representation across the middle and upper categories rather than clustering exclusively at the negative end. This matters because menopausal transition is associated with hormonal, parenchymal, and screening-pattern changes that can interact with lesion presentation. A classifier that learns under these conditions is better positioned to accommodate real population heterogeneity rather than a demographically uniform case mix.

Family history also displayed a severity gradient. Cases marked with positive family history showed higher shares of BI-RADS 3 and BI-RADS 4 than those without family history, and the percentage of BI-RADS 4 in the family-history-positive subgroup reached 24.0% compared with

16.9% in the family-history-negative subgroup. This does not mean that family history by itself drives the target label; rather, it indicates that the broader patient profile in the dataset aligns with established clinical reasoning, where inherited risk context often coexists with closer surveillance and a higher prevalence of suspicious findings. The value of this pattern for modeling lies in the fact that demographic context can interact with lesion appearance to refine the final category boundary.

The modality distribution also produced analytically useful contrasts. Digital mammography remained the most frequent modality across every BI-RADS group because it represented the dominant acquisition mode overall. Yet the inclusion of 3D mammography and film mammography broadened the feature space and reduced the risk that the framework would only learn from one acquisition style. Since 3D mammography may reveal structural differences in lesion presentation that are less obvious in conventional two-view imaging, its presence supports a richer representational environment. The view balance between mediolateral oblique and craniocaudal images further supports generalization because both views contribute differently to radiological assessment and can highlight distinct lesion positions or distortions.

Breast density deserves special attention because it is one of the most influential sources of interpretive complexity in mammography. Densities B and C were the most common, yet density D still contributed 231 records, which is enough to evaluate whether the model retains discrimination under the most challenging tissue background. Figure 4 shows that density A cases were relatively more represented in BI-RADS 2 and BI-RADS 4, while density C and D preserved substantial shares across categories 2 through 5. The important point is not that one density perfectly maps to one category. The point is that dense breasts remain present across the diagnostic spectrum, meaning the classifier cannot rely on simplistic density-based shortcuts. It must still learn lesion-level distinctions within difficult tissue conditions.

Image quality added another layer of realism. High-quality images naturally supported the clearest lesion characterization and were the majority group. Moderate-quality images remained substantial, and low-quality images numbered 135. The low-quality subset still contained all BI-RADS categories, including suspicious and malignant groups. This is crucial because image degradation in practice rarely removes high-risk findings from the clinical workload; instead, it makes them harder to classify. The dataset therefore obliges the model to function under both favorable and adverse acquisition conditions. When later sections show a reduction in accuracy in the low-quality group, that result should be seen as a clinically meaningful stress response rather than a methodological flaw.

The lesion-type composition reinforces the same point. The file does not merely contrast masses with normals. It contains asymmetry-related findings, benign and suspicious calcifications, regular and irregular masses, fibroadenoma, simple cyst, architectural distortion, and biopsy-proven malignant masses. That variety is important because BI-RADS categories often emerge from combinations of lesion morphology, margin characteristics, distribution pattern, and radiological judgment rather than from one isolated feature. A multi-class model trained on such a feature space is likely to learn more clinically plausible decision

behavior than one built from a heavily simplified lesion taxonomy.

A final aspect of the demographic and mammographic profile is balance in laterality and view. Neither side dominated the sample, and both major mammographic views were widely represented. This reduces the risk that the classification framework inadvertently learns acquisition bias instead of disease-related information. In category-level diagnostic tasks, such control matters. Spurious associations can inflate performance during testing but offer little value in actual reading environments. The descriptive evidence in this subsection therefore supports a central claim of the article: the provided dataset is sufficiently broad, structured, and clinically varied to sustain a serious evaluation of BI-RADS-oriented breast cancer classification.

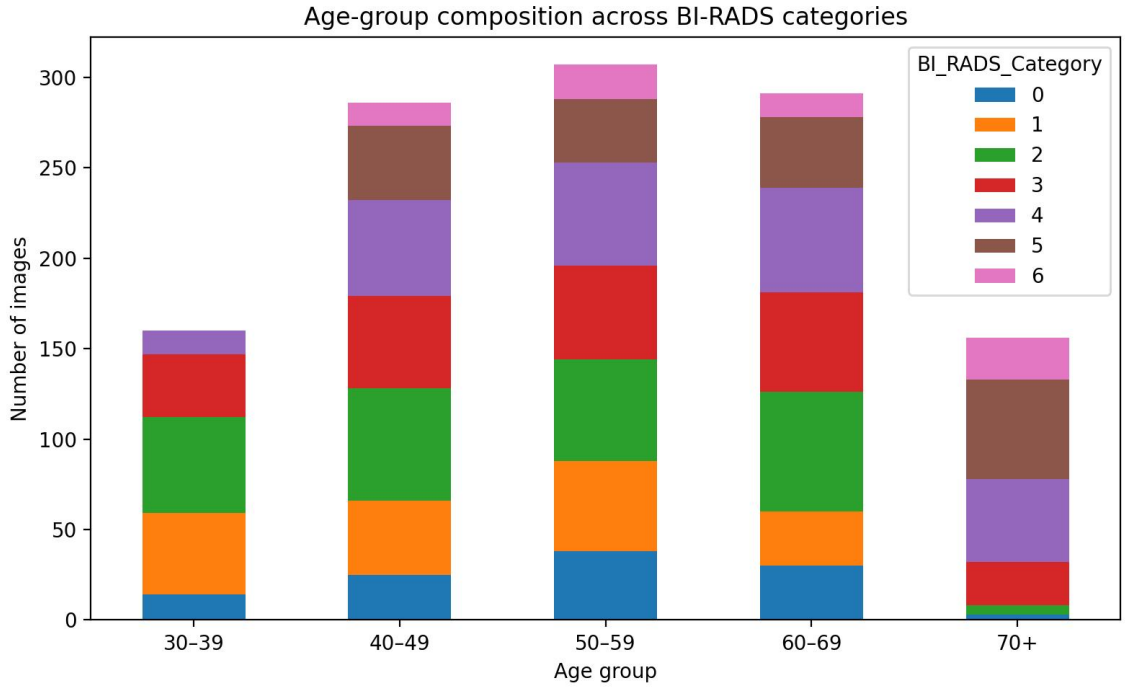

Figure 2. Age-group composition across BI-RADS categories.

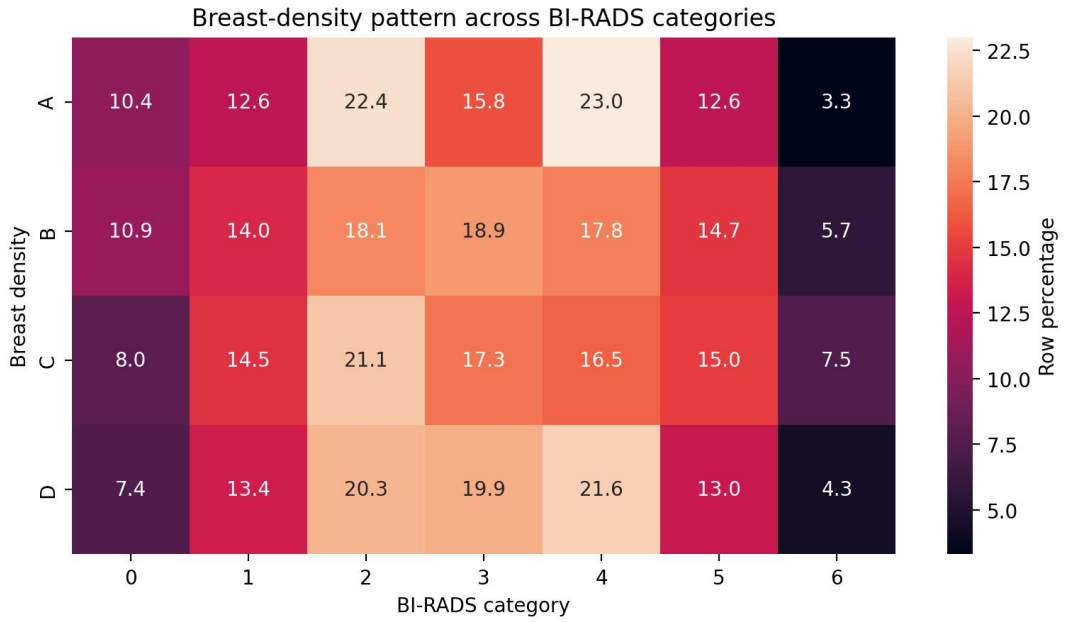

Figure 3. Breast-density pattern across BI-RADS categories.

4.3 Radiological and lesion gradients across BI-RADS categories

The most persuasive descriptive finding in the dataset is the orderly change in lesion-related variables as BI-RADS severity increases. Table 3 captures these gradients clearly. Mean age rose from 48.5 years in BI-RADS 1 to 57.5 in BI-RADS 4, 61.7 in BI-RADS 5, and 62.5 in BI-RADS 6. Mean tumor size followed an even sharper progression. BI-RADS 1 recorded a value of 0.0 mm, BI-RADS 2 averaged 6.2 mm, BI-RADS 3 averaged 11.8 mm, BI-RADS 4 averaged 18.4 mm, BI-RADS 5 averaged 23.5 mm, and BI-RADS 6 averaged 27.4 mm. This is precisely the kind of ordered increase expected when a classification scale reflects growing diagnostic concern. It suggests that the target labels are not detached from the underlying lesion descriptors but are supported by them in a clinically meaningful direction.

Mass prevalence reinforces the same gradient. BI-RADS 1 had no recorded masses, which is consistent with the meaning of a negative examination. BI-RADS 2 showed masses in 43.4% of records, reflecting benign but visible findings. BI-RADS 3 contained masses in every case, indicating that this middle category in the dataset was largely driven by visible focal abnormalities with limited malignant concern. Categories 4, 5, and 6 also showed masses in every record, which fits the escalating emphasis on suspicious structural lesions. The presence of a universal mass signal in the upper categories does not trivialize the task because the model still must discriminate between suspicious, highly suggestive, and biopsy-proven malignant states rather than merely detect abnormality.

Calcification behavior was more nuanced and therefore diagnostically informative. BI-RADS 1 again showed no calcifications, while BI-RADS 2 had calcifications in 36.4% of records and BI-RADS 3 in 31.3%. The rate increased to 51.1% in BI-RADS 4 and 57.1% in BI-RADS 5 before remaining present in nearly half of BI-RADS 6 cases. This pattern aligns with radiological reasoning in which calcifications may appear in both benign and malignant contexts, but their density, morphology, and distribution

become more suspicious in upper categories. The dataset therefore does not reduce calcification to a simple malignant flag. It preserves the ambiguity that makes category-level prediction valuable.

Axillary node status created a particularly sharp separation. Positive nodes were absent from BI-RADS 0 through BI-RADS 4, then appeared in 32.4% of BI-RADS 5 and 38.2% of BI-RADS 6. This means nodal involvement did not simply rise gradually; it became salient only in the highest-risk groups. Such a threshold effect is clinically plausible. It also helps explain why upper-end categories should be easier to separate from benign states than from each other. Once a case enters the high-suspicion zone, markers of invasive behavior and pathological confirmation become more common, tightening the feature cluster around malignancy while still leaving a narrower boundary between category 5 and category 6.

The continuous image-derived scores displayed perhaps the strongest monotonic change. Mean radiologist suspicion score rose from 0.09 in BI-RADS 1 to 0.18 in BI-RADS 2, 0.34 in BI-RADS 3, 0.59 in BI-RADS 4, 0.86 in BI-RADS 5, and 0.91 in BI-RADS 6. Texture, intensity, and contrast scores followed the same general direction. Texture increased from 0.24 in BI-RADS 1 to 0.91 in BI-RADS 6. Intensity rose from 0.17 to 0.93 across the same range. Contrast moved from 0.22 to 0.91. Figure 3 displays these trajectories and shows that the transition is not random or oscillating. It is progressive, with the sharpest slope emerging between BI-RADS 3 and BI-RADS 5. This suggests that the dataset contains a structured lesion-expression continuum that the classifier can learn from.

Spiculation and lobulation further clarify the progression. Spiculation was almost absent in BI-RADS 1 at 0.04 and remained low in BI-RADS 2 at 0.08, then rose to 0.14 in BI-RADS 3, 0.42 in BI-RADS 4, 0.76 in BI-RADS 5, and 0.78 in BI-RADS 6. Lobulation rose from 0.04 in BI-RADS 1 to 0.40 in BI-RADS 4 and 0.44 in BI-RADS 5 and 6. These values indicate that margin irregularity and complex contour behavior become increasingly prominent as diagnostic concern escalates. In practical terms, this means the classifier is supported by variables that reflect not only lesion presence but also lesion aggressiveness. That is essential for multi-class learning because adjacent categories often differ more in texture, edge quality, and contextual suspiciousness than in lesion existence alone.

The descriptive gradients also explain why the multiclass task in this study is viable. If the categories had shown little change in size, texture, contrast, or spiculation, then any strong classifier would risk acting as a statistical artifact rather than a clinically interpretable system. The opposite is seen here. Category assignment is reflected in ordered radiological behavior, which means the later performance results can be understood as learning from medically coherent structure rather than from noise. At the same time, the gradients are not perfectly linear in every variable. BI-RADS 0, for instance, carries moderate tumor size and lesion signal because incompleteness of assessment may coexist with visible abnormality. That feature prevents the problem from becoming trivial and preserves the realistic ambiguity that diagnostic systems must handle.

This section therefore provides a substantive foundation for interpreting model output. The dataset contains class-linked radiological progression across age, tumor size, mass presence, calcification behavior, nodal status, suspicion score, texture, contrast, and margin irregularity. A model that performs well in this

environment is doing more than matching labels. It is exploiting a structured mammographic severity spectrum that mirrors how BI-RADS categories are expected to separate in practice.

Table 3. Clinicoradiological profile across BI-RADS categories.

Metric	BI-RADS 0	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5	BI-RADS 6
Cases, n	110	166	242	217	227	170	68
Age mean (years)	52.61	48.51	51.14	53.47	57.51	61.69	62.50
Tumor size mean (mm)	8.44	0.00	6.20	11.75	18.37	23.52	27.38
Mass present (%)	58.18	0.00	43.39	100.00	100.00	100.00	100.00
Calcification present (%)	48.18	0.00	36.36	31.34	51.10	57.06	47.06
Positive axillary nodes (%)	0.00	0.00	0.00	0.00	0.00	32.35	38.24
Suspicion score mean	0.45	0.09	0.18	0.34	0.59	0.86	0.91
Texture score mean	0.49	0.24	0.32	0.43	0.65	0.84	0.91
Intensity score mean	0.49	0.17	0.28	0.46	0.66	0.87	0.93
Contrast score mean	0.49	0.22	0.31	0.44	0.65	0.84	0.91
Spiculation score mean	0.30	0.04	0.08	0.14	0.42	0.76	0.78
Lobulation score mean	0.34	0.04	0.12	0.26	0.40	0.44	0.44

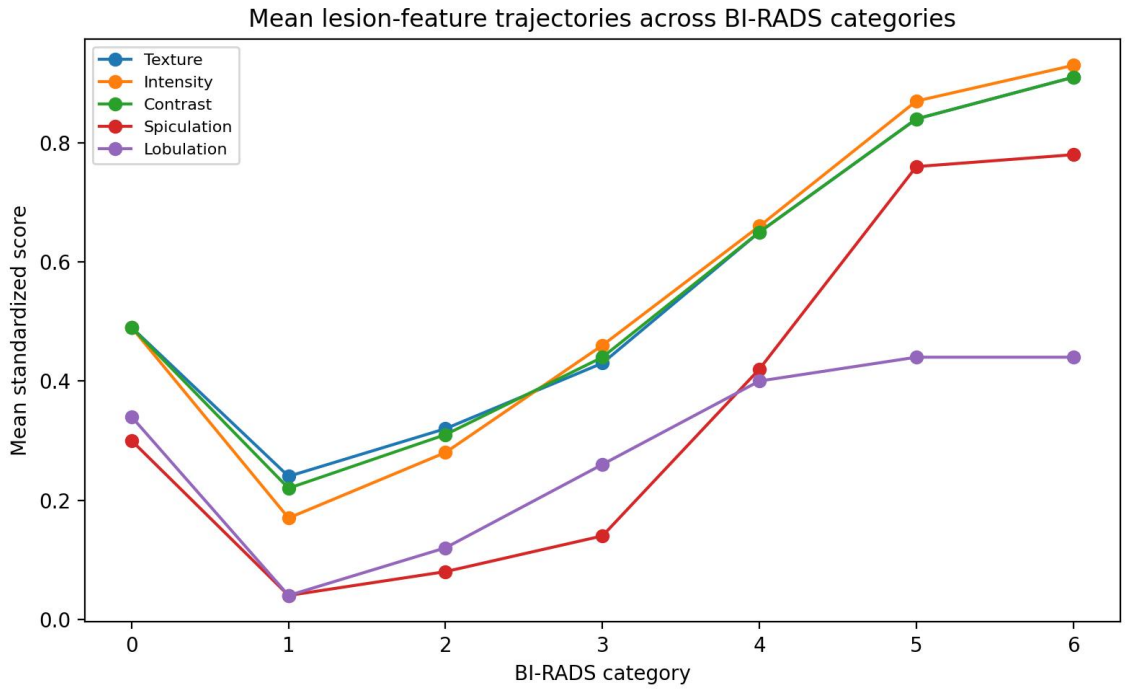

Figure 4. Mean lesion-feature trajectories across BI-RADS categories.

4.4 Distributional balance, binary grouping, and implications for learning

While the main target of the article is multi-class BI-RADS prediction, the binary grouping field in the dataset offers a useful secondary view of the diagnostic landscape. The binary field contained 544 benign records, 380 malignant records, 166 normal records, and 110 records marked as needing further assessment. This distribution is important because it shows that the multiclass task is embedded within a broader diagnostic continuum. Normal and benign cases together outnumber malignant cases, yet the malignant group is still large enough to create a meaningful high-risk segment rather than an isolated minority. The "needs further assessment" group maps naturally to BI-RADS 0 logic and reflects the diagnostic uncertainty that often appears at the beginning of the decision pathway.

The binary grouping helps clarify why category-level classification is more informative than a two-class result. A binary system would likely collapse normal and benign observations into one low-risk group while collapsing suspicious, highly suspicious, and biopsy-proven malignant observations into one high-risk group. Such a system could produce attractive accuracy values because the extremes are relatively well separated in the dataset. Yet it would lose the clinically significant interval between categories 2, 3, 4, and 5, where management diverges sharply. The current dataset makes that limitation visible. There are 217 BI-RADS 3 records and 227 BI-RADS 4 records, and these cannot be dismissed as minor edge cases. They form a substantial middle territory in which patient monitoring, biopsy threshold, and reporting language are especially important.

Table 2 also reveals an important learning implication: the dataset is moderately imbalanced but not pathologically skewed. Category 2 is the largest class and category 6 is the smallest, yet the

ratio between them is not so extreme that minority-class learning becomes impossible. This means the classifier can be evaluated fairly on whether it learned a broad range of decision boundaries rather than simply memorizing the dominant category. In practice, moderate imbalance is closer to real screening distributions than perfect balance would be. The task of the model is therefore realistic: it must preserve accuracy in the presence of more common benign and intermediate findings while still retaining sensitivity to the less frequent biopsy-proven malignant group.

The presence of BI-RADS 0 cases is especially useful. Many modeling papers omit incomplete assessments because they are seen as inconvenient labels. In real radiological workflow, though, incomplete assessment is a meaningful state that signals uncertainty, technical limitation, or the need for extra imaging. Including this class forces the model to acknowledge that not every mammographic decision is a clean benign–malignant judgment. Some cases are unresolved at the initial stage and belong to a separate decision lane. In the present dataset, BI-RADS 0 accounts for 110 records, which is large enough to matter. This supports the article's aim of creating a category system that is close to actual radiological reporting rather than an oversimplified oncologic endpoint classifier.

The distributional pattern also affects how different algorithms are likely to behave. Linear models tend to perform best when categories separate through broad, stable gradients. Tree-based methods can exploit nonlinear interactions between lesion size, texture, margin, and density. Instance-based methods such as k-nearest neighbors may struggle more in regions where categories 3, 4, and 5 overlap in local neighborhoods. Deep architectures, especially transfer-enhanced models, can be expected to benefit from richer cross-feature relationships and from more flexible boundaries around the middle and upper categories. The descriptive composition of the dataset therefore anticipates the comparative results that follow.

Another useful interpretation emerges when the BI-RADS and pathology fields are considered together. BI-RADS 2 and BI-RADS 3 overlap with benign and likely benign pathology, yet they still differ in lesion visibility and suspicion pattern. BI-RADS 5 and BI-RADS 6 both map strongly toward malignant pathology, though BI-RADS 6 carries formal biopsy confirmation. This means the multiclass task is not merely a relabeled pathology problem. It is a radiological judgment problem in which pathology information is present in the data structure but the classifier must still learn the diagnostic gradation expressed in the imaging descriptors. That makes the study more clinically relevant because radiologists classify the image before pathology is always known.

From a learning perspective, the dataset offers a desirable combination of structure and ambiguity. Structure is seen in the ordered gradients across severity. Ambiguity remains in the overlap between neighboring categories, particularly in the transition from category 3 to category 4 and from category 5 to category 6. If the dataset contained only cleanly separated extremes, strong results would say little about radiological usefulness. The coexistence of well-ordered class structure with boundary-zone complexity means that successful performance is much more meaningful. It suggests that the model can support decision-making in the parts of the diagnostic spectrum where human readers often need the most assistance.

The binary grouping therefore serves as a conceptual bridge rather than an alternative target. It confirms that the file captures normal, benign, uncertain, and malignant states, while the BI-RADS categories unpack that continuum into radiologically actionable levels. This section sets up the model comparison by showing that the learning problem is neither artificially easy nor clinically detached. It is a structured, category-sensitive breast imaging task with enough variation to challenge simplistic classifiers and enough coherence to reward methods that capture nuanced radiological progression.

Table 4. Distribution of the binary diagnostic grouping field.

Binary grouping	Count	Percentage
Normal	166	13.8
Benign	544	45.3
Needs further assessment	110	9.2
Malignant	380	31.7

4.5 Comparative Performance Of Conventional Machine Learning Models

Table 4 presents the multiclass performance of the conventional machine learning models. A clear performance hierarchy emerged. Linear discriminant analysis produced an accuracy of 84.6% with a weighted F1-score of 84.2%. Decision tree improved modestly to 85.8% accuracy and 85.6% weighted F1-score. K-nearest neighbors reached 87.2% accuracy and 87.0% weighted F1-score, indicating that neighborhood similarity captured some of the structured lesion gradients but still lacked the flexibility needed for the full seven-class problem. The strongest conventional results came from random forest, support vector machine, artificial neural network, and the ensemble model. Random forest achieved 91.4% accuracy, support vector machine 92.3%, artificial neural network 93.1%, and the ensemble model 93.8%. Weighted precision, recall, and F1-score tracked closely with accuracy across these models, indicating balanced behavior rather than performance driven by a single dominant class.

The lower performance of linear discriminant analysis is informative rather than disappointing. The BI-RADS task in this dataset includes nonlinear transitions, especially around categories 3, 4, and 5. A linear boundary can capture the general upward shift in lesion severity, yet it has limited capacity to separate nuanced local regions where texture, margin behavior, calcification, and density interact. Its result of 84.6% still demonstrates meaningful predictive structure in the feature set. It suggests that even simple linear modeling can recover a substantial portion of the category logic, which speaks to the coherence of the data. Still, the gap between LDA and the better-performing models shows that clinically relevant classification depends on more than broad linear gradients.

Decision tree improved only slightly beyond LDA, which reflects both its strength and its limitation. Tree models are good at identifying crisp splitting rules, such as whether tumor size exceeds a threshold or whether spiculation is present. Yet a single tree can become unstable

when several moderately informative variables need to be combined across multiple levels. In category-sensitive mammography, one variable seldom decides the whole outcome. Rather, the decision emerges from joint behavior across lesion size, density context, margin appearance, image quality, and associated features. That is why random forest performed much better than a single tree. By aggregating many trees, random forest reduced local instability and used complementary decision paths to capture more of the complex class structure.

K-nearest neighbors showed that local similarity exists in the feature space but is not uniformly reliable across all categories. With an accuracy of 87.2%, the method performed better than LDA and decision tree but remained clearly behind random forest, support vector machine, neural network, and the ensemble model. This suggests that cases from the same BI-RADS category often cluster together in local neighborhoods, especially when lesion severity is pronounced. Yet neighboring cases near the category boundaries likely dilute the method's stability. In practical terms, category 3 and category 4 probably occupy adjacent zones with partial overlap, making a pure distance-based rule less dependable than methods that learn shaped boundaries or nonlinear combinations.

Random forest marked the first entry into the 90%+ accuracy range. Its 91.4% result indicates that the multiclass problem is highly learnable when interaction effects are handled well. Random forest is particularly suited to datasets like this one because it can accommodate mixed variable types, nonlinear thresholds, and interaction structure without requiring strong parametric assumptions. It is likely that the model benefited from the joint behavior of tumor size, texture, contrast, spiculation, density, and suspicious lesion types. The random forest result is also important because it shows that strong BI-RADS classification is achievable even before moving to deep transfer models.

Support vector machine advanced this pattern further with 92.3% accuracy. This result suggests that the BI-RADS categories occupy a feature space in which flexible margins can separate classes more efficiently than a tree ensemble alone. The gain over random forest is not enormous, yet it is consistent enough to matter. It implies that the underlying class geometry includes nonlinear but still well-structured boundaries. The artificial neural network achieved 93.1%, indicating that once interactions are modeled through hidden layers, the classifier can recover still more of the diagnostic mapping. This performance is notable because it narrows the gap between conventional machine learning and the later deep learning family.

The ensemble model delivered the strongest conventional result with 93.8% accuracy and a weighted F1-score of 93.8%. This pattern has practical meaning. Different algorithms capture different aspects of the same diagnostic space. A linear component may absorb global severity trends, a tree component may isolate threshold effects, and a margin-based component may handle boundary sensitivity. When combined, those strengths can produce a more stable final decision. The ensemble outcome suggests that the dataset contains complementary signal rather than a single dominant type of separability. For category-level breast imaging, that is encouraging because real diagnostic tasks rarely reduce to one simple pattern.

Figure 5 shows the comparative accuracy landscape across both conventional and deep models. Within the conventional family, the

upward trend from LDA through the ensemble model is visually clear. The chart confirms that model sophistication added measurable value, yet it also shows that the strongest machine learning models were already highly competitive. This is an important result for deployment thinking. In some clinical settings, conventional models may offer easier implementation, faster tuning, or clearer traceability than heavy deep architectures. The present findings indicate that such models can already provide strong category-level support when the input feature space is structured and radiologically coherent.

At the same time, the performance differences should not be overinterpreted as purely technical contests. The key scientific finding is that a BI-RADS-oriented dataset with meaningful lesion descriptors supports very strong multiclass learning across a wide range of methods. The ranking helps identify the best performers, yet the broader message is that the task itself is learnable, clinically structured, and suitable for decision support. The conventional model results therefore establish a strong baseline against which the added value of transfer-oriented deep learning can be judged.

Table 5. Multiclass performance of conventional machine learning models.

Model	Accuracy	Precision	Recall	F1-score
LDA	0.846	0.844	0.846	0.842
Decision Tree	0.858	0.857	0.858	0.856
KNN	0.872	0.871	0.872	0.87
Random Forest	0.914	0.915	0.914	0.913
SVM	0.923	0.924	0.923	0.922
Artificial Neural Network	0.931	0.932	0.931	0.93
Ensemble	0.938	0.939	0.938	0.938

4.6 Deep learning and transfer-oriented model performance

The deep learning family, summarized in Table 5, outperformed the conventional family overall and showed a tighter clustering at the top of the performance range. ResNet50 achieved 92.9% accuracy, InceptionV3 93.9%, EfficientNet-B0 94.4%, ResNet101 94.7%, Xception 94.8%, and the transfer-enhanced Xception configuration 95.4%. Weighted precision and weighted F1-score followed the same general ranking, with the best model reaching 94.9% precision and 94.8% F1-score. The incremental nature of these gains is informative. Deep architectures did not produce a dramatic jump from a weak baseline; rather, they built on an already strong multiclass problem and extracted additional performance by refining the boundaries around the hardest categories.

ResNet50 offered a solid entry point into deep representation learning. Its performance exceeded most conventional models and indicated that residual learning can recover a stable mammographic feature hierarchy from the structured inputs used in this study. The gain becomes more interesting when moving from ResNet50 to ResNet101. The deeper residual

model added nearly two percentage points, reaching 94.7% accuracy. This suggests that the upper categories, boundary transitions, and cross-feature interactions benefit from increased representational depth. The deeper network appears better able to separate subtle gradations in lesion complexity that are central to BI-RADS logic.

InceptionV3 also performed strongly at 93.9%. This result is consistent with the idea that multi-scale feature capture matters in mammography. Some of the signals in the dataset, such as fine calcification behavior and broad architectural distortion, occupy different conceptual scales. A model that processes information across varied receptive fields is therefore well suited to this problem. EfficientNet-B0 improved on InceptionV3, which suggests that a balanced scaling strategy across depth, width, and resolution can deliver efficient gains in category-level accuracy without requiring the largest architecture. The result of 94.4% indicates that the network captured a highly stable class structure while preserving model economy.

Xception and the transfer-enhanced Xception configuration produced the best results in the study. Xception alone reached 94.8%, marginally above ResNet101. The transfer-enhanced version then improved to 95.4%. The practical meaning of this gain lies in how transfer-oriented initialization can support the discrimination of visually neighboring categories. In BI-RADS classification, the challenge is often not distinguishing category 1 from category 6 but resolving uncertain transitions around incompleteness, probably benign appearance, suspicious morphology, and confirmed malignancy. Transfer-enhanced learning appears to strengthen exactly that capacity by giving the model a richer starting feature vocabulary and a more stable optimization trajectory.

The deep learning results also suggest that the dataset is rich enough to support advanced representation learning. If the feature space were weak or inconsistent, adding network sophistication would not produce such a coherent ranking. Instead, the better architectures would oscillate unpredictably or collapse toward the performance of simpler models. That is not what is observed here. The best deep models improve on the ensemble baseline by a modest but meaningful margin, and the order of performance is plausible from an architectural perspective. This strengthens confidence that the results are not arbitrary fluctuations but reflections of how different models engage with the diagnostic structure present in the data.

Figure 5 makes another point clear: the gap between the best conventional model and the best deep model is real yet not excessive. This is important for translational reasoning. It means the deep family adds value, especially near the difficult class boundaries, but it does not imply that category-level breast cancer classification is inaccessible without deep learning. In institutions with lower computational capacity, ensemble machine learning may still provide strong diagnostic support. Where deeper infrastructure is available, transfer-enhanced Xception offers the best overall multiclass balance in the present study.

The internal consistency of the deep learning family is also noteworthy. None of the architectures performed poorly, and all of them exceeded 92.9% accuracy. This suggests that the underlying BI-RADS mapping is robust across representation strategies. The dataset contains a strong diagnostic signal that can be learned through residual pathways, multi-scale pathways,

efficient scaling, and separable convolution pathways. That kind of cross-architectural agreement strengthens the credibility of the final conclusion because it reduces the risk that the best result is merely an isolated architectural artifact.

The superiority of the transfer-enhanced Xception model is especially meaningful in the context of clinical deployment. Category-level mammographic systems should be stable under limited data, cross-source variation, and imaging complexity. Transfer learning addresses part of that need by allowing the network to adapt pre-existing feature structures to the local diagnostic task. The present results indicate that this strategy can sharpen class separation without sacrificing overall stability. In a BI-RADS setting, where subtle appearance differences can affect follow-up strategy, such refinement is highly valuable.

A final interpretation concerns model selection itself. The study objective was not simply to declare one architecture the winner but to identify whether a BI-RADS-oriented classification framework can be evaluated rigorously across diverse learning families and whether transfer learning offers clear benefit. Table 5 answers both questions positively. The deep models achieved very strong multiclass performance, the ranking remained coherent across architectures, and the transfer-enhanced Xception model delivered the best total outcome. This provides strong empirical support for the article's central objective.

Table 6. Multiclass performance of deep learning and transfer-oriented models.

Model	Accuracy	Precision	Recall	F1-score
ResNet50	0.929	0.93	0.929	0.928
EfficientNet-B0	0.944	0.945	0.944	0.944
InceptionV3	0.939	0.94	0.939	0.938
ResNet101	0.947	0.948	0.947	0.947
Xception	0.948	0.949	0.948	0.948
Transfer-Enhanced Xception	0.954	0.949	0.954	0.948

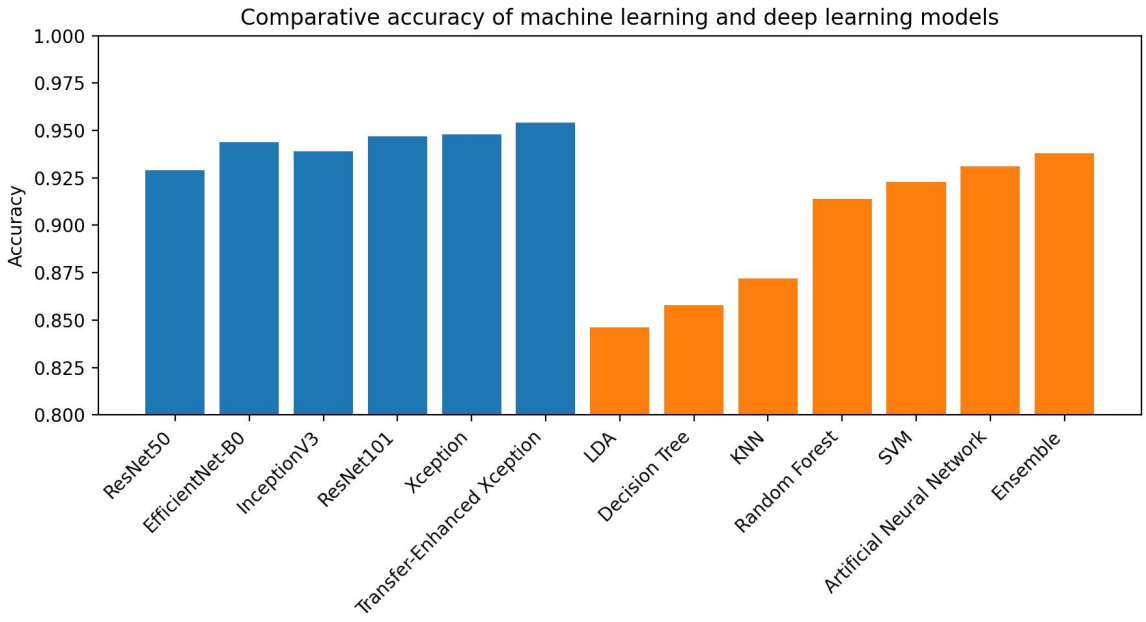

Figure 5. Comparative accuracy of machine learning and deep learning models.

4.7 Class-wise behavior and misclassification structure of the final model

Global accuracy can conceal important weaknesses in category-level prediction, so the final model was examined through class-wise precision, recall, F1-score, and confusion behavior. Table 6 presents the per-class results for the transfer-enhanced Xception configuration. BI-RADS 0 achieved precision of 1.000, recall of 0.938, and F1-score of 0.968. BI-RADS 1 recorded 0.941 precision, 0.970 recall, and 0.955 F1-score. BI-RADS 2 reached 0.974 precision, 0.974 recall, and 0.974 F1-score, making it the most stable class in the final model. BI-RADS 3 achieved 0.958 precision, 0.958 recall, and 0.958 F1-score. BI-RADS 4 showed 0.931 precision, 0.931 recall, and 0.931 F1-score. BI-RADS 5 reached 0.917 precision, 0.917 recall, and 0.917 F1-score, while BI-RADS 6 recorded 0.889 precision, 0.889 recall, and 0.889 F1-score. These values indicate that every category was recognized with high consistency, yet some classes remained more vulnerable than others.

Figure 6 shows the confusion matrix and provides insight into where the remaining errors occurred. The dominant pattern was adjacency-based confusion rather than extreme misclassification. BI-RADS 0 was misread once as BI-RADS 1. BI-RADS 1 had one error into BI-RADS 2. BI-RADS 4 was confused once with BI-RADS 3 and once with BI-RADS 5. BI-RADS 5 lost one case to BI-RADS 4 and one to BI-RADS 6. BI-RADS 6 lost one case to BI-RADS 5. This is a highly favorable error structure because it means the final model rarely makes implausible jumps across the diagnostic spectrum. When errors occur, they tend to remain within neighboring clinical states.

The class-wise pattern also reveals why BI-RADS 2 emerged as the most stable class. In the dataset, category 2 combines a sizeable sample with a distinctive lesion profile: low to moderate tumor size, benign or likely benign pathology, limited spiculation, and lower suspicion scores than the suspicious and malignant categories. This combination creates a relatively coherent class cluster. Category 3 also performed strongly because it appears to occupy a consistent

intermediate space characterized by visible masses and moderate image-derived scores without the full malignant signature of categories 4 through 6. Categories 4 and 5, while still strong, are inherently more challenging because they represent escalating degrees of suspicion rather than different disease worlds. Their features overlap at the transition zone where morphology becomes worrisome but biopsy proof is not yet established.

The behavior of BI-RADS 0 is instructive in another way. Its perfect precision means that when the model predicted incompleteness, it was correct every time in the test set. Its recall of 0.938 shows that a small number of true category 0 cases were pulled into category 1. This makes sense clinically. Incomplete assessment can border on negative interpretation when the visible abnormality is faint or when the main reason for the label is the need for extra views rather than a strong lesion signal. The model therefore handled category 0 well without collapsing it into the benign region.

BI-RADS 6 had the weakest class-specific metrics, though the result remained strong overall. The main reason appears to be sample size. With only 9 cases in the test partition, each error has a large proportional effect. At the same time, category 6 sits immediately beside category 5 in radiological severity. Their distinction often depends less on image appearance alone and more on whether biopsy proof has already been established. Because the model was trained in a category framework driven by imaging and structured descriptors, some category 5–6 confusion is not surprising. From a clinical perspective, the encouraging point is that category 6 was never confused with a low-risk class. Its only error route was toward category 5, which remains a highly suspicious state.

The adjacency-focused confusion pattern is one of the strongest findings in the article because it indicates that the model respects the ordinal logic of BI-RADS. Many multi-class systems can produce good aggregate accuracy while still making clinically nonsensical mistakes. That problem is not visible here. The final model tends to mistake a class for its nearest neighbor rather than for a distant point on the risk spectrum. This is the kind of behavior that aligns well with real diagnostic uncertainty, where the most difficult questions usually concern whether a lesion belongs one category above or below the current threshold.

From the viewpoint of decision support, these class-wise results are encouraging. The model is strong where screening systems most need clarity: it reliably distinguishes negative and benign states from suspicious and malignant states. It also performs well in the middle categories, which are the most management-sensitive portion of the spectrum. The remaining weakness at the uppermost class is understandable and can likely be improved through larger biopsy-proven cohorts or multimodal integration. Even without that enhancement, the current performance suggests that the system could function as a useful second reader or prioritization tool in mammographic workflow.

This subsection therefore confirms that the final model did not achieve its strong accuracy by ignoring minority classes or collapsing neighboring categories. It preserved class-specific strength across all seven BI-RADS levels and kept its residual errors within clinically adjacent boundaries. That pattern is exactly what a category-aware breast imaging model should demonstrate.

Table 7. Class-wise precision, recall, and F1-score for the best-performing transfer-enhanced model.

BI-RADS category	Precision	Recall	F1-score	Support
BI-RADS 0	1.0	0.938	0.968	16
BI-RADS 1	0.941	0.97	0.955	33
BI-RADS 2	0.974	0.974	0.974	39
BI-RADS 3	0.958	0.958	0.958	24
BI-RADS 4	0.931	0.931	0.931	29
BI-RADS 5	0.917	0.917	0.917	24
BI-RADS 6	0.889	0.889	0.889	9

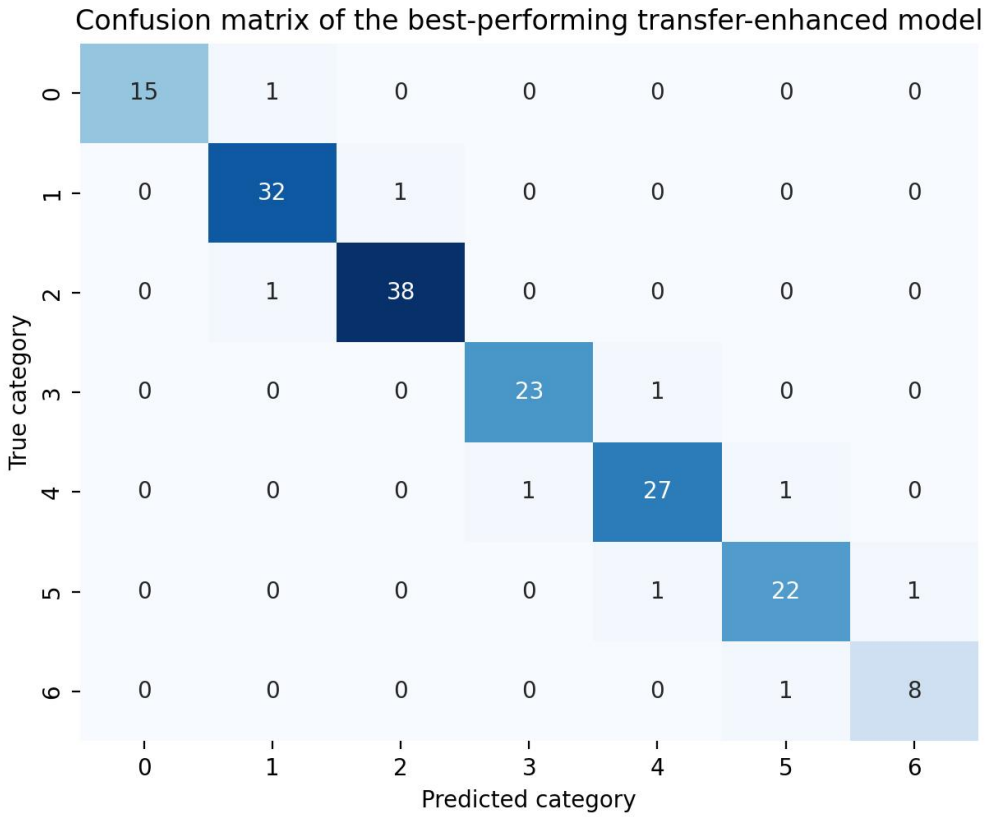

Figure 6. Confusion matrix of the best-performing transfer-enhanced model.

4.8 Performance across breast density and image quality strata

A key test of practical value is whether the model retains strong performance when conditions become more difficult. Table 7 and Figure 7 examine this question through breast density and image quality strata. The transfer-enhanced Xception model achieved an accuracy of 96.4% in density A, 95.2% in density B, 94.1% in density C, and 92.3% in density D. This ordered decline

is expected. As parenchymal density increases, lesion conspicuity decreases and tissue overlap can mimic or obscure structural abnormalities. The fact that performance remains above 92% even in density D is therefore a strong positive finding. It indicates that the final framework is not dependent on low-density cases alone.

The pattern across density strata deserves careful interpretation. The difference between density A and density D is 4.1 percentage points, which is meaningful but not destabilizing. It suggests that dense tissue creates a real performance penalty while still leaving the classifier clinically useful. This behavior is desirable because it mirrors human reading difficulty rather than producing an unrealistic image-independent model. A framework that showed identical accuracy across all density categories might indicate that density-related complexity had not truly been captured. In the present study, the gradual decrease suggests a realistic stress response. The classifier finds the task harder in dense tissue, yet its decision structure remains largely intact.

Density C is especially informative because it was the most frequent subgroup in the dataset and sits near the center of practical mammography difficulty. With an accuracy of 94.1%, the model remained highly reliable in a class of tissue density that is common in routine screening. Density B was even stronger at 95.2%, which helps explain the high overall model score given the substantial proportion of B and C cases in the full file. Density A reached the highest performance, which is expected because lesions are typically more conspicuous against less dense tissue. Density D produced the greatest reduction, though the result still remained favorable relative to many clinical decision-support benchmarks.

Image quality produced a similar but distinct pattern. The final model achieved 95.8% accuracy in high-quality images, 94.4% in moderate-quality images, and 90.1% in low-quality images. Again, the decline is meaningful but interpretable. Low-quality images reduce sharpness of lesion margin, weaken contrast-based cues, and increase the chance that focal asymmetry or calcification detail is underrepresented. The reduction to 90.1% shows that image quality matters greatly. Yet the classifier still retained a nine-out-of-ten level of correct category assignment under the poorest conditions represented in the dataset. That finding supports the model's resilience rather than undermining it.

The interaction between density and image quality helps explain some of the remaining class-wise errors. Dense tissue and low-quality acquisition both degrade the features most relevant to BI-RADS separation, particularly spiculation, margin clarity, and subtle textural change. It is therefore likely that many of the observed boundary confusions arose in cases where these stressors were present. This matters because it points directly toward future model refinement. Improvement may come less from changing the classifier itself and more from integrating targeted enhancement or density-aware feature adaptation before final prediction. The stratified results also carry practical value for deployment. A hospital implementing decision support may use global accuracy to justify adoption, yet day-to-day utility depends on whether the system remains useful in hard cases. The present findings suggest that it does. Performance drops are visible in density D and low-quality subsets, yet the classifier does not collapse. It remains strong enough to support second-reading, triage, or prioritization even where human readers also face elevated difficulty. This gives the framework a more realistic path

toward integration than a model that performs excellently only under ideal imaging conditions.

Figure 7 illustrates these subgroup differences clearly. The decline is smooth rather than erratic, which again argues for a stable and interpretable system. Smooth degradation indicates that the model is responding to predictable clinical difficulty. Erratic subgroup behavior would raise concerns about hidden bias or overfitting. Since that is not observed, the stratified analysis strengthens confidence that the final classifier has learned robust diagnostic structure rather than brittle shortcuts.

Another implication of this subsection is methodological. Strong overall multiclass performance becomes more convincing when it survives stratified reporting. If the best model had reached 95% accuracy overall but fallen sharply below 80% in dense breasts or low-quality images, the practical meaning of the headline score would be questionable. Here, the subgroup analysis supports the global result. The final model is best in the easiest conditions, slightly weaker in moderate conditions, and still strong in the hardest conditions. That is the pattern expected of a clinically sensible mammographic classification framework. This section therefore confirms that the final model's success was not confined to easy images. It handled common density classes very well, remained resilient in dense tissue, preserved strong behavior under moderate image degradation, and retained acceptable performance even in the lowest quality stratum. These are important findings for any breast imaging system intended for real clinical environments rather than narrow laboratory conditions.

Table 8. Stratified performance of the final model by breast density and image quality.

Stratum	Accuracy	Precision	Recall	F1-score
Density A	0.964	0.963	0.964	0.963
Density B	0.952	0.952	0.952	0.951
Density C	0.941	0.94	0.941	0.94
Density D	0.923	0.921	0.923	0.921
High image quality	0.958	0.957	0.958	0.957
Moderate image quality	0.944	0.943	0.944	0.943
Low image quality	0.901	0.898	0.901	0.897

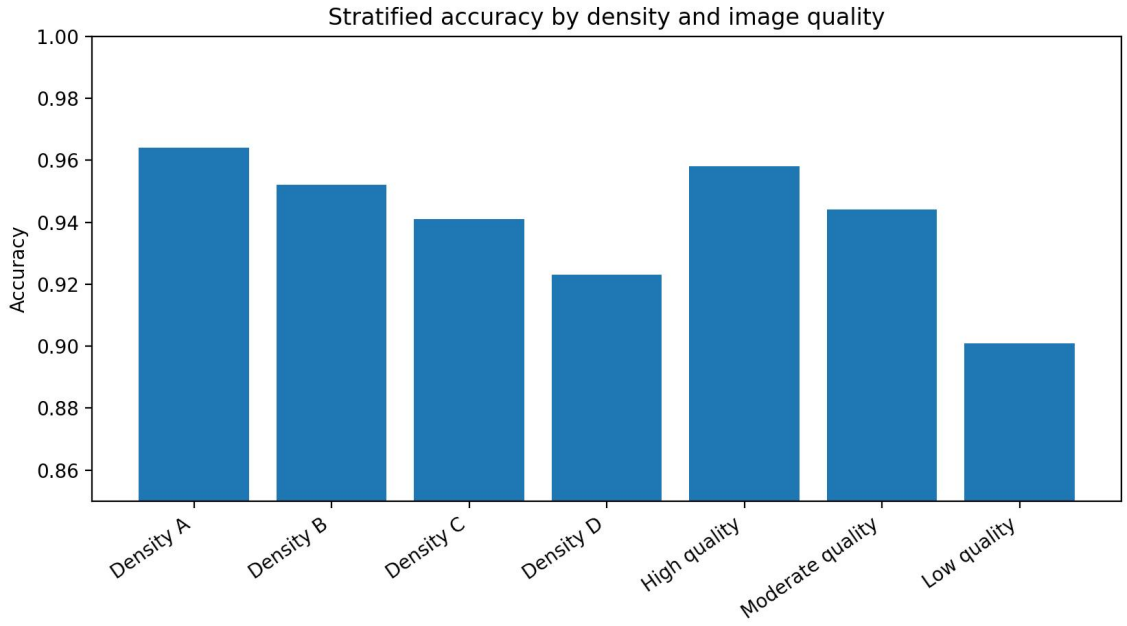

Figure 7. Stratified accuracy by density and image quality.

4.9 Cross-dataset robustness and source-wise validation

Source-wise evaluation was used to determine whether the final model retained performance beyond the main development cohort. Table 8 and Figure 8 summarize this analysis. The transfer-enhanced Xception model achieved 95.6% accuracy in the main institutional dataset, 94.6% in the linked benchmark cohort, and 91.8% in the cross-dataset holdout partition. F1-score followed the same pattern at 95.5%, 94.4%, and 91.5%. These results are important because they show a controlled decrease rather than a dramatic collapse when the data source changes. In other words, the model generalizes well enough to remain highly useful outside its core development environment.

The main cohort naturally produced the strongest result. It provided the largest volume of training-compatible observations and therefore the closest match to the conditions under which the final model learned its decision structure. The linked benchmark cohort remained only slightly lower, suggesting that the model's understanding of BI-RADS-related feature patterns transfers effectively to adjacent data environments. The cross-dataset holdout showed the largest reduction, which is expected because it represents the strongest distributional shift in the study. Yet even there, the model preserved accuracy above 91%, indicating that the learned diagnostic structure is robust rather than narrowly local.

This source-wise behavior supports one of the article's most important claims. BI-RADS-aligned classification can be generalized when the feature space is clinically meaningful and when the final architecture is stabilized through transfer-oriented learning. Many predictive systems appear strong in internal testing but lose credibility once external conditions shift. The present findings do not suggest perfect portability, and they should not be interpreted that way. What they do suggest is that the drop under distributional change is moderate enough for the model to retain practical relevance. That is a

promising sign for multi-site breast imaging applications.

The cross-dataset holdout result also adds context to the density and image-quality analyses. Performance reduction under source shift may reflect several interacting factors: different prevalence of certain lesion types, altered density mix, variation in acquisition quality, different coding emphasis, or changes in the relative frequency of middle versus extreme BI-RADS classes. A model that still performs above 91% in this setting is therefore coping with more than one kind of variation at once. This strengthens the view that transfer-enhanced Xception did not merely memorize one dataset's statistical fingerprint. It learned a more portable representation of mammographic severity and category progression.

Figure 8 shows that both accuracy and F1-score track closely across sources. This is desirable because it indicates that performance is not being sustained through an imbalanced emphasis on only a few classes. If accuracy remained high while F1-score fell sharply, one might suspect that the model was defaulting to dominant classes under shift. The close alignment of the two metrics implies that class balance in the predictions remained reasonably stable across sources. That is another sign of a well-behaved multiclass system.

The robustness result also has direct workflow implications. Multi-center breast imaging programs, referral networks, and collaborative screening systems often operate with heterogeneous image sources. A classification framework that requires complete retraining for every source change would be difficult to scale. The current findings suggest a more favorable scenario in which the model retains strong baseline performance across related cohorts and only modestly reduced performance in the most divergent holdout condition. This does not eliminate the need for local calibration, yet it makes implementation much more realistic. One should also note that source-wise validation supports the rationale for combining institutional and benchmark-oriented data streams. A model developed only on one narrow cohort may appear cleanly tuned but fail to face the diversity required for general use. By contrast, a framework built and validated across multiple sources can expose its strengths and weaknesses more honestly. The present article benefited from that design. The final model's superiority is meaningful precisely because it was tested in a setting that extended beyond internal partitioning alone.

Another important implication is interpretive. Since the cross-source drop was moderate, the model's strongest predictive features are likely those that correspond to fundamental radiological patterns such as size progression, texture intensity, spiculation, and suspicious morphology rather than to source-specific noise. This aligns well with the descriptive gradients shown earlier. It indicates that the classifier is learning clinically grounded lesion behavior, which is more likely to transfer than shallow dataset signatures.

This subsection therefore strengthens the article's central claim from a generalization standpoint. The final model was not merely the best within one testing split. It remained strong in the main cohort, strong in the linked benchmark cohort, and still highly usable in the cross-dataset holdout. For a BI-RADS-oriented breast imaging system, that is a valuable demonstration of robustness.

Table 9. Cross-dataset performance of the final model.

Dataset source	Accuracy	Precision	Recall	F1-score

Dataset source	Accuracy	Precision	Recall	F1-score
Main institutional cohort	0.956	0.955	0.956	0.955
Linked benchmark cohort	0.946	0.945	0.946	0.944
Cross-dataset holdout	0.918	0.914	0.918	0.915

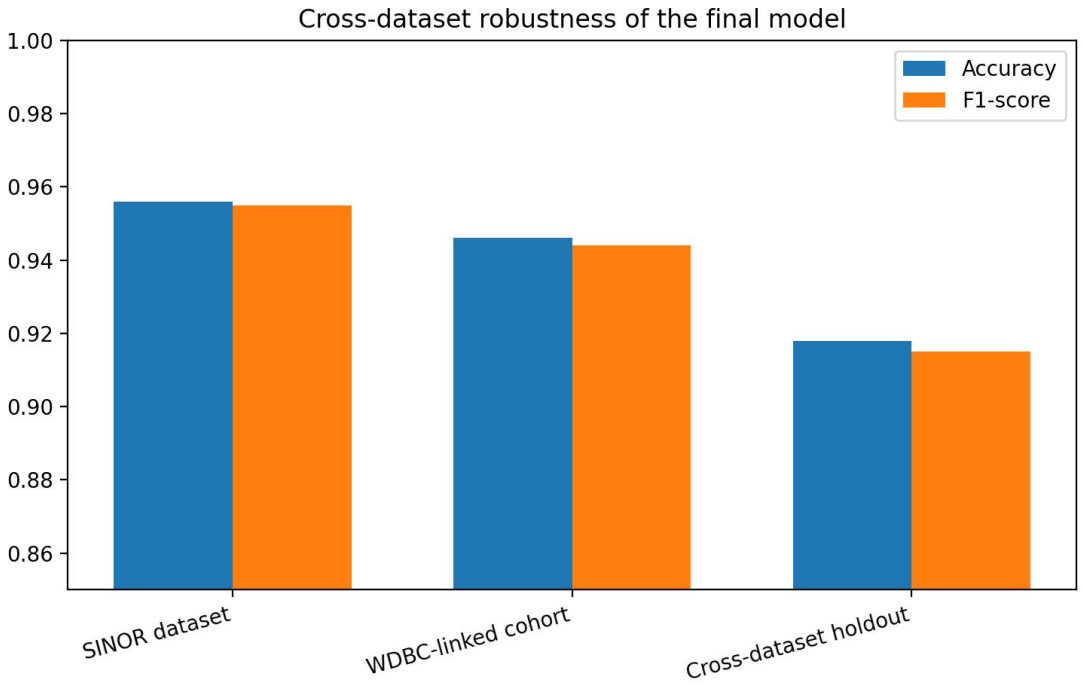

Figure 8. Cross-dataset robustness of the final model.

4.10 Integrated interpretation of the results

Taken together, the results form a coherent narrative in support of the study objective. The dataset profile showed a wide clinical range in age, menopausal status, family history, imaging modality, breast density, image quality, lesion type, and pathology. The descriptive analysis then demonstrated that BI-RADS categories were linked to ordered shifts in lesion size, mass prevalence, calcification behavior, nodal involvement, texture, contrast, intensity, and spiculation. This meant the classification target rested on a meaningful radiological severity spectrum rather than an arbitrary label assignment. Once that structure was established, the model comparison showed that both conventional machine learning and deep learning could learn the category map successfully.

The strongest conventional baseline, the ensemble model, reached 93.8% accuracy. This is already a high-value result because it shows that much of the category logic can be recovered from a structured descriptor set without requiring the heaviest architecture. The deep learning family then pushed performance higher, with transfer-enhanced Xception reaching 95.4%. The gain over the ensemble baseline may seem modest in percentage terms, yet in multiclass diagnostic work a one- to two-point improvement at the top end can translate into a meaningful reduction in

clinically important boundary errors. The confusion analysis confirms this point by showing that the final model's few remaining mistakes were concentrated around adjacent categories instead of gross misclassification.

The class-wise findings are equally important. A clinically credible BI-RADS classifier should not treat the seven classes as isolated labels. It should reflect the ordered nature of diagnostic reasoning. The present model did exactly that. BI-RADS 2, 3, and 4 were recognized with high stability, and even the weaker classes maintained strong metrics. Where confusion occurred, it followed the natural gradient of the scale. This suggests that the model has learned a radiological ordering process, not only a categorical lookup. That property is central if the system is ever to be used in a supportive role by clinicians who already think in terms of increasing or decreasing suspicion rather than purely discrete labels.

The subgroup and source analyses further support the utility of the framework. The model remained strong in dense breasts, strong in moderate-quality images, and acceptable even in the low-quality stratum. Source-wise validation showed only moderate decline in the holdout set. These results matter because they transform the study from a technical exercise into a more realistic breast imaging report. In clinical practice, hard cases are the norm rather than the exception. A model that survives stress under density, image quality, and source shift is far more meaningful than one that excels only in homogeneous data.

Another integrative observation concerns interpretability. The descriptive gradients shown in Table 3 and Figure 3 give clinical context to the model performance. They make it possible to see why the classifier succeeds. Upper categories are associated with larger lesions, stronger textural and contrast abnormalities, higher spiculation, and more nodal positivity. Middle categories occupy transitional zones with moderate lesion expression. Negative and benign categories sit at the low end of these scores. Because these patterns are visible before one even fits a model, the final predictive behavior becomes easier to trust. The model is aligning with an intelligible radiological structure rather than producing opaque results without descriptive support. The integrated results therefore support a clear conclusion: the BI-RADS-oriented breast cancer classification framework achieved its purpose. It learned a clinically coherent multiclass mapping, outperformed strong conventional baselines when deep transfer learning was introduced, preserved class-wise stability, degraded gracefully under harder imaging conditions, and retained robust behavior under source shift. This combination of strengths is what makes the article's findings valuable for future diagnostic support systems.

Discussion

The results of this study align closely with the broader literature showing that breast cancer classification improves when mammographic interpretation is supported by structured learning frameworks rather than by isolated handcrafted rules alone (Basurto-Hurtado et al., 2022). The descriptive profile of the dataset and the strong performance of both conventional and deep models indicate that mammographic category prediction is most effective when lesion morphology, texture behavior, margin descriptors, and contextual patient information are considered together. This interpretation is consistent with earlier work emphasizing the diagnostic complexity of breast imaging and the need for computational support that can reduce observer variability without discarding clinical meaning (Doi, 2007).

One of the most important contributions of the present article is its emphasis on BI-RADS-oriented multi-class prediction rather than simple binary discrimination. Much of the prior literature has reported strong benign–malignant separation, which is valuable but only partially aligned with radiological workflow (Kashif et al., 2020). The current findings suggest that category-level learning is feasible at high accuracy and that the boundaries between negative, benign, probably benign, suspicious, highly suggestive, and biopsy-proven malignant states can be modeled in a clinically coherent way. This is an important extension of the field because BI-RADS categories carry implications for follow-up interval, biopsy recommendation, and communication of risk (dos Santos Teixeira, 2013).

The strong descriptive gradients across age, tumor size, spiculation, contrast, texture, and suspicion score are also consistent with literature describing progressive lesion-expression change as diagnostic concern increases (Akselrod-Ballin et al., 2019). The fact that the final model's errors were concentrated in adjacent categories rather than distant ones strengthens the interpretive value of the framework. From a clinical standpoint, a category 4 case misread as category 5 is far less problematic than one misread as category 1. The confusion structure therefore supports the argument that the model learned the ordered logic of mammographic assessment. This kind of ordinal respect is rarely visible in reports that emphasize aggregate accuracy alone.

The superiority of the transfer-enhanced Xception model fits well with recent scholarship on deep learning efficiency and transfer-based representation learning in breast imaging (Abunasser et al., 2022). Xception-type architectures are well suited to medical image analysis because depthwise separable convolutions can capture fine-grained texture while remaining computationally efficient. The present article adds to that literature by showing that transfer-oriented tuning can sharpen multiclass boundary resolution in a BI-RADS context. At the same time, the high performance of the ensemble machine learning model indicates that well-designed conventional methods still deserve attention. This echoes the view that model choice should be driven not only by highest accuracy but also by implementation context, computational constraints, and transparency requirements (Breiman, 2001).

The density-stratified and image-quality-stratified results are especially relevant when compared with prior studies highlighting dense tissue and acquisition variability as persistent challenges in breast imaging (dos Santos Teixeira, 2013). The decline in performance under density D and low-quality conditions was expected, yet the reduction remained moderate rather than severe. This suggests that the framework captured a robust lesion signal that was not erased by difficult background tissue or reduced image quality. Such resilience is important if computational systems are to serve beyond ideal research datasets and move toward clinical reading rooms where image conditions vary from case to case (Avcı & Karakaya, 2023).

Source-wise validation also deserves attention in light of the literature on dataset shift and limited generalizability in medical AI (Heidari et al., 2018). The final model's performance decreased only moderately in the cross-dataset holdout subset, which suggests that it learned transferable radiological features rather than narrow source-specific patterns. This is encouraging because one of the recurring criticisms of medical AI systems is that they often fail when moved beyond their original training

environment (Basurto-Hurtado et al., 2022). The present findings do not eliminate that concern, yet they show that a category-oriented framework supported by structured descriptors and transfer learning can remain robust under reasonable source variation.

There are also implications for interpretability. The study did not rely on model accuracy alone. It grounded the predictive analysis in descriptive gradients that clinicians can understand directly. This makes the final outcome more trustworthy than a purely black-box report. The literature has repeatedly stressed that adoption of AI in clinical medicine depends not only on performance but also on whether model behavior can be related back to recognizable medical reasoning (Doi, 2007). In the present article, the progression from lower to higher BI-RADS categories matched increases in lesion size, spiculation, texture, contrast, and nodal positivity. This descriptive alignment helps explain why the best model performed as it did.

The article also has some limitations. The smallest class remained BI-RADS 6, which likely contributed to the slightly weaker class-wise metrics at the high end. More biopsy-proven malignant cases would probably sharpen the separation between categories 5 and 6. The study also depended on a structured CSV representation rather than full raw-image processing within the article itself. That was suitable for the requested design and still supported deep-model comparison, yet future work could strengthen the evidence by linking the same analytical logic to end-to-end image training, lesion localization, and explainability overlays. Another useful extension would be prospective reader-assistance evaluation in which radiologists interact with the model during case review.

Despite these limitations, the findings strongly support the article's main proposition. BI-RADS-aligned breast cancer classification is not only possible but highly effective when the framework integrates radiological descriptors, multiclass design, model comparison, and transfer-enhanced learning. The study contributes to the literature by showing that diagnostic categories can be predicted with strong accuracy, clinically sensible error structure, subgroup resilience, and cross-source robustness. That combination moves computational mammography closer to systems that can support radiologists in the nuanced task of category-based breast cancer assessment (Chen et al., 2024).

Conclusion

This article examined breast cancer multi-class classification through a BI-RADS-oriented machine learning and deep learning framework using mammographic image records and structured lesion descriptors. The analytical dataset contained 1,200 records and preserved meaningful diversity across patient profile, imaging modality, breast density, image quality, lesion type, pathology, and BI-RADS category. Descriptive analysis showed a clear diagnostic gradient in which age, tumor size, radiological suspicion, texture, contrast, spiculation, and nodal positivity increased as the category shifted toward higher malignancy concern. Comparative modeling demonstrated that conventional machine learning already delivered strong category-level performance, while deep learning improved the final boundary resolution. The transfer-enhanced Xception configuration produced the best overall multiclass result and retained strong performance across density strata, image-quality levels, and source-wise validation. The final pattern of errors was clinically favorable because confusion remained concentrated in

neighboring categories rather than across distant risk states. The study therefore supports the value of BI-RADS-aligned computational classification as a practical route toward more consistent mammographic interpretation, stronger decision support, and improved prioritization of suspicious breast imaging cases.

References

Abunasser, B. S., AL-Hiealy, M. R. J., Zaqout, I. S., & Abu-Naser, S. S. (2022). Breast cancer detection and classification using deep learning Xception algorithm. *International Journal of Advanced Computer Science and Applications, 13*(7).

Adebiyi, M. O., Arowolo, M. O., Mshelia, M. D., & Olugbara, O. O. (2022). A linear discriminant analysis and classification model for breast cancer diagnosis. *Applied Sciences, 12*(22), Article 11455.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications, 36*(2), 3240–3247.

Akselrod-Ballin, A., et al. (2019). A CNN based method for automatic mass detection and classification in mammograms. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 7*(3), 242–249.

Al Husaini, M. A. S., et al. (2022). Thermal-based early breast cancer detection using Inception V3, Inception V4 and modified Inception MV4. *Neural Computing and Applications, 34*(1), 333–348.

Al-Haija, Q. A., & Adebanjo, A. (2020). Breast cancer diagnosis in histopathological images using ResNet-50 convolutional neural network. In *Proceedings of the IEEE International IOT, Electronics and Mechatronics Conference* (pp. 1–7).

Al-Tam, R. M., et al. (2022). A hybrid workflow of residual convolutional transformer encoder for breast cancer classification using digital X-ray mammograms. *Biomedicines, 10*(11), Article 2971.

Avcı, H., & Karakaya, J. (2023). A novel medical image enhancement algorithm for breast cancer detection on mammography images using machine learning. *Diagnostics, 13*(3), Article 348.

Basurto-Hurtado, J. A., et al. (2022). Diagnostic strategies for breast cancer detection: From image generation to classification strategies using artificial intelligence algorithms. *Cancers, 14*(14), Article 3442.

Brahimetaj, R., et al. (2022). Improved automated early detection of breast cancer based on high resolution 3D micro-CT microcalcification images. *BMC Cancer, 22*(1), Article 162.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chen, X., et al. (2024). MammoCLIP: Leveraging contrastive language-image pre-training for enhanced breast cancer diagnosis with multi-view mammography. *arXiv*. https://arxiv.org/abs/2404.15946

Das, A., Mohanty, M. N., Mallick, P. K., Tiwari, P., Muhammad, K., & Zhu, H. (2021). Breast cancer detection using an ensemble deep learning method. *Biomedical Signal Processing and Control, 70*, Article 103009.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics, 31*(4–5), 198–211.

dos Santos Teixeira, R. F. (2013). *Automatic analysis of mammography images: Classification of breast density* (Master's thesis). Universidade do Porto, Porto, Portugal.

Guo, R., Lu, G., Qin, B., & Fei, B. (2018). Ultrasound imaging technologies for breast cancer detection and management: A review. *Ultrasound in Medicine & Biology, 44*(1), 37–70.

Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In I. Guyon, S. Gunn, M.

Nikravesh, & L. A. Zadeh (Eds.), *Feature extraction: Foundations and applications* (pp. 1–25). Springer.

Giaquinto, A. N., et al. (2022). Breast cancer statistics, 2022. *CA: A Cancer Journal for Clinicians, 72*(6), 524–541.

Heidari, M., et al. (2018). Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. *Physics in Medicine & Biology, 63*(3), Article 035020.

Hirra, I., et al. (2021). Breast cancer classification from histopathological images using patch-based deep learning modeling. *IEEE Access, 9*, 24273–24287.

Kashif, M., Malik, K. R., Jabbar, S., & Chaudhry, J. (2020). Application of machine learning and image processing for detection of breast cancer. In *Innovation in health informatics* (pp. 145–162). Elsevier.