Int. J. Appl. Math. Comput. Sci., 2025, Vol. 35, No. 2, 329–339
DOI: 10.61822/amcs-2025-0023

EXPLORING DATA PREPARATION STRATEGIES: A COMPARATIVE ANALYSIS OF VISION TRANSFORMER AND CONVNEXT ARCHITECTURES IN BREAST CANCER HISTOPATHOLOGY CLASSIFICATION

MIKOŁAJ KACZMAREK [b], MAREK KOWAL [a,*], JÓZEF KORBICZ [a]

[a]Institute of Control and Computation Engineering
University of Zielona Góra
ul. Szafrana 2, 65-516 Zielona Góra, Poland
e-mail: {M.Kowal,J.Korbicz}@issi.uz.zgora.pl

[b]Doctoral School of Exact and Technical Sciences
University of Zielona Góra
al. Wojska Polskiego 69, 65-762 Zielona Góra, Poland
e-mail: 20000943@stud.uz.zgora.pl

Breast cancer remains a major global health challenge and the accurate classification of histopathological samples into benign and malignant categories is critical for effective diagnosis and treatment planning. This study offers a comparative analysis of two state-of-the-art deep learning architectures, Vision Transformer (ViT) and ConvNeXT for breast cancer histopathology image classification, focusing on the impact of data preparation strategies. Using the BreakHis benchmark dataset, we investigated six distinct preprocessing approaches, including image resizing, patch-based techniques, and cellular content filtering, applied across four magnification levels (40×, 100×, 200×, and 400×). Both models were fine-tuned and evaluated using multiple performance metrics: accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC). The results highlight the critical influence of data preparation on model performance. ViT achieved its highest accuracy of 95.6% and an F1 score of 96.8% at 40× magnification with randomly generated patches. ConvNeXT demonstrated strong robustness across scenarios, attaining a precision of 98.5% at 100× magnification using non-overlapping patches. These findings emphasize the importance of customized data preprocessing and informed model selection in improving diagnostic accuracy. Optimizing both architectural design and data handling is essential to enhancing the reliability of automated histopathological analysis and supporting clinical decision-making.

Keywords: vision transformer, ConvNeXT, BreakHis, data preparation, image classification.

1. Introduction

Breast cancer is one of the most prevalent cancers that affect women globally, representing a significant public health challenge. Early and accurate diagnosis is crucial for effective treatment and improving patient survival rates. Histopathological examination of breast tissue remains the gold standard for diagnosing breast cancer, involving the microscopic analysis of tissue samples by expert pathologists. However, this manual process is time-consuming, subject to inter-observer variability, and limited by the availability of specialized expertise.

The advancement of digital pathology and computational techniques offers the potential to enhance diagnostic accuracy and efficiency through automated image analysis. Deep learning, in particular, has emerged as a powerful tool for medical image classification, capable of learning complex patterns and features directly from raw data. Despite significant progress, several challenges persist in applying deep learning to histopathological images, including variability in staining procedures, differences in imaging equipment, and the inherent heterogeneity of biological tissues.

In addition, the high resolution and large size of histopathological images present computational

*Corresponding author

challenges, necessitating efficient data handling and processing strategies. The selection of appropriate neural network architectures and the optimization of data preparation techniques are critical steps in developing robust and accurate classification models.

In this study, we conduct a comprehensive comparative analysis of two state-of-the-art neural network architectures, the Vision Transformer (ViT) and ConvNeXT for breast cancer histopathology image classification using the BreakHis dataset (Spanhol *et al*., 2016). We explore six distinct data preparation scenarios to assess their impact on model performance across multiple magnifications (40×, 100×, 200×, 400×). By investigating the interplay between model architectures, data handling strategies, and image magnification levels, we aim to identify optimal approaches for enhancing diagnostic accuracy in breast cancer histopathology.

Our contributions are threefold:

- *Evaluation of ViT and ConvNeXT architectures.* We assess the performance of ViT and ConvNeXT models on the BreakHis dataset, providing insights into their suitability and effectiveness for histopathological image classification tasks.
- *Investigation of data preparation strategies.* We introduce and systematically evaluate six data preparation scenarios, including advanced patch-based methods and filtering techniques based on cellular content, to determine their impact on model performance.
- *Analysis of Image Magnification Effects.* We analyze how different magnification levels affect classification accuracy, offering guidance on selecting appropriate magnification for histopathological analysis.

The findings of this research could guide the development of more effective and reliable diagnostic tools, ultimately enhancing patient outcomes in breast cancer treatment.

2. Related works

The field of histopathological image classification has witnessed significant advancements in recent years, largely driven by the adoption of deep learning techniques. In this section, we review the state-of-the-art methods, highlight studies that have conducted similar comparisons between different neural network architectures, and discuss works that emphasize the importance of data preparation in histopathological image analysis. Recent advancements also include self-supervised learning and domain adaptation techniques, which further enhance the robustness of classification models.

Deep learning has revolutionized image classification by enabling models to automatically learn hierarchical features from raw data (Ara *et al*., 2023). Convolutional neural networks (CNNs) remain a backbone of many high-performing solutions due to their capability to capture localized spatial hierarchies. Early influential architectures such as AlexNet (Krizhevsky *et al*., 2012), VGGNet (Simonyan and Zisserman, 2015), ResNet (He *et al*., 2016), and DenseNet (Huang *et al*., 2017) marked pivotal points in computer vision research, demonstrating strong performance on large-scale image datasets.

In histopathological imaging, CNNs have been widely used for tumor classification and region-of-interest detection tasks. However, these images often exhibit complex tissue structures, high intra-class variability, and subtle morphological features, posing significant challenges for traditional CNN-based approaches. To address these limitations, recent advances in computer vision have shifted attention toward Transformer-based architectures—originally developed for natural language processing. The Vision Transformer (ViT) and its variants leverage self-attention mechanisms to model long-range dependencies, enabling the extraction of global contextual information from histopathological slides (Dosovitskiy *et al*., 2021). In the medical imaging domain, Transformers have demonstrated strong potential across various tasks, including tumor segmentation and lesion classification (Wang *et al*., 2021).

Additionally, hybrid models that integrate CNN layers with Transformer modules have emerged, aiming to fuse the strengths of local feature extraction (from CNNs) and global context modeling (from Transformers) (Graham *et al*., 2021). This approach has gained significant attention in recent literature, with specialized architectures processing gigapixel whole-slide images by dividing them into patches and aggregating their embeddings for downstream analysis.

Several works have conducted comparative evaluations of diverse network architectures in the histopathological domain. Deininger *et al*. (2022) compared standard CNN-based methods with Transformer-based models for breast cancer histopathology, finding that Transformers offered improved global reasoning and marginally higher accuracies on tasks requiring contextual awareness. Takahashi *et al*. (2024) extended this comparison to include various medical imaging modalities and reported that ViT significantly outperformed CNNs whenever large-scale datasets or patch-based approaches were employed, highlighting the benefit of self-attention in capturing long-range correlations.

ConvNeXT, proposed as a modernized CNN architecture, retains convolutional structures but incorporates modifications inspired by Transformer

design (e.g., layer normalization, inverted bottleneck structures) to enhance efficiency and performance (Liu *et al*., 2022). Notably, ConvNeXT has demonstrated competitive results against Transformer-based models in computer vision benchmarks. In the histopathology domain, its ability to extract detailed local features makes it particularly promising for classification tasks that depend heavily on local texture patterns. Its performance can be further boosted with advanced data handling, as local textural information is often highly relevant for accurate classification in pathology.

Our work distinguishes itself by not only assessing the performance of ViT and ConvNeXT architectures on the BreakHis dataset but also by providing a detailed analysis of how different data preparation scenarios such as patch extraction, random transformations, and background/cell filtering affect classification outcomes. This granular perspective is crucial for practitioners to tailor model and data strategy decisions based on diagnostic requirements and hardware constraints.

Data preparation remains a cornerstone in achieving high accuracy and reliability in histopathological image classification. As Tellez *et al*. (2019) point out, variations in staining protocols, scanner types, and tissue characteristics can introduce significant distribution shifts, hindering the generalization of trained models. Data augmentation methods ranging from geometric transformations and color normalization to more sophisticated generative approaches help mitigate these shifts by increasing the effective variety of training samples.

Campanella *et al*. (2019) showed that careful selection of patches and balanced sampling strategies can substantially improve a model's recognition capabilities in whole-slide analysis, enabling the training process to focus on diagnostically relevant regions. Furthermore, Komura and Ishikawa (2018) demonstrated that filtering out uninformative patches (e.g., background-only regions) and focusing on cells or stroma can help reduce noise in training and enhance classification performance.

Recent studies have begun exploring self-supervised learning (SSL) frameworks to pre-train models on large collections of unlabeled histopathological slides, obtaining robust feature representations before task-specific fine-tuning. Such methods can be combined with domain adaptation techniques to reduce the performance gap between different institutions or scanner types. Although these methods are not the primary focus of our current work, their integration with the architectures and data preparation methods discussed here may further improve classification performance and facilitate more efficient annotation workflows.

Overall, the literature underscores that data handling strategies including patch-based extraction, selective filtering, and thorough augmentation are pivotal in medical image analysis. Our research builds upon this foundation by systematically investigating how such strategies can be tailored to maximize the potential of both Transformer-based and next-generation CNN models for breast cancer histopathology classification.

3. Materials and methods

3.1. Data. We utilize the Breast Cancer Histopathological Image Classification (BreakHis) dataset, designed to support the automated classification of breast tumor histopathological images as benign or malignant (Spanhol *et al*., 2016). Compiled by the P&D Laboratory, Department of Pathology, Parana, Brazil, and meticulously annotated by pathologists, BreakHis is a valuable resource for advancing cancer diagnostic technologies using machine learning and artificial intelligence.

The BreakHis dataset contains 7,909 microscopic images from 82 patients, captured at four different magnifications (40×, 100×, 200×, and 400×), with each magnification level comprising roughly 2,000 images. The dataset reflects real-world clinical distributions, with *31.36% benign and 68.64% malignant* cases.

3.2. Model architectures. In our study, we utilize two state-of-the-art neural network models, the Vision Transformer (ViT) and ConvNeXT, pioneering the analysis of complex image data such as histopathological slides. These models were selected due to their advanced architectural features and differing approaches to image recognition, which are pivotal in pushing the boundaries of classification accuracy in medical image analysis.

The *Vision Transformer (ViT)* architecture (Fig. 1) introduces a paradigm shift from traditional convolutional networks to a mechanism driven by self-attention, adapting methodologies that have excelled in natural language processing to the realm of image analysis (Dosovitskiy *et al*., 2021). ViT decomposes an image into a series of fixed-size patches, converts them into linear embeddings, and processes these through multiple layers of self-attention combined with positional encodings to maintain the spatial hierarchy of the image pixels. Architectural details include:

- *Patch embedding:* ViT applies a simple linear projection that turns the flattened RGB values of each patch into the input token embeddings.
- *Positional encodings:* Adds location information to the input embeddings.
- *Transformer encoder:* Consists of alternating layers of multi-headed self-attention and MLP blocks (with GELU non-linearity), each followed by layer normalization.

- *Classification head:* Uses the output of the transformer applied to the [CLASS] token for classification tasks.

This architecture allows ViT to focus on global dependencies between any parts of the image, enabling it to potentially identify complex patterns important for classifying medical images. However, the dependency on vast amounts of training data and extensive compute resources can limit its applicability in constrained environments.

The *ConvNeXT* model (Fig. 2), introduced by Liu *et al*. (2022), represents a modern evolution of traditional convolutional neural network (CNN) design. Drawing inspiration from Transformer architectures, it incorporates a series of structural refinements aimed at improving both performance and scalability. While preserving the spatial hierarchies that are fundamental to CNNs, ConvNeXT reimagines the architectural framework to better accommodate the demands of contemporary deep learning tasks. Architectural innovations include:

- *Modified ResNet design:* Adopts a ResNet-like design with simplified layer norms and restructured blocks to enhance training dynamics and feature extraction.
- *Inverted bottleneck:* Expands convolutional layers' input features before squeezing them back to reduce computational load.
- *Depthwise separable convolutions:* Lowers parameter count and computational complexity by separating convolution into two layers.
- *GELU non-linearity and layer normalization:* Stabilizes the learning process and enhances non-linear capabilities.

ConvNeXT excels in extracting localized features due to its evolved convolutional approach, making it highly effective for tasks requiring detailed feature analysis. However, its ability to capture long-range dependencies might not be as robust as that of ViT, which can impact its performance in scenarios where the global image context is crucial.

The choice of these models allows us to explore how different neural network architectures process and learn from highly detailed and complex histopathological image data. By analyzing the strengths and limitations of each model across various data preparation scenarios, we aim to identify optimal strategies for enhancing classification performance in the medical imaging domain.

3.3. Fine-tuning and hyperparameters. For this study, we employed two pretrained models: the *Vision Transformer Base* with $85.8M$ parameters and the *ConvNeXT Base* with $87.6M$ parameters, maintained in the *Hugging Face*'s Infrastructure. These models were initially trained on *ImageNet-21k*, making them well-suited as starting points due to their robust initial feature-detection capabilities. The fine-tuning process was tailored to adapt these models to the specialized domain of histopathological image analysis, focusing specifically on the BreakHis dataset.

The fine-tuning of both models was guided by careful selection and optimization of hyperparameters, primarily focusing on the learning rate and weight decay as critical factors influencing model convergence and performance. We executed a grid search (not exhaustive for every model) to identify optimal hyperparameters that balance training efficiency and model accuracy. Through this process, we determined that employing the Adam optimizer with a learning rate initially set at 3×10^{-5} and a weight decay rate of 5×10^{-3} yielded the best results for our tasks. These parameters were selected based on preliminary experiments indicating their effectiveness in achieving rapid convergence while mitigating the risk of overfitting.

3.4. Stratified and grouped splitting. In medical image analysis, ensuring the robustness and generalizability of predictive models is paramount. One of the key challenges in training such models lies in the proper structuring of the training, validation, and testing sets, particularly when dealing with patient-specific data. The integrity of the dataset split is critical not only for the validity of the model's performance metrics but also for its applicability in real-world scenarios. To address these challenges, we implemented a stratified and grouped splitting strategy, tailored to maintain the integrity of patient data.

Our stratified and grouped splitting algorithm ensures:

1. *Patient integrity*: All images from a single patient are grouped together, meaning that if one image of a patient is assigned to a particular set (e.g., training), all other images of the same patient are also assigned to the same set. This grouping is crucial for diseases like cancer, where patient-specific characteristics can significantly influence the diagnosis.
2. *Class balance*: The algorithm maintains a balanced representation of classes (e.g., benign vs. malignant) across all splits. This stratification is essential to prevent model bias toward the more prevalent class, which could skew the diagnostic performance.

The implementation involves first identifying and segregating the data according to patient IDs, ensuring that each patient's data is treated as a single unit. The algorithm then distributes these units into training,

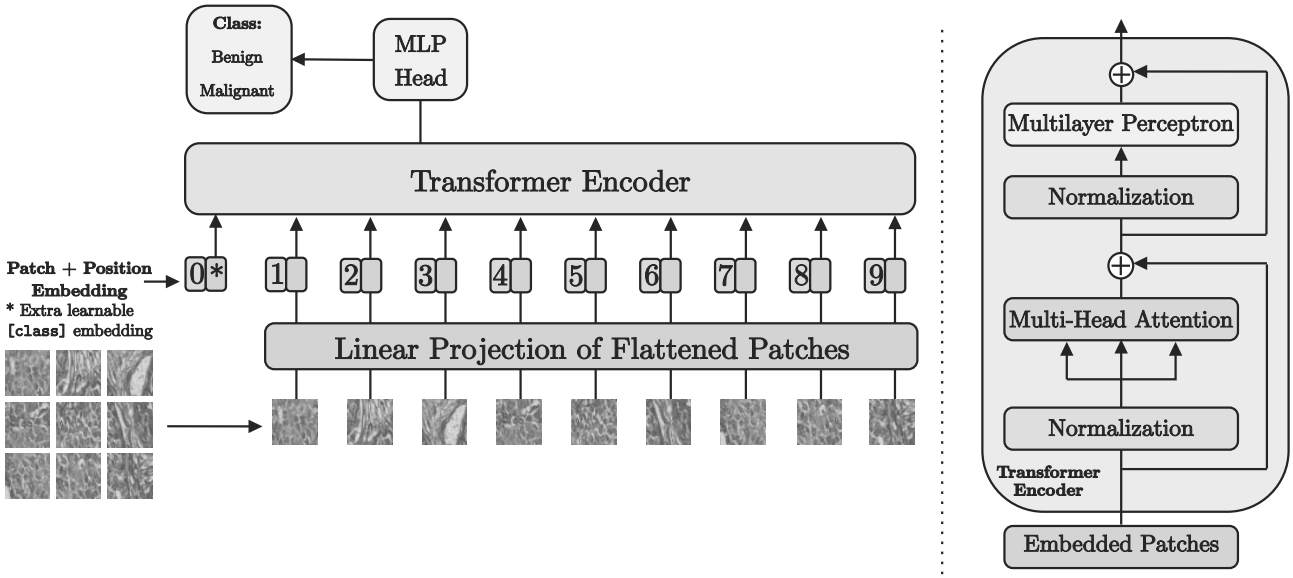

Fig. 1. Vision transformer architecture overview.

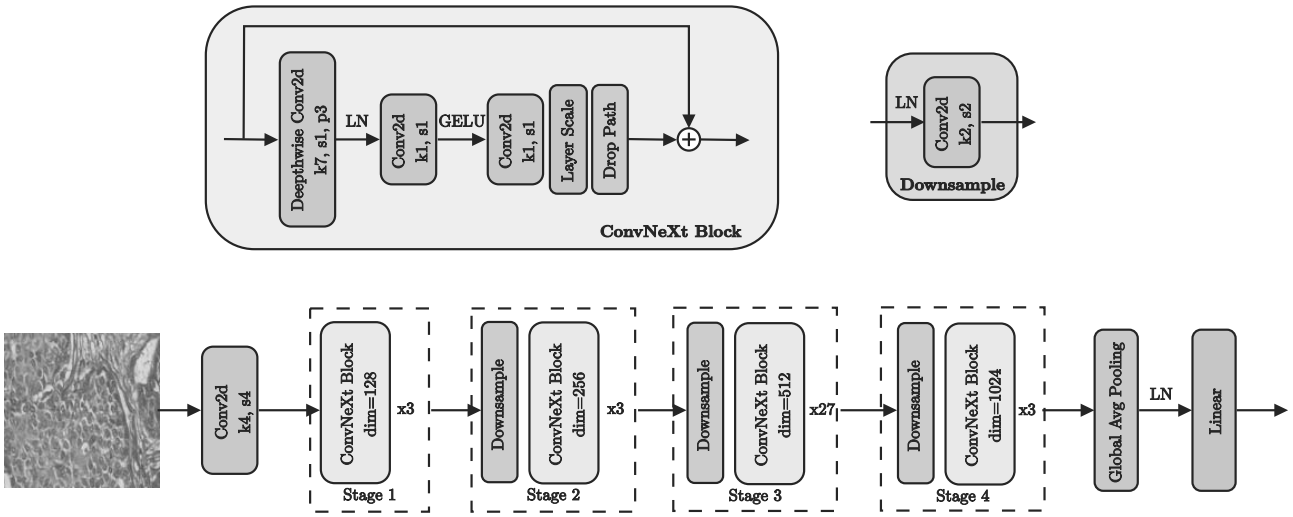

Fig. 2. ConvNeXT architecture overview.

validation, and testing sets while maintaining a balanced distribution of diagnostic categories within each set.

To ensure the integrity of our model evaluations, the BreakHis dataset was split into distinct sets for training and testing. Initially, 20% of the data was reserved strictly for testing, ensuring that this set was not seen by the models during training or validation phases. The remaining 80% of the data was used for training and validation purposes. This data was further divided into five folds, following a cross-validation approach that allows for comprehensive assessment and robustness checks of the trained models. An example data split is shown in Table 1.

In conclusion, the stratified and grouped splitting algorithm is not just a technical necessity but a fundamental aspect of our research methodology that aligns with the ethical considerations and practical requirements of medical image analysis. This approach significantly mitigates the risk of data leakage and ensures that the models developed are both scientifically valid and practically viable in real-world medical applications.

3.5. Data preparation. To accommodate the unique demands of histopathological image analysis and the specific requirements of the neural networks employed, we designed and implemented a series of data preparation scenarios, labeled from A to F. These scenarios were constructed on the foundation of the initial stratified and

Table 1. Distribution of benign and malignant cases in training, validation, and test sets; magnification 400×.

Dataset	Samples	% Benign	% Malignant
All	1820	32.31%	67.69%
Test	397	35.01%	64.99%
Train Fold 0	1103	31.91%	68.09%
Val Fold 0	320	30.31%	69.69%
Train Fold 1	1233	32.85%	67.15%
Val Fold 1	190	23.16%	76.84%
Train Fold 2	1110	30.18%	69.82%
Val Fold 2	313	36.42%	63.58%
Train Fold 3	1130	31.06%	68.94%
Val Fold 3	293	33.45%	66.55%
Train Fold 4	1116	31.63%	68.37%
Val Fold 4	307	31.27%	68.73%

grouped split, ensuring that all images from a single patient remained grouped together while maintaining a balanced representation of classes across each scenario. This rigorous approach prevents data leakage and supports the integrity of the training process.

Each scenario was carefully crafted to explore the impact of different image processing techniques on the model's ability to accurately classify images from the BreakHis dataset. By preserving the class balance and group integrity in each scenario, we aim to provide a robust examination of how variations in the input data affect the learning and generalization capabilities of the models.

Scenario A: Baseline (Original Dataset). In this scenario, we processed the original images from the BreakHis dataset, which typically measure 700×460 pixels, by resizing them directly to 224×224 pixels, the input size required by our models. This resizing was performed using bilinear interpolation. While this method preserves a significant amount of the original image content, it inevitably introduces some degree of data loss and distortion, particularly affecting fine textural details that are crucial for accurate cancer diagnosis.

Scenario B: Non-overlapping patches. To mitigate the loss of detail inherent in resizing large images to a much smaller size, we implemented a patch-based approach. Each original image was divided into non-overlapping patches of 224×224 pixels. This method ensures that no data are discarded during the resizing process, as each patch retains the original resolution. However, this technique may lose contextual information since each patch only contains a portion of the larger image, potentially omitting important diagnostic features located at the boundaries between patches.

Scenario C: Randomly generated patches. Building on the fixed patch approach, we introduced randomness in the patch extraction process to enhance model robustness and prevent overfitting. Along with fixed patches, additional patches were generated by randomly selecting their positions within the original images. This randomness may help the model learn to recognize pathological features from various parts of the image, regardless of their location, thus improving the model's ability to generalize across different image compositions.

Scenario D: Randomly generated patches with applied transformations. To further increase the diversity of the training data and simulate various real-world conditions under which histopathological images might be taken, we applied random transformations to each patch. These transformations included rotations, flips, color jittering (adjusting brightness, contrast, saturation, and sharpness), and noise addition. Such augmentations are intended to make the models resilient to variations in image acquisition and processing conditions that occur in clinical settings.

Scenario E: Randomly generated patches with filtered cells. We applied a filtering process to select patches based on cellular content. First, patches were converted to grayscale and then binarized using Otsu's method, dividing pixels into two classes: white (cells) and black (background). Patches were evaluated based on the proportion of white pixels (cellular content). Only patches containing more than 50% white pixels were selected, ensuring that training data primarily consisted of areas rich in cellular content. Alternative thresholds (e.g., 80%) were considered, but this resulted in excessive data loss due to insufficient patches meeting this stringent criterion. The selected threshold (50%) provided a balanced dataset size and scenario specificity, though some overlap near this boundary was inevitable and is acknowledged.

Scenario F: Randomly generated patches with filtered background. Conversely, this scenario focused on evaluating the impact of non-cellular (background) regions. Using the same binarization procedure with Otsu's method (white pixels representing cells and black pixels representing background), we selected patches predominantly consisting of background areas. Specifically, only patches with less than 50% white pixels (cell content) were included. Similar to scenario E, more extreme thresholds (such as 20%) significantly reduced available data, leading to inadequate representation. The chosen threshold (50%) allowed a sufficient dataset size and a meaningful investigation into the potential diagnostic information contained within tissue backgrounds. Potential ambiguity from patches near the 50% threshold boundary is recognized and discussed in the results and interpretations.

Image preprocessing. Before each training session, all images were preprocessed using model-specific

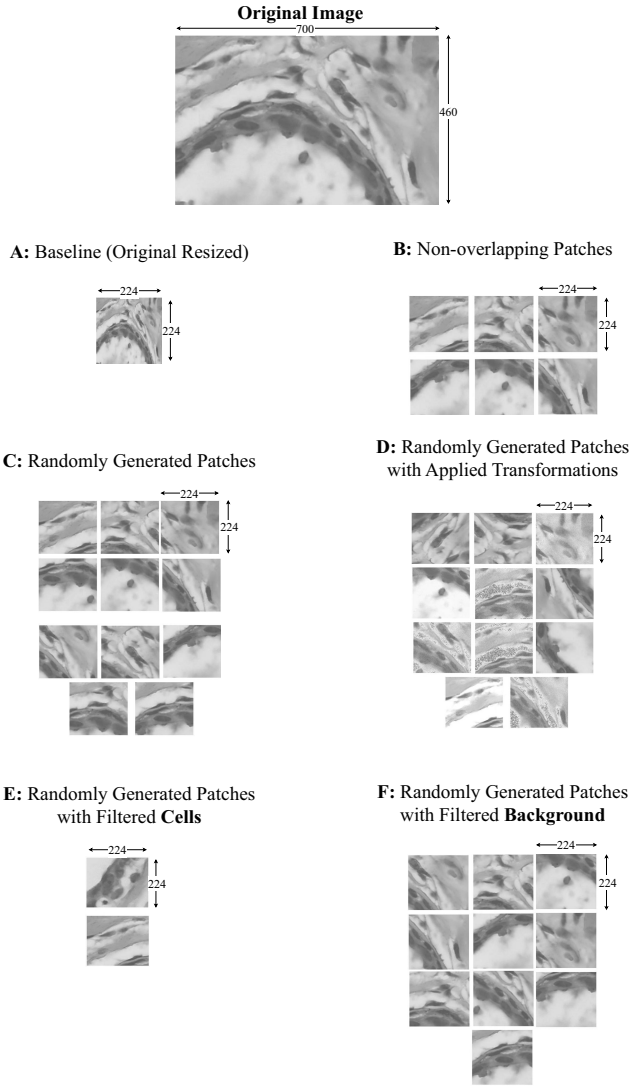

Fig. 3. Data preparation scenarios depicted on a sample.

configurations. For ConvNeXT models, preprocessing steps included cropping, normalization, and rescaling to enhance image contrast. Vision Transformer models were processed similarly, including resizing, normalization, and rescaling. These operations were applied during the actual loading of images into the model, ensuring that each image was optimally prepared for the specific requirements of the network architecture. All data preparation scenarios are depicted in Fig. 3.

3.6. Evaluation metrics. To assess the performance of the models on the BreakHis dataset, we implemented a suite of evaluation metrics that capture various aspects of model accuracy and prediction quality. These metrics provide a robust framework for evaluating the effectiveness of the models in classifying histopathological images into benign and malignant categories.

Our evaluation computes several key metrics: *accuracy*, *precision*, *recall*, *F1 score*, *ROC AUC*, and *precision-recall AUC*.

During training, model performance is continuously monitored using an *early stopping* mechanism configured to halt training if the validation loss does not improve beyond a threshold of 0.001 for 6 consecutive evaluation calls. Additionally, the validation set is evaluated every 250 steps during training, with each batch consisting of 16 samples. The model instance that achieves the lowest validation loss is preserved for further validation and testing.

3.7. Evaluation on test dataset. Following the training phase, each iteration yields five models due to 5-fold cross-validation. The best models, as determined by the lowest validation loss from each fold and scenario, are further evaluated on the test dataset. This evaluation is performed across all magnifications and data preparation scenarios for both the Vision Transformer (ViT) and ConvNeXT models. We record and analyze the performance metrics on the test dataset, providing a detailed landscape of how each model adapts to different data characteristics and magnifications.

3.8. Hardware and software configuration. The experiments were conducted on a system with an NVIDIA GeForce RTX 3070 Mobile GPU (8 GB VRAM), an AMD Ryzen 5 5600H CPU, and 32 GB RAM. This accessible yet capable hardware configuration ensures scalability and reproducibility, making it suitable for research in resource-constrained environments. The deep neural network models were implemented and evaluated using the Python programming language and the PyTorch deep learning library.

4. Results

The performance of the Vision Transformer (ViT) and ConvNeXT models across different data preparation scenarios and magnifications was thoroughly analyzed. The results demonstrate significant variability in model performance, underscoring the impact of architectural differences and the nuances of data preparation. Notably, the ViT model excelled in scenarios involving randomness and patches, particularly at the $40\times$ magnification where it reached peak performance metrics in accuracy, precision, and F1 score. In contrast, ConvNeXT showed robustness across various scenarios, with consistently high performance in transformed data scenarios, especially noted at the $100\times$ magnification with the highest recorded precision.

Attention was focused on the best-performing models from each fold. For instance, at 100× magnification, the ViT model trained on patches with randomly filtered cells demonstrated exceptional diagnostic accuracy, precision, and F1 scores, suggesting a strong alignment between model capabilities and data characteristics. Similarly, the ConvNeXT model achieved remarkable precision and specificity in the original and transformation-augmented scenarios at 400×, indicating its adeptness at integrating a broad image context.

Graphical representations such as ROC curves and precision-recall curves were employed to visually summarize the models' performance. These visualizations highlighted the strengths and weaknesses of each model across different scenarios, providing intuitive insights into their efficacy and reliability. The ROC and PR curves for ViT and ConvNeXT models consistently demonstrated higher areas under the curve in scenarios where patch-based approaches and randomness were introduced.

The detailed performance of each model across various magnification levels and data preparation strategies is summarized in Tables 2–5. They present a comprehensive overview of model behavior under varying experimental conditions, highlighting specific strengths and limitations in histopathological feature recognition. This analysis underscores the interplay between data preparation strategies and model architecture, offering guidance on optimizing performance for clinical deployment.

Both the Vision Transformer (ViT) and ConvNeXT models demonstrate rapid convergence during the fine-tuning process, typically achieving optimal validation performance within just 2–3 epochs. Beyond this point, training accuracy often approaches 100%, indicating minimal subsequent weight updates and suggesting that the models have effectively adapted to the task. This swift convergence reflects the strength of their pre-training, particularly on large-scale datasets such as ImageNet-21k, and the efficiency of transfer learning in histopathological image classification. To mitigate the risk of overfitting and enhance generalizability, we apply early stopping based on validation performance. This approach ensures that the models maintain strong predictive power on unseen data, reinforcing their suitability for practical applications.

Table 2. Performance metrics for the Vision Transformer (ViT) and ConvNeXT models at **40**× magnification.

Model	Acc.	Prec.	Rec.	F1	ROC	PR
ViT-A	89.8	89.1	95.3	92.1	0.947	0.964
ViT-B	89.6	87.5	98.9	92.9	0.905	0.940
ViT-C	**95.6**	**96.8**	96.9	**96.8**	**0.986**	**0.993**
ViT-D	93.1	93.0	97.5	95.2	0.961	0.979
ViT-E	83.8	85.0	95.1	89.8	0.909	0.969
ViT-F	88.0	94.3	83.7	88.7	0.941	0.960
CN-A	82.7	80.8	94.9	87.3	0.943	0.967
CN-B	85.1	82.4	**99.4**	90.1	0.899	0.925
CN-C	94.2	93.5	98.5	96.0	0.975	0.987
CN-D	94.2	94.9	96.9	95.9	0.980	0.990
CN-E	82.5	84.6	93.6	88.9	0.858	0.937
CN-F	81.2	89.9	74.9	81.7	0.908	0.930

Table 3. Performance metrics for the Vision Transformer (ViT) and ConvNeXT models at **100**× magnification.

Model	Acc.	Prec.	Rec.	F1	ROC	PR
ViT-A	80.3	88.6	83.6	86.0	0.842	0.924
ViT-B	87.0	86.2	96.5	91.1	0.948	0.973
ViT-C	85.1	89.4	89.9	89.7	0.928	0.971
ViT-D	82.2	92.8	81.5	86.8	0.915	0.965
ViT-E	**93.0**	94.0	**97.7**	**95.8**	0.974	**0.994**
ViT-F	74.0	76.9	73.3	75.0	0.816	0.784
CN-A	83.4	89.8	86.9	88.3	0.853	0.912
CN-B	87.8	89.0	93.8	91.3	0.945	0.975
CN-C	84.8	88.0	91.2	89.6	0.895	0.953
CN-D	89.5	**98.5**	86.6	92.2	**0.979**	0.990
CN-E	85.5	96.1	85.7	90.6	0.923	0.982
CN-F	82.2	80.1	88.6	84.1	0.920	0.937

Table 4. Performance metrics for the Vision Transformer (ViT) and ConvNeXT models at **200**× magnification.

Model	Acc.	Prec.	Rec.	F1	ROC	PR
ViT-A	85.1	84.2	94.8	89.2	0.868	0.906
ViT-B	86.0	87.1	92.9	89.9	0.924	0.955
ViT-C	82.1	88.3	84.8	86.5	0.824	0.835
ViT-D	85.7	88.3	90.8	89.6	0.858	0.893
ViT-E	**91.3**	92.2	**97.3**	**94.7**	0.901	0.970
ViT-F	90.5	**95.8**	87.2	91.3	**0.973**	**0.980**
CN-A	81.9	79.5	97.2	87.5	0.821	0.807
CN-B	85.9	90.0	88.9	89.4	0.928	0.958
CN-C	89.8	89.5	96.3	92.8	0.857	0.882
CN-D	81.4	88.3	83.6	85.9	0.832	0.861
CN-E	90.8	93.3	95.3	94.3	0.920	0.974
CN-F	88.1	90.5	88.5	89.5	0.955	0.966

5. Discussion

The primary objective of this study was to evaluate and compare the performance of the Vision Transformer (ViT) and ConvNeXT architectures in classifying breast cancer histopathology images from the BreakHis dataset under various data preparation scenarios. Our findings reveal that data preparation strategies significantly impact model performance and that both architectures have distinct advantages depending on the scenario.

5.1. Comparison with previous studies. Our results align with the findings of Deininger *et al*. (2022), who demonstrated that Transformer-based models outperform CNN-based models in capturing global contextual

Table 5. Performance metrics for the Vision Transformer (ViT) and ConvNeXT models at **400**× magnification.

Model	Acc.	Prec.	Rec.	F1	ROC	PR
ViT-A	84.9	91.2	84.9	88.0	0.920	0.953
ViT-B	84.2	84.1	92.6	88.1	0.887	0.908
ViT-C	89.2	86.3	**99.3**	**92.4**	0.916	0.939
ViT-D	85.3	85.0	94.2	89.4	0.830	0.856
ViT-E	82.8	92.7	84.4	88.4	0.915	0.971
ViT-F	69.9	69.4	82.4	75.4	0.747	0.746
CN-A	**89.7**	**99.1**	84.9	91.4	**0.978**	**0.989**
CN-B	84.0	87.4	87.3	87.4	0.882	0.895
CN-C	86.7	90.8	88.8	89.8	0.909	0.939
CN-D	88.2	86.8	96.7	91.5	0.915	0.945
CN-E	80.9	84.6	92.1	88.2	0.862	0.952
CN-F	78.0	80.2	80.5	80.4	0.755	0.701

information in breast cancer histopathology images. Specifically, we observed that the ViT model excelled in scenarios involving random patches (Scenario C) at 40× magnification, achieving a peak accuracy of 95.6%, which suggests that ViT effectively captures global patterns even when trained on localized patches. This ability to model long-range dependencies is consistent with the observations made by Takahashi *et al*. (2024), who found that ViT models are advantageous in histopathological analysis due to their proficiency in handling global context.

Conversely, the ConvNeXT model demonstrated robustness across various data preparation scenarios, particularly excelling in Scenario B (Non-overlapping Patches) at 100× magnification, where it achieved a precision of 98.5%. This observation supports the notion that CNN-based architectures, like ConvNeXT, are highly effective at capturing local features, as highlighted in prior studies (He *et al*., 2016; Liu *et al*., 2022). Our findings suggest that ConvNeXT's evolved convolutional design allows it to efficiently process histopathological images where fine-grained local features are critical for classification.

5.2. Impact of data preparation strategies. Our study extends the work of Tellez *et al*. (2019) and Campanella *et al*. (2019) by systematically evaluating the impact of various data preparation strategies on model performance. We found that scenarios involving random patch generation and the application of transformations (Scenarios C and D) significantly enhanced the performance of both models. These results underscore the importance of data augmentation and preprocessing techniques in mitigating issues related to variability in staining, illumination, and tissue morphology, which are common challenges in histopathological image analysis.

Moreover, the use of cellular content filtering (Scenario E) improved the ViT model's performance at 100× magnification, achieving an F1 score of 95.8%. This aligns with the findings of Komura and Ishikawa (2018), who emphasized that selecting informative regions is vital for training effective models on histopathological images. By focusing on patches rich in cellular content, the models could better learn relevant features associated with malignancy.

The chosen threshold of 50% for distinguishing cellular content (scenario E) from background (scenario F) provided an optimal compromise, maximizing dataset retention while maintaining distinctiveness between scenarios. Although a moderate overlap near this threshold was inevitable, the experimental outcomes validate the usefulness of this strategy, highlighting a potential diagnostic value in both cell-rich and background-dominated regions.

An interesting observation from our experiments was that scenario F, focused primarily on background areas with minimal cellular presence, sometimes achieved superior results according to certain metrics (particularly ROC AUC and precision-recall AUC). This aligns with feedback from histopathologists who indicated that non-cellular background regions might still provide valuable diagnostic information due to subtle stromal features or tissue structures. Thus, scenario F demonstrates the potential diagnostic significance of non-cellular regions and warrants further research despite not conclusively validating this hypothesis.

5.3. Implications for model selection and data handling. Our comparative analysis highlights that the choice of neural network architecture should be informed by the specific characteristics of the dataset and the objectives of the classification task. ViT models are particularly effective when the global context is essential, and when data preparation strategies enhance the diversity and representativeness of training samples. On the other hand, ConvNeXT models offer robustness and excel in scenarios where local feature extraction is paramount.

Based on our comparative analysis, we recommend the following guidelines for practitioners developing automated diagnostic systems:

- Vision Transformer models are particularly suited for scenarios where capturing global context is critical. Specifically, we recommend using ViT at lower magnifications (40×–100×) and with data preparation strategies involving randomly generated patches, particularly those focusing on cellular-rich regions (scenario E).

- ConvNeXT models offer robustness in scenarios emphasizing local textural features. They excel

particularly at medium to higher magnifications (100×–400×), especially when structured, non-overlapping patch extraction methods (scenario B) are employed. We emphasize, however, the importance of validating these recommendations using additional and diverse datasets beyond BreakHis to ensure generalizability, as clearly discussed in the limitations (Section 5.4).

5.4. Limitations and future work. While our study provides comprehensive insights, it is limited by the scope of architectures evaluated and the dataset used. We focused on two advanced architectures; however, exploring additional models, including hybrid architectures that combine CNNs and Transformers (Graham *et al*., 2021), could provide further improvements.

Additionally, the BreakHis dataset, although widely used, represents a specific set of imaging conditions and patient demographics. Future research should validate these findings on larger and more diverse datasets to enhance the generalizability of the results.

5.5. Conclusion. In conclusion, our study confirms the critical role of data preparation in histopathological image classification and highlights the strengths of ViT and ConvNeXT architectures in different contexts. By aligning our findings with existing literature, we contribute to the broader understanding of how advanced neural network architectures can be optimized for medical image analysis.

References

Ara, R. K., Matiolanski, A., Grega, M., Dziech, A. and Baran, R. (2023). Efficient face detection based crowd density estimation using convolutional neural networks and an improved sliding window strategy, *International Journal of Applied Mathematics and Computer Science* **33**(1): 7–20,k DOI: 10.34768/amcs-2023-0001.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S. and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, *Nature Medicine* **25**(8): 1301–1309, DOI: 10.1038/s41591-019-0508-1.

Deininger, L., Stimpel, B., Yuce, A., Abbasi-Sureshjani, S., Schönenberger, S., Ocampo, P., Korski, K. and Gaire, F. (2022). A comparative study between vision transformers and cnns in digital pathology, *arXiv:* 2206.00389 .

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale, *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Vienna, Austria*.

Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H. and Douze, M. (2021). LeViT: A vision transformer in ConvNET's clothing for faster inference, *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, Canada*, pp. 12259–12269, DOI: 10.1109/ICCV48922.2021.01204.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA,* pp. 770–778, DOI: 10.1109/CVPR.2016.90.

Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K.Q. (2017). Densely connected convolutional networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA,* pp. 4700–4708, DOI: 10.1109/CVPR.2017.243.

Komura, D. and Ishikawa, S. (2018). Machine learning methods for histopathological image analysis, *Computational and Structural Biotechnology Journal* **16**: 34–42, DOI: 10.1016/j.csbj.2018.01.001.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks, *in* F. Pereira *et al.* (Eds), *Proceedings of the 26th Annual Conference on Neural Information Processing Systems, NIPS 2012*, Curran Associates, Inc., Lake Tahoe, NV, USA, pp. 1106–1114.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. and Xie, S. (2022). A ConvNet for the 2020s, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA,* pp. 11966–11976, DOI: 10.1109/CVPR52688.2022.01167.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition, *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA. DOI: 10.48550/arXiv.1409.1556.

Spanhol, F.A., Oliveira, L.S., Petitjean, C. and Heutte, L. (2016). A dataset for breast cancer histopathological image classification, *IEEE Transactions on Biomedical Engineering* **63**(7): 1455–1462, DOI: 10.1109/TBME.2015.2496264.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M. and Hamamoto, R. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review, *Journal of Medical Systems* **48**(84): 1–22, DOI: 10.1007/s10916-024-02105-8.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Maksai, A., Ciompi, F. and van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational

pathology, *Medical Image Analysis* **58**: 101544, DOI: 10.1016/J.MEDIA.2019.101544.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W. and Han, X. (2021). Transpath: Transformer-based self-supervised learning for histopathological image classification, *in* M. de Bruijne *et al.* (Eds), *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2021*, Springer, Cham , pp. 186–195. DOI: 10.1007/978-3-030-87237-3_18.

Mikołaj Kaczmarek received his BS degree in technical physics from the Wrocław University of Science and Technology and his MSc degree in computer science from the University of Zielona Góra. Currently, he is pursuing his PhD in technical informatics and telecommunications at the University of Zielona Góra. Since 2018, he has been professionally involved in software engineering. His research interests include computer vision and deep learning.

Marek Kowal received his PhD in electrical engineering from the University of Zielona Góra, Poland, in 2004, and his DSc (habilitation) in computer science from the Częstochowa University of Technology in 2020. He is currently an associate professor at the Institute of Control and Computation Engineering at the University of Zielona Góra. His research interests focus on deep neural networks and self-supervised learning, particularly applied to multiple object tracking and medical image analysis.

Józef Korbicz has been a full professor of automatic control at the University of Zielona Góra, Poland, since 1994. In 2007 he was elected a corresponding member and 2020 an ordinary member of the Polish Academy of Sciences. His current research interests include fault detection and isolation, control theory and computational intelligence. He has published over 490 scientific publications, authored or co-authored 8 books and coedited 28 books. He is currently a senior member of IEEE, a member of the IFAC SAFEPROCESS TC and the Chair of the Committee on Computer Science and Automation of the Polish Academy of Sciences in Poznań.

Received: 10 January 2025
Revised: 21 March 2025
Accepted: 27 March 2025