Current Methodology Towards a Standardized and Automated Background Parenchymal Enhancement Categorization: A Literature Review

C. Miguel Palao, Utrecht University

***Abstract*—Background parenchymal enhancement (BPE) is defined as the amount of enhancement observed in normal fibroglandular tissue (FGT) after contrast administration in breast MR protocols. This phenomenon has been associated to sensitivity reductions, as a consequence of lesion occlusion, and most recently described as a potential imaging biomarker for breast cancer prediction. Thus, a standardized automated methodology for its categorization may be of utmost importance. Nonetheless, the lexicon utilized as a gold-standard by radiologists is associated to large variability and susceptibility in BPE assessment between readers. This study aims to perform a review that recollects the most recent methodologies in the literature for BPE categorization, which entails three different frameworks: BPE quantification, radiomics and deep learning models. Findings indicated that, while the former is widely utilized, its applicability for four-way categorization by discretization is associated with lower correlations to the reference than that of radiologists, thus limiting its applicability to the task under study. Furthermore, machine learning (ML) approaches provided substantially better results towards standardized automated categorization. The potential superiority of ML techniques is shown to be associated with higher correlation coefficients than those achieved by quantification techniques. Within this group, radiomic architectures, which rely on the manual selection of features for proper representation of the tissue, presented an advantage against deep learning architectures. However, their natural dependence on segmentation techniques must be taken into account, as it may be significant of error propagation and a reduction in BPE estimation and categorization. Therefore, although the goal of standardized BPE categorization is still far, this review presents some important observations that may guide future efforts. Additionally, the relevance of balanced data in the development of ML techniques, and the limitation attributed to the wide variability of breast MR protocols among institutions, is highlighted. Future outlooks may focus on differentiations between symmetrical or asymmetrical BPE breasts, and two-way classifications towards more clinically relevant BPE categorization methodologies.**

***Index Terms*—breast MR, BPE, categorization, classification, quantification, ML, radiomics, DL**

I. Introduction

Breast cancer is one of the most prominent types of cancer worldwide as stated by the World Health Organization in 2024, who attributed over 670.000 deaths solely to the disease [1]. Deductions in incidence and mortality rely on screening and imaging diagnosis, commonly performed by mammography or breast magnetic resonance (breast MR), among others [2].

Breast MR relies on the administration of a contrast agent and the subsequent follow-up on the dynamic enhancement patterns of the tissue for lesion identification. Malignant regions present rapid peak enhancement after the contrast agent enters the bloodstream, which allows its kinetic differentiation from benign lesions and normal tissue [3]. Nonetheless, normal fibroglandular tissue (FGT) also suffers partial enhancement, a phenomenon known as background parenchymal enhancement (BPE); it typically appears on the periphery of the breast, it is characterized by symmetric or asymmetric patterns and it has been associated in the literature with menstrual cycles [4].

Moreover, while higher BPE values have been attributed to reductions in sensitivity of breast MR due to a possibility of lesion occlusion, it has also been increasingly associated with a higher risk for breast cancer development. Therefore, its identification and proper classification are of utmost importance. BPE is usually assessed by radiologists on the first T1-weighted post-contrast series of dynamic contrast-enhanced MR (DCE-MR) protocols, which corresponds to the first acquisitions after contrast administration [4]. This is the standard practice for an improved categorization, as the delayed enhancement of BPE (210-320 seconds) to that of malignant lesions (60-90 seconds) allows for better differentiation of contrast uptake from regions of normal FGT enhancement [3], [4]. Moreover, the assessment is commonly performed following the Breast Imaging Report and Data System (BI-RADS) by the American College of Radiology lexicon, based on increasingly ordered enhancement levels: minimal, mild, moderate and marked.

This protocol provides a series of guidelines for its proper interpretation. Still, it is limited by inter and intraobserver variability as a consequence of the intrinsic subjectivity of radiologists during qualitative clinical assessments. Although training has been shown to significantly improve agreement between radiologists, research on the matter has shown it to be characterized by 'fair' accordance [4]. To this end, automatic and semi-automatic approaches have been studied in the literature, which aim to provide standardized methodologies for longitudinal investigations between institutions. Such approaches may include techniques that involve the delineation of a region of interest (ROI), FGT segmentation, BPE quantification or the utilization of machine learning (ML) techniques. Due to the significant variability in methodologies found in the literature, automated standardization for BPE categorization has not yet been achieved in the clinical practice.

This review aims to collect the latest publications on BPE categorization in the literature to perform a comprehensive analysis that describes the specifications and requirements of each approach. With that, create a discussion that differentiates the advantages and limitations of every methodology and concludes on future outlooks that should be further investigated towards an automated and standardized categorization. Such methodologies will include both traditional approaches and ML techniques in between 2020-2024.

II. Background

A. BI-RADS Lexicon

The BI-RADS lexicon is a guideline originally developed for standardized reporting of mammography imaging patterns, and extended to other modalities such as breast MR [5]. The lexicon includes several aspects for exam acquisition and interpretation; for instance, indications for reporting, MR imaging techniques, and descriptors of overall breast composition. These are utilized to assign to each exam a final assessment category using a scoring system that ranges from 0 to 6; 0 - incomplete assessment in need of further imaging, 1 - normal, 2 - benign, 3 - probably benign, 4 - suspicious, 5 - highly suggestive of malignancy and 6 - known biopsy-proven malignancy [6].

More specifically, the BPE categorization descriptor is visually selected based on the quantity and intensity of the FGT enhancement on the breast volume. Its interpretation must be performed on the breast with the highest apparent FGT quantity in case of asymmetry between the two breasts, and an exam label should then be assigned, categorizing the BPE as minimal, mild, moderate or marked, in increasing order. [6]. An example representing the four subcategories is shown in Fig. 1.

B. Methodologies for BPE Categorization

The inter-reader variability associated with BPE rating has been previously described in the literature. Consequently, several studies have been published for the purpose of automating this classification task and towards an improved assessment of BPE.

In a review by Liao et al. [4], a summary of the most commonly utilized approaches for BPE quantification is described up to its publication, in 2020. Such methodologies utilize in a great majority ROI delineation approaches or segmentation techniques. On the one hand, the former involves the manual definition of ROIs of small-dimensions on pre and post-contrast series; the approach is consequently limited by the accuracy of the reader, as the definition of the ROI is inherently subjective. On the other hand, segmentation methodologies aim to reduce human error through the development of algorithms that perform the delineation task automatically. These typically involve several processing steps, including skin, chest wall, breast or FGT segmentation (to differentiate from fatty tissue in the latter case), enabling the identification of enhanced voxels/pixels [4].

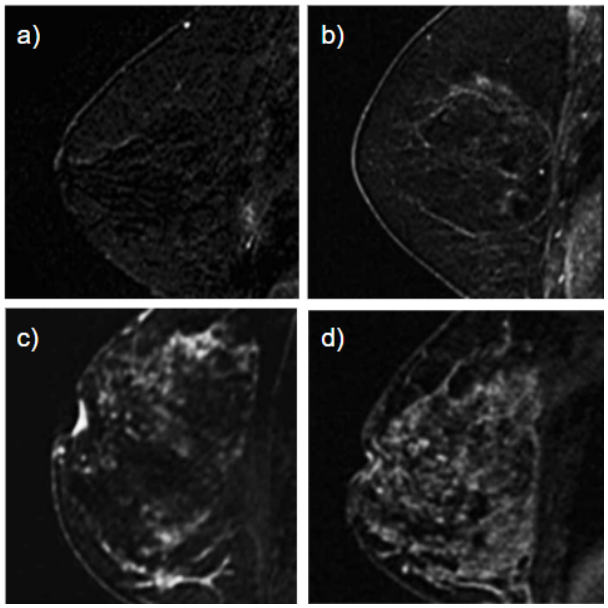

Fig. 1: BPE categories (a) minimal, (b) mild, (c) moderate, and (d) marked based on BI-RADS lexicon assessment. From [7].

After FGT segmentation, BPE quantification is carried through the application of mathematical computations related to percentage enhancement or signal enhancement ratios, as described in the aforementioned review by Liao et al. [4]. Following this computation technique, the BPE becomes quantifiable and may be discretized into each of the BI-RADS categories. This is a practice previously performed in a work presented in 2017 by Pujara et al. [8], who utilized BI-RADS lexicon enhancement percentages as the cut-off values of their four-way categorization; minimal: 25%, mild: 26-50%, moderate: 51-75% and marked: 75%. Nonetheless, it is important to note the statement in the most recent BI-RADS lexicon (5th edition), which advises against utilizing these values for BPE assessment [6].

Moreover, the applicability of artificial intelligence in the medical field is constantly growing. Consequently, several techniques are to be enclosed within this review, more specifically regarding radiomics and deep learning (DL) algorithms. Radiomics consists of a series of processing techniques that allow the recognition of features that are not visible to the common eye. These techniques rely on mathematical computations that extract characteristics retained within the intrinsic spatial distribution of intensities within an image. Nonetheless, processing of these quantifiable properties requires image segmentation in all cases; radiomic studies are uniquely performed within the delineated regions under study. Furthermore, it has been proven the importance of processing steps such as normalization or interpolation for a more robust output [9]

Other ML approaches rely on the utilization of DL frameworks, a technique that makes use of characteristics of the data to perform tasks such as segmentation or classification

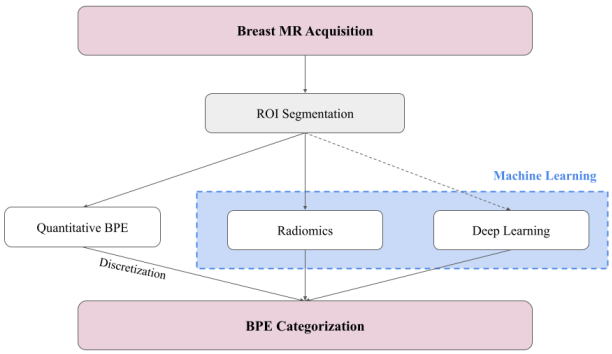

Fig. 2: Workflow representing commonly applied frameworks for BPE categorization enclosed in this review. ROI segmentation of skin/chest/breast/FGT structures is a procedure typically required for quantitative BPE and radiomic methodologies (solid arrow), whereas it establishes as a potential option for deep learning approaches (dotted arrow).

through convolutional neural networks (CNNs). A series of convolutional layers, pooling and fully connected layers are commonly organized in structured architectures in an order specific to the tasks to perform; in the case of classification, the focus relies on a final decision-making layer to perform the categorization of the data [10]. VGG models are commonly utilized classifiers. These are described as transfer learning-based deep models, a type of shallow architecture that utilizes pre-trained parametric values to address new tasks, such as medical image classification [11].

A workflow representing the previously described frameworks for BPE categorization is shown in Fig. 2.

III. Methods

This review aimed to compile all automated or semi-automated methodologies described in the literature the recent years (2020-2024) for BPE categorization. More specifically, the study entails a search for all papers including direct classification of the DCE-MR data or quantification and discretization of BPE levels after image processing. Thus, the following elements were utilized as search query in Scopus, Pubmed and Google Scholar: (("background parenchymal enhancement" OR "BPE") AND automat* AND "MR*")[Title, Abstract, Keywords] and (("background parenchymal enhancement" OR "BPE") AND quantif*)[Title] AND (("background parenchymal enhancement" OR "BPE") AND ("machine learning" OR "deep learning")[Title].

From this search, only articles detailing BPE quantification/classification methodologies were included, requiring either a full description of the process or its application within a broader research context. Ineligible studies included those based on other imaging techniques, literature reviews or categorization based on clinical reports rather than on the imaging data. It is important to note that no filtering was performed based on patient characteristics, breast MR protocols or imaging parameters.

IV. Results

The described search query led to a collection of 18 papers for the subsequent review. These were differentiated based on the methodologies applied in their work for BPE categorization, as shown in Fig. 2. Thus, the consequent sections will be divided based on the nature of the applied approaches: BPE quantification and ML, the latter consisting of radiomics and DL architectures.

The applicability of breast and FGT segmentation or image correction are preprocessing methodologies commonly utilized for BPE categorization, among others. Nonetheless, these not being the main focus of the paper, the applicability of such techniques will be mentioned only for context purposes.

A. BPE Quantification

The first results presented relate to BPE quantification and discretization frameworks. A collection of twelve papers included such methodologies in their research, which were categorized into signal intensity-based and volume-based models, following the structure presented by Müller et al. [12]. The aforementioned model differentiation will serve to define the base equation, while specifications from each publication will be presented for a more comprehensive study. Moreover, out of the complete set, six papers described methodologies based on the former, four utilized the latter and two papers described both quantification algorithms in their research [13], [14]. A compilation of all publications in this section is shown in TABLE I, which provides general information about the utilized approaches, input data and other specifications.

It must be noted, that most of the methodologies required the combination of DCE-MR series with FGT or breast masks for simplification of subsequent procedures. For instance, BPE segmentation relies on FGT to apply varying contrast enhancement cut-offs within the defined ROI, that differentiate the BPE pixels/voxels from the rest of the tissue. This is a practice applied in four out of the twelve enclosed publications [13], [15], [16], including that by Niell et al. [17], who investigated multiple thresholds across intensity-ranges between 0-100% of the FGT volume for BPE estimation.

- **Signal Intensity Based**: Different approaches were observed in this quantification framework [13], [14], [17]–[20], [23], [24]. The baseline equation mostly relied on the following expression:

$$BPE = \frac{S_1 - S_0}{S_0} \tag{1}$$

where S_0 corresponds to the signal in the pre-contrast series, and S_1 to that in the post-contrast. The time point utilized relative to the post-contrast series varied depending on the publication under research. While a grand majority utilized all DCE-MR series [13], [14], [17], [19], others focused on specific time-phases, such as that at 2.5 minutes [18] or the first acquisition after contrast injection [23]. One of the enclosed publications performed the computations directly over the 2nd post-contrast subtraction maximum intensity projection (MIP)

TABLE I: Summary on BPE quantification literature differentiated between signal intensity (SIB) and volume (VB) based models, as described by Müller-Franzes et al. [14]. Each column describes the specifications per paper; in those papers where no unilaterality was stated, bilateral examinations were assumed as input. LR: lesion removal, MTP: maximum enhancement time point.

Paper	Approach	Purpose	Segmentation	Laterality	LR	Input Data
Nguyen et al. [18]	SIB	Neoadjuvant therapy response	FGT	Contralateral	No	DCE-MR
Hu et al. [13]	SIB/VB	Risk of breast cancer	FGT	Bilateral	No	DCE-MR
Niell et al. [17]	SIB	Risk of breast cancer	FGT	Bilateral	No	DCE-MR
Wei et al. [15]	VB	BPE quantification	FGT	Bilateral	No	Pre- and 1st post-contrast
Zhang M. et al. [19]	SIB	Tumor behaviour prediction	FGT	Contralateral	No	DCE-MR
Goodburn et al. [20]	SIB	Risk of breast cancer	FGT	Unilateral	No	Pre-contrast and MTP
Nowakowska et al. [21]	VB	BPE quantification	Breast	Bilateral	No	DCE-MR subtraction
Zhang J. et al. [22]	VB	BPE quantification	FGT	Bilateral	No	Pre- and 1st post-contrast
Zhang B. et al. [23]	SIB	BPE quantification	FGT	Bilateral	No	Pre- and 1st post-contrast
Douglas et al. [24]	SIB	BPE with lesion removal	Breast	Unilateral/Bilateral	Yes	2nd subtraction MIP
Müller-Franzes et al. [14]	SIB/VB	Comparative study	FGT/Breast	Bilateral	No	DCE-MR
Arefan et al. [16]	VB	Risk of recurrence	FGT/Breast	Unilateral	Yes	DCE-MR

volume, a quantification precessed by a normalization of pixel intensities between 0-1 [24]. It must be noted that this is also the only publication in the signal intensity-based quantification category that does not utilize FGT segmentation as an input for subsequent computations on BPE. Another approach made use of the so-called maximum enhancement time point (MTP), described as the post-contrast series in the DCE-MR protocol with the highest mean FGT pixel value [20].

BPE computations were often accompanied by averaging over the total intensity of pixels in the ROI [18], [24], or by providing a description based on percentages [17], [20], [23]. Furthermore, Zhang B. et al. [23] included a BPE integral calculation, shown in equation 2, a computation later on associated with the BI-RADS BPE categories and considered the quantitative imaging biomarker.

$$BPE_{integral} = \sum_{i=3}^{8} \left| \frac{S_1 - S_0}{S_0} \right| \times |\text{FGT area ratio}| \quad (2)$$

The $BPE_{integral}$ was calculated from the BPE histogram obtained after a BPE identification procedure, that utilized time-intensity curves in a similar manner to thresholding. Moreover, i=3 referred to 30-40% enhancement intensities, as i=8 corresponded to an 80-90%.

- **Volume Based:** Six papers are included within this section [13]–[16], [21], [22], the baseline equations consisting of:

$$BPE = \frac{V_{enhanced}}{V_{ROI}} \quad (3)$$

V_{ROI} is the volume of interest, determined by the FGT mask or the breast during the preprocessing segmentation step. $V_{enhanced}$ corresponds to the BPE, identified through thresholding of the ROI. The utilized series ranged between the entire DCE-MR protocol [13]–[16], to subtraction [21] or pre-/post-contrast series [22]. Once again, the formula was utilized directly as a BPE ratio [16], [22], whereas some other cases obtained a percentage over the final result [15], [21].

A large variability in the utilization of unilateral or bilateral examinations was observed among the listed studies enclosed in TABLE I. More commonly, bilateral assessments were performed in papers that investigated BPE as an imaging biomarker for breast cancer or its quantification, whereas unilateral/contralateral studies applied their methodologies on cancer-patient datasets. On this regard, two studies conducted ipsilateral lesion removal techniques, replacing malignant pixels in the affected breast. While Arefan et al. [16] described a semiautomatic approach based on radiologist supervision, Douglas et al. [24] applied fuzzy c-means clustering techniques for region identification and subsequent substitution of the malignant pixels with average intensity values. The investigation developed by the latter demonstrated an overestimation of BPE in exams with malignant regions of about 1.48% to 3.83%, a phenomenon that mostly affected those examinations with larger lesions and lower BPE levels [24].

Hu et al. [13] utilized both methodologies, signal intensity and volume based computations, in their study for cancer prediction with BPE as biomarker. Although not aligned with the focus of this review, they found a stronger correlation with volume-based than signal intensity-based quantifications.

Müller-Franzes et al. [14] performed a study analyzing the correlation between the two main quantitative methodologies previously presented, which were attributed to the most common frameworks. Moreover, it created a comparative study that performed discretization of BPE estimated results to investigate the classification capabilities of such quantification approaches.

The study encompassed 5773 examinations from 4886

TABLE II: Müller-Franzes et al. [14] results per methodology and approach. Literature refers to papers enclosed within BPE quantification that utilized similar approaches to each specific method. Quantitative Correlation (QC) is referred to the Spearman rank correlation coefficient and the Classification Agreements (CA) to the linear weighted Cohen kappa coefficient. Both of these computations are performed per methodology with respect to the expert-annotated reference.

Method	Approach	Literature	QC	CA
1	SIB	[18], [20], [24]	0.56 ± 0.01	0.47 ± 0.01
2	SIB	[17], [23]	0.55 ± 0.01	0.46 ± 0.01
3	VB	[16], [22]	0.52 ± 0.01	0.46 ± 0.01
4	VB	[15], [16], [21]	0.50 ± 0.01	0.38 ± 0.01

women, selected after eligibility assessments based on data quality and sufficient FGT and breast volume. Although segmentations were performed through previously evaluated CNNs, this approach aimed to minimize error propagation from mask outliers. Moreover, women with implants were excluded, and only those patients without a history of breast cancer were recruited to prevent BPE result estimation bias from malignant lesions. Imaging was performed at a single institution, with two different 1.5T scanners, and based of gradient-echo DCE-MR protocols, which included a single pre-contrast and up to four post-contrast series. Finally, FGT segmentation was performed before quantification, and original BPE assessments by expert radiologists, obtained during clinical interpretation, were utilized to compare qualitative and automated methodologies.

Results are presented in TABLE II, where each paper enclosed in TABLE I is connected to its counterpart in the study by Müller-Franzes et al. [14]. It consists of four methods, two of which were based on signal-intensity computations (1 and 2) and the other two (3 and 4) on volume-based formulas. The results showed a moderate correlation between quantitative and qualitative assessments by radiologists, regardless of the approach used, with the highest Spearman rank coefficient of 0.56 ± 0.01 attributed to method 1 and the lowest of 0.50 ± 0.01 to method 4. Furthermore, discretization of examinations per BI-RADS category (minimal/mild/moderate/marked) was performed utilizing ROC curves and an optimization process based on maximal agreement. The subsequent linear Cohen kappa coefficients showed once again moderate correlation to the assessments performed by radiologists; the highest results corresponding to a 0.47 ± 0.01, and the lowest to a 0.38 ± 0.01 for methods 1 and 4, respectively.

B. Machine Learning

Machine learning methodologies included both the utilization of radiomics and DL for BPE classification, shown in TABLE III.

Firstly, the work by Nam et al. [26] applied radiomics on a dataset enclosing 794 patients between the ages of 26 and 89, who had undergone preoperative MR. Bilateral imaging was performed with 3 types of MR scanners, and the utilized protocol was based on DCE-MR acquisitions. The pipeline considered several automated applications of ML architectures; V-Net segmentation of FGT, followed by radiomics for BPE classification. 59 features were processed for radiomics classification based on the intensity, shape and texture properties of the images, a process that required prior BPE voxel identification from the FGT masks applied to the subtraction series of the DCE-MR protocols. Moreover, three classification models were analyzed, which aimed to classify the exams either based on the four categories of the BI-RADS lexicon or creating a model that would differentiate minimal versus mild/moderate/marked or minimal/mild versus moderate/marked. Results showed statistical analysis with similar results in classification between the manual and the DL segmentation under study, which ensured good agreement and low error propagation for subsequent steps. More importantly, it showed the highest performing model to be that differentiating between minimal/mild versus moderate/marked classes, with an accuracy of 91.5% for the manual segmentation, and 90.5% for the deep learning approach. In the case of the four BI-RADS categories, the corresponding accuracies were 66% and 67%, respectively, while the remaining model retrieved values of 72.5% and 72%.

Another work by Wang et al. [28] equally utilized radiomics for their BPE classification study. The dataset included examinations from 3705 patients originally recruited for the DENSE trial, consisting of women between the ages of 50-75 with extremely dense breasts. In this case, imaging acquisitions at 8 different institutions made use of 3.0T scanners, and DCE-MR series with presence and absence of fat-suppression techniques. FGT segmentation was similarly performed through a nnU-Net architecture, and the subsequent mask was applied onto the radiomics algorithm for feature extraction. In this case, volumetric density, morphologic, and enhancement characteristics were used as such, which summed up to a total of 15 features, and a combination of random forest, Naïve bayes and k-NN classifiers were utilized for BPE categorization. As this was an inter-institutional study, accuracy results were computed per hospital, leading to majority voting values that ranged between 56% to 84%. More importantly, a lack of significant differences in results between institutions was observed, despite the use of a variety of scanners and imaging parameters. Moreover, while sensitivities were found to be higher for the minimal BPE category (95%) when compared to the rest (mild: 16%; moderate: 13%; marked: 36%), its corresponding specificity showed to be significantly lower than the other 3 categories (minimal: 46%; mild: 95%; moderate: 96%; marked: 97%).

Furthermore, two publications did not make use of FGT segmentation for BPE classification, but rather deployed CNN architectures to perform BPE categorization from the DCE-MR inputs. In 2020, Borkwoski et al. [25] utilized two VGG-16 models with densely connected layers; one to perform an initial breast-slice recognition from the complete volume, and

TABLE III: Summary on the literature for BPE categorization, differentiated between radiomics and deep learning (DL) based models. Dataset describes the number of patients per study. DA: Data Augmentation.

Paper	Approach	Segmentation	Laterality	Input Data	Dataset	DA	Institutions
Borkwoski et al. [25]	DL	-	Bilateral	1st subtraction	124	Yes	1
Nam et al. [26]	Radiomics	FGT	Contralateral	DCE-MR	794	No	1
Eskreis-Winkler et al. [27]	DL	Breast	Unilateral	1st subtraction MIP	3705	Yes	1
Wang et al. [28]	Radiomics	FGT	Bilateral	DCE-MR	4553	No	8

another for a direct BPE classification into minimal, mild, moderate and marked. For that, a data collection of 9902 examinations from 794 patients, was utilized for training, validation and testing. Moreover, their methodology described the utilization of T1-weighted DCE-MR subtraction series differentiated into 2D images as an input, obtained on a 3T MR scanner at a single institution, and to which cropping and normalization preprocessing was applied. The training of the network on more than a thousand examinations led to an accuracy of 75% on the test set for BPE classification. These results were found to be related to class activation maps, which imaged the model's decision making in those areas with the most relevant information, thus strengthening the model's performance. Furthermore, a majority of misclassifications were found to occur between adjacent categories, and it was discovered that the inter-rater reliability expressed through Cohen's Kappa coefficient was highest for the model (0.815) than for the assessment of either of the two experienced readers (reader 1: 0.68, reader 2: 0.78) obtained for clinical validation. Limitations such as lower recurrence of low BPE categories were averted through transfer learning and data augmentation.

Eskreis-Winkler et al. [27] similarly utilized a VGG-19 for the same purpose. While Borkwoski et al. [25] utilized a slice-wise subtraction input approach, this most recent work aimed to compare the performance of two different approaches of their model on subtraction series (T1-weighted pre-contrast and first post-contrast). A collection of 5224 images were obtained from 3705 patients, in imaging acquisition protocols that obtained at their institution with 3T and 1.5T MR scanners. Each DCE-MR was normalized, divided into unilateral breasts, and processed either as standard 2D MIPs or within a slab framework (Slab AI), which divided the breast tissue into upper, middle and lower 2D MIPs. Furthermore, while the model was trained in a four-way classification, testing was performed in pooled groups that divided the categories into high (moderate/marked) and low (minimal/mild) BPE label. Their results showed that, the so-called, Slab AI methodology managed to outperform the standard 2D MIP throughout all statistical analyses. Similarly, area under the curve (AUC) analysis retrieved values of 0.84 for the former, while that of the latter was 0.79. Moreover, they underscore the model's 5.8% greater tendency to assign exams with a higher probability of malignancy (as indicated in the radiology reports by the BI-RADS assessment value) to moderate or marked labels, when compared to radiologists. On the contrary, those high BPE categories were 6.9% less likely to be assigned to non-suspicious exams, suggesting the clinical relevance of the study towards breast cancer prediction.

V. Discussion

The importance of standardized and automated categorization for BPE has been previously highlighted, with a primary focus relying on its impact on decreased breast MR sensitivity and its hypothesized connection to breast cancer prediction [4]. In the case of the latter, for instance, comparable BPE categorization methodologies may allow clinically relevant associations between BPE and breast cancer risk among various institutions, and thus validate its applicability [14]. This paper intends to perform an exhaustive study of the most recent frameworks in the literature, that generates a comprehensive overview of these developments, and provides an understanding of the benefits associated with each methodology.

In the selected time frame, 2020-2024, three main BPE categorization pipelines were investigated; BPE quantification, radiomics and DL. BPE quantification methodologies stand out for its simplicity in development and low computational demands when compared to ML techniques, attributed to complex architectures and large dataset requirements. This may be considered a key point in the selection of the most appropriate methodology towards standardization, however important findings were described in this review.

BPE quantification techniques may be described through signal intensity and volume based equations, which typically make use of the complete DCE-MR protocol or the first post-contrast and pre-contrast series. Müller-Franzes et al. [14] provided a comparative study on the capacities of both approaches, highlighting the significant limitations associated of this categorization pipeline. Correlation with ground truth was moderate in all four automated examined, for both quantitative and discretized results. The low variability in the qualitative assessment by radiologists, rules out poor-quality references as a factor to the inferior performance of the methodologies. Thus, it indicated the limited capability of BPE quantification techniques to perform highly accurate and comparable assessments. Moreover, no major differences were found in the number of publications that applied one method or the other, reflecting the lack of consensus on the superior approach. This is further emphasized by the absence of guideline for methodology selection when considering the purpose of the study, as shown in TABLE I.

Furthermore, the investigation on ML frameworks included both radiomics and DL architectures. Two publications applied radiomic frameworks to their data after FGT segmentation, Nam et al. [26] and Wang et al. [28]. Such algorithms extract features from the images, instead of solely analyzing voxel/pixel intensities, that include volumetric morphology, dissimilarity or enhancement properties of the data, among others [9]. This approach directly deals with the complications in classification associated with the large heterogeneous distribution of BPE [26], a limitation observed in quantification methodologies as described by Müller-Franzes et al. [14].

The main differences between the two studies lay in the number of features collected for each study, 59 by Nam et al. [26] against 15 by Wang. et al [28], and the classification approaches; each providing different insights into the applicability of radiomics. To this end, the curse of dimensionality commonly observed in radiomics must be noted, associated with overfitting and frequent use of irrelevant features that reduce its efficiency [29]. Comparing the performance of the two reports is challenging due to the intrinsic differences in protocols and architecture; however, the highest accuracies achieved by each study offers some valuable insight. While that by Nam et al. [26] achieved a 67% accuracy utilizing a four-way categorization model, that by Wang et al. [28] reported values ranging from 56% to 84% across 8 institutions, which might reflect the greater efficiency reached by the latter based on the selected features. These results highlight the importance of proper feature selection to represent properly the characteristics of the breast when considering radiomics for BPE classification.

Other ML approaches utilized VGG classifiers for BPE categorization. The work by Borkowski et al. [25] provided an initial insight into the capabilities of DL, as demonstrated by the higher correlation coefficient of the model compared to the two radiologists, when referenced against the dataset obtained through their consensus. The utilization of such technologies is recommended for standardized assessments of BPE, as it limits the subjectivity intrinsic to human interpretation [25]. Furthermore, it introduces the utilization of 2D inputs, in contrast to the 4D volumes typically acquired during breast MR. This is explained by computational efficiency; 2D inputs allow the selection of slices where the breast is depicted and focuse the attention maps of the models into the desired ROI, instead of wasting resources on the entire field-of-view. For this purpose, Eskreis-Winkler et al. [27] proposes the utilization of MIPs, against the development of slice-wise architectures, which would ensure retention of all the spatiotemporal information into computationally efficient forms of the otherwise dense 4D DCE-MR series [27]. Therefore, regardless of the approach used to translate the 4D volumes into 2D images, this presents as a desired characteristic for future ML architectures in the development of standardized methodologies for BPE categorization.

Following the aforementioned advantages and disadvantages for each approach, the path towards automated standardization of BPE categorization shows to be directed towards the utilization of ML techniques. While BPE quantification plus discretization shows slight advantages, the potential superiority of ML techniques for BPE categorization is observed in this review, as demonstrated by the greater accuracy and agreement in results. While the agreement showed by Müller-Franzes et al. [14] peaked at 0.47, that reported in the study by Borkowski et al. [25] reached an 0.82 in its VGG-16 classification model. The comparative study between ML techniques becomes complicated as a consequence of the large differences in the applicability of their methods and statistical analysis. Nonetheless, the increased performance achieved by the Slab AI architecture described by Eskreis-Winkler et al. [27] enhances the importance of categorization models that retrieve significant information from relevant areas of the breast. To this matter, radiomics may present an advantage to DL techniques, as it allows a selection of features tailored to the requirements of the task. Still, this statement must consider the need for FGT segmentation in radiomics, and the intrinsic dependence and potential error propagation it might be linked to.

A large majority of the techniques discussed in this paper commonly required FGT segmentation, regardless of the nature of the approach. Automated and validated frameworks for this purpose may be of utmost importance; BPE is described by the regions with an enhancement of normal FGT after contrast administration, consequently becoming intrinsically dependent on the accuracy of the segmentations. Some publications proposed pipelines that circumvented the utilization of the FGT masks, such as in case of Nowakowska et al. [21], that might present wide advantages in its methodology towards improved quantifications. Nonetheless, eliminating this processing step becomes complicated in the case of radiomics, which is dependent of such ROI definitions. Therefore, the importance of an appropriate selection of FGT segmentations frameworks must be noted to ensure sufficient accuracy for subsequent ML pipelines in the future of standardization.

Furthermore, the utilization of the first post-contrast has been noted as the gold-standard for BPE assessment by radiologists following the BI-RADS lexicon. While the utilization of the aforementioned series allows a superior differentiation between malignant lesions and normal tissue, Goodburn et al. [20] hypothesized the associated quantitative error in clinical assessment of cancer-free patients. Some women present steady FGT enhancement in their examinations; this phenomenon is considered to be an indicative of bias and thus a potential cause of BPE underestimation. Consequently, they considered the utilization of MTP series, against the common standard, to ensure maximal enhancement for the assessment in all types of patients. Standardization of BPE assessment should be clinically validated for both cancerous and non-cancerous women. Thus, employing this approach may offer advantages compared to the alternative, which relies on the assessment of entire DCE-MR series and is limited by the increased workload it presents for radiologists.

Finally, because of limitations commonly encountered among the three approaches investigated, automated standardization may be subjected to the variability in breast MR protocols between institutions. BPE enhancement is dependent on the pulse sequences, scanner or timing of the examination [21]. Thus, a tailored protocol would be required per institution to account for this variations, which might limit the longitudinal comparison of BPE categorization. In the case of BPE quantification, the task becomes significantly complex, as enhancement and discretization thresholds would need to be investigated per establishment. Still, the limitation may be averted through larger datasets for radiomics and DL, that would include examinations from a wide variety of institutions and acquisition protocols.

Furthermore, ML approaches are restricted to the intrinsic bias attributed to the distribution of BPE levels in the population, commonly described by minimal or mild BPE levels [30]. Subsequently, data collections are expected to be unbalanced if retrieved from normal clinical settings. The effect of this limitation may be observed in the study by Wang et al. [28], which showed the superiority in sensitivity for lower BPE categories, 95% minimal versus 36% marked, and an opposite effect with respect to the specificity, being that for marked breasts the highest. These results may be explained by a higher probability of the classifier to assign examinations with lower BPE labels, as these were the categories most represented in their dataset; a 75% of the examinations corresponded to minimal BPE while that for marked was of 4%. The unbalanced characteristics of their data potentially derived in a bias in their classifier towards the lower BPE categories. Nam et al. [26] averted the limitation through a direct classification utilizing the radiomic features in a ML algorithm to prevent DL bias towards dominant classes, this was described as one of the main limitations in the work by Wang et al. [28]. Though the possibility of substitutes to the application of data augmentation techniques is possible, this should be considered of outermost importance to reduce possible bias and ensure architectures are generalizable to the entire population.

Future outlooks could also explore determining whether examinations exhibit symmetric or asymmetric BPE, for example, by assigning differentiated BPE categories to each breast. Slight variations in FGT quantity are expected with moderation, nonetheless, the clinical relevance of highly asymmetric breasts and its potential association with suspicious findings has been stated in the literature [30]. Throughout this review, it was observed that publications investigating the applicability of their algorithms on datasets with cancer patients typically focused on either unilateral or contralateral examinations. The work by Douglas et al. [24] confirmed the association of BPE overestimation in ipsilateral breasts as a result of the higher intensity values attributed to tumors. New algorithms may implement some of the discussed lesion removal approaches in malignant exams to ensure accurate BPE estimation, followed by BPE categorization per breast for a more comprehensive assessment. It must be noted the standard practice described by the BI-RADS lexicon, which recommends differentiating between breasts and characterizing an examination with the highest BPE result in asymmetric breasts [6]. Nonetheless, the proposed outlook could consider this guideline through the definition of two separate labels per exam; one that represents the presence or absence of breast symmetry, while another provides a definite label based on BI-RADS assessments.

Moreover, an important observation was made related to the classification performance per BPE category. Nam et al. [26], investigated the performance of its model in two manners; category pooling (minimal/mild versus moderate/marked) and (minimal versus mild/moderate/marked) against a four-way categorization. The highest performing algorithm was that differentiating between "high" and "low" categories, which surpassed that for all BI-RADS categories by a 18.5%. Moreover, the work by Borkowski et al. [25] showed that most misclassifications were attributed to adjacent categories, what highlights the results by Nam et al. Therefore, a future outlook towards automated standardized BPE categorization may lie in two-way classifiers. This allows ML classifiers to prioritize the identification of exams with a higher probability of breast cancer risk and lesion occlusion. Radiologists would ensure a more extensive assessment of such cases, following the described characteristics of BPE as an imaging biomarker, thereby effectively distributing their workload. Additionally, more accurate models that simplify BPE categorization could be developed, leading the way towards more reliable models for a standardized and automated methodology.a

VI. Conclusion

The current state of BPE categorization is marked by the development of a wide variety of methodologies in the recent years, making the path towards automated standardization in the clinical practice more complex. Though far from the goal, some important observations from this study could guide future efforts.

The superior accuracy and correlation coefficients of ML techniques for BPE categorization highlights their potential as the primary focus for future developments. More importantly, radiomics, with its feature-selection framework, may allow BPE categorization to be tailored to tissue characteristics of the breast, thereby enhancing model accuracy. Still, emphasis should be remain on validated segmentation techniques that ensure accurate BPE assessments. Moreover, future research towards standardized methodologies should highlight the importance of balanced datasets and adaptability across variable breast MR protocols, to achieve full generalization. Other aspects, such as the utilization of 2D input frameworks and variable post-contrast series time points, are discussed as potential areas for improvements. Finally, future studies could explore BPE symmetry between breasts and the adaptability of two-way categorizations towards a fully comprehensive and more efficient application of BPE categorization in the clinic.

REFERENCES

[1] World Health Organization, "Breast cancer," 2024. Accessed: 2024-17-11.

[2] D. Barba, A. León-Sosa, P. Lugo, D. Suquillo, F. Torres, F. Surre, L. Trojman, and A. Caicedo, "Breast cancer, screening and diagnostic tools: All you need to know," *Critical reviews in oncology/hematology*, vol. 157, p. 103174, 2021.

[3] R. M. Mann, N. Cho, and L. Moy, "Breast mri: state of the art," *Radiology*, vol. 292, no. 3, pp. 520–536, 2019.

[4] G. J. Liao, L. C. Henze Bancroft, R. M. Strigel, R. D. Chitalia, D. Kontos, L. Moy, S. C. Partridge, and H. Rahbar, "Background parenchymal enhancement on breast mri: a comprehensive review," *Journal of Magnetic Resonance Imaging*, vol. 51, no. 1, pp. 43–61, 2020.

[5] C. D'Orsi, L. Bassett, S. Feig, *et al.*, "Breast imaging reporting and data system (bi-rads)," *Breast imaging atlas, 4th edn. American College of Radiology, Reston*, 2018.

[6] A. C. of Radiology, "BI-RADS Poster," 2023. Retrieved from https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS-Poster.pdf.

[7] A. C. of Radiology, "MRI Reporting," 2023. Retrieved from https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/MRI-Reporting.pdf.

[8] A. C. Pujara, A. Mikheev, H. Rusinek, Y. Gao, C. Chhor, K. Pysarenko, H. Rallapalli, J. Walczyk, M. Moccaldi, J. S. Babb, *et al.*, "Comparison between qualitative and quantitative assessment of background parenchymal enhancement on breast mri," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 6, pp. 1685–1691, 2018.

[9] J. E. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection," *Insights into imaging*, vol. 11, no. 1, p. 91, 2020.

[10] M. Kim, J. Yun, Y. Cho, K. Shin, R. Jang, H.-j. Bae, and N. Kim, "Deep learning in medical imaging," *Neurospine*, vol. 16, no. 4, p. 657, 2019.

[11] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.

[12] G. Müller-Franzes, F. Müller-Franzes, L. Huck, V. Raaff, E. Kemmer, F. Khader, S. T. Arasteh, T. Lemainque, J. N. Kather, S. Nebelung, *et al.*, "Fibroglandular tissue segmentation in breast mri using vision transformers: a multi-institutional evaluation," *Scientific Reports*, vol. 13, no. 1, p. 14207, 2023.

[13] X. Hu, L. Jiang, C. You, and Y. Gu, "Fibroglandular tissue and background parenchymal enhancement on breast mr imaging correlates with breast cancer," *Frontiers in Oncology*, vol. 11, p. 616716, 2021.

[14] G. Müller-Franzes, F. Khader, S. Tayebi Arasteh, L. Huck, M. Bode, T. Han, T. Lemainque, J. N. Kather, S. Nebelung, C. Kuhl, *et al.*, "Intraindividual comparison of different methods for automated bpe assessment at breast mri: a call for standardization," *Radiology*, vol. 312, no. 1, p. e232304, 2024.

[15] D. Wei, N. Jahani, E. Cohen, S. Weinstein, M.-K. Hsieh, L. Pantalone, and D. Kontos, "Fully automatic quantification of fibroglandular tissue and background parenchymal enhancement with accurate implementation for axial and sagittal breast mri protocols," *Medical physics*, vol. 48, no. 1, pp. 238–252, 2021.

[16] D. Arefan, M. L. Zuley, W. A. Berg, L. Yang, J. H. Sumkin, and S. Wu, "Assessment of background parenchymal enhancement at dynamic contrast-enhanced mri in predicting breast cancer recurrence risk," *Radiology*, vol. 310, no. 1, p. e230269, 2024.

[17] B. L. Niell, M. Abdalah, O. Stringfield, N. Raghunand, D. Ataya, R. Gillies, and Y. Balagurunathan, "Quantitative measures of background parenchymal enhancement predict breast cancer risk," *American Journal of Roentgenology*, vol. 217, no. 1, pp. 64–75, 2021.

[18] A. A.-T. Nguyen, V. A. Arasu, F. Strand, W. Li, N. Onishi, J. Gibbs, E. F. Jones, B. N. Joe, L. J. Esserman, D. C. Newitt, *et al.*, "Comparison of segmentation methods in assessing background parenchymal enhancement as a biomarker for response to neoadjuvant therapy," *Tomography*, vol. 6, no. 2, p. 101, 2020.

[19] M. Zhang, M. Sadinski, D. Haddad, M. S. Bae, D. Martinez, E. A. Morris, P. Gibbs, and E. J. Sutton, "Background parenchymal enhancement on breast mri as a prognostic surrogate: correlation with breast cancer oncotype dx score," *Frontiers in Oncology*, vol. 10, p. 595820, 2021.

[20] R. Goodburn, E. Kousi, C. Sanders, A. Macdonald, E. Scurr, C. Bunce, K. Khabra, M. Reddy, L. Wilkinson, E. O'Flynn, *et al.*, "Quantitative background parenchymal enhancement and fibro-glandular density at breast mri: Association with brca status," *European Radiology*, vol. 33, no. 9, pp. 6204–6212, 2023.

[21] S. Nowakowska, K. Borkowski, C. M. Ruppert, A. Landsmann, M. Marcon, N. Berger, A. Boss, A. Ciritsis, and C. Rossi, "Generalizable attention u-net for segmentation of fibroglandular tissue and background parenchymal enhancement in breast dce-mri," *Insights into Imaging*, vol. 14, no. 1, p. 185, 2023.

[22] J. Zhang, Z. Cui, L. Zhou, Y. Sun, Z. Li, Z. Liu, and D. Shen, "Breast fibroglandular tissue segmentation for automated bpe quantification with iterative cycle-consistent semi-supervised learning," *IEEE Transactions on Medical Imaging*, 2023.

[23] B. Zhang, J. Zhu, P. Zhang, Y. Wei, Y. Li, A. Xu, Y. Zhang, H. Zheng, X. Dong, K. Yang, *et al.*, "A background parenchymal enhancement quantification framework of breast magnetic resonance imaging," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 12, p. 8350, 2023.

[24] L. Douglas, J. Fuhrman, Q. Hu, A. Edwards, D. Sheth, H. Abe, and M. Giger, "Computerized assessment of background parenchymal enhancement on breast dynamic contrast-enhanced-mri including electronic lesion removal," *Journal of Medical Imaging*, vol. 11, no. 3, pp. 034501–034501, 2024.

[25] K. Borkowski, C. Rossi, A. Ciritsis, M. Marcon, P. Hejduk, S. Stieb, A. Boss, and N. Berger, "Fully automatic classification of breast mri background parenchymal enhancement using a transfer learning approach," *Medicine*, vol. 99, no. 29, p. e21243, 2020.

[26] Y. Nam, G. E. Park, J. Kang, and S. H. Kim, "Fully automatic assessment of background parenchymal enhancement on breast mri using machine-learning models," *Journal of Magnetic Resonance Imaging*, vol. 53, no. 3, pp. 818–826, 2021.

[27] S. Eskreis-Winkler, E. J. Sutton, D. D'Alessio, K. Gallagher, N. Saphier, J. Stember, D. F. Martinez, E. A. Morris, and K. Pinker, "Breast mri background parenchymal enhancement categorization using deep learning: outperforming the radiologist," *Journal of Magnetic Resonance Imaging*, vol. 56, no. 4, pp. 1068–1076, 2022.

[28] H. Wang, B. H. van der Velden, E. Verburg, M. F. Bakker, R. M. Pijnappel, W. B. Veldhuis, C. H. van Gils, and K. G. Gilhuijs, "Automated rating of background parenchymal enhancement in mri of extremely dense breasts without compromising the association with breast cancer in the dense trial," *European Journal of Radiology*, vol. 175, p. 111442, 2024.

[29] W. Zhang, Y. Guo, and Q. Jin, "Radiomics and its feature selection: A review," *Symmetry*, vol. 15, no. 10, p. 1834, 2023.

[30] C. S. Giess, E. D. Yeh, S. Raza, and R. L. Birdwell, "Background parenchymal enhancement at breast mr imaging: normal patterns, diagnostic challenges, and potential for false-positive and false-negative interpretation," *Radiographics*, vol. 34, no. 1, pp. 234–247, 2014.