Int. J. Advance Soft Compu. Appl, Vol. 18, No. 2, July 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Beyond Accuracy: Protocol-Dependent Robustness and Reliability in Patch-Based Breast Histopathology Classification

Suleyman A. AlShowarah[1, 2], Wael Alzyadat[2], Ameen Shaheen[2], Aysh Alhroob[2]

[1]Department of Software Engineering, Faculty of Information Technology, Mutah University, Jordan
[2]Department of Software Engineering, Al-Zaytoonah University of Jordan, Amman, Jordan

*Corresponding Author: showarah@mutah.edu.jo

Abstract

Breast cancer is a leading cause of death among women worldwide. Early diagnosis can save many lives and this will lead to a proper treatment. Histopathological image analysis is an important diagnostic method for detecting breast cancer. Computer-aided diagnosis of breast images helps radiologists do the task more efficiently and appropriately. The reliable evaluation of deep learning models in computational pathology depends strongly on how the data is split during validation. In patch-based histopathology classification, random image-level splitting can obscure inter-patient heterogeneity and produce overly stable performance estimates. This study investigates the effect of the evaluation protocol on Invasive Ductal Carcinoma (IDC) detection using the Breast Histopathology Images dataset. A lightweight Convolutional Neural Network (CNN) was trained using balanced sampling and evaluated under two 5-fold cross-validation schemes: strict patient-level separation and stratified random image-level splitting. Under patient-level evaluation, the model achieved a mean accuracy of 0.7826 ± 0.0295 and a mean AUC of 0.8691 ± 0.0257, with a Brier score of 0.1582 ± 0.0159. Random splitting produced similar mean accuracy (0.7867 ± 0.0096) and AUC (0.8652 ± 0.0053), with slightly improved Brier score (0.1509 ± 0.0047). While paired statistical testing showed no significant differences in mean performance ($p > 0.05$), patient-level evaluation exhibited substantially higher fold-wise variability. Calibration analysis indicated moderate reliability with mild overconfidence, and error inspection revealed consistent morphological failure patterns. Overall, the evaluation protocol mainly affects perceived robustness rather than average performance, highlighting the importance of patient-level validation and variability reporting.

Keywords: *Deep Learning, Breast Cancer, Histopathology, Convolutional Neural Networks, Evaluation protocol, Computational Pathology*

1 Introduction

Breast cancer remains the most frequently diagnosed cancer among women worldwide and a leading cause of cancer-related mortality [1]. Histopathological examination of Hematoxylin and Eosin (H&E) stained tissue slides remains the gold standard for definitive diagnosis and grading [2]. However, manual microscopic evaluation is time-consuming, subjective, and prone to inter-observer variability, particularly in borderline or heterogeneous tissue regions [3]. Deep learning, especially convolutional neural network (CNN), has recently shown strong performance in histopathological image classification

Received 22 January 2026; Accepted 15 March 2026

tasks, such as detecting Invasive Ductal Carcinoma (IDC) [4]. Additionally, publicly available datasets, such as the Breast Histopathology Images (IDC) dataset, enable to reproduce benchmarked results for the automated cancer detection [5]. Most reported results are also measured based on accuracy and Area Under Curve (AUC) values that suggest good potential clinically [6].

Despite there are many advancements in detection systems for the histopathology-based machine learning. But, there are still number of studies that concerns about the very important methodologies used in the breast cancer detection, which are under-investigated. A major problem with using histopathology using machine learning is data leakage as a result of improperly divided images [7]. In cases where multiple image patches from the same patient appear in both the training set and the testing set (image-level split), the model learns the specific textures from the individual patient instead of the morphological characteristics that define the disease. Thus, this results in overly optimistic estimates of performance that do not accurately represent how well the model will perform in real-world use [8].

Although multiple studies such as [9-11] have documented the need for a high degree of separation at the level of individual patients. As stated, there is still a very limited number of systematic and quantitative studies based on patient-level data that were compared to those using randomly selected images especially when considering IDC dataset. Although, beyond the classification accuracy, model reliability, calibration quality, and prediction stability across the different folds are important aspects for the successful translation into a clinical setting, they are often inadequately reported.

This study provides a leakage-aware and reliability-centered assessment of deep learning architectures to detect IDC using the Breast Histopathology Images. The focus is not only on achieving the greatest possible accuracy in the classification of images, but also on assessing the effects for the evaluation method that has effects on both the variability and probabilistic calibration using deep learning models.

Specifically, this study makes the following contributions:

- Leakage-Aware Evaluation Framework: To maintain consistent experimental conditions, a strict 5-fold patient-level cross-validation was performed, comparing our results to a stratified random image-level split.

- Quantitative Assessment of Protocol Sensitivity: The study compares the results of fold-wise accuracies, AUC, and Brier scores. Whereas the study further uses paired statistical testing to determine if different methods of evaluating models have a significant impact on estimates of model performance.

- Reliability and Calibration Analysis: In addition to the analysis of discrimination ability, the calibration of the model using reliability curves and brier scores were evaluated; therefore, the study can assess the quality of the confidence provided by the model.

- Error Pattern Characterization: The study performed an examination of the qualitative aspects of errors (both false positive and false negative) and identified the morphological characteristics of the errors, which indicate where

the model fails, including low-cellularity malignant areas and structurally complex benign tissue.

- Variance and Stability Analysis Across Folds: The study demonstrated that the method used to evaluate the model does not only affect the average performance of the model, but also affects the variance of the model's performance across folds, which is important for determining the amount of information required when reporting about the stability of a model in the medical AI research.

Findings from this study reveal that the average performance difference between patient-level splits and random splits is often not statistically significant when sampling is balanced. However, random splits produce lower variance estimates, which may mask patient-level heterogeneity.

This research is designed into five sections. The overview of related studies in literature is introduced in Section II. Section III discusses the methodology for the proposed model and the experiment design. Then, the results analysis is presented in Section IV. Whilst Section V discussed the results of the study. Finally, Section VI presents the conclusion.

2 Related Work

Deep learning is currently the most prevalent method of computational pathology to enable automatic identification and classification of malignant tissue in digitized histopathology slides [12]. Such research in this area was based on the use of handcrafted features and traditional machine learning methods; however, both were limited due to the lack of expressiveness in the development of features and due to the color variability from one type of staining to another. Numerous research has shown significant improvement in the performance of multiple tasks related to histopathology using representation learning with CNNs [13].

2.1 Deep Learning for Breast Histopathology and IDC Detection

Histopathology images from breast cancer have been widely analyzed as a result of their clinical relevance in the detection of early-stage cancer and the availability of public image databases [14]. The authors of [15] are notable for introducing what is likely the most commonly used dataset for IDC patch-level breast histopathology, along with an evaluation of both classical and CNN-based approaches for the detection of IDC. Since then, studies have shown that utilizing a large-scale pre-trained CNN can enhance the ability of the model to generalize across new datasets, especially for small medical datasets [16, 17].

While several works studied deeper architectures like ResNet and Inception variants and showed strong discrimination performance for patch-based IDC classification [18, 19], the majority of the works focused mainly on improving the accuracy or AUC performance and provided limited analyses on robustness, uncertainty, or leakage in the evaluations.

2.2 Evaluation Protocols and Data Leakage in Medical Imaging

A number of studies have shown how a variety of methods used for evaluation design directly affect the results produced by medical AI. One major issue is improper splitting methods, which may lead to data leakage where inflated metrics and lower

reproducibility result [20]. This type of leakage occurs within patch-based histopathology data because the split can be done at the level of the individual slide or the patient-level. When patches that come from the same patient or slide are assigned to both the training and testing groups, this creates patient-specific memorization and reduces the potential for the model to generalize well to different cases of the disease [21].

The issue of data leakage and biased evaluation is also found in all types of medical imaging techniques, such as radiology and pathology [22]. In their paper [23] point out that many of the published medical AI systems do not perform at the same level when they are tested using an external validation, and this can be explained by several methodological shortcomings, among them the lack of representativeness in the test sets used and the presence of biases within the datasets. Hidden stratification was reported in [24] as another major cause of clinically significant failure; these are models that perform well overall but fail in very specific subgroups. Therefore, these findings reinforce that there is a need for methods of assessment that take into account potential data leaks and thus reflect the actual conditions in clinical practice.

2.3 Reliability, Calibration, and Uncertainty in Clinical AI

In addition to their classification accuracy, a reliable decision-support system must also offer calibrated probability estimates of its output. Most neural networks are poorly calibrated and produce overly confident probability estimates that are not reflective of the actual likelihood. This issue is currently a growing area of study; researchers have proposed methods, including temperature scaling and reliability diagrams, to measure and improve the probabilistic quality of models [25]. Clinicians may make decisions based upon the output confidence of predictive models in a clinical setting, which could be particularly dangerous since they are relying upon the model's confidence to determine case triage or workload priority [26].

Despite the limited study of calibration and quantifying uncertainty in computational pathology compared to optimizing model performance, recent research indicates that evaluating models with uncertainty awareness is more interpretable and represents a safer path toward integrating into the clinical setting [27], and both Brier score and ECE are increasingly being suggested as additional reliability metrics to report alongside AUC and accuracy [28].

2.4 Error Analysis and Model Failure Characterization

A limitation in many research on deep learning for pathology is a complete lack of systematic evaluation of errors in these models [29]. As such, reporting only average statistics hides clinically important failures like borderline cases, artifacts from staining, and ambiguous morphology, which lead to misclassifications [30].

Recent work in interpretable pathology has shown that false positives often arise from benign tissue with high cellular density or inflammation, while false negatives may occur in low-tumor-content regions or heterogeneous slides [31]. Such patterns motivate the inclusion of qualitative error analysis as a complement to quantitative evaluation.

In summary, previous research has demonstrated that CNN-based approaches can effectively be used in patch-based methods of detecting breast cancer. Nonetheless, there are significant gaps that have not been sufficiently addressed in the context of the IDC dataset:

- Systematized comparative evaluations of patient-level and randomly selected image-level leakage awareness.
- Calibration and reliability reporting beyond traditional measures such as accuracy and AUC.

- Structural analysis of failure patterns associated with visual features of errors.

This work will address these gaps through the use of an empirically grounded and reliability-centered evaluation methodology (cross-validated at the patient-level), including comparative protocol assessments, calibration assessments, and qualitative analyses of errors.

3 Methodology

3.1 Dataset

This study utilizes the publicly available Breast Histopathology Images (IDC) dataset that includes Hematoxylin and Eosin (H&E) -stained patches of breast tissue derived from the digitization of whole slide images of the breast tissue. These patches are classified into one of two groups based on whether they contain Invasive Ductal Carcinoma (IDC positive) or non-IDC (benign). As a result, the IDC dataset forms a binary classification task. The dataset is organized in a manner that is consistent with patient identification in that each patient's directory contains two subdirectories that correspond to the two different classes [32].

The dataset from the experimental design of this study includes 279 patients with no duplicate patients. The patient-level structure of the dataset was used to enable a strict cross-validation model that prevented overlap between test and train sets. This is important for preventing information leaking into the models during patch-based histopathology classification experiments. An overview of the data set characteristics and experimental design is described in Table 1.

Table 1. Dataset Summary and Experimental Configuration

Property	Value
Number of patients	279
Task	Binary classification (IDC vs Non-IDC)
Nominal patch size	50 × 50 pixels
Color space	RGB
Cross-validation scheme	5-fold
Sampling strategy	Balanced per class
Training samples per fold	4000 per class
Testing samples per fold	1000 per class

3.2 Preprocessing and Sampling

All patches were loaded into RGB format and scaled to the interval [0, 1]. The dataset is generally referred to as having 50 × 50 patch sizes; however, there was a small number of images with some minor size irregularities. Those patches that had size irregularities were uniformly resized to 50 × 50 prior to model training. Only 285 images were resized across all of the patient-level 5-fold evaluations, which represents an exceptionally consistent dataset.

In order to have comparable data for all folds of the cross-validation as well as for different evaluation protocols, a stratified sampling method (balanced sampling) was chosen. From each partition for every fold, 1000 test samples and 4000 train samples were randomly selected from each class. Stratified sampling is one of the most frequently used methods for controlling the class-imbalance problem in the field of medical image classification when the amount of available computing resources is limited, and a stable estimation of the model's performance has to be achieved [33].

3.3 Evaluation Protocols

A central objective of this study is to assess how evaluation design influences reported performance. Therefore, two distinct cross-validation strategies were implemented under controlled experimental conditions.

Firstly, a strict patient-level 5-fold cross-validation procedure was implemented. A total of five groups of patients were randomly selected using an exact same random number generator seed, to ensure results could be reproduced. In every iteration of the 5-fold, all image patches of a single patient were assigned either to the training group, or the test group; this method assures that there will be no patient specific morphological characteristics that would be shared by both the training and testing images, thus reflecting a real-world situation in which a trained model would encounter completely new patients at the time of deployment. Cross-validation on a per-patient basis is recommended for clinical AI evaluations and has been cited as a critical need for the successful development of medical machine learning [34].

A second approach to evaluate the model is by using a stratified random image-level 5-fold cross-validation protocol. The patches were split randomly in this case, independent of the patients that they came from. Therefore, patches coming from the same patient can be found in either the train set or the test set. This method reflects standard methods of evaluating patches at the level of the patch; it also allows to quantify the amount of bias due to leakage. To keep the experiments computationally feasible, yet representative of all cases, up to 50 patches of each class per patient were sampled, resulting in a total number of 27,259 images in the random-split experiment.

3.4 Model Architecture and Training

A reproducible model for a three-layered CNN has been developed. The number of filters in each convolutional layer is increased, while a pooling operation follows after each convolutional layer. With this approach, the increasingly high-level features such as textures and morphology are captured. Then global average pooling is performed. Finally, a fully connected layer follows, which reduces overfitting with dropout. The last layer had a sigmoid activation function to produce probability outputs for the binary classification task [35], because CNNs have shown excellent results for the classification of patches from histopathology images in the past, and this is due to the ability of the CNNs to capture discriminant texture and spatial relationships.

All models were trained using the Adam optimizer as well as a learning rate of 0−310^ {-3}10−3, binary cross entropy loss, and a batch size of 64 [36]. Each fold was trained for 5 epochs, and each epoch used 20% of the training data as an internal validation set. A fixed random seed (42) was also used to be able to reproduce experiments. We did all of our training on a CPU-only environment utilizing TensorFlow 2.20.

3.5 Performance Metrics and Reliability Assessment

To conduct an overall assessment of the model, both discrimination and reliability metrics were evaluated. The discrimination of the model was evaluated by calculating accuracy and the area under the receiver operating characteristic curve (AUC) as measures of how well the model can distinguish malignant patches from benign patches. However, neither of these metrics measures the quality of the predicted probability values.

To evaluate probabilistic reliability, we report the Brier score [37], defined as in Eq. (1)

$$Brier = \frac{1}{N} \sum_{i=1}^{n} (p_i - y_i)^2 \qquad (1)$$

Beyond numbers, calibration plots are used to show how the predicted confidence levels for an image compare to the actual number of times that the same type of image is correct. Further, a qualitative error assessment was performed through examination of false positive and false negative patch examples to establish patterns in morphology related to the misclassification of images.

3.6 Statistical Analysis

To assess whether the variations observed between patient-level and random image-level protocols at the statistical level were significant, the authors performed a paired t-test of the metrics they analyzed for each of the five splits or folds [38]. This paired approach was applied to account for the potential variability from the different fold levels that are often used in machine learning comparative analysis to provide a statistical framework for comparing protocols.

4 Result Analysis

This This section will present the experimental data collected on the Baseline CNN model using two different assessment methods: (i) patient-level 5-fold cross-validation strictly, and (ii) stratified 5-fold cross-validation randomly at an image-level. Both discrimination (accuracy and AUC) will be assessed, as well as reliability and stability (calibration metrics, Brier score, and calibration plots) through the study of fold-by-fold variability and qualitative visual error analysis. By providing these three levels of analysis for the same model, the results provide a more clinically relevant description of how a model behaves than that which would be provided by accuracy measures alone.
The experiment design has kept the same model architecture, training configuration, and the way images were sampled for each dataset so as to isolate the effect of the test protocol on the performance of the models. Therefore, the results from the two protocols are comparable.

4.1 Patient-Level 5-Fold Cross-Validation Performance

The main assessment utilized a very strict patient-level separation to limit the possibility of leakages. In Table 2, the accuracy, AUC, and Brier score for each fold were reported. Table 2 shows that the model had an average accuracy of 0.7826 with an associated standard deviation of 0.0295; similarly, the model had an average AUC of 0.8691 ± 0.0257, indicating strong ranking ability, considering both the compact structure of the model and the few training epochs performed on the data.
The variance between folds was significant. The accuracy varied from 0.7400 in Fold 5 to 0.8170 in Fold 2. These fold-to-fold variations reflect a real clinical variation that exists with the heterogeneity of breast morphology in patients, as some patients have more difficult benign patterns to detect, slight malignant infiltration, or staining differences. The clinical realism of this variance further supports why patient-specific analysis must be performed in order to produce reliable reports.

The Brier scores for calibration ranged from 0.1444 to 0.1834, which had an average of 0.1582 ± 0.0159. As the lowest AUC was observed in the least reliable fold, this indicates reliability loss is possible along with discrimination loss. Therefore, reporting calibration related metrics in addition to AUC is important. Table 2 shows fold-wise performance under strict patient-level separation.

Table 2. Patient-Level 5-Fold Cross-Validation Results

Fold	Accuracy	AUC	Brier
1	0.7955	0.8750	0.1444
2	0.8170	0.8958	0.1511
3	0.7680	0.8649	0.1637
4	0.7925	0.8817	0.1482
5	0.7400	0.8278	0.1834
Mean ± Std	**0.7826 ± 0.0295**	**0.8691 ± 0.0257**	**0.1582 ± 0.0159**

4.2 Random Image-Level 5-Fold Cross-Validation Performance

To assess the impact of the evaluation approach on classification performance, the experiment was duplicated with a stratified random image-level cross-validation procedure, as shown in Table 3. In addition to similar mean accuracy values (mean = 0.7867 ± 0.0096), the model also had similar mean AUC values (mean = 0.8652 ± 0.0053). Since both were only marginally different, the comparison suggests that the method used to split data into training and test sets, random vs. patient-based, did not lead to large increases in discriminatory ability when using balanced sample sizes as the constraint for this study.

However, an important difference in terms of variability is observed when the variability of the different methods is compared. The standard deviation of accuracy under random splitting (.0096) is significantly less than under patient-level splitting (.0295), and there is a significant reduction in the variance of AUC. Therefore, random splitting will provide a more consistent performance estimate between splits; however, this could mislead as to the amount of patient-level variability that is seen with patient-wise splitting. In terms of reliability, the Brier score was consistently lower and more stable across folds (mean 0.1509 ± 0.0047). This suggests that random splitting may also provide more optimistic calibration estimates, potentially because patient-specific characteristics shared between training and testing reduce uncertainty.

Table 3. Random Image-Level 5-Fold Cross-Validation Results

Fold	Accuracy	AUC	Brier
1	0.7935	0.8686	0.1471
2	0.7773	0.8653	0.1560
3	0.7821	0.8629	0.1525
4	0.7804	0.8576	0.1541
5	0.8000	0.8714	0.1450
Mean ± Std	**0.7867 ± 0.0096**	**0.8652 ± 0.0053**	**0.1509 ± 0.0047**

4.3 Direct Protocol Comparison and Statistical Significance

To provide a direct quantitative comparison between patient-level and random image-level evaluation protocols, fold-wise paired differences were computed and statistical significance was assessed using paired t-tests. Table 4 summarizes the mean paired differences and p-values for accuracy, AUC, and Brier score.

The results demonstrate that the observed differences in performance metrics were not statistically significant ($p>0.05$ for all performance metrics). In particular, the random split of the data increased the mean accuracy by a mere 0.0041, it reduced the mean AUC by 0.0039, and reduced the average Brier score by 0.0072. These changes, therefore, represent minor impacts on model performance due to leakage inflation under an equal sample size sampling scheme.

Although there are similarities in average performance between both methods, the evaluation process significantly impacts fold-wise variability of the model's performance. The higher dispersion of performance at the individual-patient-level reflects the true difficulty of developing a model that will be generalizable to different unseen patients. The lower variability of estimates developed using a random split may potentially fail to capture clinical variation. Differences like this have significant implications for translating results into clinical settings, as they will need to be able to generalize to many patient populations.

Table 4. Patient-Level vs Random-Level Comparison (Paired Across Folds)

Metric	Δ Mean (Random − Patient)	p-value
Accuracy	+0.0041	0.8170
AUC	−0.0039	0.7794
Brier	−0.0072	0.4375

4.4 Calibration and Reliability Analysis

Beyond the measurement of discrimination, the model's confidence ratings were assessed against actual correctness. Fig.1 provides a calibration plot based on the individual-patient-level. This calibration curve indicates that while the model appears to be relatively well-calibrated at lower levels of confidence when the model predicts less than a 0.3 chance of the patient having breast cancer, it is also fairly accurate at higher confidence levels when the model predicts more than an 0.8 chance.

However, slight over-confidence exists for the medium probability range approximately (0.4-0.6) as the predictive probabilities slightly exceed the corresponding true positive rates. This can also be seen to reflect typical calibration properties of present-day artificial neural networks, which are well-documented to show some level of over-confidence in the middle region of their decisions. Values for the Brier Score in Table 2 also illustrate that while there is a moderate degree of probabilistic reliability, the reliability is not complete.

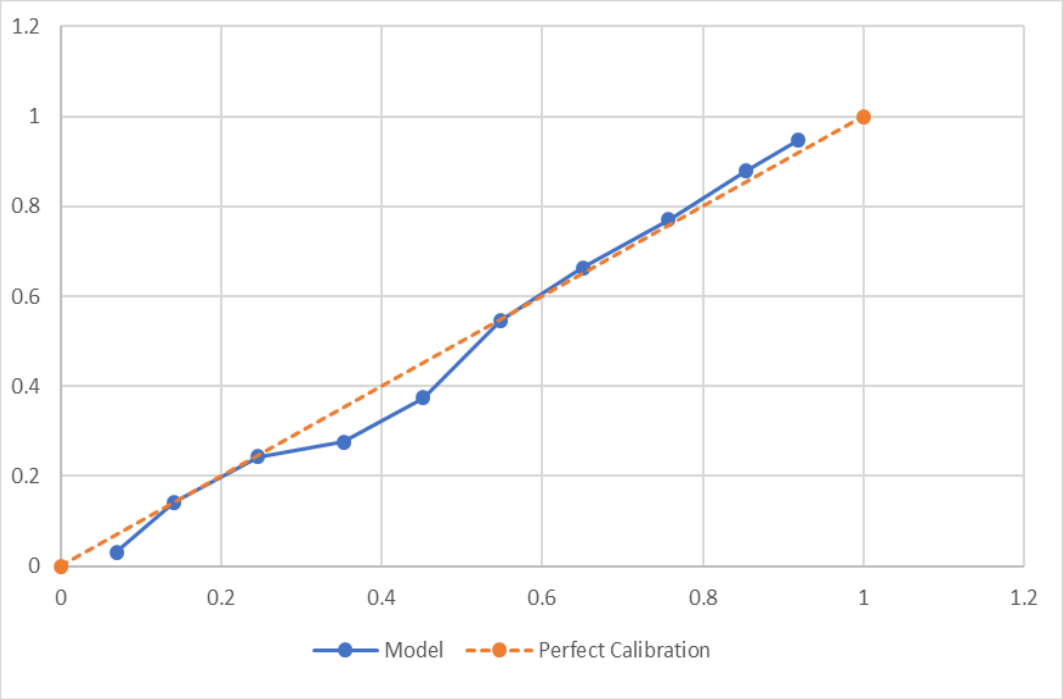

Fig. 1: Calibration Curve Under Patient-Level Evaluation

Fig. 1 demonstrates that the model exhibits good calibration at both low and high predicted probabilities, with predictions closely aligning with observed outcomes in these regions. However, mild deviations from perfect calibration are observed in the mid-probability range, indicating that predictions around 0.4–0.6 should be interpreted with caution in clinical decision-support settings.

4.5 Confusion Matrix and Error Distribution

A confusion matrix was created to see how the classification errors were distributed for a representative patient-level split. Fig. 2 shows the confusion matrix resulting from the patient-level evaluation with strict application of the rules. A comparison of the number of false positives and false negatives can be made in this matrix. This indicates that the baseline model does not have a large imbalance in favor of one of the two classes when using the configuration used for the balanced sampling.

The error distribution also indicates that both benign and malignant patches presented equivalent classification challenges. Thus, the balance in the distribution of errors highlights the need to analyze the modes of failures in addition to the overall accuracy, because a single performance metric cannot fully describe all types of errors that may be committed by a classification model.

The confusion matrix further supports the motivation for subsequent qualitative error analysis, which examines representative false positive and false negative patches to identify underlying morphological patterns associated with incorrect predictions.

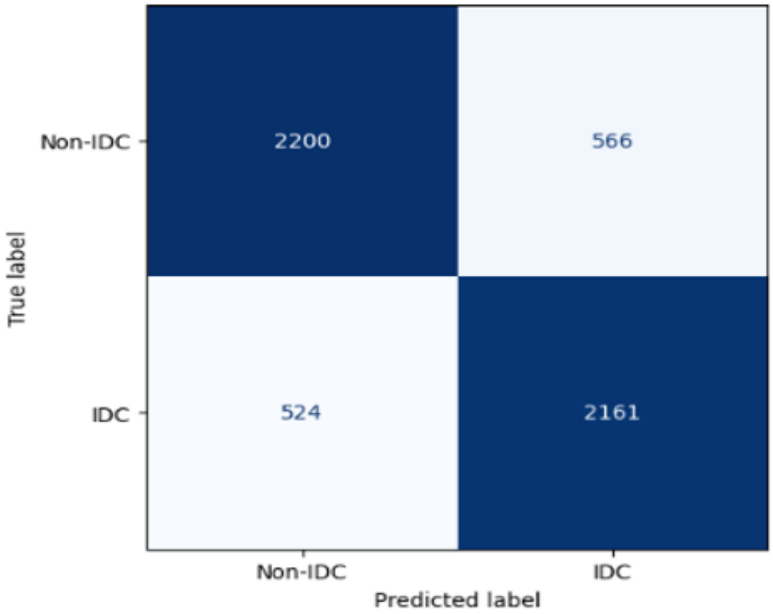

Fig. 2 Confusion matrix (Representative Patient-level Fold)

Fig. 2 shows that both false positives and false negatives contribute substantially to overall error, motivating deeper qualitative analysis of failure patterns rather than assuming errors are dominated by one class.

4.6 Qualitative Error Analysis (False Positives and False Negatives)

To characterize failure modes in a clinically interpretable manner, representative false positive and false negative patches were visualized. Fig 3 presents a montage of false positives and false negatives sampled from the patient-level evaluation.

False positives are commonly associated with benign tissues in areas of dense nuclei, intense staining, or highly complex stroma. At a patch level, such characteristics can mimic those seen in cancerous morphology, especially if there is some degree of disruption to normal tissue architecture. False negatives typically arise from cancers having low cellularity, minimal invasion, or cases in which the tumor only covers a limited area within a patch. The observations above suggest that the baseline CNN uses predominantly textural and density features and therefore does not capture the subtle structural malignancies present in the images.

The results of this error analysis provide evidence that misclassifications occur based on a non-random pattern, in the form of morphological characteristics. These patterns also represent areas where new approaches to clinical AI models will need to be developed to address current limitations, such as multi-scale contextualization and stain normalization.

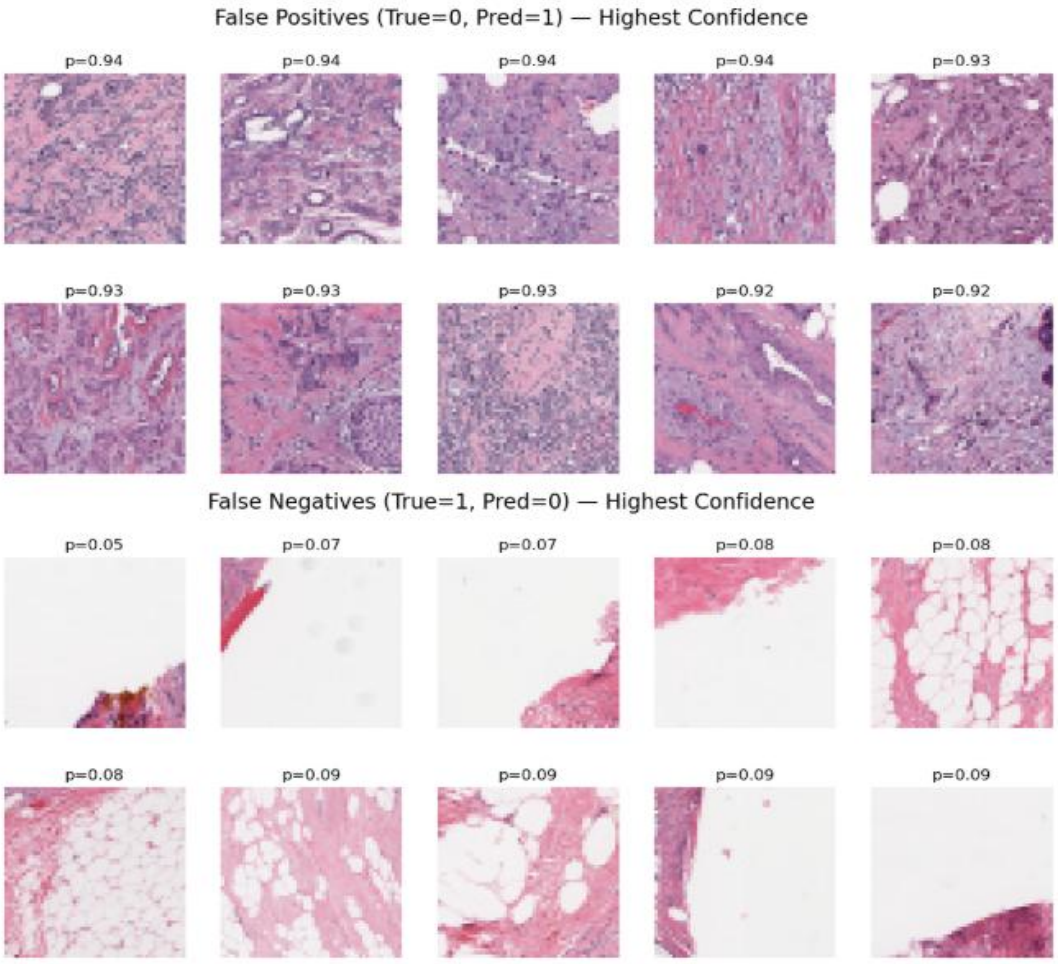

Fig. 3: Representative False Positives and False Negatives (Patient-Level Evaluation)

False positives commonly occur in benign regions with dense nuclei and complex textures, whereas false negatives tend to appear in low-cellularity malignant regions or patches where tumor occupies only a small fraction of the tissue.

4.7 Fold-Wise Stability and Variance Visualization

While Tables 2 and 3 report mean performance values, fold-wise variability is critical for understanding model robustness. To visualize this, Fig 4 compares the distribution of fold-wise accuracy under patient-level and random image-level evaluation protocols.

Fig 4 illustrates the substantial increase in variance at the patient-level when evaluating the model rather than a random split. Random split results in accuracy values distributed over a narrower area than patient-level evaluation. This narrowness is an indicator of the stability of the model's accuracy; this is especially relevant when using datasets related to pathology, where each patient has unique staining characteristics, and each dataset may be acquired under different conditions.

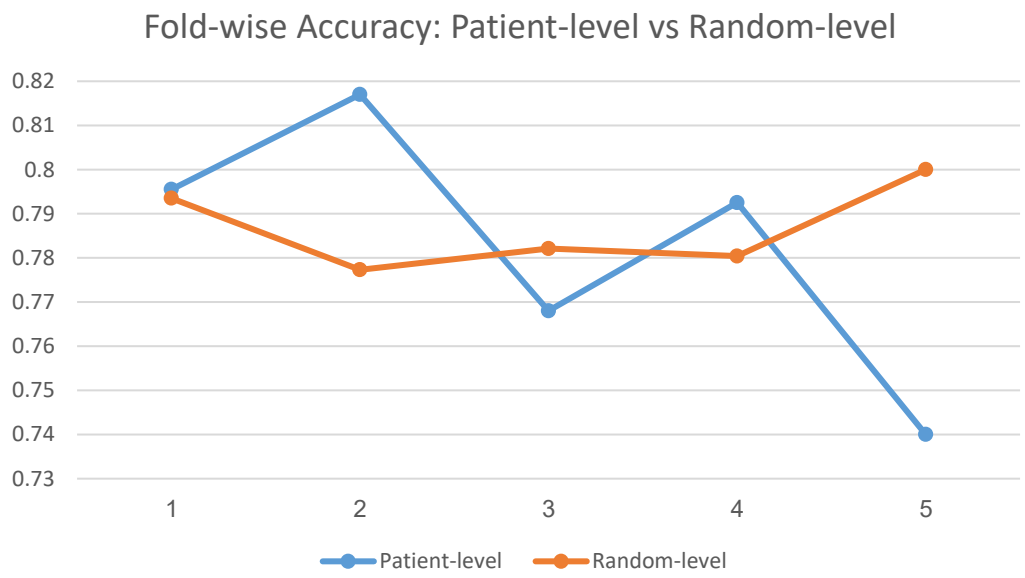

Fig. 4: Fold-wise accuracy comparison between patient-level and random-level cross-validation protocols.

Fig 4 highlights substantially higher variability under the patient-level protocol compared to the random image-level protocol. This dispersion suggests sensitivity to patient heterogeneity, whereas the random split yields more stable estimates that may underestimate real-world uncertainty.

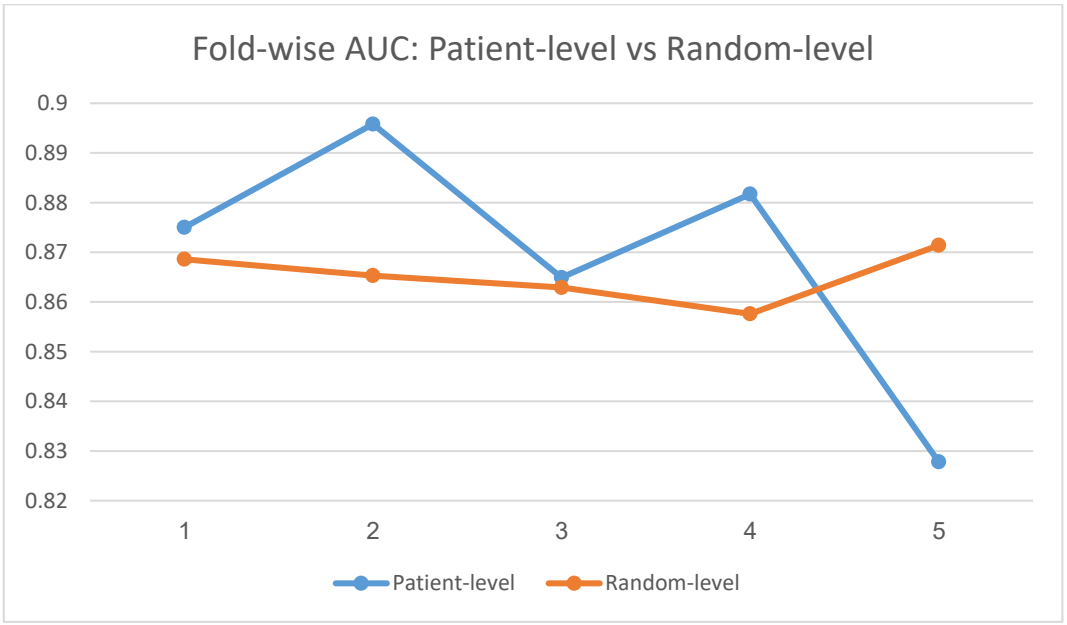

Fig. 5: Fold-wise AUC comparison between patient-level and random-level cross-validation protocols.

A similar pattern is observed for AUC, where patient-level evaluation exhibits wider fold-to-fold fluctuations. This reinforces the importance of reporting dispersion measures and fold-level performance rather than relying on a single point estimate.

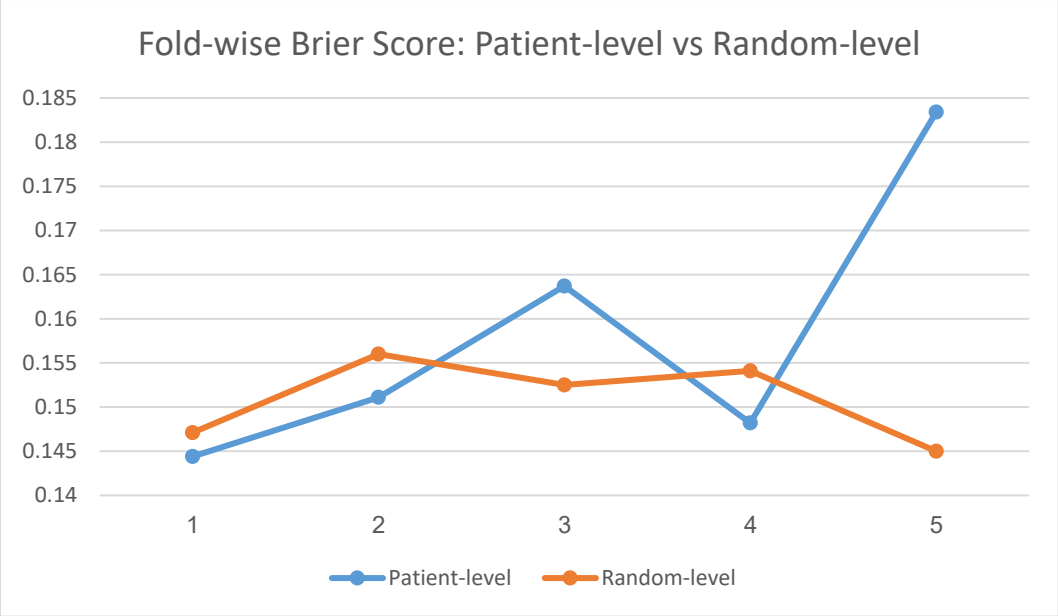

Fig. 6: Fold-wise Brier score comparison between patient-level and random-level cross-validation protocols.

Image-level evaluation of random images results in a greater stability in the fold-wise comparison of the Brier score as opposed to the patient-level evaluation, where there is a higher degree of variability. This finding indicates that the choice of protocol affects not only the metric of discrimination but also the probabilistic reliability of the system, which is important for applications of medical decision support.

5 Discussion

This research explored how an evaluation methodology would influence a patch-based breast cancer histopathology classification system using the IDC dataset. The results demonstrate that the baseline CNN produced excellent discrimination performance for both patient-level and random image-level cross-validation, but provide important methodological observations extending beyond the average accuracy and AUC values.

One of the main findings was that mean performance measures cannot accurately indicate model robustness. Mean accuracy and Brier scores for patient-level and image-level splits, with a slight increase and slight decrease, respectively, were not significantly different among protocols. The results suggest that with a controlled and balanced method of random sampling, there will be minimal inflation of average performance in this study design. However, the major difference between these methods was fold-to-fold variation. The use of patient-level evaluations showed much larger variability from one-fold to another due to patient heterogeneity. Random image splitting provided more consistent and tighter cluster results.

Although the method shows stable results with random splitting, such stability may overestimate overall generalizability. In fact, because patches of the same patient are included in both the train and test partitions, there will inevitably be some degree of patient-specific morphology that will be present in the data splits for training and testing. Although sharing patient-specific features will likely have only a marginal effect on average classification accuracy, it could effectively decrease variance and uncertainty around the accuracy estimates, thus obscuring the real challenge of generalizing to new patients. Clinical relevance lies in the ability of a model to perform robustly in diverse patient populations; therefore, the potential marginal gain in average classification accuracy is less relevant than the ability to achieve this result in multiple different patient populations.

Another major contribution of this research was the incorporation of reliability analysis in an explicit manner. Many studies reporting on discrimination metrics for histopathology have done so without incorporating calibration metrics or Brier scores. Calibration curves indicated moderate quality, with some evidence of mild overconfidence at intermediate probabilities. Perhaps more importantly, calibration variability exhibited a pattern that was very similar to that seen with discrimination variability: as with discrimination variability, patient-level evaluation exhibited larger variance in Brier scores than did slide-level evaluation. Therefore, it appears that the evaluation methodology employed to assess classifier performance does not merely influence the accuracy of classifier-based classifications but also impacts the reliability of predicted probabilities, which are relevant when classifier-based predictions will be used to guide clinical action.

The qualitative error analysis also provides additional evidence to support the interpretation of how the models behave. Typically, the false positives that occurred as a result of the CNN misclassifying malignant tissues as benign resulted from having tissue with either high nuclear density or with complex stromal texture characteristics, both of which can be visually similar to malignant morphology at the patch level. Conversely, false negative results typically occurred in the malignant regions with low levels of cellularity or mild infiltration into normal tissue, where there are limited discriminative characteristics. These results suggest that the baseline CNN is utilizing predominantly local textural density characteristics to classify tissue, which can be inadequate to provide sufficient detail of either subtle or context-dependent tumor architecture. This type of finding corresponds with the larger body of research indicating that multi-scale models and integrating information over space and time are both critical to providing improved accuracy in classifying histopathologic images.

Although these studies have provided new understanding of the field, there are many limitations that must be recognized. First, the model's architecture was developed as a lightweight version in order to allow for processing on a CPU-based machine. This design provides better reproducibility; however, better model designs and longer training times can provide higher performance. Second, the study focused on patch-level classifications and did not address whole-slide aggregations. Therefore, the results reported in this study cannot be directly correlated with slide-level diagnostic accuracy. Third, balanced sampling has been applied across the data sets in each fold to make direct comparisons possible; however, balanced sampling does not reflect the typical unbalanced nature of clinical datasets.

Future work should focus on developing multiscale architectural designs that will include a larger amount of spatial information, in addition to utilizing stain normalization and domain adaptation methodologies to reduce variability among patients. In addition, focused on the development of uncertainty-based methods and explicit calibration techniques, such as temperature scaling, to increase the reliability of the probabilistic nature of the model. Furthermore, future work should expand the current evaluation to assess the model at the level of individual slides rather than the whole glass-slide image to further add to the clinical relevance of the model.

Overall, this study demonstrates that evaluation protocol design significantly influences perceived model robustness in patch-based histopathology classification. While mean performance differences between patient-level and random-level splitting may be small under controlled sampling, variability analysis reveals substantial protocol-dependent effects. These findings emphasize the necessity of reporting fold-wise dispersion, calibration metrics, and qualitative error analysis in addition to traditional discrimination scores when developing AI systems for medical imaging.

6 Conclusion

This study presented a reproducible baseline evaluation of patch-based invasive ductal carcinoma (IDC) detection using the Breast Histopathology Images dataset, with a specific focus on how evaluation protocol design influences reported performance and robustness. Using a lightweight CNN trained under controlled balanced sampling, we compared strict patient-level 5-fold cross-validation against stratified random image-level splitting while maintaining identical model architecture and training configuration.

Across both protocols, the model achieved competitive discrimination performance, with mean AUC values around 0.86–0.87 and mean accuracy close to 0.78–0.79. Although average performance differences between protocols were small and not statistically significant, the analysis revealed an important methodological distinction: patient-level evaluation produced substantially higher fold-wise variability, reflecting clinically realistic heterogeneity across individuals. In contrast, random patch-level splitting yielded more stable performance estimates that may underestimate real-world uncertainty and generalization difficulty.

Beyond discrimination metrics, this work incorporated calibration assessment through Brier score and calibration curves, demonstrating moderate reliability and highlighting that protocol design also affects probabilistic stability. Furthermore, qualitative inspection of false positives and false negatives identified consistent morphological error patterns, suggesting that the baseline model relies heavily on local density and texture cues and may fail in subtle infiltration or low-cellularity malignant regions.

Overall, the results emphasize that reliable evaluation in computational pathology requires more than reporting mean accuracy and AUC. Patient-level separation, fold-wise dispersion reporting, calibration analysis, and interpretable error inspection should be considered essential components of robust benchmarking. Future work should extend this framework to multi-scale architectures, slide-level aggregation, and uncertainty-aware modeling to further improve clinical relevance and reliability.

References

[1] Hamadneh, B. Abu-Irmaileh, M. Al-Majawleh, Y. Bustanji, Y. Jarrar, and T. Al-Qirim, “Doxorubicin–paclitaxel sequential treatment: insights of DNA methylation and gene expression changes of luminal A and triple negative breast cancer cell lines,” Molecular and Cellular Biochemistry, vol. 476, no. 10, pp. 3647–3654, May 2021, doi: https://doi.org/10.1007/s11010-021-04191-5

[2] K. F. Tehrani, J. Park, E. J. Chaney, H. Tu, and S. A. Boppart, “Nonlinear Imaging Histopathology: A Pipeline to Correlate Gold-Standard Hematoxylin and Eosin Staining With Modern Nonlinear Microscopy,” IEEE journal of selected topics in quantum electronics, vol. 29, no. 4: Biophotonics, pp. 1–8, Jul. 2023, doi: https://doi.org/10.1109/jstqe.2022.3233523

[3] D. Humaran et al., “Towards a New Standard: Prospective Validation of Ex Vivo Fusion Confocal Microscopy for Intraoperative Margin Assessment in Breast-Conserving Cancer Surgery,” Cancers, vol. 17, no. 23, p. 3848, Nov. 2025, doi: https://doi.org/10.3390/cancers17233848

[4] I. Gupta et al., “A deep learning based approach to detect IDC in histopathology images,” Multimedia Tools and Applications, vol. 81, Apr. 2022, doi: https://doi.org/10.1007/s11042-021-11853-5

[5] I. Ellis et al., “Dataset for reporting of the invasive carcinoma of the breast: recommendations from the International Collaboration on Cancer Reporting (ICCR),” Histopathology, vol. 85, no. 3, pp. 418–436, May 2024, doi: https://doi.org/10.1111/his.15191

[6] N. White, R. Parsons, G. S. Collins, and A. Barnett, "Evidence of questionable research practices in clinical prediction models," BMC Medicine, vol. 21, no. 1, Sep. 2023, doi: https://doi.org/10.1186/s12916-023-03048-6
[7] F. Abdullakutty, Y. Akbari, S. Al-Maadeed, A. Bouridane, I. M. Talaat, and R. Hamoudi, "Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis," Frontiers in Medicine, vol. 11, Sep. 2024, doi: https://doi.org/10.3389/fmed.2024.1450103
[8] D. J. Rumala, "How You Split Matters: Data Leakage and Subject Characteristics Studies in Longitudinal Brain MRI Analysis," Lecture notes in computer science, vol. 14242, pp. 235–245, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-45249-9_23
[9] S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," PLOS ONE, vol. 17, no. 1, p. e0262838, Jan. 2022, doi: https://doi.org/10.1371/journal.pone.0262838
[10] A. S. Sambyal, U. Niyaz, N. C. Krishnan, and D. R. Bathula, "Understanding calibration of deep neural networks for medical image classification," Computer Methods and Programs in Biomedicine, vol. 242, p. 107816, Dec. 2023, doi: https://doi.org/10.1016/j.cmpb.2023.107816
[11] V. Ivanov et al., "Use of AI Histopathology in Breast Cancer Diagnosis," Medicina, vol. 61, no. 10, pp. 1878–1878, Oct. 2025, doi: https://doi.org/10.3390/medicina61101878
[12] Y. Wu et al., "Recent Advances of Deep Learning for Computational Histopathology: Principles and Applications," Cancers, vol. 14, no. 5, p. 1199, Feb. 2022, doi: https://doi.org/10.3390/cancers14051199
[13] F.-Z. Nakach, A. Idri, and E. Goceri, "A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification," Artificial Intelligence Review, vol. 57, no. 12, Oct. 2024, doi: https://doi.org/10.1007/s10462-024-10984-z
[14] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: a comprehensive review," Multimedia Tools and Applications, vol. 84, Aug. 2024, doi: https://doi.org/10.1007/s11042-024-20011-6
[15] P. Singh, R. Kumar, M. Gupta and A. J. Obaid, "A Critical Analysis and Classification of Breast Cancer Using Histopathology Images," 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chikkaballapur, India, 2024, pp. 1-7, doi: https://doi.org/10.1109/ICKECS61492.2024.10617282
[16] A. W. Salehi et al., "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," Sustainability, vol. 15, no. 7, p. 5930, Jan. 2023, doi: https://doi.org/10.3390/su15075930
[17] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," BMC Medical Imaging, vol. 22, no. 1, Apr. 2022, doi: https://doi.org/10.1186/s12880-022-00793-7
[18] R. E. Shawi, K. Kilanava, and S. Sakr, "An interpretable semi-supervised framework for patch-based classification of breast cancer," Scientific Reports, vol. 12, no. 1, p. 16734, Oct. 2022, doi: https://doi.org/10.1038/s41598-022-20268-7
[19] A. E.-S. Mansour and S. Martinez, "Automated Pulmonary Nodule Detection In LDCT Using 3D ResNet And Adaptive Patch Strategy," Frontiers in Medical and Clinical Sciences, vol. 2, no. 1, pp. 1–6, Jan. 2025, doi: https://doi.org/10.64917/fmcs-001
[20] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, pp. 100804–100804, Aug. 2023, doi: https://doi.org/10.1016/j.patter.2023.100804.
[21] G. Irmak and A. Saygılı, "Deep learning-based histopathological classification of breast tumors: a multi-magnification approach with state-of-the-art models," Signal, Image and Video Processing, vol. 19, no. 7, May 2025, doi: https://doi.org/10.1007/s11760-025-04122-7
[22] F. Hussein et al., "Hybrid CLAHE-CNN Deep Neural Networks for Classifying Lung Diseases from X-ray Acquisitions," Electronics, vol. 11, no. 19, p. 3075, Sep. 2022, doi: https://doi.org/10.3390/electronics11193075
[23] N. Sourlos et al., "Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology," Insights into Imaging, vol. 15, no. 1, Oct. 2024, doi: https://doi.org/10.1186/s13244-024-01833-2
[24] A. Bissoto, T.-D. Hoang, T. Flühmann, S. Sun, C. F. Baumgartner, and L. M. Koch, "Subgroup Performance Analysis in Hidden Stratifications," Lecture Notes in Computer Science, vol. 15973, pp. 594–603, Sep. 2025, doi: https://doi.org/10.1007/978-3-032-05185-1_57
[25] S. Ao, S. Rueger, and A. Siddharthan, "Two Sides of Miscalibration: Identifying Over and Under-Confidence Prediction for Network Calibration," in Proc. 39th Conf. Uncertainty in Artificial Intelligence (UAI), PMLR, vol. 216, pp. 77–87, 2023.

[26] R. Blythe et al., “Clinician perspectives and recommendations regarding design of clinical prediction models for deteriorating patients in acute care,” BMC Medical Informatics and Decision Making, vol. 24, no. 1, Sep. 2024, doi: https://doi.org/10.1186/s12911-024-02647-4.
[27] M. Salvi et al., “Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare,” International Journal of Medical Informatics, vol. 197, p. 105846, Feb. 2025, doi: https://doi.org/10.1016/j.ijmedinf.2025.105846.
[28] L. Famiglini, A. Campagner, and F. Cabitza, “Towards a Rigorous Calibration Assessment Framework: Advancements in Metrics, Methods, and Use,” Frontiers in artificial intelligence and applications, vol. 2, Sep. 2023, doi: https://doi.org/10.3233/faia230327.
[29] M. Unger and J. N. Kather, “A systematic analysis of deep learning in genomics and histopathology for precision oncology,” BMC Medical Genomics, vol. 17, no. 1, Feb. 2024, doi: https://doi.org/10.1186/s12920-024-01796-9.
[30] H. Evans and D. Snead, “Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact,” npj Digital Medicine, vol. 7, no. 1, pp. 1–6, Apr. 2024, doi: https://doi.org/10.1038/s41746-024-01093-w
[31] M. Ritter et al., “Spatially resolved transcriptomics and graph-based deep learning improve accuracy of routine CNS tumor diagnostics,” Nature Cancer, vol. 6, no. 2, pp. 292–306, Jan. 2025, doi: https://doi.org/10.1038/s43018-024-00904-z.
[32] P. T. Mooney, “Breast Histopathology Images,” Kaggle Dataset, 2018.
[33] A. Paproki, O. Salvado, and C. Fookes, “Synthetic Data for Deep Learning in Computer Vision & Medical Imaging: A Means to Reduce Data Bias,” ACM Computing Surveys, vol. 56, no. 11, pp. 1–37, May 2024, doi: https://doi.org/10.1145/3663759.
[34] M. A. Shahbazi, A. Baheri, and N. Azadeh-Fard, “A hierarchical conformal framework for uncertainty-aware length of stay prediction in multi-hospital settings,” Scientific Reports, vol. 16, no. 1, Jan. 2026, doi: https://doi.org/10.1038/s41598-026-37450-w.
[35] K. M. Alzhrani, “From Sigmoid to SoftProb: A novel output activation function for multi-label learning,” Alexandria Engineering Journal, vol. 129, pp. 472–482, Jun. 2025, doi: https://doi.org/10.1016/j.aej.2025.06.013.
[36] J. Ghosh and S. Gupta, "ADAM Optimizer and CATEGORICAL CROSSENTROPY Loss Function-Based CNN Method for Diagnosing Colorectal Cancer," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 470-474, doi: https://doi.org/10.1109/CISES58720.2023.10183491
[37] W. Yang, J. Jiang, E. M. Schnellinger, S. E. Kimmel, and W. Guo, “Modified Brier score for evaluating prediction accuracy for binary outcomes,” Statistical Methods in Medical Research, vol. 31, no. 12, pp. 2287–2296, Aug. 2022, doi: https://doi.org/10.1177/09622802221122391.
[38] B. Langenberg, M. Janczyk, V. Koob, R. Kliegl, and A. Mayer, “A tutorial on using the paired t test for power calculations in repeated measures ANOVA with interactions,” Behavior Research Methods, vol. 55, no. 5, pp. 2467–2484, Aug. 2022, doi: https://doi.org/10.3758/s13428-022-01902-8.