Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 33, Num. 2 (2026) 260-268

RESEARCH ARTICLE

Assessing Deep Hybrid Representations for Colorectal Polyp Classification in Histopathology Images

Avaliação de Representações Profundas Híbridas para Classificação de Pólipos Colorretais em Imagens Histopatológicas

Luan B. Guerra[1], Ana Beatriz S. Zerati[1], Ana Catarina M. Vicentim[1], Lucas C. Ribas[1]*

Abstract: This study investigates the use of Convolutional Neural Networks (CNNs) and *Vision Transformers* (ViT) backbones for feature extraction from histopathological images, and further evaluates the effectiveness of combining features from different depths of both architectures. The backbones were employed to extract features from colorectal polyp images in the MHIST dataset. Features obtained from each model individually, across different deep layers, were also combined and subjected to a dimensionality reduction process. Subsequently, both the individual and combined feature sets, with and without dimensionality reduction, were evaluated using linear classification models. The combined features achieved an accuracy of 85.55%, surpassing all results obtained from individual models and other approaches. These findings show that features from the first and middle layers, and feature fusion enhances robustness by leveraging the complementary strengths of each architecture.
Keywords: feature extraction — computer vision — histopathological images — convolutional neural networks — vision transformers

Resumo: Este estudo investiga o uso de backbones de Redes Neurais Convolucionais (CNNs) e *Vision Transformers* (ViT) para extração de características de imagens histopatológicas e avalia ainda mais a eficácia da combinação de características de diferentes profundidades e de ambas as arquiteturas. Os backbones foram empregados para extrair características de imagens de pólipos colorretais no conjunto de dados MHIST. Características obtidas de cada modelo individualmente, em diferentes camadas profundas, também foram combinadas e submetidas a um processo de redução de dimensionalidade. Posteriormente, os conjuntos de características individuais e combinados, com e sem redução de dimensionalidade, foram avaliados usando modelos de classificação linear. As características combinadas alcançaram uma acurácia de 85,55%, superando todos os resultados obtidos de modelos individuais e outras abordagens. Esses resultados mostram que características das camadas iniciais, intermediárias e a fusão de características aumentam a robustez, alavancando os pontos fortes complementares de cada arquitetura.
Palavras-Chave: extração de características — visão computacional — imagens histopatológicas — redes Neurais artificiais — vision transformers

[1] *Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista (Unesp), Brazil*
***Corresponding author**: lucas.ribas@unesp.br
DOI: http://dx.doi.org/10.22456/2175-2745.150896 • **Received:** 13/10/2025 • **Accepted:** 27/10/2025

1. Introduction

Cancer is known to be a highly serious disease and becomes incurable even when treatment is started at the time of diagnosis, due to the occurrence of a high number of cancer cells. In this context, colorectal cancer is the third most common type of cancer in the world and the fourth leading cause of cancer deaths [1, 2]. Therefore, the early classification of colorectal polyps is a critical task that must be performed with high accuracy to prevent these polyps from progressing into malignant tumors. However, in some countries, the number of pathologists has declined over the years [3]. This shortage, combined with the fact that colorectal polyp treatment is among the most time-consuming and labor-intensive tasks in pathology [4], has made the workload increasingly burdensome.

AI models that perform image processing automatically have been gaining considerable interest from medical scientists in recent years [5]. This is mainly due to the emergence of deep networks, which have achieved impressive results in various tasks such as classification and segmentation [6].

More specifically, in the Computational Pathology (CPath), an interdisciplinary science focused on developing computational methods to analyze histopathological images in order to facilitate cancer diagnosis and treatment [7], deep neural networks dominate as base models for feature encoding tasks, with Convolutional Neural Networks (CNNs) being the most widely used. In addition, Vision Transformers (ViTs) [8] are also emerging as a growing trend in this field, demonstrating relative success in various tasks [7, 9].

CNNs perform strongly in tasks involving histopathological images because their convolutional operations provide an inductive bias well suited to the textures and local patterns characteristic of such images [7, 10]. However, these models have a limited receptive range, causing them to capture fewer long-range relationships [11]. As an alternative to CNNs, ViTs can model global dependencies via attention mechanisms, which is useful when the diagnostic signal depends on long-range spatial relationships between regions [12]. Nevertheless, ViTs also present certain limitations, notably their high computational cost and their reduced ability to capture low-level features compared to CNNs [12].

In addition, a common limitation of both models is their reliance on large datasets for effective training, which poses a major challenge in medical imaging applications—particularly in histopathology, where labeled data are scarce [12, 7]. As a result, many studies choose to use Self-Supervised Learning (SSL) methods, since labeling medical images tends to be expensive [12], or perform intensive fine-tuning on pre-trained models in order to improve model efficiency [11].

In this context, this work aims to evaluate the capability of pre-trained CNN and ViT models to serve as feature extractors for colorectal polyp classification in histopathology images. The representations obtained from each model are combined to leverage their strengths while mitigating their weaknesses within a single deep hybrid feature set. In addition, for CNNs, features from different layers are also analyzed, along with the impact of feature reduction techniques on their quality. Therefore, the contribution of this work lies in analyzing the ability of models pre-trained on general datasets to extract high-quality features from colorectal polyp images, as well as in evaluating how different layers of the CNN architecture influence feature extraction and which layer is best suited for this application. Finally, the study of hybrid features, combined with the use of feature reduction algorithms, enables an investigation of how convolutional and attention-based mechanisms complement each other, leading to more discriminative and compact representations for classification tasks in pathology.

2. Related Works

Several studies have explored the use of deep learning models of computer vision and image processing techniques in the classification task of histopatological images, including colorectal polyp images, which are a significant task in gastrointestinal pathology. Wei et al. [4] introduced the MHIST dataset, used in our work, as a compact benchmark for colorectal polyp classification. The authors showed that convolutional architectures (ResNets) are efficient baselines thanks to their inductive bias, which naturally captures local patterns and textures characteristic of tissue. Thus, this article demonstrated that convolutional networks remain a pragmatic and robust choice for histopathological image classification tasks.

Nergiz et al. [13] evaluate the ability of variants of the ConvNext model (Tiny, Small, Base, and Large), a modernized CNN, to classify colorectal polyps in the MHIST dataset, testing various scenarios such as “full data” and “k-shot.” The study showed that ConvNext architectures achieve state-of-the-art performance on the proposed task, in addition to having good generalization in few-shot scenarios (few example images) and in subsets with gradual difficulty.

Kang et al. [11] conducted a large-scale study comparing CNNs (ResNet) and ViTs architectures in histopathological image classification and segmentation tasks. Four self-supervised learning methods were used to perform a more robust comparison. The authors demonstrated that CNNs and ViTs achieve high performance on multiple benchmarks, including the MHIST dataset, when fine-tunned with histopatological images. The results reinforce that both model families are capable of learning effective discriminative representations for applications in digital pathology.

Wang et al. [12] presented a new method for classifying histopathological images called TransPath, which combines the Transformer architecture with CNN models. The method consists of using a CNN as a low-level feature extractor and then using these features as inputs to a Transformer backbone in order to capture long-range dependencies between regions of the image. When evaluated on multiple pathology benchmarks, TransPath demonstrated that combining CNNs and Transformers for feature extraction resulted in competitive or superior performance compared to purely convolutional models, showing that both architectures are effective in classifying histopathological images.

The state of the art demonstrates the growing importance of applying CNNs and ViTs in the task of classifying histopathological images, since these models have the ability to recognize patterns that differentiate malignant and benign polyps. Thus, this work proposes a hybrid method for feature extraction in histopathological images, where CNNs and ViTs will be employed as feature extractors and the obtained representations will be combined in order to utilize the best of both worlds.

3. Materials and Methods

3.1 Colorectal polyps dataset

The Minimalist Histopathology (MHIST) [4] hematoxylin and eosin (H&E)-stained image analysis dataset is composed of 3,152 fixed-size images with 224 by 224 pixels of colorectal polyps from the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (DHMC). The classification task of this dataset focuses on the binary

distinction between hyperplastic polyps (HPs) and sessile serrated adenomas (SSAs), which is a challenging problem because of the inter-pathologist high variability when comparing the two types of polyps [14, 15]. HPs are, in general, benign, while SSAs are lesions that can turn in cancer if left untreated, making early treatment necessary [4].

The dataset comprises 328 FFPE whole-slide images of colorectal polyps (HPs and SSAs) collected from patients at the Dartmouth-Hitchcock Medical Center. These images were scanned at 40x resolution using an Aperio T2 scanner [4]. Accordingly, image tiles of size 224×224 pixels were extracted from each of the 328 whole-slide images, resulting in a total of 3,152 tiles representing diagnostically relevant regions of interest for HPs and SSAs. These images were confirmed by the pathologists of the DHMC to be high-quality with few artifacts. Finally, the authors scrambled and anonymized all images by removing all personal data from patients.

Image labels (HP or SSA) were determined through the consensus of seven board-certified gastrointestinal pathologists at the Dartmouth-Hitchcock Medical Center. This way, each pathologist, independently from the others, classified each images of the MHIST as HP or SSA. After the classification of all images by the pathologists, the labels were collected and the gold standard label for each image was set based on the majority vote of seven board-certified pathologists, resulting in 2,162 images with the HP label and 990 images with the label SSA. Some samples of each class can be seen in the Figure 1.

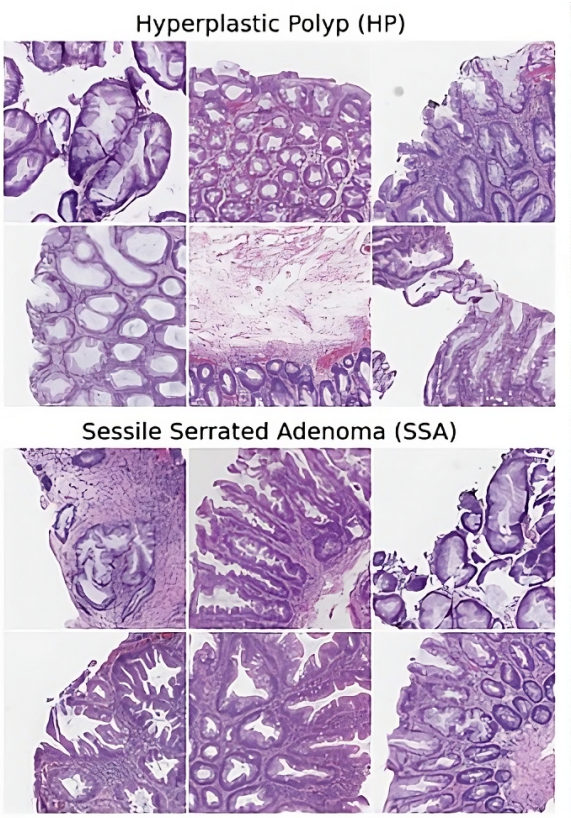


Figure 1. Some samples of the MHIST dataset [4]. The first two rows are samples of the Hyperplastic Polyp (HP), which are benign in general, and the two last rows represents images from a Sessile Serrated Adenoma (SSA), a precancerous lesion.

3.2 Selected Backbones

In this work, we choose two families of CNN architectures, namely ConvNext [16] and ResNet [17], and the original family of the ViTs [8] for feature extraction of the MHIST dataset. All the models were pre-trained using the IN-21k [18] dataset and the source code of each one is freely accessible [1]. All of these models belong to the pytorch timm library, so they are well known in the literature and have already been extensively tested by the Deep Learning community. The general details of each backbone can be found in the Table 1 and a more detailed description of the models families is right below:

Table 1. Details of each backbone explored in this work. The ViT models do not have values for the size of the first and intermediary layers features because, in this case, we extract only the features of the last layer.

backbone	param. (M)	Layers features size		
		First	Mid	Last
ConvNext-S [16]	50	96	384	768
ConvNext-B [16]	88	128	512	1024
ConvNext-L [16]	197.8	192	768	1536
ResNet-50 [17]	25.6	64	512	2048
ResNet-101 [17]	44.5	64	512	2048
ResNet-152 [17]	60.2	64	512	2048
ViT-S/16 [8]	22	-	-	382
ViT-B/16 [8]	85.8	-	-	768
ViT-L/16 [8]	304.2	-	-	1280

1. **ResNet** [17]: Very deep networks suffer from vanishing gradients and accuracy degradation when stacking many layers, so the solution was to introduce shortcuts (skip connections) that "skip" one or more layers, allowing the gradient to flow directly. Thus, ResNets are CNNs with residual blocks, which are a set of layers where the processed features are added to the features from the skip connections, preventing gradient degradation. In general, ResNets have a 7x7 pixels convolutional filter, a stride of 2 and are made up of 4 block stages. The main benefits of this model are the possibility of training very deep networks without gradient degradation and the fact that its blocks are simple and modular, providing a basis for various subsequent works.

2. **ConvNext** [16]: In order to modernize CNN architecture, the authors of [16] adapted lessons from Vision Transformers to CNNs, maintaining convolution efficiency but using contemporary design practices. Thus,

[1] https://github.com/huggingface/pytorch-image-models

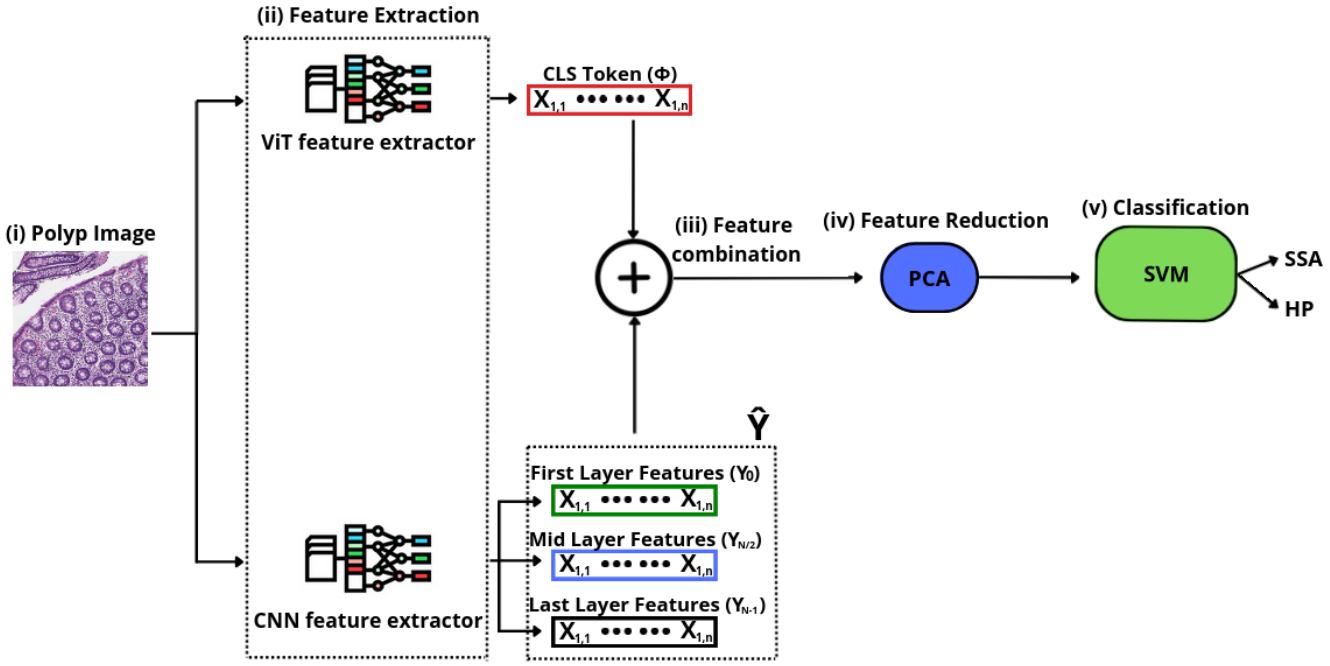


Figure 2. Illustration of the methodology of this work that led to the best results. The features of each model are extracted and saved individually and then used to perform the combinations.

in addition to the 7x7 pixel convolutional filter used in ResNets, ConvNexts also use a 4x4 pixel convolutional filter with a stride of 4 in order to "patch" the image. In addition, these models also have 4 stage blocks, just like ResNets. From this, ConvNexts have the ability to capture local contexts while mimicking ViTs in attempting to capture global contexts from images.

3. **ViT** [8]: Derived directly from the original Transformers [19], the Vision Transformer adapts the transformer architecture to process images. Unlike the local convolutions of CNNs, ViTs use attention mechanisms to extract global patterns from the image. In conjunction with this mechanism, these models use patch division, dividing the image into small sub-images and using these as tokens. In this way, these models are able to understand the relationship between each part of the image and, consequently, capture long-range dependencies of the images.

It is important to note that the variations (S, B, L) of each model family share the same fundamental architecture, with the only difference being their size. In general, the main variations are in the models depth, with an increase in the number of layers, and in the feature vectors dimension, where larger models produce larger vectors, as can be seen in the Table 1.

3.3 Feature Extraction

3.3.1 Single-Model Features

To employ the selected backbones as feature extractors, we removed the classification layer and froze the pre-trained parameters. The general feature extraction process is illustrated in Figure 2. The images were first normalized and then used as inputs to the models. In the case of CNNs, feature extraction was explored across different layers.

So we considered the features of the first, mid, and last layers, where we basically performed a Global Average Pooling (GAP) of the output of these layers and saved these results as our features: We considered the features from the first Υ_0, middle $\Upsilon_{N/2}$, and last Υ_{N-1} layers, applying Global Average Pooling (GAP) to the outputs of these layers, resulting in the vectors:

$$\Upsilon_0 = GAP(output_0), \tag{1}$$

$$\Upsilon_{N/2} = GAP(output_{N/2}), \tag{2}$$

$$\Upsilon_{N-1} = GAP(output_{N-1}), \tag{3}$$

where $output_i$ represents the output of the i-th layer of the CNN, Υ_i represents the features of the i-th layer and N represents the total number of layers. Thus, the features from each CNN model can be combined as:

$$\hat{\Upsilon} = [\Upsilon_0, \Upsilon_{N/2}, \Upsilon_{N-1}], \tag{4}$$

where $\hat{\Upsilon}$ represents all the features from different layers.

For the ViTs, the CLS token was extracted from the last layer of the model. This layer produces a matrix $X \in \mathbb{R}^{(n+1)\times d}$, where n is the number of patches and d is the hidden dimension of the model. The CLS token corresponds to the first row of X, which is considered the features for the ViTs:

$$\phi = X[0][0,...,d-1]. \tag{5}$$

The process of feature extraction for each model is illustrated by step (ii) in the Figure 2.

3.3.2 Combined features

After extracting features from each model separately, we investigated whether combining CNN and ViT features could provide complementary information about image patterns. Thus, we performed all possible pairwise combinations between the selected ViT and CNN models. For each ViT–CNN pair, the feature vector ϕ from the ViT was concatenated horizontally with all elements of $\hat{\Upsilon}$, as follows:

$$\lambda_0 = [\phi; \Upsilon_0], \tag{6}$$

$$\lambda_{N/2} = [\phi; \Upsilon_{N/2}], \tag{7}$$

$$\lambda_{N-1} = [\phi; \Upsilon_{N-1}], \tag{8}$$

where the symbol [;] denotes concatenation, and λ_i represents the combination of the ViT CLS token with the feature from the i-th CNN layer. This process is illustrated in step (iii) of Figure 2. In this manner, we evaluate how features from different layers influence the final result and whether they provide additional information about the images.

3.4 Feature Reduction and Classification

Feature reduction consists of decreasing the dimension of a feature set while attempting to retain as much information as possible from the original set. This way, we remove redundancies and noise from our features, while reducing the computational cost. In this work, we also explore the application of this reduction on the combined features (step (iv) from Figure 2) in order to improve the results by removing the redundancies produced by the combination of models. Thus, we use Principal Component Analysis (PCA) [20], a linear dimensionality reduction technique that transforms data into a new coordinate system based on variance, to perform feature reduction. We applied Principal Component Analysis (PCA) [20] to reduce feature dimensionality by projecting the data onto variance-defined principal components. Therefore, after extracting the combined features λ_i as explained in section 3.3.2, we reduced the dimension of these features by dividing their sizes by 2, 4, 8, 16, and 32 using PCA. It is important to note that each size division generated a new feature that was saved separately.

After extracting all features, we used the training features to train the linear classifier Support Vector Machine (SVM) [21] and then used the trained model to classify the test features, as indicated by step (v) in the Figure 2. By using a linear model, we were able to verify the quality of the features extracted by the backbones, and much less training data is required compared to deep learning models. The SVM is configured with a linear kernel and a penalty parameter C = 1.

4. Results and Discussion

4.1 Backbone deep features analysis

It is interesting to analyze the results of the features extracted from each model individually, in order to verify, in the future, whether there is an improvement in the results of the combined features. Thus, the accuracies obtained from each individual feature extracted by the models can be observed in the Table 2.

Table 2. Accuracies considering each feature extracted by the backbones alone. In the case of ViTs, we have only the results from the last layer, since we only extract the features from that layer.

backbone	Layers features accuracies (%)		
	First	Mid	Last
ConvNext-S [16]	65.98	81.97	80.23
ConvNext-B [16]	64.74	81.86	79.71
ConvNext-L [16]	66.5	80.5	79.0
ResNet-50 [17]	63.11	76.74	72.75
ResNet-101 [17]	63.11	74.59	76.23
ResNet-152 [17]	63.11	76.95	75.41
ViT-S/16 [8]	-	-	81.25
ViT-B/16 [8]	-	-	81.05
ViT-L/16 [8]	-	-	79.3

The ViTs presented results comparable to those of CNNs, with ViT-S showing the third highest accuracy (81.25%), with only a 0.72% difference when compared to the highest accuracy (81.97%). This shows that the CLS token from the last layer of ViTs has great potential to be used as a feature, even though it mostly contains global image patterns, in contrast to the features generated by CNN convolutions, which emphasize local patterns.

ConvNexts obtained better results overall when compared to CNNs, which can be explained by the fact that ConvNexts are newer models and, therefore, their development was done using more advanced technology, which gives them an advantage in feature extraction.

The first layer features did not obtain such satisfactory results, since the accuracies were 9% to 17% lower compared to the features of other layers. This occurs because, in addition to the fact that the MHIST dataset presents a complex classification problem, the features of the first layers lack precise

patterns and, therefore, do not have the same robustness as the features of later layers.

The best results are found when using the features of mid and last layers, as these are more robust features because they have undergone more processing. In any case, the features of the mid layers stand out, with accuracies up to 4% higher than the features of the last layers. Based on this, in the context of the MHIST dataset application, for the CNNs used, we find that the mid layer features are of the highest quality and that the ViTs CLS token also has a large number of important patterns for colorectal polyp classification.

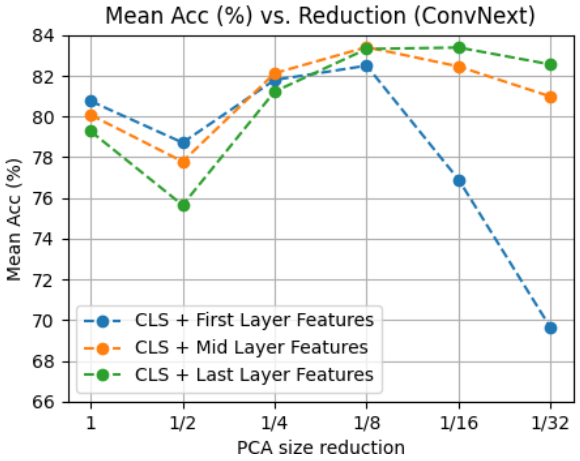
4.2 Reduction analysis

After analyzing the models individually, it is necessary to analyze the results of feature combinations of these models and how the reductions in features impact the final result. Thus, we calculate the average of the combinations results without reduction and with the reductions defined in section 3.4. We performed this process considering the features of different layers of the CNNs separately. The results of these calculations can be seen in Figures 3a and 3b for the ConvNexts and ResNets models, respectively.

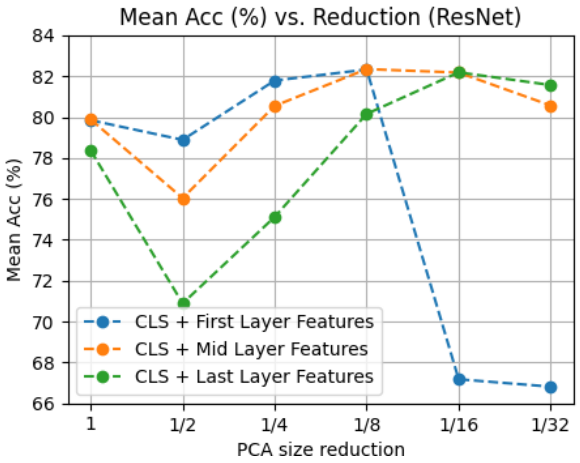
We separated the results of ConvNext and ResNet models to enable a more robust comparison between the two model families. When performing the combinations, it is possible to verify that the features of ConvNexts add more information in the combinations, since, in general, the average accuracies of the ConvNext-ViT pairs were better than those of the ResNet-ViT pairs. This is because, as verified in section 4.1, the features of ConvNexts, regardless of the layer, obtained better results, that is, these features are more robust than those of ResNets, which is to be expected, since ConvNexts are more recent CNN models.

Despite the difference in values, the results of the two families behave similarly in relation to feature reduction. The results decrease when we halve the size of the combined features, but they start to increase again as we reduce their sizes even further. Furthermore, after a drastic reduction, the results start to fall again. This occurs because, as features are reduced to smaller sizes, there is an increasing removal of noise and redundancies, causing these features to retain only the most important data, which consequently increases their quality and improves the performance of linear model training. However, at a certain point, if the reduction is too intense, the reduction algorithm may end up sacrificing important information, causing a drop in the overall quality of the feature.

Regarding the features of different layers of CNNs, we find that a pattern remains independent of reductions, where combinations with features from earlier layers obtain worse results than combinations with features from deeper layers as reductions are made. In this case, the combination of CLS + Last Layer features with a reduction of 1/16 presents the best average accuracy (83.49%). However, we decided to choose the combinations of CLS + Mid Layer features with a reduction of 1/8 to perform a more complete analysis of



(a) Average of the ConvNext-ViT pairs for each type of CNN features.



(b) Average of the ResNet-ViT pairs for each type of CNN features.

Figure 3. Relationship between the average accuracy of the combined features and the reduction in the size of these features using the PCA algorithm. Each line represents the combination of the CLS token with the features of a specific layer of the CNNs.

the combinations, since, although the average accuracy was not the best (83.48%), the difference between the two results is minimal and the 1/8 reduction may contain more valuable information than a 1/16 reduction.

4.3 Combined features analysis

Based on what was discussed in the previous subsection, we chose the CLS + Mid Layer features from ConvNexts with PCA reduction = 1/8 configuration to perform a more indepth analysis of the feature combinations. The results from each ViT-ConvNext combinated features can be seen in the Figure 4.

It is clear that as the size of the feature vector increases, accuracy also tends to increase. In any case, it can be observed that the combination of CNN and ViT features improves the final result, since, except for the result of the pair (ConvNext-

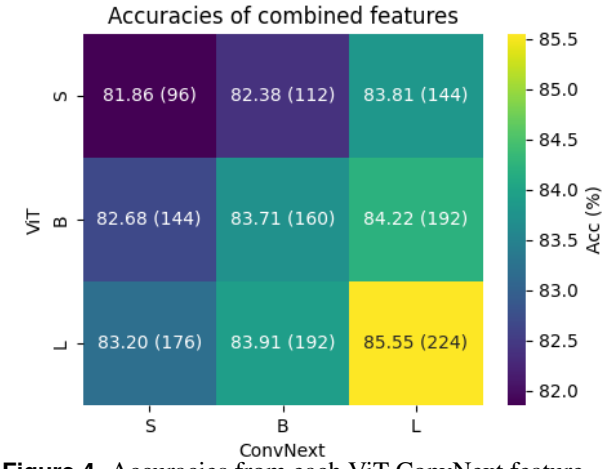


Figure 4. Accuracies from each ViT-ConvNext feature combination. In parentheses is the size of the combined feature vector.

Table 3. Comparison of accuracies between other literature methods for the task of classification of colorectal polyps on the MHIST dataset. The first four methods are Non-Deep Learning, while all the others use CNN or ViT models. Results marked with a dash (–) were not reported in the original paper. The last three rows represents the features proposed in this work. Bold denotes the best results.

Method	Accuracy	Precision	Recall	F1-score
LCP integrative [22]	68.4	65.7	62.7	63.0
LETRIST (gray-level) [23]	68.5	65.7	63.6	64.0
SSR [24]	68.6	66.1	65.7	65.9
VCTex [25]	73.4	71.4	70.6	70.9
ResNet-18 [17]	73.0	70.9	70.3	70.6
ResNet-20 [26]	76.0	-	-	-
ResNet-34 [4]	85.1	-	-	-
ResNet-50 (BT) [11]	78.81	-	-	-
VGG11 [27]	73.0	70.9	70.3	70.6
DenseNet121 [28]	72.5	70.5	70.6	70.5
ConvNext-S (Mid Layer)	81.97	81.0	79.0	80.0
ViT-S	81.25	79.0	77.0	78.0
(ConvNext-L, ViT-L)	**85.55**	**83.0**	**81.0**	**82.0**

S, ViT-S), all results were equal to or superior to the results seen in table 2, with the pair (ConvNext-L, ViT-L) obtaining the best result (85.55%).

Furthermore, it is noteworthy that, due to the application of reduction, these features are smaller in size than the features of the models alone. Even so, the reduced features achieve better results, highlighting an advantage of applying reduction. Based on this, we chose the result of the pair (ConvNext-L, ViT-L) to compare our method with other methods in the literature.

4.4 Comparison

Finally, we conducted a brief comparison between our method and other approaches reported in the literature. Table 3 presents the results of different methods on the MHIST classification task, along with our own results. The first four methods in Table 3 are non–deep learning approaches, while the remaining ones employ deep neural networks. The last three entries correspond to the features proposed in this work.

The features proposed in this study achieved overall better results compared to other methods reported in the literature. In general, deep learning (DL) approaches outperformed non–deep learning methods, demonstrating the strong capability and effectiveness of DL techniques for colorectal polyp classification. Our study outperformed approaches that rely solely on DL models, demonstrating that combining CNN and ViT features yields superior results by leveraging the strengths of each architecture.

Finally, our features present an effective way to deal with the classification problem. Our work demonstrates that the combination of ViTs and CNNs features can complement each other when combined and that feature reduction is an important step, as it decreases the dimension of these features, i.e., it decreases the classification computational cost and improves the overall results. These findings suggest that our method could be further explored for other histological image problems, offering a simple and cost-effective solution.

5. Conclusion

In this work, we investigated the ability of CNN and ViT models to extract features from colorectal polyp images, while also analyzing the effectiveness of combining features from both models and evaluating how feature reduction algorithms influence the quality of these combinations. In addition, we analyzed the features extracted from different layers of the CNNs.

The results demonstrated the strong potential of deep learning computer vision models for feature extraction in the MHIST dataset, while also highlighting the importance of considering features from different CNN layers. In addition, the results demonstrated that combining features leverages the strengths of both models, producing more robust representations of colorectal polyps. They also highlighted the importance of exploring feature reduction, which enhances feature quality while reducing the computational cost of classification.

For future work, it would be valuable to explore additional CNN and ViT architectures, as well as alternative feature reduction algorithms. Further investigation could also consider combining other ViT representations, such as Global Average Pooling (GAP) of spatial tokens, and analyzing features from different ViT layers. Finally, the feature fusion strategy proposed in this study could be extended to other types of histopathological images.

Acknowledgements

The authors acknowledge support from the São Paulo Research Foundation (FAPESP), Brazil. Process N., #2024/07241-8, #2025/01057-3, and #2023/04583-2. This study was also

financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

Contribuição dos Autores

Luan B. Guerra was responsible for conceptualization, methodology, software, investigation, validation, writing (original draft), and writing (review and editing); Ana Beatriz S. Zerati contributed to methodology, investigation, validation, writing (original draft), and writing (review and editing); Ana Catarina Vicentim contributed to methodology, investigation, validation, writing (original draft), and writing (review and editing); Lucas C. Ribas contributed to conceptualization, methodology, validation, resources, writing (original draft), writing (review and editing), supervision, project administration, and funding acquisition.

References

[1] FAMITHA, S.; MOORTHI, M. Intelligent and novel multi-type cancer prediction model using optimized ensemble learning. *Computer Methods in Biomechanics and Biomedical Engineering*, Taylor & Francis, 2022.

[2] LABIANCA, R. et al. Colorectal carcinoma: A general overview and future perspectives in colorectal cancer. *International Journal of Molecular Sciences*, v. 18, n. 1, p. 197, 2017.

[3] METTER, D. M. et al. Trends in the us and canadian pathologist workforces from 2007 to 2017. *JAMA Network Open*, v. 2, n. 5, p. e194337, 2019.

[4] WEI, J. et al. A petri dish for histopathology image analysis. In: SPRINGER. *International Conference on Artificial Intelligence in Medicine*. [S.l.], 2021. p. 11–24.

[5] FICHTINGER, G.; RUECKERT, D.; ZHOU, S. K. (Ed.). *Handbook of Medical Image Computing and Computer Assisted Intervention*. London, United Kingdom; Boston, Massachusetts, USA: Academic Press, Elsevier, 2022. ISBN 9781787859289.

[6] LI, J. et al. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, v. 85, p. 102762, 2023.

[7] HOSSEINI, M. S. et al. Computational pathology: A survey review and the way forward. *Journal of Pathology Informatics*, v. 15, p. 100357, Jan 2024.

[8] DOSOVITSKIY, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. Disponível em: ⟨https://arxiv.org/abs/2010.11929⟩.

[9] SCABINI, L. et al. A comparative survey of vision transformers for feature extraction in texture analysis. *Journal of Imaging*, MDPI, v. 11, n. 9, p. 304, 2025.

[10] RIBAS, L. C.; CASACA, W.; FARES, R. T. Conditional generative adversarial networks and deep learning data augmentation: A multi-perspective data-driven survey across multiple application fields and classification architectures. *AI*, MDPI, v. 6, n. 2, p. 32, 2025.

[11] KANG, M. et al. Benchmarking self-supervised learning on diverse pathology datasets. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2023. p. 3344–3354.

[12] WANG, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, v. 79, p. 102430, 2022.

[13] NERGIZ, M. Classification of precancerous colorectal lesions via convnext on histopathological images. *Balkan Journal of Electrical & Computer Engineering*, v. 11, n. 2, p. 129–135, April 2023. Disponível em: ⟨https://dergipark.org.tr/en/pub/bajece/issue/77274/1240284⟩.

[14] FARRIS, A. et al. Sessile serrated adenoma: challenging discrimination from other serrated colonic polyps. *The American Journal of Surgical Pathology*, v. 32, n. 1, p. 30–35, 2008.

[15] KHALID, O. et al. Reinterpretation of histology of proximal colon polyps called hyperplastic in 2001. *World Journal of Gastroenterology*, v. 15, n. 30, p. 3767–3770, 2009.

[16] LIU, Z. et al. A convnet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 11976–11986.

[17] HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778.

[18] RIDNIK, T. et al. *ImageNet-21K Pretraining for the Masses*. 2021. Disponível em: ⟨https://arxiv.org/abs/2104.10972⟩.

[19] VASWANI, A. et al. *Attention Is All You Need*. 2023. Disponível em: ⟨https://arxiv.org/abs/1706.03762⟩.

[20] JOLLIFFE, I. T. *Principal Component Analysis*. 2nd. ed. [S.l.]: Springer, 2002. (Springer Series in Statistics). ISBN 978-0387954424.

[21] CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995.

[22] GUO, Y.; ZHAO, G.; PIETIKÄINEN, M. Texture classification using a linear configuration model based descriptor. In: CITESEER. *BMVC*. [S.l.], 2011. p. 1–10.

[23] SONG, T. et al. Letrist: Locally encoded transform feature histogram for rotation-invariant texture classification. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 28, n. 7, p. 1565–1579, 2018.

[24] RIBAS, L. C. et al. Color-texture classification based on spatio-spectral complex network representations. *Physica A: Statistical Mechanics and its Applications*, p. 129518, 2024. ISSN 0378-4371.

[25] FARES, R. T.; RIBAS, L. C. Volumetric color-texture representation for colorectal polyp classification in histopathology images. In: *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - VISAPP (Vol. 3)*. [S.l.]: SciTePress, 2025. p. 210–221.

[26] ALALI, M. H. et al. Enabling intelligent iots for histopathology image analysis using convolutional neural networks. *Micromachines*, v. 13, n. 8, p. 1364, 2022. Disponível em: ⟨https://www.mdpi.com/2072-666X/13/8/1364⟩.

[27] SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. 2014.

[28] HUANG, G. et al. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 4700–4708.