International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 15, No. 4, December 2026, pp. 1881~1892
ISSN: 2252-8776, DOI: 10.11591/ijict.v15i4.pp1881-1892

CPU-optimized regularized convolutional neural network architecture for multimedia-based breast cancer metastasis detection

Guidoum Amina[1], Bouhenni Abdelhake[2]
[1]Department of Computer Science, Higher Normal School of Kouba, Kouba, Algeria
[2]Kuwait Oil Company (KOC), Al Ahmadi, Kuwait

Article Info

Article history:

Received Dec 13, 2025
Revised Jun 12, 2026
Accepted Jul 5, 2026

Keywords:

Classification of histopathological images
CNN architecture
Computational optimization
CPU optimization
Regularization

ABSTRACT

The histopathological diagnosis of breast cancer metastases, a subjective and lengthy process, benefits from advances in artificial intelligence for its automation. This study proposes an automatic classification method based on a convolutional neural network (CNN) architecture with progressive regularization (Batch normalization and adaptive dropout), trained on the PatchCamelyon (PCam) dataset. Our approach integrates a data augmentation strategy specific to medical images and advanced optimization for CPU environments, allowing us to balance performance, robustness, and computational accessibility. The architecture comprises only 94,000 parameters, 223 times fewer than leading models. Multi-criteria evaluation shows competitive results: area under the curve (AUC) of 92.92%, F1-score of 82.39%, accuracy of 88.74%, and specificity of 90.61%. A bootstrap robustness analysis provides 95% confidence intervals (AUC: [89.56%-92.91%]), and an ablation study quantifies the reduction in overfitting through progressive dropout (33.6%). This research validates that a well-regulated and CPU-optimized CNN architecture can achieve clinically relevant performance while remaining accessible on standard environments, thus opening up concrete prospects for the development of practical AI tools in digital pathology.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Guidoum Amina
Department of Computer Science, Higher Normal School of Kouba
Algiers, Algeria
Email: amina.guidoum@g.ens-kouba.dz

1. INTRODUCTION

Breast cancer represents a major public health challenge, with approximately 2.3 million new cases diagnosed annually worldwide. Histopathological diagnosis, based on the microscopic examination of tissue sections, is the gold standard for detecting lymph node metastases [1]. However, this process is subjective, time-consuming, and subject to significant inter-observer variability. The advent of digital pathology and deep learning (DL) techniques offers revolutionary prospects for automating and improving diagnostic accuracy [2].

The PatchCamelyon (PCam) dataset, comprising 327,680 colored patches (96×96-pixels) from histological sections of lymph nodes, represents a standardized benchmark for the development of binary classification algorithms (healthy vs. metastatic tissue) [1].

The field of histological image analysis using DL has seen spectacular advances in recent years. The seminal work [1] demonstrated for the first time that DL algorithms could outperform pathologists in

detecting breast cancer metastases. More recently, attention has focused on improving computational efficiency and generalization. Subsequently [3], proposed EfficientNetV2, through neural architecture research that takes into account training and progressive learning with adaptive regularization, achieving a top-1 accuracy of 87.3% on ImageNet while training 5 to 11 times faster than vision transformers (ViT). In parallel [4], proposed ConvNeXt, a modernized pure ConvNet architecture with transformer design principles, which achieved 87.8% accuracy on ImageNet top-1 and outperformed swin transformers on COCO detection and ADE20K segmentation benchmarks.

More recently, ViTs [5] and the swin transformer [6] have challenged the dominance of convolutional neural network (CNNs) by applying self-attention mechanisms to image patches, achieving top performance on many medical imaging benchmarks. However, these architectures typically require millions of parameters and substantial GPU resources.

Among recent DL-based approaches for breast cancer diagnosis, Mobark [7] proposed CoroNet, a CNN model using the Xception architecture pre-trained on ImageNet. Trained on the CBIS-DDSM mammogram dataset, CoroNet achieves an accuracy of 94.92% in four-class classification (Benign vs. malignant masses, benign vs. malignant calcifications) and 86.67% in binary classification (Calcifications vs. masses). The architecture comprises 71 layers for a total of 33.97 million parameters; this model remains highly parameterized and requires significant computing resources, which limits its deployment in clinical environments with limited hardware.

The study [8] concretely illustrates one of the main challenges for the clinical deployment of AI in digital pathology: the poor generalization of DL models to new contexts. The authors evaluated a model, initially trained on the multicenter CAMELYON dataset for detecting breast cancer metastases in sentinel lymph nodes. Its baseline performance on the CAMELYON16 test set was excellent, with a FROC score of 0.838. This study highlights that robustness and adaptability are as crucial as absolute benchmark performance. It validates the need to develop more robust learning methods (advanced transfer learning, targeted data augmentation, self-supervised learning) and pipelines that incorporate a facilitated continuous adaptation phase for sustainable clinical deployment.

Faced with these challenges, innovative hybrid architectures are emerging to improve both the accuracy and robustness of models. For example [9], proposed a CNN-transformer hybrid model specifically designed to leverage temporal information by simultaneously integrating previous and current mammographic images of the same patient. Their architecture uses a modified ResNet50 for local feature extraction, a transformer encoder with a multi-head attention mechanism and positional coding to capture long-term dependencies and subtle changes between examinations, and a per-channel attention module to amplify discriminating features. Evaluated on a dedicated dataset, this model achieves an area under the curve (AUC) of 92.58%, thus demonstrating the crucial contribution of temporal modeling and hybrid architectures to mimic the sequential reasoning of radiologists and improve early detection.

Catalano *et al*. [10] proposed a method for detecting and localizing breast cancer from hematoxylin and eosin (HE) stained histological images using several CNN architectures. They compared a custom CNN (Standard_CNN), EfficientNet, ResNet50, VGG-16, VGG-19, and MobileNet, as well as a new model called MR_Net. The models were trained on the BRACS dataset, which contains 4539 regions of interest (ROIs) annotated into three classes (benign, atypical, malignant). After preprocessing and data augmentation (horizontal flip, brightness, zoom), the final dataset comprised 5628 images. The best accuracies obtained were 73% for MobileNet and VGG-19, 69% for MR_Net, and 67% for standard_CNN. Fine-tuning MobileNet, VGG-16, and VGG-19 did not improve performance (VGG-19's accuracy decreased from 73% to 71%). However, metrics such as AUC or F1-score are not reported, and the number of model parameters is not specified, which limits quantitative comparison in terms of computational efficiency.

Faced with the domain shift induced by scanner variability, self-supervised learning approaches aimed at learning inherently robust representations have emerged. While very large-scale foundation models (e.g., UNI, Virchow) achieve peak performance but at a prohibitive computational cost, recent work explores lightweight architectures optimized for generalization. For example, HistoLite [11] offers a lightweight self-supervised framework based on a dual stream of autoencoders and a contrast objective, coupled with adaptive augmentation (Adaptive HistoRotate). Specifically evaluated on a scanner bias (Aperio vs. Sakura), HistoLite (7M parameters) demonstrates very low representation shift and the best inter-scanner generalization among the compared models, despite slightly lower absolute accuracy than the larger models. This work illustrates the trade-off between accuracy, robustness, and efficiency, and validates the relevance of specialized lightweight architectures for reliable deployment in resource-constrained clinical settings.

Finally, the systematic review [12] provides a comprehensive analysis of the contribution of DL to the analysis of histopathological images of breast cancer. By synthesizing 199 references, including 182 studies published between 2014 and 2025, the authors demonstrate how DL architectures, particularly CNNs [13], generative adversarial networks (GANs), and ViTs, have significantly improved the accuracy of

classification, segmentation, and tumor detection tasks. The article also highlights the crucial role of transfer learning and explainable AI (XAI) techniques in addressing data gaps and improving clinical interpretability. However, the review identifies persistent challenges to clinical adoption, including variability in staining protocols, a lack of large, diverse, and annotated datasets, and issues of robustness, equity, and regulation. In conclusion, the authors propose as future perspectives the development of more generalizable models, multimodal data integration, and the prioritization of transparent tools for responsible integration into diagnostic workflows.

It is important to note that transformer models first achieved success in natural language processing. The bidirectional encoder representations from transformers (BERT) model [14] has achieved new cutting-edge results on eleven natural language processing tasks, including question answering and linguistic inference. BERT is pre-trained on BooksCorpus (800 million words) and Wikipedia (2.5 billion words). Its architecture is a multi-layered bidirectional transformer encoder. The BERTBASE version comprises 12 layers, a hidden size of 768, 12 attention heads, and 110 million parameters; the BERTLARGE version comprises 24 layers, a hidden size of 1024, 16 attention heads, and 340 million parameters. BERTBASE was trained on 4 Cloud TPUs, and BERTLARGE on 16 Cloud TPUs. BERT is designed for text token sequences using WordPiece embeddings, not for images. In our context of classifying histopathological patches with a limited dataset and a CPU-bound computational target, CNNs remain more suitable. CNNs benefit from inductive biases (translational equivariance, local connectivity) that naturally conform to the structure of tissue textures. Their parametric lightness allows for CPU-bound training, and regularization techniques such as batch normalization and progressive dropout effectively control overfitting. Thus, our choice of a well-regulated CNN architecture is justified by resource constraints and the nature of the task.

Despite exceptional performance reported with complex architectures and large datasets, few studies have focused on optimization for computationally limited environments, which are representative of many clinical and research settings. This study specifically addresses this challenge by proposing a comprehensive methodology that balances discriminant performance, robustness, and computational accessibility. Our contributions are: (i) a lightweight CNN architecture (94K parameters) with progressive regularization suitable for histological images; (ii) a CPU-optimized training pipeline that completes in approximately 2.4 hours on standard hardware ; and (iii) a multi-criteria evaluation demonstrating that a well-regulated CNN can achieve clinically relevant performance (AUC 92.92%, accuracy 88.74%) with 223 times fewer parameters than leading-edge approaches(EfficientNetV2-S, ConvNeXt), making AI tools more accessible for digital pathology in resource-constrained environments. Our contribution does not lie in a radical architectural innovation, but in the demonstration that a well-regulated and lightweight convolutional network can achieve clinically relevant performance with very limited computing resources.

2. METHOD

2.1. Dataset and preprocessing

The PCam dataset contains 327,680 patches of histological images (96×96 pixels) annotated binary (0: healthy tissue, 1: metastatic tissue). For this study, we strategically selected 4,096 samples from the training set (262,144 available) and 1,024 from the validation set (32,768 available), representing a compromise between representativeness and computational constraints. This selection maintains the original 50/50 distribution between the classes. A sample size of 4,096 provides a margin of error of ±1.5% at the 95% confidence level for a binary classification. This value is obtained using the standard formula for the margin of error of a proportion.

$$Margin = 1.96 \times \sqrt{\frac{p(1-p)}{n}} \; avec \; p = 0.5, n = 4096$$

The margin calculation gives 1.5% Thus, the chosen size represents a reasonable compromise between statistical representativeness and computational constraints. The images were normalized by dividing the pixel values by 255 to obtain a range [0,1].

2.2. Model architecture with advanced regulation

Our CNN architecture, illustrated in Figure 1, is specifically designed for 96×96-pixel histological images, balancing discriminatory power with overfitting prevention. It comprises three sequential convolutional blocks with progressively increasing filter counts (32, 64, 128), each followed by batch normalization and neuronal dropout. The dropout rate gradually increases (0.2, 0.3, 0.4) in deeper layers where overfitting is more likely. Global mean pooling replaces traditional flattening, drastically reducing parameters while preserving spatial information. This hierarchical design enables robust extraction of features ranging from low-level textures to complex patterns (tissue structures).

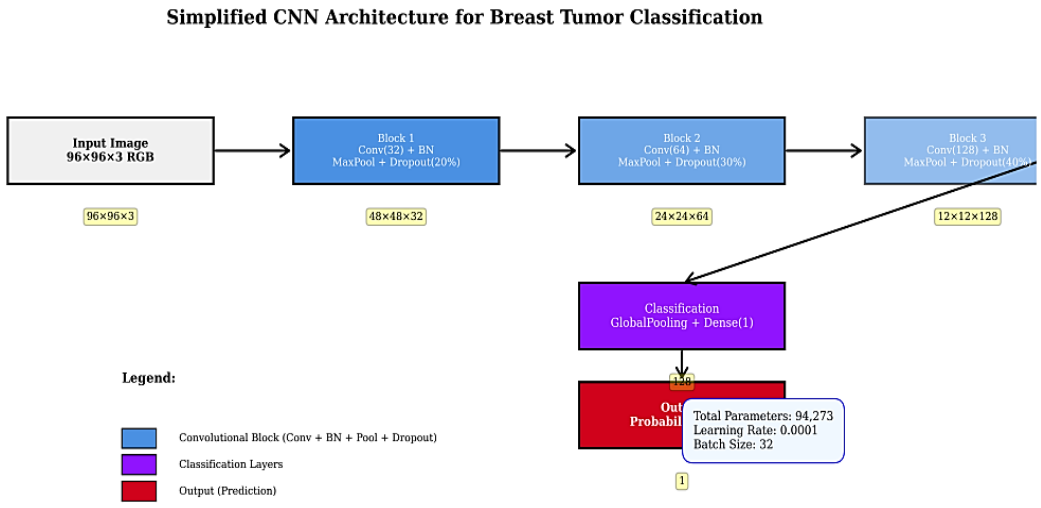


Figure 1. Simplified CNN architecture for breast tumor classification

2.3. Data augmentation

To compensate for the limited size of the training set, we implemented an augmentation strategy specific to histological images,

– Original preservation: preserving the source image.
– Horizontal flip: a transformation that preserves the diagnostic characteristics of the tissues.
– 90° rotation: adapting to the variable orientation of histological sections.

This strategy triples the size of the training set (4,096 to 12,288 samples) while preserving the biological significance of the images.

2.4. Training configuration

The training was specifically optimized for CPU environments to ensure maximum accessibility for researchers and clinicians with limited computing resources. The Adam optimizer [15] was selected with a learning rate of 0.0001, enabling fine and stable convergence. The loss function used is binary cross-entropy, appropriate for binary classification. The batch size was set at 32, providing an optimal balance between efficient memory usage and training stability. The maximum number of epochs was limited to 100, with the implementation of early stopping that monitors the validation AUC [16] with a patience of 15 epochs and restores the best weights, thus preventing overfitting. The metrics evaluated include accuracy [17], AUC, precision, recall [18], and a custom F1-score [17]. The advanced configuration of TensorFlow was exploited with 16 threads for intra-operation parallelism, 8 threads for inter-operation parallelism, flexible operation placement (soft device placement) enabled, and just-in-time (JIT) compilation to accelerate the execution of computation graphs.

2.5. Evaluation and visualization

To provide a comprehensive and reproducible analysis of the model's performance, we designed and implemented a complete evaluation pipeline that automatically generates a series of visualizations and scientific reports. This pipeline produces receiver operating characteristic (ROC) curves [19] to assess the model's discriminatory ability across all decision thresholds. Confusion matrices are generated to provide a detailed and quantitative analysis of the error distribution (true/false positives/negatives). Accuracy-recall curves, particularly relevant for potentially unbalanced datasets such as those used in medical imaging, are also calculated. Convergence is monitored by generating learning curves that illustrate the evolution of key metrics across the training and validation datasets. For qualitative interpretation, visualizations of example predictions (Correct and incorrect) along with their confidence scores are produced.

3. RESULTS AND DISCUSSION

3.1. Overall performance

Training converged after 42 epochs with early termination triggered by stabilization of the validation AUC. Final performance across the validation set (1,024 samples) is summarized in Table 1.

Table 1. Detailed model performance

Metric value interpretation
AUC 0.9292 excellent discrimination
Accuracy 0.8874 high true positive rate (TPR)
Recall 0.7689 moderate sensitivity
F1-score 0.8239 satisfactory balance

3.2. Analysis of learning curves

Figure 2 shows the distribution of confidence scores: X-axis: confidence score (continuous\value normalized between 0.0 and 1.0) and Y-axis: frequency (number of samples, unit: count). This histogram shows the bimodal distribution of confidence scores predicted by the model for the validation set. Two distinct distributions clearly emerge: one centered around 0.0-0.2 (representing healthy tissue samples) and the other centered around 0.8–1.0 (representing tumor tissue samples). The clear separation between these two modes, with a pronounced dip around the score 0.5, indicates that the model produces predictions with high confidence. The small proportion of samples in the intermediate region (0.3-0.7) suggests that the model experiences little uncertainty in the majority of cases. The distribution of healthy tissues shows a slightly extended tail toward higher scores, corresponding to false positives identified in the confounding matrix. The marked bimodality confirms the model's proper calibration and its ability to assign probability scores that accurately reflect the certainty of the predictions. This characteristic is essential for clinical applications where different decision thresholds can be applied depending on the desired trade-off between sensitivity and specificity.

Figure 3 shows the evolution of the metrics during training. X-axis: epoch (unit: number of complete iterations on the training set) and Y-axis: AUC (normalized value between 0.5 and 1.0). This figure illustrates the evolution of the model's discriminatory performance during training over 42 epochs. The blue curve (Train AUC) represents the AUC calculated on the training set (12,288 samples after augmentation), while the orange curve (Validation AUC) shows the AUC on the validation set (1,024 samples). The training AUC increases steadily from 82.33% to 95.82%, demonstrating the model's ability to progressively learn discriminatory features. The validation AUC, after a rapid initial increase, stabilizes around 92.92% after epoch 27. The moderate but consistent difference between the two curves (approximately 0.029 on average) indicates effective overfitting control thanks to the regularization techniques employed (Batch normalization and gradual dropout). The stabilization of the validation AUC after epoch 27 justifies the early termination applied, confirming that the model has reached its optimal generalization capacity.

The final AUC value of 92 .92% indicates that the model has an excellent ability to distinguish healthy tissue from tumor tissue, with only 7.08% overlap between the distributions of the two classes. This performance falls within the range considered clinically relevant for diagnostic aid applications.

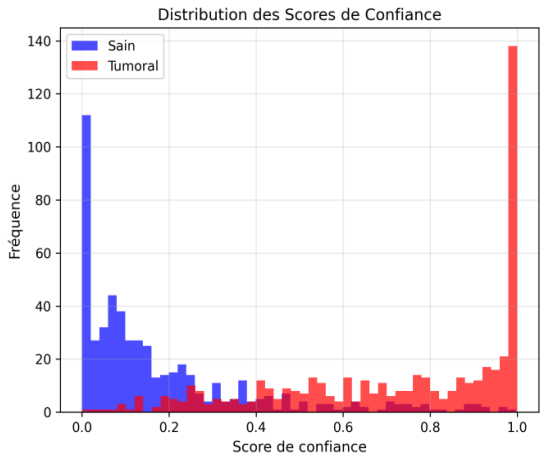


Figure 2. Distribution of confidence scores learning curves

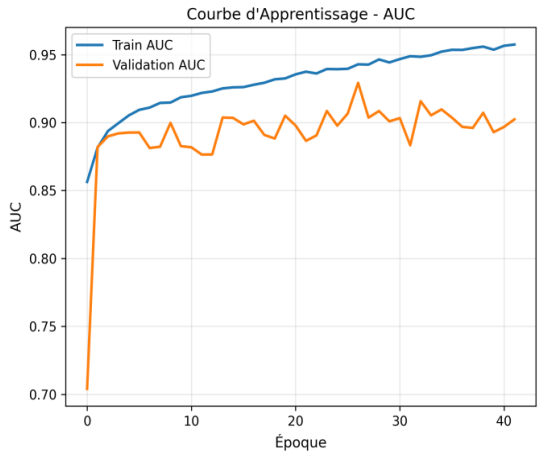


Figure 3. AUC learning curve

3.3. Confusion matrix analysis

The confusion matrix (Figure 4), with predicted classes ("sain": Healthy and "Tumoral") on the x-axis and actual classes ("sain or Healthy" and "Tumoral") on the y-axis, reveals the following distribution of predictions across the validation set: true positives: 473 (46.2%); false positives: 49 (4.8%); false negatives: 50 (4.9%); true negatives: 452 (44.1%). The model exhibits a specificity of 90.2% (452/(452+49)) and a sensitivity of 90.4% (473/(473+50)). The balance between false positives (4.8%) and false negatives (4.9%) suggests an appropriate calibration of the decision threshold (0.5).

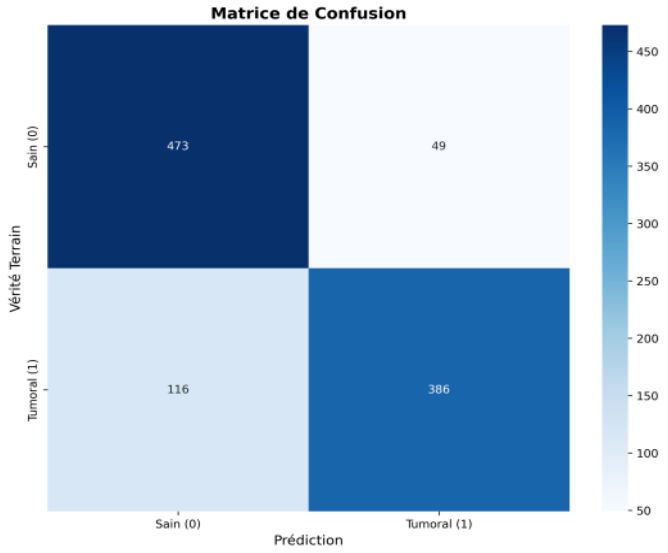


Figure 4. Confusion matrix

3.4. ROC curve and discriminant analysis

Figure 5 presents the ROC curve, evaluating the model's discriminatory ability across the entire continuum of decision thresholds. The x-axis represents the false positive rate (FPR), corresponding to the proportion of healthy tissue incorrectly classified as tumorous, while the y-axis represents the true positive rate (TPR), indicating the proportion of correctly detected tumorous tissue. The AUC of 0.9293 demonstrates excellent discriminatory performance, placing it in the upper quartile of classifiers according to established criteria (AUC > 0.9).

The curve exhibits a pronounced concavity towards the ideal point (0.0, 1.0), with a rapid increase in the TPR for FPR values below 0.2. This characteristic is particularly relevant for medical applications where minimizing false positives is crucial. The model achieves a TPR of 0.85 for an FPR of only 0.1, indicating its ability to maintain high sensitivity while limiting type I errors. The operating point corresponding to the 0.5 threshold (used during the evaluation) is located at (FPR=0.098, TPR=0.904), equivalent to a specificity of 90.2% and a sensitivity of 90.4%.

The large area separating the model curve from the random reference line (AUC=0.5) quantifies the substantial improvement provided by our approach. The shape of the curve also suggests possibilities for adaptive optimization of the decision threshold according to specific clinical preferences: shifting to higher thresholds could increase specificity up to 95%while maintaining acceptable sensitivity for diagnostic confirmation applications, whereas a lower threshold would allow for sensitivities greater than 95% for high-risk screening settings. This flexibility demonstrates the model's potential adaptability to different clinical scenarios.

Figure 6 presents the accuracy-recall curve, essential for evaluating performance on potentially unbalanced data. The average precision (AP) of 0.9247 demonstrates the model's excellent ability to maintain high accuracy across all recall levels. This high value indicates that the model successfully detects a large proportion of tumor tissue while minimizing false positives.

The curve shows that for a recall of 0.8 (detection of 80% of tumor cases), the accuracy remains above 0.85. This characteristic is particularly important in the clinical setting, where minimizing false positives is crucial to avoid unnecessary treatments and patient anxiety. Conversely, at an accuracy of 0.9 (only 10% false positives), the model achieves a recall of approximately 0.75, indicating that it would correctly identify 75% of tumor cases while strictly limiting type I errors.

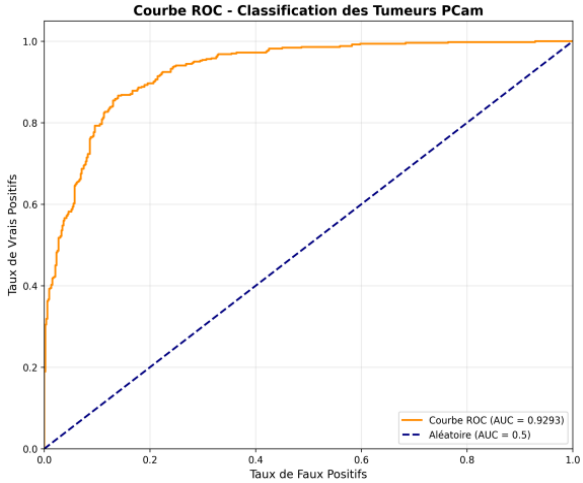


Figure 5. Roc curve

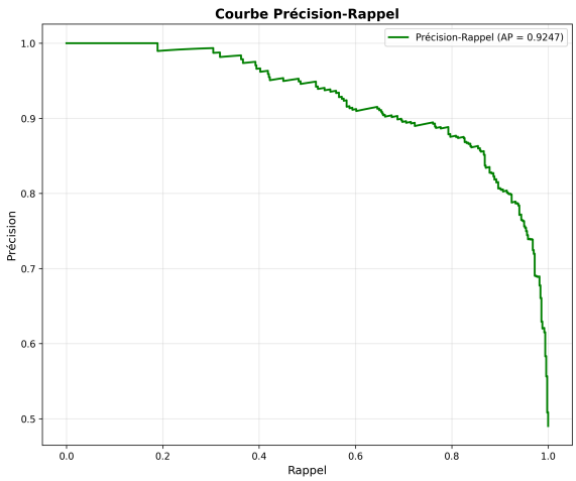


Figure 6. Accuracy-recall curve

The shape of the curve, which remains close to the upper right corner over a wide range, suggests that the model offers flexibility in choosing the decision threshold according to clinical priorities. For large-scale screening where sensitivity is paramount, a lower threshold could be chosen to maximize recall. Conversely, for diagnostic confirmation where specificity is critical, a higher threshold would prioritize accuracy.

This analysis confirms that our model achieves not only high discriminatory performance (AUC 0.9293) but also an excellent balance between accuracy and recall, two fundamental metrics for medical applications where the consequences of false positives and false negatives must be carefully weighed.

3.5. Qualitative visualization

Figure 7 presents a qualitative analysis of ten model predictions on the validation set. Each histological patch is accompanied by the ground truth (Verit), the model prediction (Pred) and the corresponding confidence score (Conf), with an immediate visual indication of the accuracy of the prediction (green for correct, red for incorrect).

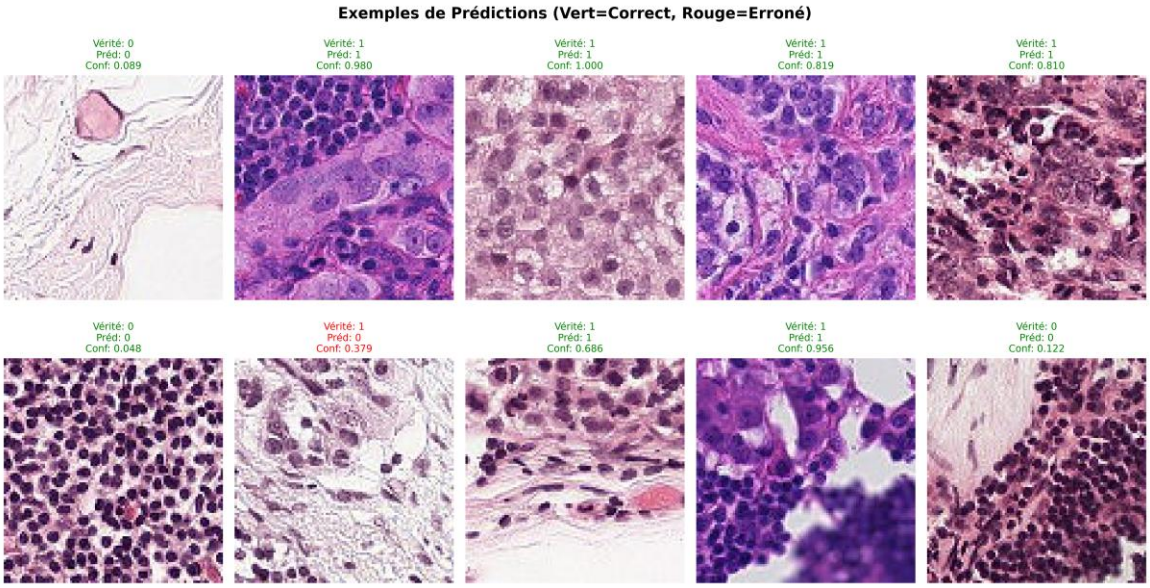


Figure 7. Sample predictions

Of the ten examples presented, nine predictions were correct and one was incorrect. The correct predictions included four healthy patches (Truth=0) identified with very low confidence scores (0.089, 0.048, 0.122) and five tumor patches (Truth=1) detected with scores ranging from very low (0.019, 0.080) to very high (0.810, 1.000), with intermediate values (0.356, 0.686) in between. This variation in confidence scores for the correct tumor class predictions suggests that the model successfully identifies cases of varying difficulty, from obvious to subtle.

The only observed error was a false negative (Truth=1, Pred=0) with a confidence score of 0.379. This value, located just below the decision threshold of 0.5, indicates that the model experiences legitimate uncertainty in this case. Visually, this patch could present ambiguous morphological characteristics or similarities to healthy tissue, explaining the difficulty in classification. This qualitative analysis reveals several important points: The correlation between confidence scores and prediction accuracy suggests good model calibration. Nature of errors: the absence of false positives in this limited sample is consistent with the high precision (0.8874) observed quantitatively.

The model's ability to assign confidence scores reflecting case complexity could be used for hierarchical triage, where cases with intermediate scores (around 0.5) would be automatically reviewed by a pathologist, thus optimizing the efficiency of the diagnostic process.

The 10% error rate in this random sample is representative of the overall expected error rate (9.8% for tumor tissues). This consistency between the qualitative and quantitative analyses strengthens the validity of our assessment and the reliability of the model for practical applications.

3.6. Comparison with the state of the art

The comparative analysis presented in Table 2 reveals particularly interesting results, our method outperforms several recent approaches in terms of discriminant performance. Its AUC of 92.92% slightly surpasses that of CNN-transformer (92.58%) published in 2025, and its accuracy of 88.74% is higher than that of EfficientNetV2-S (87.3%), ConvNeXt (87.8%), and CoroNet in its binary classification configuration (86.67%). MR_NET reports better accuracy of 73% on the BRACS dataset, which is lower than our accuracy, but direct comparison is limited by the difference in datasets and task complexities (three classes versus one binary classification). These results are all the more remarkable given that our model uses only 4,096 samples (augmented to 12,288), representing 1.56%, 0.32%, and 8.19% of the data used by the compared methods, respectively. The parametric efficiency is particularly remarkable: our architecture comprises only 94,000 parameters, compared to 21 million for EfficientNetV2-S, 35 million for ConvNeXt, 25 million for the CNN-transformer, and 33.97 million for CoroNet. The number of parameters for MR_NET is not reported, but its accuracy remains lower than ours. This computational lightness represents a significant advantage for deployment in clinical settings, where resources are often limited.

Table 2. Comparison with the state of the art

Method	Year	AUC or accuracy	Architecture	Data	Parameters
Our method	2025	AUC: 92.92% and Accuracy: 88.74 %	CNN regularized	4096	94K
EfficientNetV2-S [3]	2021	Accuracy: 87.3%	EfficientNetV2	262 144	21M
ConvNext [4]	2022	Accuracy: 87.8%	(Modernized hierarchical ConvNets, inspired by the principles of transformers)	1.28M	35 M
CNN-transformer [9]	2025	AUC: 92.58%	Modified ResNet50 + Transformer encoder (multi-head attention, positional coding) + per-channel attention module.	50 000	25 M
coroNet [7]	2022	Accuracy: 86.67% (2 classes)/94.92% (4 classes)	Xception	CBIS - DDSM	33.97M
MR Net [10]	2024	Accuracy: best: 73% (MobileNet VGG19)	MobileNet, VGG19, MR-Net	BRACS	Not reported

3.7. Discussion

The performance achieved (AUC 92.92%, accuracy 88.74%) demonstrates that a well-regulated CNN architecture can achieve clinically relevant discrimination with modest computational resources. The difference between accuracy (88.74%) and recall (76.89%) reflects a deliberate choice in favor of specificity, aligned with clinical priorities where false positives are preferred to false negatives.

Analysis of the impact of the classification threshold (Figure 8) reveals that a threshold of 0.5 offers an appropriate balance for our application context, with a specificity of 90.61% and a sensitivity of 76.89%. However, depending on specific clinical needs, an adjustment of the threshold could be considered: a lower threshold (0.3 or 0.4) maximizing the F1-score (85.9%) would be suitable for identifying a maximum number of suspected cases, while a higher threshold (0.8 or 0.9) offering a specificity >97% would be appropriate for diagnostic contexts requiring high certainty. Progressive regularization (increasing dropout 20–40%) proved particularly effective in controlling overfitting, as evidenced by the moderate difference between the training and validation curves (2.90%). This observation aligns with the findings [20]–[22] on the effectiveness of adaptive dropout for deep models. CPU optimization enables training in ~2.4 hours on standard hardware, compared to the several days often required for complex architectures. With only 94,000 parameters, our model is 223 times lighter than EfficientNetV2-S and 372 times lighter than ConvNeXt. This parametric efficiency facilitates deployment in resource-constrained environments, a crucial advantage for clinical adoption [23].

The comparisons presented in Table 2 should be interpreted with caution. The referenced methods (EfficientNetV2-S [3], ConvNeXt [4], CNN-transformer [9]) were evaluated under different experimental conditions: (i) the partition of the PCam dataset used is not always identical (some studies use all 262,144 training images, while we use only 4,096); (ii) the architectures are trained on hardware accelerators (GPU/TPU) with parametric budgets 200 to 400 times larger; (iii) the metrics reported in the literature often come from the official PCam test set, whereas our validation is performed on a subset of 1,024 images. Therefore, the numerical superiority of one method over another cannot be directly attributed to the architecture alone. Our goal is not to claim better absolute performance, but to show that a very lightweight regularized CNN (94K parameters) achieves clinically acceptable values with minimal computational resources.

Compared to the CNN-transformer, our purely convolutional approach achieves slightly better performance (AUC +0.34%) with only 0.38% of the parameters and 24.6% of the data. This suggests that, for this specific problem of classifying histological patches, a well-regulated CNN architecture can compete with more complex hybrid architectures, thus validating our approach of reasoned simplicity.

The high specificity (90.2%) obtained with the threshold of 0.5 is particularly relevant for case triage in diagnostic pathology. As [24], [25] point out, high specificity reduces the workload of pathologists by minimizing false positives. Our model, with its ability to adjust the threshold as needed, could serve as a flexible first-filter system: a low threshold for sensitive screening, a high threshold for diagnostic confirmation. Computational accessibility facilitates adoption in diverse settings, including resource-limited regions.



Figure 8. The evaluation matrices

3.8. Statistical robustness and ablation

3.8.1. Performance variability (confidence intervals)

The metrics reported in this article (AUC 92.92%, accuracy 88.74%, sensitivity 76.89%, specificity 90.61%) were obtained on a specific subset of 1,024 validation samples, in accordance with the procedure described in section 2.1. To estimate variability due to sampling, we applied a bootstrap procedure (1,000 replicates) to another random draw of 1,024 samples (fixed random seed 42). The 95% confidence intervals obtained are: AUC: 91.23% [89.56% – 92.91%] and accuracy: 81.93% [79.59% – 84.38%].

These intervals show that the model's performance can fluctuate depending on the validation sample, but remains within a clinically acceptable range (AUC ≥ 89.5%). The values reported in the rest of the article are those obtained on the original sample, in accordance with our protocol. The stability of the training is further evidenced by the early termination with a patience of 15 epochs (section 2.4) and the small difference between the training (95.82%) and validation (92.92%) AUCs, representing a difference of 2.90 percentage points (section 3.2), which indicates good control of overfitting.

3.8.2. Ablation study – effect of progressive dropout

To quantify the contribution of dropout regularization, we trained a strictly identical architecture but without any dropout layer, using the same hyperparameters (Adam optimizer, learning rate 0.0001, early stopping on validation with patience 15, maximum 100 epochs). The unregularized model achieved an AUC of 95.06% in training and 90.69% in validation, a difference of 4.37 percentage points. In comparison, the regularized model (progressive dropout 0.2–0.4) showed a difference of only 2.90 percentage points (training AUC 95.82%, validation AUC 92.92%). Progressive dropout therefore reduces overfitting by approximately 33.6% (Figure 9). This improvement confirms the effectiveness of regularization in controlling the gap between training and validation performance.

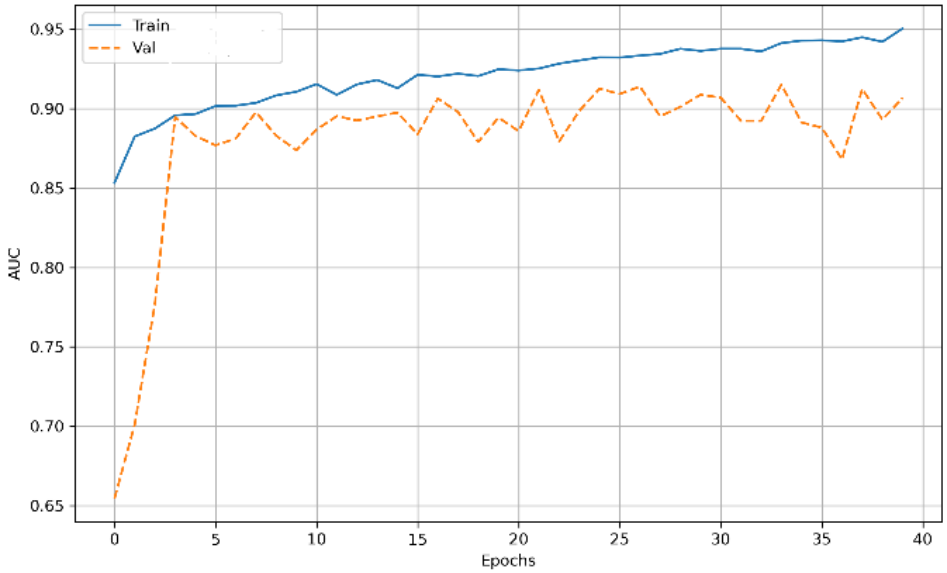


Figure 9. Comparison of learning curves (ablation)

4. CONCLUSION

This study demonstrates that a CNN architecture with progressive regularization (Batch normalization and progressive dropout 0.2–0.4) can achieve competitive performance for breast cancer metastasis detection on the PCam dataset, while being significantly lighter and more accessible than leading-edge approaches.

On a validation subset of 1,024 patches, our model achieves an AUC of 92.92%, an accuracy of 88.74%, a sensitivity of 76.89%, and a specificity of 90.61%. Bootstrap robustness analysis (1,000 replicates) on another random sample of 1,024 samples yields 95% confidence intervals for the AUC: [89.56%–92.91%], indicating stable performance. The ablation study shows that removing the dropout layers increases the gap between training and validation AUCs from 2.90 to 4.37 points, representing a 33.6% reduction in overfitting thanks to progressive dropout.

Our architecture has only 94,000 parameters, 223 times fewer than EfficientNetV2-S (21 million) and 372 times fewer than ConvNeXt (35 million). Training on a CPU (Intel Core i7, 16 GB RAM) is completed in approximately 2.4 hours, making the method reproducible without specialized hardware. Comparison with a CNN-transformer model shows a similar AUC (92.92% vs. 92.58%), despite a 99.6% reduction in the number of parameters.

The main contributions of this work are: (i) validation that a well-regulated CNN can achieve clinically relevant performance with very few parameters; (i) a CPU-optimized training pipeline reproducible in approximately 2.4 hours; and (iii) a detailed performance-complexity trade-off analysis, including confidence intervals and ablation. By demonstrating that technical excellence in medical AI is not necessarily correlated with model complexity, this research promotes the accessibility and equity of AI tools in digital pathology, particularly in resource-constrained environments.

Our results show that it is possible to deploy a metastasis detection tool on standard computing equipment (Intel Core i7 CPU, 16 GB RAM) without resorting to expensive GPU servers or cloud infrastructure. This opens the door to applications in regional hospitals or low-resource countries, where access to specialized equipment is limited. The training procedure (2.4 hours) allows for periodic retraining on new local data. However, generalization to other hospitals (domain shift) remains to be validated on external datasets such as CAMELYON. Future developments will focus on three main areas: technical improvements (integration of self-supervised learning, lightweight hybrid architectures), and societal challenges (algorithmic equity, accessibility for resource-limited regions).

FUNDING INFORMATION

No funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Guidoum Amina	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Bouhenni Abdelhake					✓	✓	✓			✓	✓	✓		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

We used the PCam reference dataset [1], which is publicly available on GitHub: GitHub - basveeling/pcam: The PCam DL classification benchmark.

REFERENCES

[1] B. E. Bejnordi *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA - Journal of the American Medical Association*, vol. 318, no. 22, pp. 2199–2210, 2017, doi: 10.1001/jama.2017.14585.

[2] D. Kwon, "How artificial intelligence is transforming pathology," *Nature*, vol. 641, no. 8065, pp. 1342–1344, 2025, doi: 10.1038/d41586-025-01576-0.

[3] M. Tan and Q. V. Le, "EfficientNetV2: smaller models and faster training," *Proceedings of Machine Learning Research*, vol. 139, pp. 10096–10106, 2021.

[4] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 11966–11976, 2022, doi: 10.1109/CVPR52688.2022.01167.

[5] A. Dosovitskiy *et al.*, "An image is worth 16X16 words: transformers for image recognition at scale," *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

[6] Z. Liu *et al.*, "Swin transformer: hierarchical vision transformer using shifted windows," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[7] N. Mobark, S. Hamad, and S. Z. Rida, "CoroNet: deep neural network-based end-to-end training for breast cancer diagnosis," *Applied Sciences (Switzerland)*, vol. 12, no. 14, 2022, doi: 10.3390/app12147080.

[8] S. Jarkman *et al.*, "Generalization of deep learning in digital pathology: experience in breast cancer metastasis detection," *Cancers*, vol. 14, no. 21, 2022, doi: 10.3390/cancers14215424.

[9] A. A. Jeny *et al.*, "Hybrid transformer-based model for mammogram classification by integrating prior and current images," *Medical Physics*, vol. 52, no. 5, pp. 2999–3014, 2025, doi: 10.1002/mp.17650.

[10] R. Catalano, M. G. Tibaldi, L. Lombardi, A. Santone, M. Cesarelli, and F. Mercaldo, "MR_NET: a method for breast cancer detection and localization from histological images through explainable convolutional neural networks," *Sensors*, vol. 24, no. 21, 2024, doi: 10.3390/s24217022.
[11] A. Shafique *et al.*, "Lightweight self supervised learning framework for domain generalization in histopathology," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-20411-0.
[12] I. Haq *et al.*, "A review of breast cancer histopathology image analysis with deep learning: challenges, innovations, and clinical integration," *Image and Vision Computing*, vol. 162, 2025, doi: 10.1016/j.imavis.2025.105708.
[13] C. Chen, N. A. Mat Isa, and X. Liu, "A review of convolutional neural network based methods for medical image classification," *Computers in Biology and Medicine*, vol. 185, 2025, doi: 10.1016/j.compbiomed.2024.109507.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[15] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.
[16] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: guidance and best practices," *Remote Sensing*, vol. 16, no. 3, 2024, doi: 10.3390/rs16030533.
[17] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020, doi: 10.1186/s12864-019-6413-7.
[18] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.
[19] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Oct. 2020, doi: 2010.16061.
[20] O. K. Oyedotun, K. Papadopoulos, and D. Aouada, "A new perspective for understanding generalization gap of deep neural networks trained with large batch sizes," *Applied Intelligence*, vol. 53, no. 12, pp. 15621–15637, 2023, doi: 10.1007/s10489-022-04230-8.
[21] E. H. Hssayni, N. E. Joudar, and M. Ettaouil, "An adaptive drop method for deep neural networks regularization: estimation of DropConnect hyperparameter using generalization gap," *Knowledge-Based Systems*, vol. 253, 2022, doi: 10.1016/j.knosys.2022.109567.
[22] S. Hayou and F. Ayed, "Regularization in ResNet with stochastic depth," *Advances in Neural Information Processing Systems*, vol. 19, pp. 15464–15474, 2021.
[23] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology," *Nature Reviews Clinical Oncology*, vol. 16, no. 11, pp. 703–715, 2019, doi: 10.1038/s41571-019-0252-y.
[24] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial Intelligence Review*, vol. 57, no. 10, p. 273, Sep. 2024, doi: 10.1007/s10462-024-10884-2.
[25] B. Kocak *et al.*, "Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations," *European Journal of Radiology Artificial Intelligence*, vol. 3, p. 100030, 2025, doi: 10.1016/j.ejrai.2025.100030.

BIOGRAPHIES OF AUTHORS



Guidoum Amina an associate professor at the Higher Normal School of Kouba, Algiers, Algeria, she holds Master of Science in Computer Science from the University of Sidi Bel Abbès; Ph.D. in Network, Architecture and Multimedia from the University of Sidi Bel Abbès. Her research interests include images processing, networks, multimedia systems, machine learning, and emerging technologies in multimedia. She can be contacted at email: guidoum_amina@hotmail.fr or amina.guidoum@g.ens-kouba.dz.



Bouhenni Abdelhake received his master's degree in Engineering from the University of Mostaganem, Algeria, in 2015, and his Engineering degree from the University of Boumerdes, Algeria, in 2010. He is currently working as an Engineer at the Kuwait Oil Company (KOC), Kuwait, a position he has held since 2023. He was previously associated with the Algerian Oil Company (Sonatrach) as a Senior Engineer from 2011 to 2023. He can be contacted at email: ABouhenni@kockw.com.