IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 May 2026, accepted 29 May 2026, date of publication 1 June 2026, date of current version 9 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3699018

RESEARCH ARTICLE

A Systematic Interaction Analysis of Distance Measures and Feature Representations in Mammography

MUHAMMAD IQBAL[1], TALAL ASHRAF BUTT[1], ZAHID SHAREEF[1], ABUBAKAR SIDDIQUE[2], (Member, IEEE), AND WILL N. BROWNE[3], (Member, IEEE)

[1]Higher Colleges of Technology, Fujairah, United Arab Emirates
[2]Auckland University of Technology, New Zealand
[3]Queensland University of Technology, Brisbane, QLD, Australia

Corresponding author: Muhammad Iqbal (miqbal1@hct.ac.ae)

ABSTRACT Distance-based classifiers such as k-nearest neighbors remain widely used in mammographic computer-aided diagnosis and image retrieval, yet distance measure selection is rarely evaluated systematically. This paper investigates how the interaction between feature representation and distance formulation shapes mammographic lesion classification, evaluating 20 distance measures across 13 feature sets (handcrafted, deep, hybrid), three datasets (MIAS, CBIS-DDSM, INbreast), and two classifiers (k-NN, nearest mean-distance), spanning 13,260 configurations under 5-fold cross-validation. Five findings emerge. First, distance choice significantly affects performance (Friedman $\chi^2 = 1901$, $p \approx 0$; best–worst AUC gap = 0.079). Second, the optimal measure depends on feature type: handcrafted features favor Learned Mahalanobis, deep features favor Adaptive Weighted Euclidean, and hybrid features favor the proposed CCFCD (AUC = 0.712). Third, measure rankings are dataset-specific (cross-dataset ρ as low as -0.10). Fourth, k-NN and NMD agree at the configuration level ($\rho = 0.94$) but not at the measure level ($\rho = 0.077$). Fifth, increasing k improves AUC while reducing sensitivity from 0.40 to 0.14; 11.6% of configurations achieve accuracy above 80% with sensitivity below 10%, showing that AUC and accuracy alone are insufficient for clinical assessment. A computational complexity analysis shows that per-pair costs range from $O(d)$ for most measures to $O(d^2)$ for Mahalanobis, with the top-performing adaptive measures remaining tractable. These findings demonstrate that distance-based mammographic classification is governed by a coupled interaction among feature type, distance formulation, classifier, neighborhood size, and evaluation metric. Adaptive and class-aware measures consistently outperform conventional defaults when matched to the feature space.

INDEX TERMS Mammography, breast cancer classification, distance measures, similarity measures, nearest neighbor classification, k-nearest neighbors, feature representation, hybrid features, deep features, handcrafted features, class imbalance, balanced accuracy.

I. INTRODUCTION

Breast cancer remains the most frequently diagnosed cancer among women worldwide, with an estimated 2.3 million new cases in 2022 [1]. Early detection through mammographic screening remains the primary strategy for reducing mortality. Computer-aided diagnosis (CAD) systems can support radiologists by automatically classifying mammographic lesions as benign or malignant [2], [3], [4]. Such systems may reduce inter-observer variability [5] and serve as a second reader by flagging suspicious cases that might otherwise be missed.

Many mammographic CAD and content-based image retrieval (CBIR) systems rely on distance-based classification,

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

in which similarity between a query image and reference cases is quantified by applying a distance function to extracted feature vectors [6], [7]. The k-nearest neighbor (k-NN) classifier [8] is the most widely used distance-based method in mammography. It classifies a query by identifying the k training samples with the smallest distance and assigning the majority class among those neighbors. A complementary approach is the Nearest Mean-Distance (NMD) classifier [9], which assigns a query to the class with the smallest average distance to the samples of that class. In this study, k-NN and NMD represent two distinct aggregation strategies over the same pairwise distances: k-NN uses local neighbor-based aggregation, whereas NMD uses global class-conditional aggregation. Distance-based approaches remain attractive in clinical decision support because predictions can be related to similar prior cases, their decision mechanism aligns naturally with how radiologists compare lesions against prior experience, and new reference cases can be incorporated without retraining. In CBIR systems, the same distance function also directly determines retrieval quality.

The effectiveness of any distance-based system depends jointly on two components: the *feature representation* and the *distance measure*. Feature representations for mammography have evolved substantially. Earlier CAD systems relied on handcrafted descriptors such as the Gray-Level Co-occurrence Matrix (GLCM) [10], Local Binary Patterns (LBP) [11], and Histogram of Oriented Gradients (HOG) [12]. More recently, deep features extracted via transfer learning from pretrained Convolutional neural networks (CNNs) [13] have become dominant, and hybrid representations that concatenate handcrafted and deep features are increasingly common [14], [15]. These feature types differ markedly in dimensionality, sparsity, scale, distributional structure, and semantic interpretation. Consequently, the appropriateness of a distance measure cannot be assessed independently of the feature space in which it operates.

Despite this dependence, distance selection in mammography has received far less attention than feature design. Many studies adopt Euclidean distance by default, often without systematic comparison against alternatives [16], [17], [18], [19]. This practice warrants scrutiny because different measures encode fundamentally different notions of similarity. Euclidean distance emphasizes magnitude differences, cosine distance emphasizes angular separation, Mahalanobis distance accounts for feature correlations, and histogram-based measures compare distributional profiles. There is no *a priori* reason to expect that one measure will remain optimal across low-dimensional handcrafted descriptors, high-dimensional CNN activations, and heterogeneous hybrid vectors.

This issue becomes more pronounced in high-dimensional settings. Beyer et al. [20] and Aggarwal et al. [21] showed that the discriminative power of distance measures degrades in high-dimensional spaces, with the rate of degradation depending on the choice of norm. In mammography, feature dimensionalities in our study range from 30-dimensional morphological descriptors to hybrid vectors exceeding 6,000 dimensions. This range spans a regime in which distance concentration can materially affect both neighbor ranking and class assignment. Hybrid feature spaces are particularly challenging because handcrafted and deep subvectors coexist with different scales, variances, and statistical properties.

Several studies have benchmarked distance measures for k-NN on general-purpose datasets. Alfeilat et al. [22] compared 54 measures on 28 UCI datasets and found that no single measure dominated across tasks. Hu et al. [23] reported similar conclusions for medical tabular datasets. Ehsani and Drabløs [24] showed that Manhattan and Hassanat distances outperformed Euclidean on cancer data, again indicating that the default choice is often suboptimal. These studies provide important evidence that distance selection matters. However, their scope remains limited from the perspective of mammographic image analysis: they use tabular data with comparatively homogeneous features, do not include deep or hybrid representations, evaluate only a single classifier (k-NN), and do not address the imaging characteristics of mammograms. More importantly, they do not examine whether the ranking of distance measures changes systematically with feature type. This feature–measure interaction is central in mammography because handcrafted, deep, and hybrid representations occupy substantially different geometric and statistical spaces. To the best of our knowledge, this interaction has not been studied systematically in mammographic lesion classification across a broad set of feature types and distance measures.

A second gap concerns the robustness of conclusions drawn from any single distance-based classifier. A distance measure that performs well under the local aggregation mechanism of k-NN may not remain equally effective under a global aggregation rule such as NMD. Distinguishing between these cases is important because agreement across classifiers would indicate that observed performance differences arise primarily from the underlying feature–measure pairing rather than from a particular decision rule. Similarly, conclusions drawn from a single dataset may not generalize across mammography collections that differ in imaging modality, sample size, and class distribution.

A third gap concerns evaluation methodology. Mammography datasets are often imbalanced, with benign cases substantially outnumbering malignant cases. In this setting, accuracy can be misleading: a classifier may achieve a high overall accuracy while failing to detect malignant cases reliably. AUC remains informative, but because it summarizes performance across thresholds, it may not fully reflect the clinically relevant trade-off at the operating point used in practice. In screening and diagnosis, the central objective is to detect as many malignant cases as possible while keeping false alarms within an acceptable range. For this reason, sensitivity, specificity, balanced accuracy,

and F_1-score must be considered alongside accuracy and AUC when assessing the practical value of a distance-based classifier.

To address these gaps, this paper investigates how mammographic lesion classification performance is shaped by the interaction among feature representation, distance measure, classifier aggregation strategy, dataset, and evaluation metric. The study is organized around six research questions. The first three examine whether distance-measure effectiveness depends systematically on feature type and whether the resulting feature–measure pairings remain consistent across datasets. The remaining three assess the robustness of these conclusions with respect to classifier aggregation strategy, neighborhood size, and evaluation under class imbalance.

Primary: Feature–Measure Interaction

RQ1. *Does the choice of distance measure significantly affect mammographic classification performance, and if so, which measure families perform best?*
This question establishes whether distance selection materially influences mammographic classification rather than acting as a secondary implementation detail.

RQ2. *Does the optimal distance measure depend on the feature type (handcrafted, deep, or hybrid), and if so, how?*
This is the central question of the study. Handcrafted, deep, and hybrid representations differ substantially in dimensionality, scaling, and statistical structure, which may favor different notions of similarity. We note that feature type serves as a practical proxy for the underlying data characteristics that govern distance behavior: handcrafted descriptors tend to be low-dimensional with interpretable axes, deep features are high-dimensional with correlated activations, and hybrid vectors combine heterogeneous statistical regimes. Although a purely theoretical formulation might ask which distance suits a given distributional geometry (e.g., linear separability or covariance structure), feature type is the design variable that practitioners actually control when building a mammographic pipeline, making it the most actionable framing for empirical guidance.

RQ3. *Are the best-performing feature–measure combinations consistent across mammography datasets with different imaging characteristics?*
This question examines whether conclusions about effective feature–measure pairings remain stable across datasets that differ in modality, class distribution, and sample size.

Secondary: Methodology and Clinical Relevance

RQ4. *Are distance measure rankings consistent across different aggregation strategies: local (k-NN) versus global (NMD)?*
Agreement across these two classifiers would indicate that the observed rankings primarily reflect properties of the feature–measure pairing rather than only a specific decision rule.

RQ5. *How does the neighborhood size k interact with the distance measure, and what are the consequences for minority-class detection under class imbalance?*
Since k controls the locality of the k-NN decision rule, this question examines the trade-off between local sensitivity and the tendency of larger neighborhoods to favor the majority class.

RQ6. *Do standard evaluation metrics such as accuracy and AUC adequately capture classifier performance on imbalanced mammography data, or do sensitivity, specificity, balanced accuracy, and F_1 provide a more clinically meaningful assessment?*
The importance of sensitivity and F_1 in medical applications is well established in principle. However, the empirical question of *how much* these metrics disagree with AUC and accuracy in distance-based mammography — and, critically, whether they change *which distance measure* is recommended — has not been quantified. RQ6 does not propose new metrics; it examines whether the choice of evaluation criterion alters the substantive conclusions drawn from RQ1–RQ5, thereby determining whether single-metric reporting is sufficient or whether multi-metric assessment is necessary for this design space.

To answer these questions, we conduct a large-scale empirical study spanning 20 distance measures, 13 feature sets, 3 mammography datasets, and 2 distance-based classifiers (k-NN and NMD). The evaluated measures include 18 baseline distances from six families and two proposed class-conditional measures, CCFCD and PDP. The feature sets comprise handcrafted, deep, and hybrid representations, enabling direct analysis of feature–measure interaction across heterogeneous mammographic feature spaces. The complete evaluation covers 13,260 configurations under 5-fold stratified cross-validation. Statistical differences among measures are assessed using the Friedman test with Holm-corrected Wilcoxon post-hoc analyses and critical difference diagrams. Results are reported across multiple metrics, including AUC, accuracy, F_1, sensitivity, specificity, and balanced accuracy, to support the metric-level analysis required by RQ6.

The main contributions of this paper are threefold. First, it provides a systematic empirical analysis of how distance-measure effectiveness depends on feature representation in mammography, covering handcrafted, deep, and hybrid feature spaces across a broad set of candidate measures. Second, it evaluates whether distance-measure rankings remain stable across two distinct aggregation strategies, local neighbor-based aggregation and global class-conditional aggregation, thereby separating effects attributable to the underlying feature–measure pairing from those induced by the classifier. Third, it examines the robustness and practical value of these conclusions across multiple mammography datasets and clinically relevant

evaluation metrics, showing how performance interpretation changes when class imbalance and minority-class detection are treated explicitly.

II. RELATED WORK

This section reviews three bodies of literature that converge in our study: distance measures for nearest-neighbor classification, distance-based classifiers in mammography, and feature representations for mammographic image analysis.

A. DISTANCE MEASURES FOR NEAREST-NEIGHBOR CLASSIFICATION

The performance of nearest-neighbor classification depends critically on how similarity is quantified in the underlying feature space. Although the k-NN rule is simple, its behavior is inseparable from the distance measure used to rank training samples. Different measures encode different geometric assumptions: Euclidean and Manhattan distances compare pointwise coordinate differences, cosine distance emphasizes angular similarity, Mahalanobis-based approaches account for covariance structure, and histogram-oriented measures capture distributional resemblance. Consequently, distance selection is not a minor implementation choice but a primary determinant of classifier behavior, particularly when feature representations differ substantially in scale, correlation structure, and dimensionality.

A substantial literature has therefore examined distance measures for nearest-neighbor classification. Cha [25] provided a comprehensive taxonomy of dissimilarity measures and grouped them into major families, including Minkowski, L_1, intersection, fidelity, squared chord, inner product, and Shannon entropy based measures. This taxonomy remains useful because it highlights that many measures are designed for specific data characteristics, such as non-negative histogram vectors, normalized distributions, or correlated numerical attributes. Within the nearest-neighbor setting, however, the practical question is not only how measures differ formally, but how those differences translate into classification performance.

Several empirical studies have shown that the answer is strongly data-dependent. Alfeilat et al. [22] compared 54 distance measures on 28 UCI datasets using k-NN and found that no single measure consistently dominated across tasks. Chomboon et al. [26] reached similar conclusions on a smaller set of benchmarks. Hu et al. [23] reported similar variability on medical datasets, again showing that the effectiveness of a measure depends on the structure of the data being classified. Ehsani and Drabløs [24] evaluated multiple measures on cancer datasets and observed that Manhattan and Hassanat distances often outperformed Euclidean distance, indicating that the common default choice may be suboptimal even in biomedical applications. Taken together, these studies support the premise of RQ1: distance choice can materially affect nearest-neighbor performance and should be treated as an empirical question rather than a fixed design decision. At the same time, prior benchmarking results provide only partial guidance for mammography. Most existing studies use tabular datasets with relatively homogeneous attributes, whereas mammographic CAD increasingly relies on heterogeneous feature spaces that include handcrafted descriptors, deep CNN activations, and hybrid concatenations of both. These representations differ not only in dimensionality but also in scale, sparsity, redundancy, and statistical distribution. A measure that performs well on low-dimensional tabular data or on a single feature family cannot be assumed to remain effective on high-dimensional mammographic image descriptors. This limitation motivates RQ2, which asks whether the preferred distance measure depends systematically on feature type.

This question is reinforced by the geometry of high-dimensional spaces. Beyer et al. [20] showed that in high dimensions the contrast between nearest and farthest neighbors can diminish, reducing the usefulness of distance as a discriminative signal. Aggarwal et al. [21] further demonstrated that different norms degrade differently as dimensionality increases, so the choice of distance measure can alter neighbor ranking even when the underlying classifier remains unchanged. These findings are directly relevant to mammography, where feature vectors may range from compact morphological descriptors to deep or hybrid representations with thousands of dimensions. They also suggest that feature–measure interaction is not merely an implementation artifact, but a consequence of the geometry induced by the feature space itself.

Distance metric learning can improve classification by adapting similarity to the data structure instead of using a fixed metric such as Euclidean distance [27]. Xing et al. [28] introduced a metric learning framework that learns a Mahalanobis distance from pairwise constraints, and Weinberger and Saul [29] developed large-margin nearest-neighbor learning to improve k-NN classification. More recent work has considered image-specific similarity learning, including metric learning for mammographic mass classification [30]. Jiao et al. [31] further demonstrated the value of class-specific distance adaptation by decomposing a global metric into pairwise class-conditional distance metrics within a belief-function framework, showing improved k-NN performance on small datasets. These approaches confirm that similarity is task-dependent and can be learned to improve discrimination. However, they address a different problem from the one considered here. The present study does not seek to learn a new metric. Instead, it evaluates a broad family of plug-in distance measures under controlled conditions in order to determine how far careful distance selection alone can improve mammographic classification across different feature spaces.

Most prior studies also evaluate distance measures only within a single classifier, almost always k-NN [22], [23], [24], [26]. This leaves open an important methodological question: whether an observed ranking reflects the intrinsic suitability of a feature–measure pairing, or merely the aggregation rule used by the classifier. In k-NN, classification depends on

a small local neighborhood, whereas in class-conditional mean-distance methods the decision depends on aggregated distances to an entire class. A measure that is effective for local neighbor voting may not retain the same ranking under global aggregation. This motivates RQ4, which examines whether distance-measure rankings remain consistent across local and global distance-based classifiers.

B. DISTANCE-BASED CLASSIFIERS IN MAMMOGRAPHY

Distance-based classification has long been used in mammographic analysis, both for lesion classification and for content-based image retrieval [32], [33]. Among these methods, the k-NN classifier remains the most commonly used because of its conceptual simplicity, competitive empirical performance, and natural interpretability [34]. In mammography, a k-NN decision can be explained in terms of similarity to previously observed lesions, which aligns well with case-based clinical reasoning. This property has made k-NN a recurring baseline and, in many studies, a primary classifier for distinguishing benign from malignant findings.

A number of mammography studies have employed k-NN with different feature representations. Early work often combined handcrafted lesion descriptors with distance-based classification. For example, Papadopoulos et al. [16] used texture and shape features with k-NN for microcalcification classification, showing that nearest-neighbor decisions could produce competitive diagnostic performance. Similar use of distance-based classification appears in studies based on handcrafted texture descriptors, local structural patterns, and region-based feature extraction, including works such as [17] and [35]. These studies established the practical relevance of nearest-neighbor methods in mammography, but they typically evaluated only a limited number of distance measures, most often Euclidean distance, and were tied to a narrow feature design.

Importantly, k-NN remains actively used in recent mammographic classification research, contrary to the perception that it has been entirely supplanted by end-to-end deep learning. Singh et al. [18] applied k-NN with texture and geometric features on INbreast, Chakravarthy et al. [36] proposed a metaheuristic-weighted k-NN for severity classification, Yan et al. [19] used ensemble k-NN with feature weighting on MIAS and CBIS-DDSM, Sridevi [37] combined Gabor features with a shrunk-kernel k-NN on CBIS-DDSM, and Attallah [38] combined wavelet-guided CNN features with k-NN on MIAS and INbreast. Galván-Tejada et al. [39] compared multiple classifiers including random forest, nearest centroid, and k-NN on mammographic image descriptors, while Uddin et al. [40] benchmarked six k-NN variants on medical datasets. These recent studies confirm that distance-based classification remains a relevant and actively investigated paradigm in mammography, particularly when interpretability, case-based retrieval, or lightweight deployment is prioritized over end-to-end deep learning.

With the shift toward deep learning, distance-based methods have not disappeared. Instead, they have often been combined with deep representations for either classification or retrieval. For instance, recent CBIR and similarity-based mammography systems retrieve clinically analogous prior cases using distances computed on learned or hybrid feature spaces [6], [7]. In such systems, the distance function is not merely a component of the classifier; it is the central mechanism that determines which historical cases are deemed most similar to the query. This makes distance selection especially consequential in mammography, where retrieved exemplars may influence both automated predictions and clinician trust in those predictions.

Although k-NN is the dominant distance-based classifier, it is only one way to aggregate pairwise distances. Its decision rule is *local*: only the k nearest training samples contribute directly to the class label. This locality can be advantageous when class boundaries are irregular, but it also makes the classifier sensitive to the neighborhood size, to label noise in the local region, and to class imbalance when the minority class is sparsely represented [41]. In mammography, where malignant cases are often less frequent than benign ones, these effects can materially influence performance, especially sensitivity.

An alternative family of distance-based classifiers uses *class-level* or *prototype-level* aggregation rather than neighbor voting. Prototype methods such as nearest-centroid classification [42] assign a query to the class whose representative point is closest in the feature space. These methods are computationally attractive and often more stable than local neighbor rules, but they compress each class into a single summary vector and may therefore lose information about multimodality or local structure. Related approaches replace the single centroid by other class summaries or by average class-conditional dissimilarities [43], [44]. The Nearest Mean-Distance (NMD) classifier considered in this paper belongs to this broader class of global class-conditional aggregation methods: rather than selecting a few nearest samples, it computes the average distance from the query to all samples in each class and predicts the class with the smaller mean distance.

This distinction is important for the present study. k-NN and NMD operate on the same pairwise distances but aggregate them differently. As a result, they allow two complementary questions to be separated. The first concerns the quality of the underlying feature–measure pairing: does a given distance function induce a useful notion of similarity in a particular feature space? The second concerns the decision mechanism: are the resulting rankings of distance measures specific to local neighbor voting, or do they persist under global class-conditional averaging? Including NMD therefore provides a more informative test bed than evaluating distance measures with k-NN alone.

The mammography literature has only partially explored this distinction. Most published studies focus on k-NN or use distance primarily within retrieval systems, without explicitly

comparing local and global aggregation strategies under the same feature and distance settings. Even when centroid- or prototype-style methods appear, they are usually introduced as isolated classifiers rather than as tools for analyzing the robustness of distance-measure rankings. Consequently, the literature does not yet clarify whether performance differences attributed to a distance measure reflect the measure itself or the aggregation rule through which it is used.

This gap is particularly relevant in the context of heterogeneous representations. A distance function that yields strong local neighborhoods in a handcrafted feature space may not produce equally informative class-level averages in a deep or hybrid space, and conversely a measure that stabilizes global class separation may not be optimal for local neighbor selection. For mammographic lesion classification, where both interpretability and robustness are important, it is therefore insufficient to evaluate distance measures within a single distance-based classifier.

C. FEATURE REPRESENTATIONS FOR MAMMOGRAPHY

Feature representation has undergone a substantial evolution in mammographic image analysis [45], [46]. Earlier CAD systems relied primarily on handcrafted descriptors designed to capture radiologically meaningful properties such as shape irregularity, margin sharpness, intensity variation, and texture. These representations were motivated by the observation that benign and malignant lesions often differ in morphology and internal structure, and that such differences can be quantified through carefully engineered image features.

A large body of mammography research has therefore used handcrafted texture, shape, and morphological descriptors. Gray-Level Co-occurrence Matrix (GLCM) features [10] have been widely employed to characterize spatial intensity dependencies, while Local Binary Patterns (LBP) [11] capture local micro-texture and Histogram of Oriented Gradients (HOG) [12] represents edge orientation and local structural variation. Related mammographic CAD studies have also explored handcrafted texture, structural, and region-based representations in conjunction with classical machine-learning classifiers, as illustrated by works such as [16], [17], and [35]. Their main advantages are interpretability, compact dimensionality, and close alignment with radiological semantics. However, handcrafted features are constrained by design assumptions and may fail to capture more abstract visual patterns that contribute to lesion discrimination.

The rise of deep learning shifted the focus from manually engineered descriptors to learned representations. CNNs can extract hierarchical image features that encode increasingly abstract patterns, and transfer learning has made such representations accessible even when mammography datasets are relatively small [13]. Deep features derived from pretrained architectures have shown strong performance in mammographic classification, often outperforming purely handcrafted pipelines by capturing high-level lesion appearance more effectively [47], [48]. These representations, however, are typically high-dimensional, correlated, and less directly interpretable than handcrafted descriptors. Their geometry therefore differs markedly from that of classical radiomic or morphological feature spaces.

More recently, hybrid representations have emerged as a pragmatic attempt to combine the complementary strengths of both paradigms. In these approaches, handcrafted descriptors preserve interpretable low-level and morphological information, while deep features contribute rich abstract representations learned from large image corpora. Several studies have reported that such fusion can improve medical image analysis, including pathology [14] and, most relevant to the present work, mammographic lesion classification [15], [49]. Hybridization is appealing because handcrafted and deep features capture different aspects of lesion appearance, and their combination can therefore yield a more informative representation than either component alone.

From the perspective of distance-based classification, however, this evolution creates a new challenge. Handcrafted, deep, and hybrid feature spaces differ substantially in dimensionality, variance structure, redundancy, scale, and semantic composition. A compact handcrafted vector may be well matched to coordinate-wise distances, whereas high-dimensional deep activations may favor measures that are less sensitive to magnitude and more sensitive to angular or distributional similarity. Hybrid vectors are more complex still, because they concatenate heterogeneous subvectors that may obey different statistical regimes within the same representation. As a result, the notion of ''closeness'' induced by a distance measure is not invariant across feature types.

Despite the importance of this issue, most mammography studies evaluate a single feature family or compare feature extractors while holding the distance measure fixed, often at Euclidean distance [37], [38], [47]. Even when multiple representations are considered, the analysis usually emphasizes classifier or feature performance rather than the interaction between feature geometry and distance behavior. Consequently, the literature does not yet establish whether the most effective distance measure for mammographic classification varies systematically across handcrafted, deep, and hybrid representations.

This limitation is especially important because conclusions drawn from one feature family do not necessarily transfer to another. A measure that performs well on handcrafted descriptors may do so because those features are low-dimensional and semantically structured. The same measure may become less effective on deep embeddings with high redundancy or on hybrid concatenations with mixed scales and feature semantics. Conversely, a distance that is robust for high-dimensional deep features may not provide the same discrimination in compact handcrafted spaces. Therefore, feature representation is not merely an input to distance-based classification; it determines the geometry within which the distance function operates.

Overall, the literature shows that distance choice matters in nearest-neighbor learning [22], [24], that distance-based classifiers have long been relevant to mammographic classification [34], [46], and that feature representations have shifted from handcrafted descriptors to deep and hybrid embeddings with substantially different structural properties. However, these three strands have rarely been examined together. Existing mammography studies usually assess classifier variants or feature extractors in isolation, often under a fixed distance measure, leaving unresolved whether the most suitable distance function depends systematically on the representation space and whether such effects remain consistent across alternative distance-based decision rules. This unresolved interaction defines the gap addressed in the present study. Table 1 summarizes key differences between prior benchmarking studies and our work.

III. METHODOLOGY

This section details the experimental framework of our study. We begin by formulating the classification problem and the role of the distance measure (Section III-A), then introduce the two classifiers used to evaluate distance measure quality (Section III-B). We describe the 18 baseline distance measures organized by family (Section III-C), followed by our two proposed class-conditional measures (Section III-D). The three mammography datasets and preprocessing pipeline are presented in Section III-E, the 13 feature representations in Section III-F, and the evaluation protocol — including cross-validation, metrics, and statistical testing — in Section III-G.

A. PROBLEM FORMULATION

Given a labeled training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector extracted from a mammographic region of interest (ROI) and $y_i \in \{0, 1\}$ denotes benign or malignant pathology, a distance-based classifier assigns a query sample $\mathbf{x}_q$ to a class based on the distances $d(\mathbf{x}_q, \mathbf{x}_i)$ to the training samples. The choice of distance function $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ fundamentally determines which training samples are considered ''close'' and therefore directly impacts classification performance. Different distance measures impose different geometric notions of similarity, which may or may not align with the underlying class structure in the feature space.

This paper systematically evaluates 20 distance measures (18 established baselines and two proposed) across two classifier families (k-NN and Nearest Mean-Distance) to quantify the impact of this choice and to determine whether the findings are specific to a single classification algorithm or reflect intrinsic properties of the feature–measure interaction.

B. CLASSIFIERS

We evaluate two distance-based classifiers that share the same precomputed distance matrix, isolating the effect of the distance measure from the classification strategy. Both classifiers operate directly on pairwise distances $d(\mathbf{x}_q, \mathbf{x}_i)$ without modifying the feature space, ensuring that any performance differences are attributable solely to how distances are aggregated for classification: locally (k-NN, using the k nearest samples) versus globally (NMD, using all training samples per class).

1) K-NEAREST NEIGHBOR (K-NN)

The k-NN classifier [8] assigns a query sample $\mathbf{x}_q$ to the majority class among its k nearest neighbors in $\mathcal{D}$:

$$\hat{y}_q = \arg\max_{c \in \{0,1\}} \sum_{i \in \mathcal{N}_k(\mathbf{x}_q)} \mathbb{1}[y_i = c] \tag{1}$$

where $\mathcal{N}_k(\mathbf{x}_q)$ denotes the indices of the k training samples with the smallest distance $d(\mathbf{x}_q, \mathbf{x}_i)$. Probabilistic predictions are obtained as $P(y_q = 1) = \frac{1}{k} \sum_{i \in \mathcal{N}_k} \mathbb{1}[y_i = 1]$, enabling threshold-independent evaluation via the area under the ROC curve (AUC). The neighborhood size k is evaluated over 16 values from $k = 1$ to $k = 131$, spanning from purely local decisions to neighborhoods encompassing up to half the training set (see Section III-G2).

2) NEAREST MEAN-DISTANCE (NMD) CLASSIFIER

While k-NN classifies based on the k closest individual neighbors, the Nearest Mean-Distance (NMD) classifier considers the average distance to *all* training samples in each class. Specifically, it assigns a query $\mathbf{x}_q$ to the class with the smallest mean distance:

$$\hat{y}_q = \arg\min_{c \in \{0,1\}} \bar{d}_c(\mathbf{x}_q) \tag{2}$$

where

$$\bar{d}_c(\mathbf{x}_q) = \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} d(\mathbf{x}_q, \mathbf{x}_i) \tag{3}$$

and $\mathcal{D}_c = \{i : y_i = c\}$ denotes the training samples of class c. Unlike the standard Nearest Centroid Classifier, which computes the distance from the query to a single class centroid $\boldsymbol{\mu}_c = \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \mathbf{x}_i$, NMD computes the mean of all pairwise distances to class members. This distinction matters because NMD is well-defined for *any* distance function, including non-metric and distributional measures such as KL-Divergence and Chi-Square. In such cases, a class centroid may not exist or may not be meaningful. Moreover, NMD operates directly on the same precomputed distance matrix as k-NN and therefore adds negligible computational overhead.

NMD has no hyperparameters, producing a single classification result per (dataset, feature set, distance measure) combination. Including NMD alongside k-NN transforms the study from a k-NN benchmarking exercise into a broader analysis of how distance measures interact with feature representations across two fundamentally different aggregation strategies: local (k-NN) versus global (NMD).

For AUC computation, a decision score for class 1 is derived from the relative mean distances:

$$s(\mathbf{x}_q) = \frac{\bar{d}_0(\mathbf{x}_q)}{\bar{d}_0(\mathbf{x}_q) + \bar{d}_1(\mathbf{x}_q)} \tag{4}$$

TABLE 1. Comparison with prior distance measure benchmarking studies.

Study	Measures	Datasets	Domain	Classifiers	Deep Feat.	Hybrid Feat.	Supervised	Statistical Tests
Haneen et al. [22]	54	28 (UCI)	General	1 (k-NN)	No	No	No	Partial
Chomboon et al. [26]	9	4	General	1 (k-NN)	No	No	No	No
Hu et al. [23]	9	6 (medical)	Medical (tabular)	1 (k-NN)	No	No	No	Partial
Ehsani & Drabløs [24]	8	4 (cancer)	Cancer (tabular)	1 (k-NN)	No	No	No	No
Galván-Tejada et al. [39]	1	2 (mammo)	Mammography	3 (RF, NC, k-NN)	No	No	No	No
Uddin et al. [40]	1	5 (medical)	Medical (tabular)	6 (k-NN variants)	No	No	No	Partial
This work	**20**	**3 (mammo)**	**Mammography**	**2 (k-NN, NMD)**	**Yes (3 CNNs)**	**Yes (3)**	**Yes (4)**	**Friedman + Wilcoxon/Holm**

This score increases when the query is, on average, farther from benign training samples and closer to malignant ones. It serves as a monotonic decision score for threshold-independent evaluation rather than a calibrated posterior probability.

C. BASELINE DISTANCE MEASURES

We evaluate 18 baseline distance measures drawn from the literature, organized into six families: Minkowski, inner product, statistical, information-theoretic, histogram-based, and adaptive/supervised. Table 2 provides a summary of all 20 measures. The following subsections present each family with its mathematical formulation.

1) MINKOWSKI FAMILY

The Minkowski L_p norms form the most widely used distance family. We evaluate five measures from this family, including two element-wise measures (Lorentzian and Canberra) that, while not strict L_p norms, share the property of summing per-dimension contributions of $|x_j - y_j|$.

Euclidean distance (L_2) is the default in most k-NN implementations and serves as the principal reference:

$$d_{\text{Euc}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{d} (x_j - y_j)^2} \tag{5}$$

Manhattan distance (L_1) sums absolute differences and is more robust to outliers, as individual feature deviations are not squared:

$$d_{\text{Man}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} |x_j - y_j| \tag{6}$$

Chebyshev distance (L_∞) considers only the single feature with the largest disagreement, making it sensitive to extreme deviations in any one dimension:

$$d_{\text{Cheb}}(\mathbf{x}, \mathbf{y}) = \max_{j=1,\ldots,d} |x_j - y_j| \tag{7}$$

Lorentzian distance applies a logarithmic dampening to each element-wise absolute difference, compressing the contribution of large deviations and providing robustness to outlier features. It was identified by Alfeilat et al. [22] as a top performer on tabular datasets:

$$d_{\text{Lor}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} \ln(1 + |x_j - y_j|) \tag{8}$$

Canberra distance normalizes each feature's contribution by the magnitude of the values, making it sensitive to differences near zero. Also identified among the top performers by Alfeilat et al. [22]:

$$d_{\text{Can}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} \frac{|x_j - y_j|}{|x_j| + |y_j| + \epsilon} \tag{9}$$

2) INNER PRODUCT FAMILY

Inner-product-based measures operate on vector orientation rather than magnitude, which can be advantageous for high-dimensional deep features where direction is often more informative than scale [50].

Cosine distance quantifies the angular separation between two feature vectors:

$$d_{\text{Cos}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|} \tag{10}$$

Correlation distance operates similarly but on mean-centered vectors, measuring their Pearson correlation:

$$d_{\text{Corr}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_j (y_j - \bar{y})^2}} \tag{11}$$

3) STATISTICAL MEASURE

Mahalanobis distance accounts for correlations between features through the sample covariance matrix $\mathbf{S}$, computed from the training data:

$$d_{\text{Mah}}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top \mathbf{S}^{-1} (\mathbf{x} - \mathbf{y})} \tag{12}$$

When the feature dimensionality d exceeds the number of training samples N (as occurs with deep features), the covariance matrix becomes singular. We apply Tikhonov regularization $\mathbf{S}_{\text{reg}} = \mathbf{S} + \lambda \mathbf{I}$ with $\lambda = 10^{-4}$ to ensure invertibility [9].

4) INFORMATION-THEORETIC MEASURES

Information-theoretic divergences quantify distributional differences between feature vectors. To apply these measures, feature values are first shifted to ensure strict positivity and

TABLE 2. Summary of all 20 distance measures evaluated in this study.

#	Measure	Family	Category	Supervised	Pair Complexity
1	Euclidean	Minkowski	Baseline	No	$O(d)$
2	Manhattan	Minkowski	Baseline	No	$O(d)$
3	Chebyshev	Minkowski	Baseline	No	$O(d)$
4	Lorentzian	Minkowski	Baseline	No	$O(d)$
5	Canberra	Minkowski	Baseline	No	$O(d)$
6	Cosine	Inner Product	Baseline	No	$O(d)$
7	Correlation	Inner Product	Baseline	No	$O(d)$
8	Mahalanobis	Statistical	Baseline	No	$O(d^2)$
9	KL-Divergence	Info-Theoretic	Baseline	No	$O(d)$
10	Jeffrey Divergence	Info-Theoretic	Baseline	No	$O(d)$
11	Chi-Square	Histogram	Baseline	No	$O(d)$
12	Histogram Intersection	Histogram	Baseline	No	$O(d)$
13	Earth Mover's (EMD)	Histogram	Baseline	No	$O(d \log d)$
14	Exp. Chi-Square	Histogram	Baseline	No	$O(d)$
15	AWE	Adaptive	Baseline	Yes	$O(d)$[a]
16	FGA	Adaptive	Baseline	No[b]	$O(d)$
17	Learned Mahalanobis	Adaptive	Baseline	Yes	$O(d^2)$[c]
18	RWC	Ensemble	Baseline	No	$O(3d)$
19	**CCFCD**	**Proposed**	**Proposed**	**Yes**	$O(Kd)$[a]
20	**PDP**	**Proposed**	**Proposed**	**Yes**	$O(d)$[a]

[a]Training phase: $O(Nd)$. [b]Uses feature naming conventions, no label-based training. [c]Training phase: $O(T \cdot n_{\text{batch}} \cdot d^2)$ for T iterations; PCA reduces d to 200 when $d > 200$. K = number of classes (2 in this study); N = training samples; d = feature dimensionality.

then normalized to unit sum, so that each vector can be interpreted as an empirical probability distribution.

KL-Divergence (symmetrized) measures the information lost when one distribution is used to approximate the other:

$$d_{\text{KL}}(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_{j=1}^{d} \left[x_j \log \frac{x_j}{y_j} + y_j \log \frac{y_j}{x_j} \right] \tag{13}$$

Jeffrey divergence is a symmetrized variant that is more numerically stable in practice:

$$d_{\text{Jeff}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} (x_j - y_j) \log \frac{x_j}{y_j} \tag{14}$$

A smoothing constant $\epsilon = 10^{-10}$ is added to prevent division by zero.

5) HISTOGRAM AND DISTRIBUTION-BASED MEASURES

These measures were originally developed for comparing probability distributions in image retrieval and are thus a natural fit for feature vectors derived from histogram-based descriptors such as LBP and HOG.

Chi-Square distance weighs each bin's squared difference by its magnitude:

$$d_{\chi^2}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} \frac{(x_j - y_j)^2}{x_j + y_j + \epsilon} \tag{15}$$

Histogram Intersection distance measures the overlap between two non-negative vectors:

$$d_{\text{HI}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{j=1}^{d} \min(x_j, y_j)}{\sum_{j=1}^{d} y_j + \epsilon} \tag{16}$$

Earth Mover's Distance (EMD). The classical EMD between d-dimensional histograms requires solving a linear program in $O(d^3)$ [51], which is prohibitive for repeated pairwise computation. We therefore adopt a lightweight surrogate: each d-dimensional feature vector is treated as an empirical distribution over d values, and the Wasserstein-1 distance is computed between the sorted feature vectors:

$$d_{\text{EMD}}(\mathbf{x}, \mathbf{y}) = \frac{1}{d} \sum_{j=1}^{d} |x_{\pi(j)} - y_{\sigma(j)}| \tag{17}$$

where π and σ are permutations that sort $\mathbf{x}$ and $\mathbf{y}$ in ascending order, respectively. This is equivalent to the closed-form Wasserstein-1 distance between two one-dimensional empirical distributions and runs in $O(d \log d)$. Note that this formulation discards the original feature ordering; it measures how different the overall value distributions of two feature vectors are, rather than comparing corresponding dimensions.

Exponential Chi-Square kernel distance applies a radial basis function to the Chi-Square distance, mapping it into a kernel space:

$$d_{\text{Exp}\chi^2}(\mathbf{x}, \mathbf{y}) = 1 - \exp\left(-\gamma \, d_{\chi^2}(\mathbf{x}, \mathbf{y})\right) \tag{18}$$

where γ is set to the inverse of the median pairwise Chi-Square distance in the training set.

6) ADAPTIVE AND SUPERVISED MEASURES

The following four baseline measures exploit training labels to learn task-specific distance functions, requiring a training phase before distance computation.

Adaptive Weighted Euclidean (AWE) weights each feature dimension by its Fisher Discriminant Ratio (FDR),

amplifying dimensions that best separate the two classes [52]:

$$d_{\text{AWE}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{d} w_j (x_j - y_j)^2} \tag{19}$$

where $w_j = (\mu_j^{(1)} - \mu_j^{(0)})^2 / ((\sigma_j^{(1)})^2 + (\sigma_j^{(0)})^2 + \epsilon)$ and $\mu_j^{(c)}$, $\sigma_j^{(c)}$ are the mean and standard deviation of feature j in class c. Weights are normalized to unit sum.

Feature-Group-Aware (FGA) distance applies different base metrics to different feature subgroups, acknowledging that handcrafted and deep features may have fundamentally different statistical properties:

$$d_{\text{FGA}}(\mathbf{x}, \mathbf{y}) = \alpha \cdot d_{\text{Corr}}(\mathbf{x}_{\text{HC}}, \mathbf{y}_{\text{HC}}) + (1 - \alpha) \cdot d_{\text{Cos}}(\mathbf{x}_{\text{D}}, \mathbf{y}_{\text{D}}) \tag{20}$$

where $\mathbf{x}_{\text{HC}}$ and $\mathbf{x}_{\text{D}}$ denote the handcrafted and deep feature subvectors, respectively, and $\alpha = 0.5$ assigns equal weight to both groups. This equal-weight default is used because neither feature group is *a priori* more discriminative, and tuning α per dataset would introduce an additional hyperparameter that confounds the comparison of distance measures. Preliminary experiments with $\alpha \in \{0.3, 0.5, 0.7\}$ showed that performance was not highly sensitive to this parameter (AUC variation < 0.01). Feature subgroups are identified automatically from feature naming conventions. When all features belong to a single group, FGA reduces to the corresponding base metric.

Learned Mahalanobis (L-Maha) learns a positive semi-definite transformation matrix $\mathbf{L} \in \mathbb{R}^{d \times d}$ following the Large Margin Nearest Neighbor (LMNN) framework [29]:

$$d_{\text{L-Maha}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{L}(\mathbf{x} - \mathbf{y})\|_2 \tag{21}$$

The matrix $\mathbf{L}$ is optimized via stochastic gradient descent over 100 iterations with a target neighborhood of $k_{\text{target}} = 3$ and a margin of 1.0. Because LMNN learns an $O(d^2)$ transformation matrix, applying it directly to deep features or hybrid features would require estimating millions of parameters from training sets of only a few hundred samples, leading to severe overfitting and prohibitive memory costs. When $d > 200$, PCA is therefore applied to reduce dimensionality to 200 components before learning $\mathbf{L}$. The threshold of 200 was chosen as a practical compromise: it retains over 95% of the variance in all feature sets while keeping the parameter count ($200^2 = 40{,}000$) tractable for the smallest training fold ($N \approx 264$ for MIAS). This preprocessing is standard practice in metric learning with LMNN [29].

Rank-Weighted Composite (RWC) is an ensemble approach that combines three base measures (Euclidean, Manhattan, Cosine) using accuracy-derived weights. Each base measure's distance matrix is first min–max normalized to [0, 1], and the final distance is a weighted sum:

$$d_{\text{RWC}}(\mathbf{x}, \mathbf{y}) = \sum_{m \in \mathcal{M}} w_m \, \hat{d}_m(\mathbf{x}, \mathbf{y}) \tag{22}$$

where $\hat{d}_m$ denotes the min–max normalized distance under base measure m, and the weights w_m are determined by softmax-scaled leave-one-out 1-NN accuracy on the training set: $w_m = \exp(5 \cdot a_m) / \sum_{m'} \exp(5 \cdot a_{m'})$, with a_m being the 1-NN classification accuracy achieved by measure m. The scaling factor of 5 sharpens the weight distribution so that more accurate base measures receive substantially higher weight. When training labels are unavailable, equal weights ($w_m = 1/3$) are used.

D. PROPOSED DISTANCE MEASURES

We introduce two class-conditional distance measures that leverage training labels to adaptively weight feature-level distances. These measures build on a long tradition of incorporating class-conditional statistics into nearest-neighbor distance functions [53], [54], adapting the principle to the specific challenges of mammographic feature spaces where handcrafted and deep features coexist. Both measures share a common preprocessing step of computing class-conditional Gaussian statistics from the training data, but differ fundamentally in how they use this information.

For each feature dimension j and class $c \in \{0, 1\}$, the training phase estimates the class-conditional mean $\mu_j^{(c)}$ and standard deviation $\sigma_j^{(c)}$. The class posterior for a feature value x_j is then:

$$P(c \mid x_j) = \frac{\mathcal{N}(x_j; \mu_j^{(c)}, \sigma_j^{(c)}) \cdot \pi_c}{\sum_{c'} \mathcal{N}(x_j; \mu_j^{(c')}, \sigma_j^{(c')}) \cdot \pi_{c'}} \tag{23}$$

where $\mathcal{N}(\cdot; \mu, \sigma)$ is the Gaussian density and π_c is the class prior. In practice, we compute the log-likelihoods and apply the softmax function for numerical stability. The posteriors are computed independently for each feature dimension, producing a posterior vector $\mathbf{P}(x_j) = [P(c = 0 \mid x_j), P(c = 1 \mid x_j)]$ for each feature.

1) CLASS-CONDITIONAL FEATURE CONCORDANCE DISTANCE (CCFCD)

The core insight behind CCFCD is that two samples should be considered more dissimilar when their features *disagree* on which class they point toward. A small numerical difference in a feature that straddles the benign/malignant boundary should be penalized more heavily than a large difference where both values clearly indicate the same class.

For a pair of samples $(\mathbf{x}, \mathbf{y})$, the concordance for feature j measures the agreement between their posterior vectors via the dot product:

$$c_j(\mathbf{x}, \mathbf{y}) = \sum_k P(c_k \mid x_j) \cdot P(c_k \mid y_j) \tag{24}$$

When both samples agree on class assignment for feature j, $c_j \approx 1$; when they disagree, $c_j \approx 0.5$ (binary case).

The CCFCD distance amplifies disagreeing features and down-weights concordant ones:

$$d_{\text{CCFCD}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} w_j^{\text{FDR}} \cdot |x_j - y_j| \cdot (2 - c_j(\mathbf{x}, \mathbf{y})) \quad (25)$$

where w_j^{FDR} is the normalized Fisher Discriminant Ratio weight for feature j, and the factor $(2 - c_j)$ ranges from 1 (full concordance) to 2 (maximum discordance), acting as a per-pair, per-feature amplification.

Three properties characterize CCFCD relative to existing measures. First, the concordance weighting is *pairwise*: it depends on both $\mathbf{x}$ and $\mathbf{y}$, not just global feature statistics. Second, the L_1 base distance (absolute difference) makes it more robust to outliers than L_2-based measures. Third, the FDR weighting ensures that class-discriminative features dominate regardless of the concordance pattern.

The training phase requires $O(N \cdot d)$ operations for computing class-conditional statistics. Distance computation requires $O(d)$ per pair after the posteriors are precomputed, though the pairwise concordance computation uses $O(K \cdot d)$ where K is the number of classes. We implement block-wise computation with chunk sizes of 50 to manage memory for large distance matrices.

2) POSTERIOR DISTANCE PROFILE (PDP)

While CCFCD operates in the original feature space with concordance-modulated weighting, PDP takes a fundamentally different approach: it transforms each sample into an ''evidence space'' and then applies a standard distance metric in the transformed space.

For each feature j, the log-odds transformation maps the raw feature value to a diagnostic evidence score:

$$e_j(x_j) = \log \frac{P(c = 1 \mid x_j) + \epsilon}{P(c = 0 \mid x_j) + \epsilon} \quad (26)$$

Positive values indicate evidence for malignancy; negative values indicate evidence for benign.

The vector $\mathbf{e}(\mathbf{x}) = [e_1(x_1), \ldots, e_d(x_d)]$ is the sample's *diagnostic evidence profile*. Log-odds values are clipped to the range $[-10, 10]$ for numerical stability. The distance between two samples is then the FDR-weighted cosine distance in evidence space:

$$d_{\text{PDP}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{(\sqrt{\mathbf{w}} \odot \mathbf{e}(\mathbf{x})) \cdot (\sqrt{\mathbf{w}} \odot \mathbf{e}(\mathbf{y}))}{\|\sqrt{\mathbf{w}} \odot \mathbf{e}(\mathbf{x})\| \, \|\sqrt{\mathbf{w}} \odot \mathbf{e}(\mathbf{y})\|} \quad (27)$$

where $\mathbf{w}$ is the normalized FDR weight vector and $\odot$ denotes element-wise multiplication. The $\sqrt{\mathbf{w}}$ scaling ensures that features with high discriminative power contribute proportionally more to the cosine similarity computation.

CCFCD and PDP differ in the level at which class information is used. CCFCD operates in feature space and modulates each pairwise distance via concordance, asking ''do $\mathbf{x}$ and $\mathbf{y}$ agree on which class each feature points toward?'' PDP transforms to evidence space and applies a global metric, asking ''do $\mathbf{x}$ and $\mathbf{y}$ have similar diagnostic fingerprints?'' PDP also offers a computational advantage: after the log-odds transformation, distance computation reduces to a single call to cosine distance, making it substantially faster than CCFCD on large datasets.

TABLE 3. Summary of the three mammography datasets.

Dataset	ROIs	Benign	Malignant	Modality
MIAS	330	276	54	Digitized film
CBIS-DDSM	1,696	912	784	FFDM
INbreast	410	310	100	FFDM
Total	**2,436**	**1,498**	**938**	

FFDM = Full-Field Digital Mammography.

For $K > 2$ classes, PDP generalizes via the centered log-ratio transform of the posterior vector: $e_{j,k}(x_j) = \log P(c_k \mid x_j) - \frac{1}{K} \sum_m \log P(c_m \mid x_j)$, with evidence vectors concatenated across classes.

E. DATASETS

We evaluate on three publicly available mammography datasets that are widely used in the breast cancer classification literature. Table 3 summarizes their characteristics.

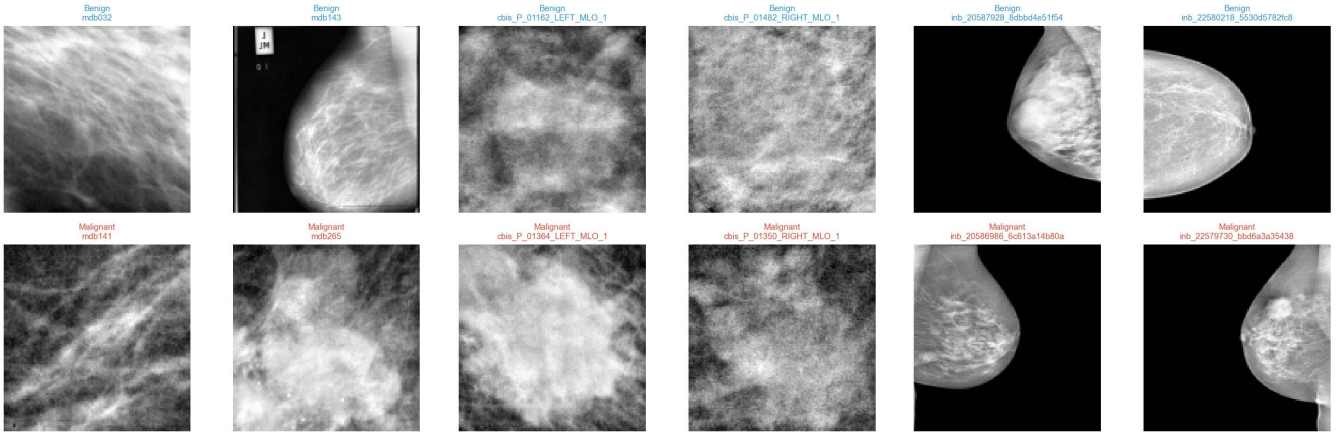
MIAS (Mammographic Image Analysis Society) [55] contains digitized film mammograms from the UK National Breast Screening Programme, digitized at a resolution of 200 μm pixel edge. We extract mass ROIs based on the provided center coordinates and radius annotations, yielding 330 ROIs classified as benign or malignant.

CBIS-DDSM (Curated Breast Imaging Subset of DDSM) [56] is a curated and standardized subset of the Digital Database for Screening Mammography. It provides full-field digital mammograms with updated ROI segmentation masks. We extract 1,696 mass ROIs by applying the provided binary masks to the corresponding full mammograms, merging the ''benign'' and ''benign without callback'' categories as benign, and treating ''malignant'' as positive.

INbreast [57] is a fully annotated full-field digital mammography dataset acquired using a MammoNovation Siemens system at a breast center in Porto, Portugal. Annotations are provided in XML format with precise contour coordinates. We extract 410 ROIs from mass annotations.

Fig. 1 shows representative benign and malignant ROIs from each dataset. The three datasets differ substantially in imaging modality and quality: MIAS mammograms were digitized from film and exhibit higher noise levels and lower contrast, while CBIS-DDSM and INbreast provide full-field digital mammograms (FFDM) with finer tissue detail. CBIS-DDSM ROIs are extracted using segmentation masks applied to full mammograms, whereas INbreast annotations use precise XML contour coordinates. This diversity ensures that our findings are not specific to a single imaging pipeline or annotation protocol.

All images are preprocessed with CLAHE (Contrast Limited Adaptive Histogram Equalization [58], clip



(a) MIAS (b) CBIS-DDSM (c) INbreast

FIGURE 1. Sample ROIs from each dataset after preprocessing (CLAHE contrast normalization). Top row: benign cases; bottom row: malignant cases. The three datasets exhibit different imaging characteristics: (a) MIAS contains digitized film mammograms with lower resolution and higher noise; (b) CBIS-DDSM provides full-field digital mammograms with ROIs extracted via segmentation masks applied to full mammograms; (c) INbreast offers high-quality digital mammograms with precise contour-based annotations. These differences in image quality, resolution, and annotation style motivate our multi-dataset evaluation.

limit $= 2.0$, tile grid $= 8 \times 8$) for contrast normalization. ROIs are resized to 128×128 pixels for handcrafted feature extraction and 224×224 pixels for deep feature extraction. Because the inputs are already cropped lesion ROIs rather than full mammograms, resizing preserves the internal morphological structure of the mass (margins, shape, texture) while standardizing the spatial dimensions required by each feature extractor. The 224×224 target matches the default ImageNet input size expected by the pretrained CNN architectures, and 128×128 is a standard resolution for handcrafted descriptor extraction in mammographic studies [18], [36].

F. FEATURE EXTRACTION

We extract 13 feature sets organized into three families: handcrafted (7), deep learning (3), and hybrid (3). This design allows us to evaluate whether the optimal distance measure depends on the nature of the feature representation and whether findings on hybrid features generalize across different deep learning backbones. Table 4 summarizes all feature sets.

1) HANDCRAFTED FEATURES

Six individual handcrafted feature types are extracted from each 128×128 ROI.

GLCM (F1): Gray-Level Co-occurrence Matrices are computed at three distances $\{1, 3, 5\}$ and four angles $\{0, \pi/4, \pi/2, 3\pi/4\}$ [10]. For each of six properties (dissimilarity, correlation, homogeneity, energy, contrast, angular second moment), the mean and standard deviation across angles are recorded per distance, yielding $6 \times 3 \times 2 = 36$ features.

LBP (F2): Uniform Local Binary Patterns [11] are extracted from circular neighborhoods with radius values $r \in \{1, 2, 3\}$, using $P = 8r$ sampling points at each radius. For uniform LBP, each histogram has $P + 2$ bins, where $P + 1$ bins represent uniform patterns and one bin aggregates all non-uniform patterns. Concatenating the normalized histograms therefore produces $10 + 18 + 26 = 54$ features.

HOG (F3): Histogram of Oriented Gradients [12] are extracted using 9 orientation bins, 16×16 pixel cells, and 2×2 cell blocks with L_2-Hys normalization. For a 128×128 input image, this configuration produces 1,764 features.

Gabor (F4): A Gabor filter bank [59] is constructed using four frequencies, $\{0.1, 0.2, 0.3, 0.4\}$, and four orientations. For each of the 16 resulting filters, three statistics are extracted from the response magnitude: the mean, standard deviation, and 75th percentile. This yields a total of $16 \times 3 = 48$ features.

Morphological + Intensity (F5): Morphological features include normalized area, perimeter, eccentricity, solidity, extent, compactness, and 7 log-transformed Hu moments [60] (13 features). Intensity features comprise 17 first-order statistics: mean, standard deviation, skewness, kurtosis,

TABLE 4. Feature sets used in experiments.

ID	Name	Components	Dim.	Family
F1	GLCM	GLCM	36	HC
F2	LBP	LBP	54	HC
F3	HOG	HOG	1,764	HC
F4	Gabor	Gabor	48	HC
F5	Morph+Intensity	Morph, Intensity	30	HC
F6	Texture	F1+F2+F4	138	HC
F7	All Handcrafted	F1–F5	1,932	HC
F8	ResNet-50	ResNet-50 avg-pool	2,048	Deep
F9	VGG-16	VGG-16 avg-pool	4,096	Deep
F10	EfficientNet-B0	EfficientNet-B0 avg-pool	1,280	Deep
F11	ResNet-50+All-HC	F7+F8	3,980	Hybrid
F12	VGG-16+All-HC	F7+F9	6,028	Hybrid
F13	EffNet-B0+All-HC	F7+F10	3,212	Hybrid

HC = handcrafted. Dim. = dimensionality of the feature vector.

median, min, max, range, entropy, energy, four percentiles (10th, 25th, 75th, 90th), interquartile range, coefficient of variation, and smoothness. So, in total 30 features.

Two composite handcrafted sets are also defined: *Texture* (F6) concatenates F1+F2+F4 (138 features), and *All-Handcrafted* (F7) concatenates all six types (1,932 features).

2) DEEP FEATURES

Deep features are extracted via transfer learning from three CNN architectures pretrained on ImageNet [61], without fine-tuning. These three backbones were selected because they represent distinct architectural paradigms — residual connections (ResNet-50), deep sequential convolutions (VGG-16), and compound scaling (EfficientNet-B0) — and have been widely adopted for mammographic transfer learning in the recent literature [38], [47], [48]. They also span a range of output dimensionalities (1,280 to 4,096), enabling the study to assess whether distance measure rankings are sensitive to the embedding size. Features are not fine-tuned because the goal is to evaluate distance measures on fixed representations; fine-tuning would confound the comparison by introducing representation changes that covary with the training procedure. Each 224×224 ROI is converted to three-channel input with ImageNet normalization ($\mu = [0.485, 0.456, 0.406], \sigma = [0.229, 0.224, 0.225]$) and passed through the network up to the global average pooling layer [13].

ResNet-50 (F8) [50]: 2,048-dimensional feature vector from the penultimate pooling layer.

VGG-16 (F9) [62]: 4,096-dimensional feature vector.

EfficientNet-B0 (F10) [63]: 1,280-dimensional feature vector.

3) HYBRID FEATURES

Three hybrid feature sets are constructed by concatenating each deep backbone's features with the full handcrafted suite (F7), testing whether distance measure performance on hybrid features generalizes across backbones.

ResNet-50 + All-HC (F11): Concatenation of F8 and F7, yielding a $2{,}048 + 1{,}932 = 3{,}980$-dimensional vector.

VGG-16 + All-HC (F12): Concatenation of F9 and F7, yielding a $4{,}096 + 1{,}932 = 6{,}028$-dimensional vector. This is the highest dimensionality in this study.

EfficientNet-B0 + All-HC (F13): Concatenation of F10 and F7, yielding a $1{,}280 + 1{,}932 = 3{,}212$-dimensional vector.

These three hybrid configurations allow us to assess whether the interaction between distance measures and hybrid features is consistent across backbones or specific to a particular deep architecture.

G. EVALUATION PROTOCOL

1) CROSS-VALIDATION

All experiments use 5-fold stratified cross-validation with a fixed random seed (42) for reproducibility. Feature normalization (z-score standardization) is applied per fold: the scaler is fit exclusively on training data and applied to both training and test partitions, preventing any data leakage. For supervised distance measures (AWE, L-Maha, CCFCD, PDP), the training phase parameters (FDR weights, class-conditional statistics, learned transformation matrices) are estimated using only the training fold.

Class imbalance handling. The three datasets exhibit different levels of class imbalance (CBIS-DDSM 1.2:1, INbreast 3.1:1, MIAS 5.1:1). We deliberately do not apply resampling techniques such as SMOTE or random oversampling, for two reasons. First, resampling modifies the training distribution and therefore changes the pairwise distance structure on which both k-NN and NMD operate. This would confound the comparison of distance measures, because the observed differences would reflect the joint effect of the distance function and the resampled distribution rather than the distance function alone. Second, a central goal of this study is to examine how distance measures behave *under* class imbalance (RQ5, RQ6), which requires preserving the natural class proportions. Instead of removing imbalance, we address it analytically: stratified cross-validation preserves the class ratio in every fold, AUC is used as the primary metric because it is robust to class priors [64], and we report sensitivity, F_1, and balanced accuracy alongside accuracy to ensure that minority-class performance is visible rather than masked.

2) HYPERPARAMETERS

For k-NN, the neighborhood size is evaluated at $k \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 37, 51, 75, 101, 131\}$ using uniform voting (unweighted k-NN). This range extends to $k = 131 \approx 0.5\, N_{\text{train}}$ for the smallest dataset (MIAS), enabling analysis of the majority-voting effect under class imbalance (see Section V). The lower end ($k = 1$) provides baseline nearest-neighbor performance, while $k = 37 \approx \sqrt{N}$ for the largest dataset (CBIS-DDSM) tests the commonly recommended $\sqrt{N}$ heuristic [22], [23]. The NMD classifier has no hyperparameters. All other hyperparameters, e.g. CLAHE parameters, feature extraction settings, and regularization constants, are fixed across all experiments to isolate the effect of the distance measure.

3) EVALUATION METRICS

We report the area under the ROC curve (AUC) as the primary metric, as it is threshold-independent and robust to class imbalance [64], [65]. Supplementary metrics include accuracy, F_1-score, sensitivity (recall), specificity, and balanced accuracy. For retrieval evaluation (k-NN only), we additionally report mean average precision (MAP), precision at k (P@k), and normalized discounted cumulative gain (NDCG@k). All metrics are averaged across the five folds.

4) STATISTICAL TESTING

We employ the Friedman test [66] to assess whether statistically significant differences exist among the 20 distance measures across all experimental configurations. Each "observation" is a unique (dataset, feature set, k, classifier) combination, and each measure's AUC serves as the response

variable. When the Friedman test rejects the null hypothesis ($p < 0.05$), we proceed with pairwise Wilcoxon signed-rank tests [67] with Holm correction [68] for multiple comparisons. Results are visualized using critical difference diagrams following the framework of Demšar [69].

5) EXPERIMENTAL SCALE

The complete evaluation spans two classifiers. For k-NN: 20 measures $\times$13 feature sets $\times$16 k-values $\times$3 datasets = 12,480 configurations. For NMD: 20 measures $\times$13 feature sets $\times$3 datasets = 780 configurations. The combined total is 13,260 configurations, each evaluated under 5-fold stratified cross-validation for a total of 66,300 model fits. Both classifiers share the same precomputed distance matrices, so the NMD evaluation adds negligible computational overhead. All experiments are implemented in Python using scikit-learn, SciPy, and PyTorch (for deep feature extraction).

6) COMPUTATIONAL COMPLEXITY

Table 2 reports the per-pair complexity of each measure in big-O notation. To make these costs concrete, we express them in FLOPs at three representative dimensionalities from our study. For compact handcrafted descriptors ($d = 30$, e.g., Morph+Intensity), all $O(d)$ measures require fewer than 200 FLOPs per pair, and even Mahalanobis ($O(d^2)$) requires only $\sim$900 FLOPs. For deep features ($d = 2{,}048$, ResNet-50), $O(d)$ measures remain inexpensive ($\sim$ 4,000–10,000 FLOPs), whereas standard Mahalanobis rises to $\sim 4.2 \times 10^6$ FLOPs per pair, and a full $N \times N$ distance matrix ($N \approx 1{,}357$ training samples for CBIS-DDSM) requires $\sim 7.7 \times 10^{12}$ FLOPs. For the largest hybrid vectors ($d = 6{,}028$, VGG-16+All-HC), standard Mahalanobis exceeds 3.6×10^7 FLOPs per pair, making it the most expensive measure by a wide margin.

Supervised measures incur a one-time training cost per fold. AWE, CCFCD, and PDP require $O(Nd)$ operations to compute class-conditional statistics, which is negligible relative to the $O(N^2 d)$ cost of the full distance matrix. Learned Mahalanobis (L-Maha) is more expensive: it requires $O(T \cdot d'^2)$ for $T = 100$ LMNN iterations after PCA reduction to $d' = 200$, yielding $\sim 4 \times 10^6$ FLOPs for the learning phase. The proposed CCFCD adds a per-pair concordance computation that doubles the effective per-pair cost from $O(d)$ to $O(Kd)$ (where $K = 2$), while PDP reduces to a single cosine distance call after the log-odds transformation, adding no overhead beyond the training phase. Overall, the top-performing adaptive measures (AWE at $O(d)$, CCFCD at $O(2d)$, L-Maha at $O(d'^2)$ with $d' = 200$) remain computationally tractable for the dataset sizes encountered in mammographic classification.

IV. EXPERIMENTAL RESULTS

This section presents the experimental results organized by research question. All reported values are averages over the five folds; cross-fold standard deviations ($\pm$) are included in the main ranking tables to indicate the precision of these estimates. Unless stated otherwise, measures are ranked by mean AUC.

Two points of context are important for interpreting the results that follow. First, the purpose of this study is systematic comparative analysis of distance measures across feature families, not the pursuit of state-of-the-art classification accuracy. The experimental design deliberately uses a simple k-NN classifier with uniform voting and transfer learning features without fine-tuning, in order to isolate the distance measure as a single controlled variable. These constraints produce moderate absolute AUC values (0.61–0.71) that are lower than those achievable by end-to-end deep learning systems on the same datasets [47], [48]. The informative content of the study lies in the *relative rankings* (i.e. which measures outperform which and how this depends on feature type, dataset, classifier, and evaluation metric) rather than in the absolute performance levels. Second, all six evaluation metrics (AUC, accuracy, F_1, sensitivity, specificity, balanced accuracy) are reported throughout, because the analysis in RQ6 shows that these metrics can disagree substantially on which measure is "best," particularly under class imbalance.

A. RQ1: DOES DISTANCE MEASURE CHOICE MATTER?

Table 5 reports the overall ranking of all 20 distance measures averaged across all datasets, feature sets, k-values, and both classifiers. The Adaptive Weighted Euclidean (AWE) measure achieves the highest mean AUC of 0.687, followed closely by the proposed CCFCD. The bottom of the ranking is occupied by Chebyshev (0.608) and Earth Mover's Distance (0.617). The gap between the best and worst measure is 0.079 AUC, a substantial difference indicating that distance measure selection materially affects classification performance.

A Friedman test confirms that the performance differences among the 20 measures are statistically significant ($\chi^2 = 1901.09$, $p < 10^{-300}$, $n = 663$ observations). Holm-corrected Wilcoxon post-hoc tests against the top-ranked measure reveal that several measures in the upper tier (ranks 1–7) are not significantly different from each other, forming a cluster of statistically equivalent top performers. In contrast, Chebyshev, Earth Mover's Distance, and Mahalanobis are significantly worse than the best measure after correction.

When measures are grouped by family, the Adaptive family (AWE, FGA, L-Mahalanobis) achieves the highest average AUC (0.686), followed by Inner Product (Cosine, Correlation; 0.684) and the proposed measures (CCFCD, PDP; 0.683). The Minkowski family (including Euclidean) averages only 0.663, placing fifth out of eight families. Euclidean distance itself ranks ninth overall (0.679), confirming that the widespread default of Euclidean distance is suboptimal for mammographic classification.

The metric breakdown further shows that AUC alone does not fully capture the differences among the top measures. Although AWE ranked first by AUC, Learned Mahalanobis achieved the strongest balanced accuracy (0.579), the highest

TABLE 5. RQ1: Overall measure ranking (all configurations). Values are means over all datasets, feature sets, classifiers, and neighborhood sizes; ± values denote average cross-fold standard deviation.

#	Measure	AUC	F_1	Sens.	Spec.	B.Acc.
1	AWE	.687±.040	.294±.051	.271±.046	.867	.569
2	CCFCD†	.687±.040	.300±.051	.276±.045	.866	.571
3	L-Mah	.686±.040	.329±.058	.310±.059	.847	.579
4	Cos	.684±.038	.306±.050	.284±.047	.863	.573
5	Corr	.684±.038	.310±.049	.286±.046	.863	.575
6	FGA	.684±.038	.309±.050	.287±.047	.861	.574
7	RWC	.683±.037	.297±.048	.273±.045	.868	.571
8	PDP†	.680±.038	.313±.056	.292±.054	.849	.571
9	Euc	.679±.038	.280±.049	.259±.046	.870	.564
10	HI	.678±.039	.285±.049	.261±.048	.879	.570
11	Man	.678±.038	.288±.049	.269±.046	.864	.567
12	ExpChi	.676±.040	.275±.051	.259±.050	.879	.569
13	Lor	.676±.038	.290±.048	.274±.046	.859	.567
14	Chi2	.676±.040	.275±.051	.259±.051	.878	.569
15	Can	.676±.042	.298±.045	.290±.044	.845	.567
16	KL	.675±.039	.278±.053	.261±.053	.876	.569
17	Jeff	.675±.039	.278±.053	.261±.053	.876	.569
18	Mah	.659±.045	.217±.048	.208±.045	.863	.536
19	EMD	.617±.042	.256±.048	.243±.048	.832	.537
20	Cheb	.608±.051	.231±.049	.217±.047	.853	.535

† Proposed measures. Cross-fold std is highest for INbreast (avg 0.053) and lowest for CBIS-DDSM (avg 0.025), reflecting dataset size differences.

sensitivity (0.310 ± 0.059), and the highest F_1 score (0.329 ± 0.058). CCFCD combined near-best AUC with slightly stronger balanced accuracy than AWE and a higher F_1, while PDP remained competitive in sensitivity and F_1 despite a lower AUC rank. Thus, the strongest measures are not identical across all evaluation criteria, but the same small group remains consistently competitive.

Fig. 2 reinforces this interpretation. The best-performing measures form a compact leading cluster, whereas the weakest measures are clearly separated at the lower end of the ranking. This indicates that the effect of distance choice is not merely statistical but also practically meaningful across the full experimental space.

B. RQ2: DOES THE OPTIMAL MEASURE DEPEND ON FEATURE TYPE?

Table 6 presents the top five measures for each feature family. While the top-performing measures are drawn from the same pool of adaptive and inner-product measures across all three families, the ranking order shifts. For handcrafted features, Learned Mahalanobis ranks first (0.671 ± 0.037), ahead of AWE and CCFCD. For deep features, AWE leads (0.707 ± 0.039), with CCFCD second and Correlation third. For hybrid features, CCFCD achieves the highest AUC of any family–measure combination (0.712 ± 0.039), followed by AWE and Learned Mahalanobis.

Notably, the Mahalanobis distance (standard, non-learned) ranks fourth on handcrafted features (0.669) but drops to 18th overall. This suggests that statistical distances that account for feature correlations are particularly suited to lower-dimensional handcrafted descriptors where the covariance matrix can be reliably estimated. On deep features, where dimensionality is high and covariance estimation is ill-conditioned, standard Mahalanobis degrades while inner-product measures (Cosine, Correlation) remain strong.

TABLE 6. RQ2: Top 5 measures by feature family (Mean AUC ± cross-fold standard deviation).

	HC (7 sets)		Deep (3 sets)		Hybrid (3 sets)	
#	Meas.	AUC	Meas.	AUC	Meas.	AUC
1	L-Mah	.671±.037	AWE	.707±.039	CCFCD†	.712±.039
2	AWE	.669±.040	CCFCD†	.706±.039	AWE	.710±.039
3	CCFCD†	.669±.040	Corr	.703±.036	L-Mah	.707±.044
4	Mah	.669±.040	Cos	.702±.037	FGA	.704±.036
5	Cos	.668±.041	FGA	.702±.037	Cos	.704±.034

† Proposed measures.

Spearman rank correlations between families are moderate: $\rho = 0.72$ between HC and Deep, $\rho = 0.73$ between HC and Hybrid, and $\rho = 0.77$ between Deep and Hybrid. These values indicate that while the broad trend is preserved (adaptive measures tend to outperform Minkowski measures regardless of feature type), the specific ordering within the top tier does depend on the feature family. The HC–Deep correlation is the weakest, consistent with the expectation that these feature types have the most different statistical properties.

The initial hypothesis that bin-comparison measures (Chi-Square, Histogram Intersection) would be natural for handcrafted features and angular measures (Cosine) for deep features is *not confirmed*. Instead, supervised/adaptive measures (AWE, CCFCD, L-Mahalanobis) dominate all three families. This suggests that the advantage of incorporating label information into the distance computation outweighs any family-specific geometric preference.

A second important result is that the highest absolute AUC values are achieved in the hybrid feature space, where the top measures all exceed 0.704 and the two strongest, CCFCD and AWE, exceed 0.710. Deep features also perform strongly, whereas handcrafted features remain clearly lower in absolute AUC. This suggests that hybrid representations provide the richest structure for distance-based classification, but also require a measure that can respond appropriately to their mixed statistical properties.

Fig. 3 makes this interaction visually explicit. The handcrafted column is led by L-Mah, the deep column by AWE, and the hybrid column by CCFCD. This directly supports the central claim of the study: the choice of distance measure cannot be separated from the type of feature representation. Feature families with different dimensionality, covariance structure, and statistical heterogeneity favor different notions of similarity.

C. RQ3: ARE RANKINGS CONSISTENT ACROSS DATASETS?

The top-performing measure families differ markedly across the three datasets. On MIAS, histogram-based measures

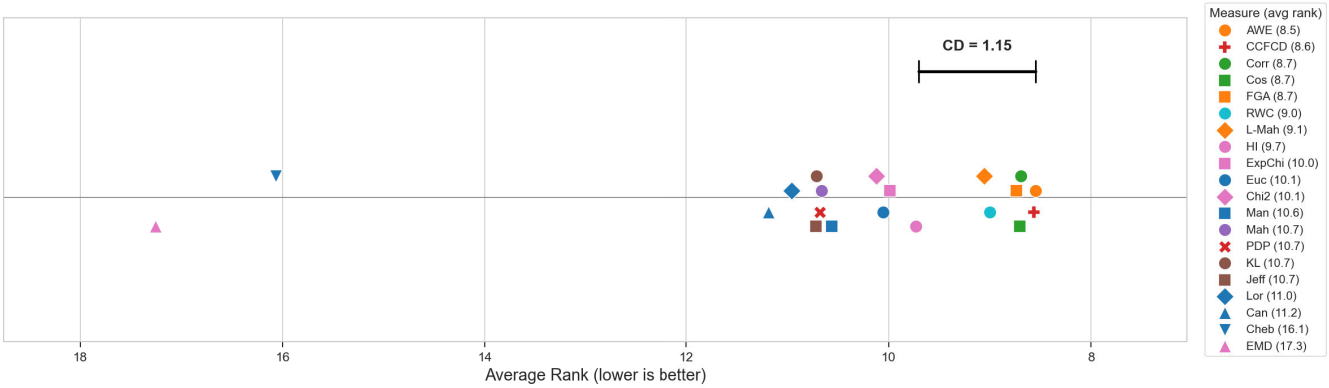


FIGURE 2. **Critical-difference diagram for RQ1 based on AUC rankings across all 663 matched experimental settings. Each measure is represented by a unique marker colored by family; lower average rank is better. The CD bar indicates the Nemenyi critical difference at $\alpha = 0.05$: measures whose average ranks differ by less than CD are not significantly different. The top group forms a compact cluster, whereas EMD and Chebyshev are clearly separated as the weakest overall performers.**

dominate: Exponential Chi-Square (0.844 ± 0.041), Chi-Square (0.843 ± 0.041), and Histogram Intersection (0.842 ± 0.039) occupy the top three positions. On CBIS-DDSM, inner-product measures lead, with Correlation (0.652 ± 0.026) and FGA (0.652 ± 0.026) at the top. On INbreast, statistical measures prevail: Mahalanobis (0.604 ± 0.049) and Learned Mahalanobis (0.580 ± 0.051) rank first and second. The cross-fold standard deviations are notably larger on INbreast (avg std $= 0.053$) and MIAS (0.042) than on CBIS-DDSM (0.025), reflecting the smaller sample sizes of these datasets (410 and 330 ROIs, respectively, vs. 1,696 for CBIS-DDSM).

Cross-dataset Spearman rank correlations are low. CBIS-DDSM vs. MIAS yields $\rho = 0.11$, and INbreast vs. MIAS yields $\rho = -0.10$, both indicating near-zero agreement. CBIS-DDSM vs. INbreast shows moderate correlation ($\rho = 0.42$). These values indicate that distance measure rankings are largely dataset-specific and do not generalize across datasets with different imaging characteristics and class distributions. A measure that ranks highly on one dataset may rank near the middle or bottom on another.

This dataset specificity likely reflects the interaction between imaging modality, class imbalance, and feature space geometry. MIAS is a digitized film dataset with a 5.1:1 benign-to-malignant imbalance and produces texture histograms where bin-comparison measures are effective. CBIS-DDSM is a full-field digital dataset with near-balanced classes (1.2:1), favoring direction-sensitive measures. INbreast, a small full-field digital dataset with moderate imbalance (3.1:1), benefits from covariance-aware measures that can adapt to limited data.

Fig. 4 quantifies this instability more directly. The scatter plots show that measures occupying the top of the ranking on one dataset can appear near the middle on another. At the same time, the results do not imply complete instability: a group including Corr, FGA, Cos, AWE, CCFCD, and L-Mah frequently appears near the upper part of the ranking, while EMD and Chebyshev remain persistently weak. Thus, cross-dataset consistency exists at the level of

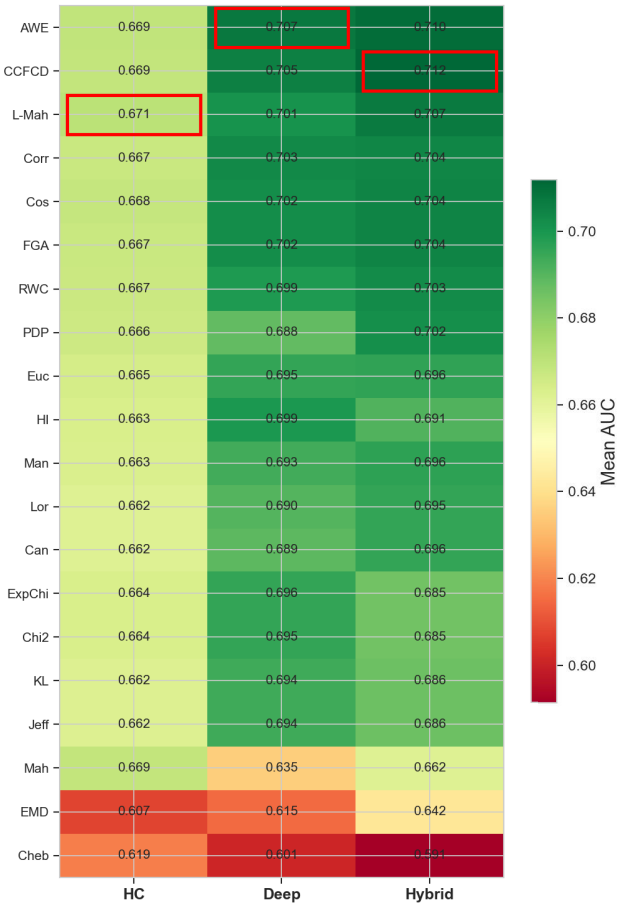


FIGURE 3. **Mean AUC for each measure within the three feature families. Red boxes indicate the best-performing measure in each family. The strongest measure changes across HC, Deep, and Hybrid features, confirming a clear feature–measure interaction.**

broad high-performing and low-performing groups, but not at the level of a single universally optimal feature–measure combination. These findings support the recommendation that the distance measure should be cross-validated on the target dataset rather than selected by convention.

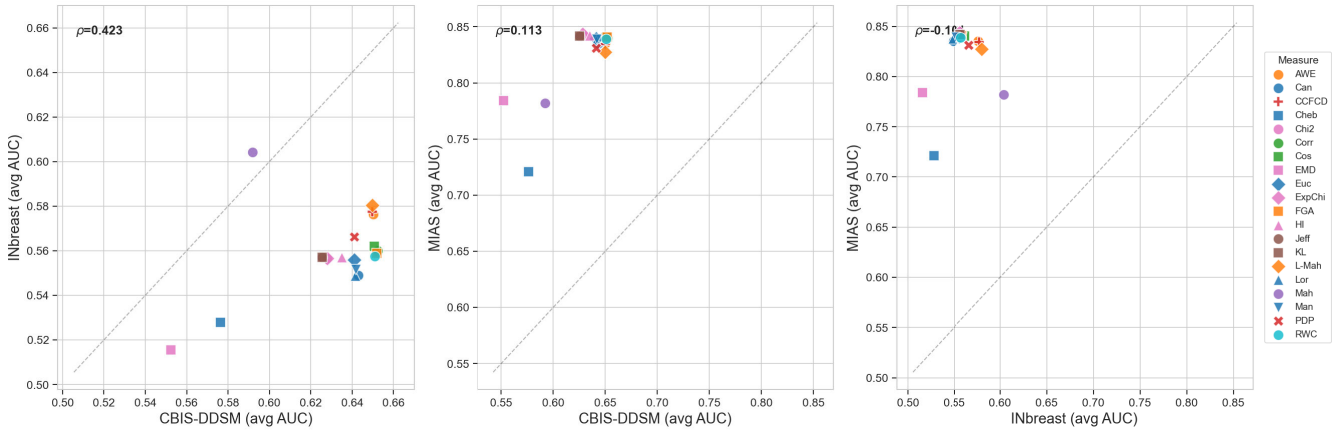


FIGURE 4. Pairwise comparison of dataset-level mean AUC values for all 20 measures. Each measure is identified by a unique marker colored by family (see legend). The Spearman ρ in each panel quantifies agreement in measure rankings between datasets. Moderate agreement is observed only between CBIS-DDSM and INbreast ($\rho = 0.42$), whereas the remaining pairs show weak or near-zero agreement.

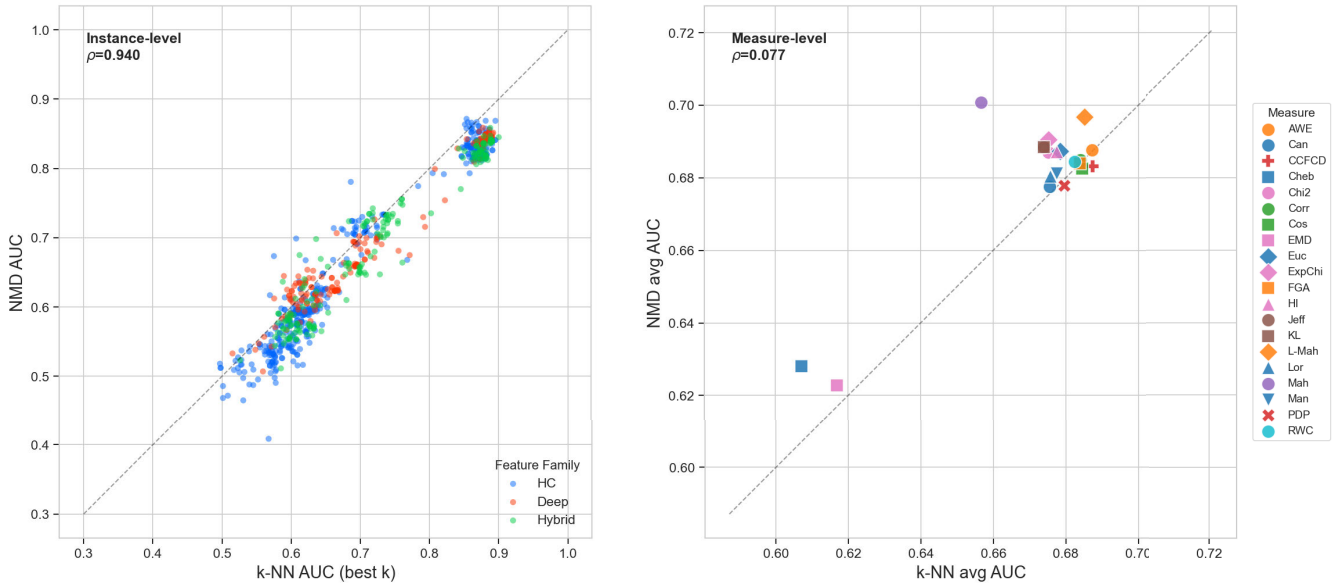


FIGURE 5. Comparison of AUC obtained by k-NN and NMD. Left: all matched configurations colored by feature family (HC = blue, Deep = red, Hybrid = green), showing strong overall alignment ($\rho = 0.94$). Right: measure-wise average AUC with unique markers per measure (see legend), showing much weaker agreement in the ranking of individual measures ($\rho = 0.077$).

Although Tables 5 and 6 and the per-dataset rankings above are marginal views, they together capture the principal joint sensitivity of performance to (feature, dataset) variations. The top-ranked measure shifts across feature families (L-Mah for HC, AWE for Deep, CCFCD for Hybrid; Table 6) and, independently, across datasets (Exp. Chi-Square for MIAS, Correlation for CBIS-DDSM, Mahalanobis for INbreast). Because these two marginal rankings disagree on which measure leads, no single measure is uniformly optimal in the underlying joint cells: the marginal divergence is itself a sufficient condition for joint heterogeneity. The cross-dataset Spearman rank correlations reported above (as low as $\rho = -0.10$) and the between-family correlations of Section IV-B ($\rho \in [0.72, 0.77]$) provide single-number summaries of this joint disagreement. We report marginal sensitivity along each axis, with cross-fold standard deviations within each marginal cell, rather than per-cell rankings: each individual (feature, dataset) cell carries only 5-fold sample support, which would make a 20-measure ranking within a single cell dominated by fold-level variance and require extensive multiple-comparison correction across 39×190 pairwise contrasts.

D. RQ4: ARE RANKINGS CONSISTENT ACROSS CLASSIFIERS?

A striking finding emerges when comparing measure rankings between k-NN and NMD. The measure-level Spearman rank correlation is $\rho = 0.077$ ($p = 0.75$), indicating

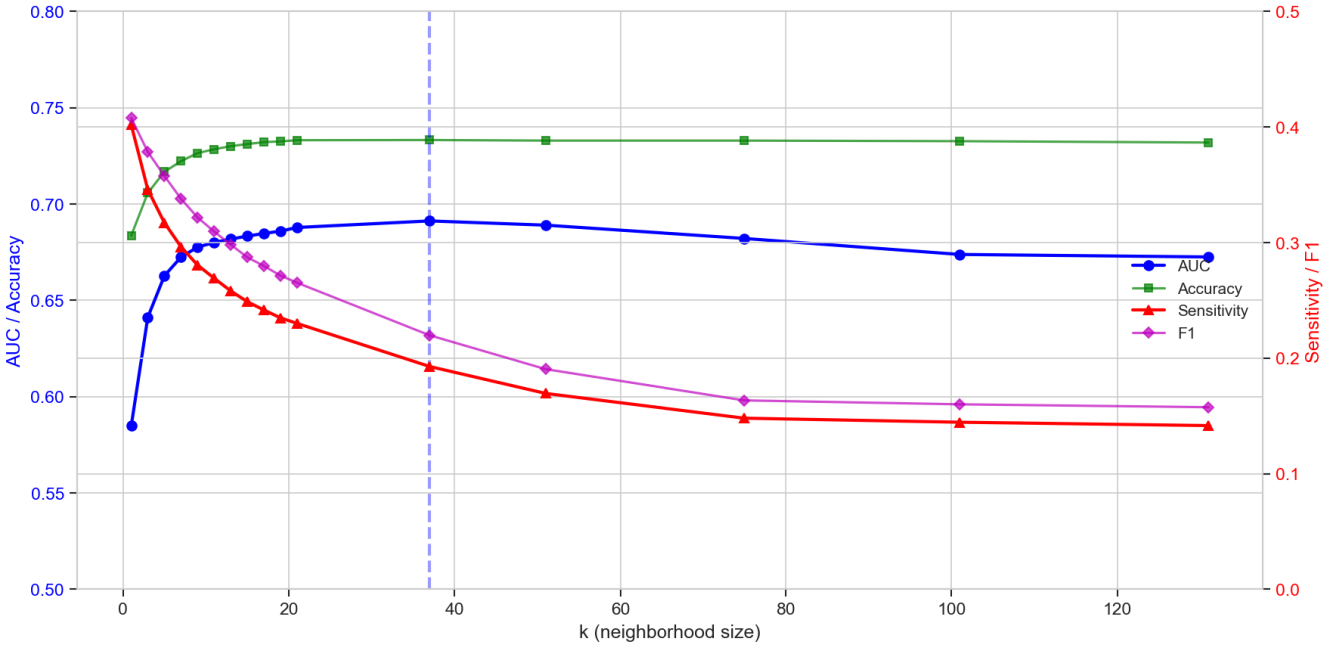


FIGURE 6. RQ5: AUC–sensitivity paradox. As neighborhood size k increases, AUC and accuracy initially improve (left axis), while sensitivity and F_1 decline steadily (right axis). The dashed vertical line marks $k = 37$ where AUC peaks; at this point, sensitivity has already fallen below 0.20.

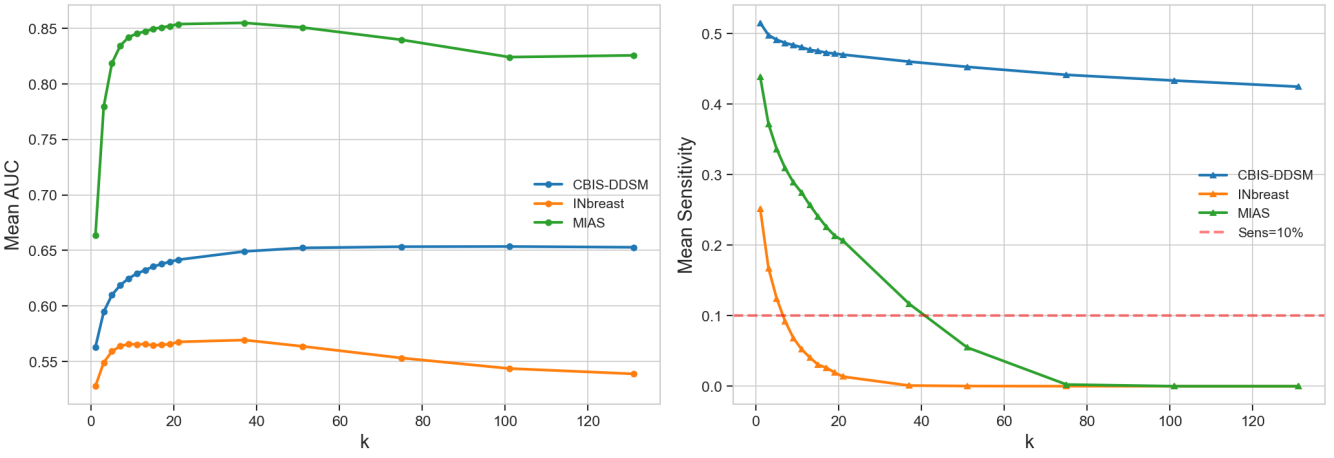


FIGURE 7. Dataset-specific interaction between k and performance. Left: mean AUC versus k for each dataset. Right: mean sensitivity versus k for each dataset. The dashed line marks sensitivity = 10%. INbreast collapses below this threshold at $k \geq 7$, while CBIS-DDSM remains above it for all evaluated k values.

essentially no agreement between the two classifiers on which measures are best. However, this correlation is computed over only $N = 20$ data points (one per measure), which limits statistical power: at this sample size, the minimum detectable $|\rho|$ at 80% power and $\alpha = 0.05$ is approximately 0.44, so moderate correlations below this threshold cannot be distinguished from zero. The result should therefore be interpreted as *no evidence of strong agreement* rather than as definitive evidence of zero correlation. The Mahalanobis distance provides the most extreme example: it ranks 18th under k-NN but 1st under NMD, a shift of 17 positions. Similarly, Jeffrey divergence shifts from 17th to 4th, CCFCD from 1st to 13th, and Exponential Chi-Square from 14th to 3rd.

This near-zero correlation at the measure level contrasts sharply with the instance-level correlation. When each (dataset, feature set, measure) triple is treated as an observation and the best-k AUC for k-NN is compared with the NMD AUC, the Spearman correlation is $\rho = 0.94$. This means that for a given dataset and feature representation, k-NN and NMD largely agree on which configurations produce high or low AUC. The disagreement is specifically about the *aggregated ranking* of distance measures: the two classifiers weight different properties of the distance function.

TABLE 7. RQ5: Effect of neighborhood size k on classification metrics (averaged over all measures, features, and datasets).

k	AUC	Acc.	F_1	Sens.	Spec.	B.Acc.
1	0.585	0.684	0.408	0.402	0.768	0.585
3	0.641	0.706	0.378	0.346	0.817	0.582
5	0.663	0.717	0.358	0.317	0.842	0.580
7	0.672	0.722	0.338	0.296	0.856	0.576
9	0.678	0.726	0.322	0.281	0.867	0.574
11	0.680	0.728	0.310	0.269	0.874	0.571
13	0.682	0.730	0.298	0.258	0.880	0.569
15	0.683	0.731	0.287	0.249	0.884	0.566
17	0.685	0.732	0.280	0.242	0.888	0.565
19	0.686	0.733	0.271	0.235	0.890	0.562
21	0.688	0.733	0.265	0.230	0.893	0.561
37	0.691	0.733	0.220	0.193	0.903	0.548
51	0.689	0.733	0.190	0.169	0.909	0.539
75	0.682	0.733	0.163	0.148	0.916	0.532
101	0.674	0.733	0.160	0.145	0.918	0.531
131	0.673	0.732	0.158	0.142	0.919	0.530

Fig. 5 makes this distinction clear. In the left panel, most points lie close to the identity line, confirming that strong and weak configurations remain broadly similar under both classifiers. In the right panel, however, the measure-wise scatter is far less aligned, which shows that the exact relative standing of the measures depends on how distances are aggregated into class predictions.

Across all evaluation metrics, k-NN (averaged over all k) achieves higher accuracy (0.725 vs. 0.640) and specificity (0.876 vs. 0.646), while NMD achieves substantially higher sensitivity (0.619 vs. 0.245), F_1 (0.445 vs. 0.275), and balanced accuracy (0.633 vs. 0.561). This pattern reflects the fundamental difference between local and global aggregation: k-NN's majority vote at typical k values is dominated by the majority (benign) class, inflating accuracy and specificity while suppressing sensitivity. NMD's global averaging distributes weight more evenly across class members, yielding a more balanced decision boundary.

E. RQ5: HOW DOES K INTERACT WITH DISTANCE MEASURE AND CLASS IMBALANCE?

Table 7 presents the effect of neighborhood size k on all classification metrics, averaged across all measures, feature sets, and datasets. The results reveal an AUC–sensitivity paradox: AUC increases monotonically from 0.585 ($k = 1$) to a peak of 0.691 ($k = 37$), while sensitivity decreases monotonically from 0.402 ($k = 1$) to 0.142 ($k = 131$). Accuracy shows a similar monotonic increase (0.684 to 0.733), masking the progressive collapse of minority-class detection. F_1-score drops from 0.408 to 0.158 over the same range. A Friedman test on k-values confirms that k significantly affects performance ($p < 10^{-300}$).

Fig. 6 makes this trade-off explicit. The AUC curve rises rapidly from small k and peaks in the mid-range before flattening and slightly declining. Accuracy behaves similarly and remains high even at large k. By contrast, sensitivity and F_1 decrease almost monotonically. This produces an AUC–sensitivity paradox: the neighborhood size that maximizes AUC does not maximize clinically meaningful minority-class detection. Instead, beyond relatively small values of k, the classifier becomes increasingly conservative, favoring specificity over recall.

The sensitivity collapse is dataset-dependent and reflects the underlying class imbalance. On INbreast (3.1:1 imbalance), average sensitivity falls below 10% at $k \geq 7$, meaning that the classifier detects fewer than 1 in 10 malignant cases for most neighborhood sizes. On MIAS (5.1:1 imbalance), the threshold is reached at $k \geq 51$. On CBIS-DDSM (1.2:1, near-balanced), sensitivity remains above 10% for all evaluated k values, confirming that the collapse is driven by the majority-class voting effect under class imbalance.

Fig. 7 shows that this interaction is strongly dataset-dependent. For CBIS-DDSM, AUC improves and then stabilizes, while sensitivity declines only moderately. INbreast behaves much more severely: AUC improves only modestly, but sensitivity collapses rapidly toward zero as k increases. MIAS shows a similar pattern, although its AUC reaches a much higher absolute level before sensitivity steadily declines. These differences indicate that the effect of k depends not only on the classifier, but also on the dataset-specific balance between local structure and class imbalance.

F. RQ6: DO STANDARD METRICS CAPTURE CLINICAL PERFORMANCE?

Different evaluation metrics identify different "best" distance measures on every dataset. On CBIS-DDSM, Correlation is the top measure by both AUC and accuracy, but Canberra is top by F_1 and sensitivity, and FGA by balanced accuracy. On MIAS, Chi-Square leads by AUC, but KL divergence by accuracy and F_1, and Jeffrey divergence by sensitivity. On INbreast, the split is between Mahalanobis (AUC, accuracy) and Learned Mahalanobis (F_1, sensitivity, balanced accuracy). In no dataset do all six metrics agree on a single best measure.

This disagreement is quantified by the Spearman rank correlation between metric-induced measure rankings. AUC and sensitivity yield a rank correlation of $\rho = 0.78$ across the 20 measures, indicating moderate agreement but with notable exceptions. AUC and accuracy are more aligned ($\rho > 0.9$), but accuracy and sensitivity are weakly correlated, confirming that optimizing for accuracy may sacrifice minority-class detection.

The accuracy trap is prevalent. Of the 12,480 k-NN configurations, 1,452 (11.6%) achieve accuracy above 80% while sensitivity falls below 10%. In these configurations, the classifier correctly identifies more than four out of five cases overall but detects fewer than one in ten malignant cases. Such configurations would appear acceptable by accuracy alone but are clinically unacceptable for cancer screening.

Fig. 8 translates this result into a more clinically interpretable form. Most measures cluster within a relatively narrow specificity band (0.84–0.90), but they differ more

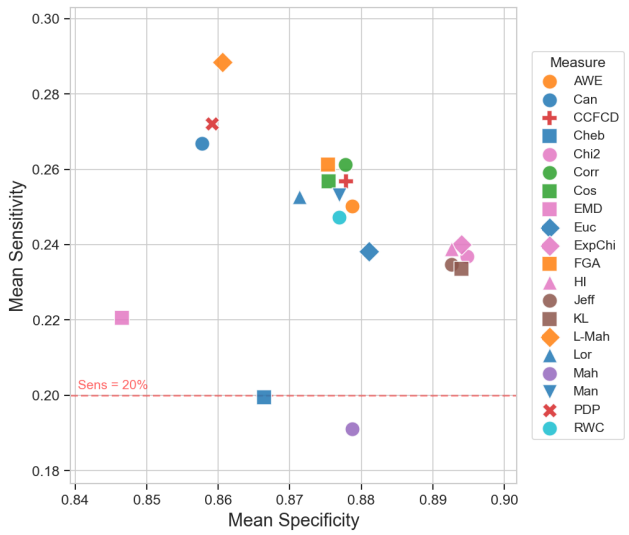


FIGURE 8. Sensitivity versus specificity by measure for k-NN. Each measure is identified by a unique marker colored by family (see legend). The dashed line marks sensitivity = 20%. Learned Mahalanobis (L-Mah) occupies the most favorable region, combining the highest sensitivity with moderate specificity. Several measures with high specificity fall near or below the 20% sensitivity threshold.

substantially in sensitivity (0.19–0.29). Learned Mahalanobis (L-Mah) occupies the most favorable region, combining the strongest sensitivity among the plotted measures with only moderate specificity. PDP and CCFCD also remain in a comparatively strong region, retaining sensitivity above 0.25 while maintaining high specificity. In contrast, standard Mahalanobis and Chebyshev occupy a weaker clinical region, with relatively high specificity but substantially lower sensitivity.

V. DISCUSSION

The results show that distance-based mammographic classification is governed not by a single dominant design factor, but by the interaction among feature representation, distance formulation, aggregation strategy, neighborhood size, dataset characteristics, and evaluation criterion. This section interprets the experimental findings, derives practical guidelines, relates the results to prior work, and acknowledges the limitations of the study.

A. METHODOLOGICAL IMPLICATIONS

The obtained results have several methodological implications for distance-based evaluation in medical image analysis.

First, broad benchmarking of distance measures is justified. The significant differences observed across the 20 measures show that restricting evaluation to one or two conventional distances risks producing conclusions that are contingent on an arbitrary implementation choice. This is especially important in mammography, where the same classifier can behave quite differently under different notions of similarity.

Second, the results support the value of testing both local and global aggregation strategies. The near-zero measure-level rank correlation between k-NN and NMD ($\rho = 0.077$) shows that distance measure rankings are partially classifier-dependent. If a measure performs well under both k-NN and NMD, confidence increases that the result reflects a genuine property of the feature–measure pairing rather than a peculiarity of one classifier. For future studies, this argues for evaluating distance functions under more than one distance-based decision rule whenever the aim is to make claims about the distance itself.

Third, hybrid feature spaces deserve separate treatment rather than being viewed merely as larger feature vectors. The hybrid results were not just numerically better; they also changed which measures performed best. This is consistent with the fact that hybrid spaces combine dimensions with different origins, scales, correlations, and semantic roles. A distance function that works well in a relatively homogeneous deep embedding may not be equally effective in a heterogeneous concatenated space.

Fourth, neighborhood size should be regarded as a metric-sensitive hyperparameter rather than a purely classifier-level setting. The k-value that appears optimal under AUC corresponds to a practically weak detector of the positive class. In imbalanced medical classification, the choice of k is not neutral with respect to clinical objectives. A conservative increase in k should be treated with caution when minority-class detection is the priority.

Fifth, evaluation methodology must include class-aware metrics. The finding that 11.6% of k-NN configurations achieve accuracy above 80% while sensitivity falls below 10% demonstrates that accuracy alone can mask clinically unacceptable failure modes. Sensitivity, F_1, and balanced accuracy should be treated as primary interpretive metrics rather than as secondary diagnostics appended after AUC.

B. CLINICAL INTERPRETATION

From a clinical perspective, the most consequential result is that stronger overall discrimination does not necessarily imply better malignant-case detection. This was visible most clearly in the behavior of large neighborhood sizes and in the sensitivity–specificity trade-offs across measures. The comparison between k-NN and NMD further highlights the clinical importance of the aggregation rule. NMD achieves more than 2.5× higher sensitivity than k-NN's because global averaging avoids the majority-voting collapse that suppresses minority-class recall at typical k values. For applications where sensitivity is the primary concern, NMD or similar global aggregation strategies may be preferable to k-NN, even if k-NN achieves higher accuracy.

A second clinical implication is that dataset-dependent validation is necessary before deployment-oriented conclusions are drawn. The observed instability of top feature–measure combinations across datasets suggests that a configuration validated on one benchmark cannot be assumed to transfer unchanged to another clinical environment. This is especially

relevant for mammography because acquisition conditions, compression, digitization, annotation conventions, and population characteristics may alter the representation geometry seen by distance-based classifiers.

C. RELATION TO PRIOR WORK

This empirical study extends the broader nearest-neighbor and distance-measure literature in two ways. Earlier benchmarking studies on general or tabular medical datasets [22], [23], [24] showed that no single measure dominates universally and that dataset characteristics influence the best choice. The present results support that broader conclusion, but they add an imaging-specific layer: the preferred distance depends not only on the dataset, but also on the feature family and the aggregation rule. Alfeilat et al. [22] found Lorentzian and Canberra among the top performers on tabular data; in our mammographic setting these measures rank 13th and 15th, while adaptive measures that were not included in their study dominate the top tier. This confirms that benchmarking results from tabular data do not transfer directly to image-based classification.

The results also complement prior mammography studies that used k-NN or retrieval-based similarity with handcrafted or deep features but often fixed Euclidean distance by default [16], [17]. The present work does not reject Euclidean distance outright; rather, it shows that Euclidean distance is only one point in a much larger design space and is not always the most effective choice. Similarly, the findings complement recent metric-learning studies [30] by showing that even within a plug-in, non-end-to-end framework, distance design still matters substantially. This is useful because plug-in distances preserve modularity and interpretability, both of which remain attractive properties for clinical decision-support pipelines.

The NMD comparison extends prior work that has typically evaluated distance measures under a single classifier. By showing that measure-level rankings change substantially between k-NN and NMD ($\rho = 0.077$) while instance-level agreement remains strong ($\rho = 0.94$), we provide evidence that part of the ranking reflects the intrinsic feature–measure compatibility and part depends on the aggregation rule. This distinction has not been systematically demonstrated in the mammography literature.

It is also instructive to compare the absolute performance levels observed in this study with those reported for non-distance-based classifiers on the same datasets. Galván-Tejada et al. [39] compared random forest, nearest centroid, and k-NN on mammographic descriptors and found that k-NN remained competitive with tree-based classifiers. A recent radiomics study comparing LDA, SVM, and k-NN on CBIS-DDSM mass classification reported AUC values of 0.63, 0.60, and 0.60, respectively, for the three classifiers [70], which closely matches our CBIS-DDSM result (best AUC = 0.652). End-to-end deep learning methods with fine-tuning and data augmentation achieve substantially higher AUC values, typically 0.80–0.90 on CBIS-DDSM [47], [48]. The gap between our results and end-to-end deep learning is expected and by design: we deliberately use frozen transfer features with simple classifiers to isolate the distance measure as the sole controlled variable (see Section IV). The relative rankings of distance measures, which are the primary contribution of this study, are independent of this absolute performance gap and are directly applicable to CBIR systems and case-based reasoning pipelines where plug-in distance functions remain the standard retrieval mechanism.

D. PRACTICAL RECOMMENDATIONS

For *handcrafted mammographic descriptors*, covariance-sensitive distances such as Learned Mahalanobis appear to be a reasonable first choice. For *deep feature embeddings*, adaptive weighted distances (AWE) or inner-product measures (Correlation, Cosine) are more reliable starting points. For *hybrid handcrafted–deep representations*, class-aware measures such as CCFCD appear especially promising, as the proposed measure achieved the highest AUC on this feature family. For computational simplicity, Correlation distance is recommended as a strong unsupervised default: it ranks fifth overall, requires no training phase, and has $O(d)$ per-pair complexity.

For comparative studies, including one local and one global aggregation strategy (e.g., k-NN and NMD) helps separate robust feature–measure effects from classifier-specific effects. Distance-based mammography experiments should not rely on a single value of k. At minimum, they should evaluate a small set of neighborhood sizes and inspect sensitivity and balanced accuracy alongside AUC.

Finally, result reporting should always include class-aware metrics. A configuration should not be recommended on the basis of AUC or accuracy alone if it exhibits poor malignant-case recall.

E. LIMITATIONS

First, although three datasets were used, they do not exhaust the range of mammographic acquisition conditions, annotation standards, and population variation encountered in practice. The observed dataset dependence is therefore informative but not definitive.

Second, the feature space was broad but still selective. The handcrafted, deep, and hybrid sets covered representative descriptors and CNN backbones, yet other radiomic formulations, self-supervised embeddings, or task-specific deep representations may produce different interactions with distance measures. In particular, deep features were extracted from ImageNet-pretrained networks without fine-tuning; fine-tuned features may have different statistical properties (better class separation, lower effective dimensionality) that could alter the optimal distance measure.

Third, the classifier scope was intentionally restricted to distance-based methods, specifically k-NN and NMD. This was appropriate for isolating the role of the distance function,

but it means that the results should not be overgeneralized to non-distance-based classifiers such as tree ensembles, SVMs, or end-to-end neural networks.

Fourth, k-NN was evaluated with uniform voting (unweighted). Distance-weighted k-NN variants, where closer neighbors contribute more to the classification, may interact differently with the evaluated measures and could mitigate some of the sensitivity collapse observed at large k.

Fifth, only z-score standardization was used for feature normalization. Alternative strategies, e.g., min–max scaling, L_2-normalization, or per-measure optimal normalization, may yield different results.

Sixth, the study focused on classification performance rather than on computational cost, calibration, or deployment robustness. Some measures may be attractive from an accuracy perspective while being less practical in resource-constrained or real-time settings.

F. FUTURE WORK

One direction is to extend the benchmark to more recent mammographic representations, including self-supervised or foundation-model embeddings, to test whether the same feature–measure interaction persists in more semantically structured deep spaces. Fine-tuned features extracted at multiple network layers would also clarify whether the observed rankings change when the feature extractor is adapted to the target domain.

A second direction is to examine whether feature normalization, subspace weighting, or learned feature grouping can further improve hybrid-distance design. Since hybrid spaces were the strongest overall but also the most heterogeneous, they are likely to benefit from more structured distance formulations.

A third direction is to evaluate cost-sensitive or recall-oriented neighborhood selection. The strong dependence of sensitivity on k suggests that a fixed global neighborhood size may be suboptimal for imbalanced clinical tasks. Adaptive, class-dependent, or distance-weighted neighborhood rules may offer a better balance between overall discrimination and malignant-case recall.

A fourth direction is to connect plug-in distance benchmarking with metric learning. Rather than treating these as competing paradigms, future work could use the present benchmark to identify strong inductive biases or initialization targets for learned similarity models in mammography.

Finally, extending the study to microcalcification classification, architectural distortion, and other imaging modalities (breast ultrasound, MRI, histopathology) would establish whether the observed feature–measure interactions generalize beyond mass-based mammographic classification.

VI. CONCLUSION

This paper presented a systematic study of the interaction between feature representations and distance measures in mammographic lesion classification. Twenty distance measures were evaluated across 13 feature sets (handcrafted, deep, and hybrid), three mammography datasets, two classifiers (k-NN and NMD), and 16 neighborhood sizes, yielding 13,260 configurations under 5-fold cross-validation.

Five principal conclusions emerge. First, distance measure choice significantly affects classification performance (Friedman $\chi^2 = 1901$, $p \approx 0$; AUC gap $= 0.079$). Euclidean distance ranks ninth, confirming that the common default is suboptimal. Second, the optimal measure depends on feature type: handcrafted features favor Learned Mahalanobis, deep features favor AWE, and hybrid features favor the proposed CCFCD, which achieves the highest AUC of any family–measure combination (0.712). Third, measure rankings are largely dataset-specific (cross-dataset ρ as low as -0.10), supporting dataset-level validation rather than universal measure selection. Fourth, k-NN and NMD agree strongly at the configuration level ($\rho = 0.94$) but weakly at the measure level ($\rho = 0.077$), showing that rankings are partially classifier-dependent. Fifth, increasing k improves AUC while degrading sensitivity (from 0.40 to 0.14), and 11.6% of configurations achieve accuracy above 80% with sensitivity below 10%, demonstrating that AUC and accuracy alone are insufficient for clinical evaluation. A computational complexity analysis shows that per-pair costs range from $O(d)$ for most measures to $O(d^2)$ for standard Mahalanobis, but the top-performing adaptive measures (AWE, CCFCD) remain computationally tractable at $O(d)$ and $O(2d)$ per pair.

These findings establish that distance-based mammographic classification is governed by a coupled interaction among feature representation, distance formulation, aggregation rule, neighborhood size, and evaluation metric. The practical implication is that these factors should be treated as a joint design problem rather than fixed independently. Within this design space, adaptive, covariance-sensitive, and class-aware measures consistently outperform conventional defaults, provided they are matched to the feature space and validated under clinically meaningful criteria.

Future work should extend this analysis to self-supervised and foundation-model representations, evaluate distance-weighted and recall-oriented neighborhood rules, and connect plug-in distance benchmarking with end-to-end metric learning for mammography.

ACKNOWLEDGMENT

The authors would like to thank the Higher Colleges of Technology (United Arab Emirates) and Volpara Health (New Zealand).

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 74, no. 3, pp. 229–263, May 2024.

[2] F. Burling-Claridge, M. Iqbal, and M. Zhang, "Evolutionary algorithms for classification of mammographie densities using local binary patterns and statistical features," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2016, pp. 3847–3854.

[3] A. Siddique, M. Iqbal, and W. N. Browne, "A comprehensive strategy for mammogram image classification using learning classifier systems," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2016, pp. 2201–2208.

[4] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. V. D. Laak, B. V. Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Sep. 2017.

[5] J. G. Elmore, C. K. Wells, C. H. Lee, D. H. Howard, and A. R. Feinstein, "Variability in radiologists' interpretations of mammograms," *New England J. Med.*, vol. 331, no. 22, pp. 1493–1499, Dec. 1994.

[6] A. Jouirou, I. Souissi, and W. Barhoumi, "Two-level content-based mammogram retrieval using the ACR BI-RADS assessment code and learning-driven distance selection," *J. Supercomput.*, vol. 80, no. 11, pp. 15690–15724, Jul. 2024.

[7] M. S. Rahman, F. Humayara, S. M. E. Rabbi, and M. M. Rashid, "Advanced multi-architecture deep learning framework for BIRADS-based mammographic image retrieval: Comprehensive performance analysis with super-ensemble optimization," *J. Imag. Informat. Med.*, Dec. 2025.

[8] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., Hoboken, NJ, USA: Wiley, 2001.

[10] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vols. SMC–3, no. 6, pp. 610–621, Nov. 1973.

[11] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.

[13] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[14] X. Huang, Z. Li, M. Zhang, and S. Gao, "Fusing hand-crafted and deep-learning features in a convolutional neural network model to identify prostate cancer in pathology images," *Frontiers Oncol.*, vol. 12, Sep. 2022, Art. no. 994950.

[15] M. A. Jones, R. Faiz, Y. Qiu, and B. Zheng, "Improving mammography lesion classification by optimal fusion of handcrafted and deep transfer learning features," *Phys. Med. Biol.*, vol. 67, no. 5, Mar. 2022, Art. no. 054001.

[16] A. Papadopoulos, D. I. Fotiadis, and A. Likas, "Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 141–150, Jun. 2005.

[17] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, "Benign and malignant breast tumors classification based on region growing and CNN segmentation," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 990–1002, Feb. 2015.

[18] H. Singh, V. Sharma, and D. Singh, "Comparative analysis of proficiencies of various textures and geometric features in breast mass classification using k-nearest neighbor," *Vis. Comput. for Ind., Biomed., Art*, vol. 5, no. 1, p. 3, Dec. 2022, doi: 10.1186/s42492-021-00100-1.

[19] F. Yan, H. Huang, W. Pedrycz, and K. Hirota, "Automated breast cancer detection in mammography using ensemble classifier and feature weighting algorithms," *Expert Syst. Appl.*, vol. 227, Oct. 2023, Art. no. 120282, doi: 10.1016/j.eswa.2023.120282.

[20] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is 'nearest neighbor' meaningful?" in *Proc. 7th Int. Conf. Database Theory (ICDT)*, vol. 1540. Cham, Switzerland: Springer, 1999, pp. 217–235.

[21] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in *Proc. 8th Int. Conf. Database Theory*, vol. 1973. Cham, Switzerland: Springer, 2001, pp. 420–434.

[22] H. A. A. Alfeilat, A. B. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. E. Salman, and V. B. S. Prasath, "Effects of distance measure choice on K-nearest neighbor classifier performance: A review," *Big Data*, vol. 7, no. 4, pp. 221–248, 2019.

[23] L.-Y. Hu, M.-W. Huang, S.-W. Ke, and C.-F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *SpringerPlus*, vol. 5, no. 1, p. 1304, Dec. 2016.

[24] R. Ehsani and F. Drabløs, "Robust distance measures for kNN classification of cancer data," *Cancer Informat.*, vol. 19, Jan. 2020, Art. no. 1176935120965542.

[25] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Models Methods Appl. Sci.*, vol. 1, no. 4, pp. 300–307, 2007.

[26] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, "An empirical study of distance metrics for k-nearest neighbor algorithm," in *Proc. 2nd Int. Conf. Ind. Appl. Eng.*, 2015, pp. 280–285.

[27] L. Li, X. Shang, J. Li, and J. Hu, "Learning distance metrics for entity resolution," *IEEE Access*, vol. 6, pp. 54900–54909, 2018.

[28] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, Jordan, 2002, pp. 521–528.

[29] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, no. 9, pp. 207–244, 2009.

[30] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A parasitic metric learning net for breast mass classification based on mammography," *Pattern Recognit.*, vol. 75, pp. 292–301, Mar. 2018.

[31] L. Jiao, X. Geng, and Q. Pan, "BP*k* NN: *K* -nearest neighbor classifier with pairwise distance metrics and belief function theory," *IEEE Access*, vol. 7, pp. 48935–48947, 2019.

[32] X. Zhou et al., "A similarity learning approach to content-based image retrieval: Application to digital mammography," *Med. Phys.*, vol. 31, no. 11, pp. 2828–2836, 2004.

[33] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Wavelet optimization for content-based image retrieval in medical databases," *Med. Image Anal.*, vol. 14, no. 2, pp. 227–241, Apr. 2010.

[34] K. Loizidou, R. Elia, and C. Pitris, "Computer-aided breast cancer detection and classification in mammography: A comprehensive review," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106554, doi: 10.1016/j.compbiomed.2023.106554.

[35] M. Muštra, M. Grgić, and K. Delač, "Breast density classification using multiple feature selection," *Automatika*, vol. 53, no. 4, pp. 362–372, Jan. 2012.

[36] S. R. Sannasi Chakravarthy, N. Bharanidharan, and H. Rajaguru, "Deep learning-based Metaheuristic weighted K-nearest neighbor algorithm for the severity classification of breast cancer," *IRBM*, vol. 44, no. 3, Jun. 2023, Art. no. 100749, doi: 10.1016/j.irbm.2022.100749.

[37] V. Sridevi, "Breast cancer examination in digitized mammograms using integrated K-means clustering with garbor filter and shrunk kernel KNN method," *Indian J. Sci. Technol.*, vol. 17, no. 23, pp. 2444–2454, May 2024, doi: 10.17485/ijst/v17i23.736.

[38] O. Attallah, "MERGE: Mammogram-enhanced representation via wavelet-guided CNNs for computer-aided diagnosis of breast cancer," *Mach. Learn. Knowl. Extraction*, vol. 8, no. 2, p. 40, Feb. 2026, doi: 10.3390/make8020040.

[39] C. Galván-Tejada, L. Zanella-Calzada, J. Galván-Tejada, J. Celaya-Padilla, H. Gamboa-Rosales, I. Garza-Veloz, and M. Martinez-Fierro, "Multivariate feature selection of image descriptors data for breast cancer with computer-assisted diagnosis," *Diagnostics*, vol. 7, no. 1, p. 9, Feb. 2017.

[40] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci. Rep.*, vol. 12, no. 1, p. 6256, Apr. 2022.

[41] S. Shilaskar, A. Ghatol, and P. Chatur, "Medical decision support system for extremely imbalanced datasets," *Inf. Sci.*, vol. 384, pp. 205–219, Apr. 2017.

[42] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 10, pp. 6567–6572, May 2002.
[43] J. S. Sánchez, F. Pla, and F. J. Ferri, "On the use of neighbourhood-based non-parametric classifiers," *Pattern Recognit. Lett.*, vol. 18, nos. 11–13, pp. 1179–1186, Nov. 1997.
[44] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A generalized mean distance-based k-nearest neighbor classifier," *Expert Syst. Appl.*, vol. 115, pp. 356–372, Aug. 2019.
[45] F. Pesapane, C. Trentin, M. Montesano, F. Ferrari, L. Nicosia, A. Rotili, S. Penco, M. Farina, I. Marinucci, F. Abbate, L. Meneghetti, A. Bozzini, A. Latronico, A. Liguori, G. Carrafiello, and E. Cassano, "Mammography in 2022, from computer-aided detection to artificial intelligence applications," *Clin. Experim. Obstetrics Gynecol.*, vol. 49, no. 11, p. 237, 2022, doi: 10.31083/j.ceog4911237.
[46] L. Wang, "Mammography with deep learning for breast cancer detection," *Frontiers Oncol.*, vol. 14, Feb. 2024, Art. no. 1281922, doi: 10.3389/fonc.2024.1281922.
[47] G. Ayana, J. Park, and S.-W. Choe, "Patchless multi-stage transfer learning for improved mammographic breast mass classification," *Cancers*, vol. 14, no. 5, p. 1280, Mar. 2022.
[48] L. Xia, J. An, C. Ma, H. Hou, Y. Hou, L. Cui, X. Jiang, W. Li, and Z. Gao, "Neural network model based on global and local features for multi-view mammogram classification," *Neurocomputing*, vol. 536, pp. 21–29, Jun. 2023.
[49] S. Vijayalakshmi, B. K. Pandey, D. Pandey, and M. E. Lelisho, "Innovative deep learning classifiers for breast cancer detection through hybrid feature extraction techniques," *Sci. Rep.*, vol. 15, no. 1, p. 22212, Jul. 2025.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[52] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.
[53] R. Short and K. Fukunaga, "The optimal distance measure for nearest neighbor classification," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 5, pp. 622–627, Sep. 1981.
[54] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. Nat. Inst. Sci. India*, vol. 2, pp. 49–55, May 1936.
[55] J. Suckling, J. Parker, S. Astley, I. Hutt, C. Boggis, I. W. Ricketts, E. A. Stamatakis, N. Cerneaz, S. Kok, P. Taylor, D. Betal, and J. Savage, "The mammographic image analysis society digital mammogram database," *Exerpta Medica*, vol. 1069, pp. 375–378, Mar. 1994.
[56] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, no. 1, Dec. 2017, Art. no. 170177.
[57] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a full-field digital mammographic database," *Academic Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.
[58] K. J. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed., New York, NY, USA: Academic, 1994, pp. 474–485.
[59] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.
[60] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.
[61] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[62] Simonyan, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–18. [Online]. Available: https://www.robots.ox.ac.uk/
[63] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.
[64] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[65] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[66] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.
[67] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, Dec. 1945.
[68] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandin. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
[69] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, 2006.
[70] N. Lauciello, G. Pasini, G. Russo, F. Marinozzi, F. Bini, and A. Stefano, "Machine learning in mammography classification: A radiomics-based evaluation and developments," in *Image Analysis and Processing—ICIAP 2025 Workshops*, E. Rodolà, F. Galasso, and I. Masi, Eds., Cham, Switzerland: Springer, 2026, pp. 199–209.

MUHAMMAD IQBAL received the Ph.D. degree from Victoria University of Wellington, New Zealand, in 2014.

He is currently an Assistant Professor with the Higher Colleges of Technology, Fujairah, United Arab Emirates, with over 20 years of academic and research experience. Previously, he has been leading the Research and Development Team with Xtracta Ltd., New Zealand, to develop artificial intelligent powered information extraction systems capable of extracting data from scanned, photographed, or digital documents, with a focus on the financial domain. His research is primarily focused on pattern recognition, with an emphasis on applying computational intelligence and transfer learning techniques. His main research interest includes computational intelligence.

TALAL ASHRAF BUTT received the Ph.D. degree in computer science from Loughborough University, U.K.

He has pioneered adaptive, context-aware service discovery protocols for the Internet of Things with Loughborough University. He is currently an Assistant Professor with the Higher Colleges of Technology, Fujairah, United Arab Emirates. Over his 15 year academic and research career, he has published extensively in high-impact venues. His interdisciplinary research spans the IoT, the Social Internet of Things, the Internet of Vehicles, artificial intelligence, blockchain, and big data analytics. He maintains active collaborations with industry partners, translating theoretical advances into practical solutions. Committed to pedagogical excellence, he employs active-learning strategies and emerging technologies in the classroom and dedicates himself to mentoring the next generation of scholars and engineers. He is a FHEA.

ZAHID SHAREEF is an educator and a researcher with more than 18 years of international academic experience across U.K., United Arab Emirates, Oman, and Pakistan. He is currently a Faculty Member with the Higher Colleges of Technology, United Arab Emirates, where he teaches quantitative and analytical courses and contributes to curriculum development, assessment design, and course coordination. He actively integrates digital technologies and computational thinking into his teaching and scholarly activities. His work has been published in several peer-reviewed international journals. His research interests include the intersection of computational methods, mathematical modeling, and data-driven problem solving. His work explores analytical and computational approaches to complex systems, including modeling biological processes and other real-world phenomena. He is particularly interested in applying computational and programming techniques to problems in mathematical biology and interdisciplinary scientific domains.

Dr. Shareef is a fellow of Advance HE, U.K.; and a Recognized Expert on United Arab Emirates Research Map in Computer, Information Sciences, and Mathematics.

ABUBAKAR SIDDIQUE (Member, IEEE) is currently a Faculty Member with Auckland University of Technology, New Zealand. Prior to joining academia, he spent nine years with Elixir Technologies, Pakistan, a California (USA)-based leading software company. Rising to the role of Principal Software Engineer, he led a team of developers in designing and delivering enterprise-level software solutions for global clients, including Xerox, IBM, and Adobe. He has shared his expertise by delivering five tutorials and talks at various forums, including the Genetic and Evolutionary Computation Conference (GECCO). His main research interests include creating novel machine learning systems, inspired by the principles of cognitive neuroscience, to provide efficient and scalable solutions for challenging and complex problems in different domains, such as biomedical, computer vision, navigation, and NLP.

Dr. Siddique has received several distinctions, including the Student of the Session Award, the VUWSA Gold Award, and the Emerging Research Excellence Medal, throughout his academic career. He serves the academic community as an author for prestigious journals and international conferences, including IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, and GECCO.

WILL N. BROWNE (Member, IEEE) is currently a Professor and the Chair of manufacturing robotics with Queensland University of Technology, Brisbane, QLD, Australia. He has co-authored the first textbook on *LCSs Introduction to Learning Classifier Systems* (Springer, 2017). His research interest includes applied cognitive systems. Specifically, how to use inspiration from natural intelligence to enable computers/machines/robots to behave usefully. This includes cognitive robotics, learning classifier systems, and modern heuristics for industrial applications. He has been the Co-Track Chair of the Genetics-Based Machine Learning (GBML) Track and the Co-Chair of the Evolutionary Machine Learning Track at the Genetic and Evolutionary Computation Conference. He has also provided tutorials on rule-based machine learning and advanced learning classifier systems at GECCO, chaired the International Workshop on Learning Classifier Systems (LCSs), and lectured graduate courses on LCSs.