

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Prezja, Fabi; Annala, Leevi; Kiiskinen, Sampsa; Lahtinen, Suvi; Ojala, Timo; Ruusuvuori, Pekka; Kuopio, Teijo

Title: Improving Performance in Colorectal Cancer Histology Decomposition using Deep and Ensemble Machine Learning

Year: 2024

Version: Published version

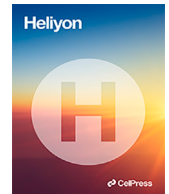
Copyright: © 2024 The Author(s). Published by Elsevier Ltd

Rights: CC BY 4.0

Rights url: <https://creativecommons.org/licenses/by/4.0/>

Please cite the original version:

Prezja, F., Annala, L., Kiiskinen, S., Lahtinen, S., Ojala, T., Ruusuvuori, P., & Kuopio, T. (2024). Improving Performance in Colorectal Cancer Histology Decomposition using Deep and Ensemble Machine Learning. *Heliyon*, 10(18), Article e37561.
<https://doi.org/10.1016/j.heliyon.2024.e37561>



Research article

Improving performance in colorectal cancer histology decomposition using deep and ensemble machine learning

Fabi Prezja^{a,*}, Leevi Annala^{b,c}, Sampsa Kiiskinen^a, Suvi Lahtinen^{a,d}, Timo Ojala^a, Pekka Ruusuvoori^{e,f}, Teijo Kuopio^{g,h}^a University of Jyväskylä, Faculty of Information Technology, Jyväskylä, 40014, Finland^b University of Helsinki, Faculty of Science, Department of Computer Science, Helsinki, Finland^c University of Helsinki, Faculty of Agriculture and Forestry, Department of Food and Nutrition, Helsinki, Finland^d University of Jyväskylä, Faculty of Mathematics and Science, Department of Biological and Environmental Science, Jyväskylä, 40014, Finland^e University of Turku, Institute of Biomedicine, Cancer Research Unit, Turku, 20014, Finland^f Turku University Hospital, FICAN West Cancer Centre, Turku, 20521, Finland^g University of Jyväskylä, Department of Biological and Environmental Science, Jyväskylä, 40014, Finland^h Hospital Nova of Central Finland, Department of Pathology, Jyväskylä, 40620, Finland

ARTICLE INFO

Dataset link: https://drive.google.com/drive/folders/1ypFyU2V6ifRkLB6hRkIqb6C2D_fvL_5Y?usp=sharing

Keywords:

Deep learning

CRC

Histopathology

Biomarkers

Hybrid model

Vision Transformer

Neural Architecture Search

Convolutional Neural Networks

ABSTRACT

In routine colorectal cancer management, histologic samples stained with hematoxylin and eosin are commonly used. Nonetheless, their potential for defining objective biomarkers for patient stratification and treatment selection is still being explored. The current gold standard relies on expensive and time-consuming genetic tests. However, recent research highlights the potential of convolutional neural networks (CNNs) to facilitate the extraction of clinically relevant biomarkers from these readily available images. These CNN-based biomarkers can predict patient outcomes comparably to golden standards, with the added advantages of speed, automation, and minimal cost. The predictive potential of CNN-based biomarkers fundamentally relies on the ability of CNNs to accurately classify diverse tissue types from whole slide microscope images. Consequently, enhancing the accuracy of tissue class decomposition is critical to amplifying the prognostic potential of imaging-based biomarkers. This study introduces a hybrid deep transfer learning and ensemble machine learning model that improves upon previous approaches, including a transformer and neural architecture search baseline for this task. We employed a pairing of the EfficientNetV2 architecture with a random forest classification head. Our model achieved 96.74% accuracy (95% CI: 96.3%-97.1%) on the external test set and 99.89% on the internal test set. Recognizing the potential of these models in the task, we have made them publicly available.

1. Introduction

Cancer comprises diseases marked by rapid, uncontrolled growth of abnormal cells that form malignant tumors. These cells can detach, spread, and form new tumors in distant body parts, a process known as metastasis, which is the primary cause of cancer-related deaths [1]. The World Health Organization reports that cancer is a leading global cause of death, responsible for one in six deaths [2]. The most common areas for cancer to initially develop are the breast, lung, colon, and prostate.

* Corresponding author.

E-mail addresses: fabif.prezja@jyu.fi, faprezja@jyu.fi (F. Prezja).

Colorectal Cancer (CRC) is the third most common yet second deadliest cancer [3]. American Cancer Society data indicates 56% of patients are diagnosed at stages where cancer has begun to metastasize [4,5]. Early detection and treatment are paramount [6]. Machine vision advancements, especially through deep neural networks [7], have improved automatic cancer and other disease classification [8–14]. Despite the technological progress, healthcare professionals still need to examine histologic samples to confirm diagnoses and assess tumor stages. Hematoxylin and Eosin (HE) staining is typically used to highlight key histopathological features in these tissues [15,16].

CRC patients are categorized into different groups to tailor their treatment and surveillance strategies. Grouping relies on multiple factors, including clinical outcomes, tumor genetics, quantitative biomarkers, clinical data, and histopathological and molecular analyses of the tumor. Many biomarkers stem from molecular and genetic tests [17–20]. Recent insights into tumor immunology have revealed the critical role of the tumor microenvironment in tumor growth. Therefore, discovering novel predictive and prognostic biomarkers that effectively identify tumor characteristics is crucial.

In recent times, the first quantitative biomarkers based on deep learning have been extracted from HE stained whole-slide images [10,8,21–24]. A clinically relevant biomarker derived from Convolutional Neural Networks is a quantifiable indicator extracted from medical images (like histology slides) using deep learning techniques. This biomarker can independently predict clinical outcomes, such as survival rates or disease progression, and may provide insights into the biological processes of a disease. Kather and colleagues [8] were the first to use deep learning to identify a biomarker for stages III and IV of CRC. This novel biomarker exhibited performance comparable to the existing gold standards for determining CRC outcomes [25,26] and could be automatically generated from images, saving time and resources. Conversely, known biomarkers (MSI, BRAF and KRAS) were also predicted [27] with deep learning transformers [28]. The approach greatly improved current methods for detecting microsatellite instability in surgical samples and achieved clinical-level accuracy in colorectal cancer biopsies, a significant finding in the field [29].

In their seminal research, Kather et al. [8] applied convolutional neural networks (CNNs) [30] to identify nine distinct tissue classes from HE-stained whole-slide images. Their methodology led to a noteworthy classification accuracy of 94.3% on their external testing data. They then compiled output layer neuron activations into a single weighted score, named ‘Deep Stroma’, and tested this new CNN-biomarker for outcome prediction in new patient cohorts. They discovered that the ‘Deep Stroma’ score was a significant prognostic factor, especially in patients with advanced tumor stages (UICC IV). The CNN-biomarker was significantly prognostic in all tumor stages, while manual pathologist annotations and cancer-associated fibroblast (CAF) scores were not. Kather and his team prioritized the model with the highest accuracy, as it directly enhanced the quality and applicability of the new prognostic CNN-biomarker.

Subsequent studies [31–38] have made notable strides in improving accuracy. Nevertheless, some have either reported results not improving upon the original Kather, et al. [8] model or faced challenges with incompatible output layer specifications and different validation methodologies. In our prior study [39], we refined the model design originally proposed by Kather et al. [8], achieving improved results to those previously reported. Building on that foundation, the current study introduces a new model that improves classification accuracy for this task and further builds upon previous solutions. Utilizing the EfficientNetV2 CNN architecture [40] which has demonstrated notable capabilities in microscope image analysis [41–44] and in conjunction with the random forest ensemble algorithm [45], we introduce a hybrid deep and ensemble model.

2. Materials and methods

In this Methodology section, we detail our research process, outlining our data handling, image augmentation techniques, and the use of Convolutional Neural Networks. As shown in Fig. 1 we focused on EfficientNetV2 for image classification, supplemented by the Random Forest method to generate the final deep ensemble model for refined predictions. These models were further evaluated against Transformer [46] and Auto-Keras baselines. The process began with the acquisition of HE-stained colorectal cancer data as the first step. In the second step, EfficientNetV2 models were trained for tissue classification. Once the best-performing EfficientNetV2 model was identified, it was frozen in the third step, and a new random forest classifier was trained using the learned features. The fourth step involved evaluating the hybrid model using both test and external datasets. Additionally, results from this model were compared with those obtained from Autokeras NAS and ViT transformers.

2.1. Data acquisition and pre-processing

The data utilized in this study were collected by the National Center for Tumor Diseases (NCT) in Heidelberg, Germany, and the University Medical Center Mannheim (UMM) in Mannheim, Germany. This comprehensive dataset has been made publicly available by Kather et al. [8] and consists of 100,000 non-overlapping image tiles derived from 86 HE-stained tissue slides [47]. The image tiles, each measuring 224 x 224 pixels, were normalized using the Macenko method [48]. These images span nine distinct classes: 1) adipose tissue (ADI); 2) background (BACK); 3) debris (DEB); 4) lymphocyte (LYM); 5) mucus (MUC); 6) smooth muscle (MUS); 7) normal colon mucosa (NORM); 8) cancer-associated stroma (STR); 9) CRC Epithelium (TUM). A detailed breakdown of the class distribution can be found in Fig. 2, and representative images for each class are displayed in Fig. 3.

The dataset was partitioned into a training, validation, and testing set containing 69996, 14995, and 15009 images, respectively. This distribution corresponded to 70% of the original data for training, 15% for validation, and 15% for testing. We also incorporated the external testing set from the original work by Kather et al. [8], consisting of 25 CRC HE slides from the NCT biobank, providing an additional 7180 image patches, code named (CRC-VAL-HE-7 K) [47]. Fig. 3 visualizes the number of images in each class for the training and external testing data.

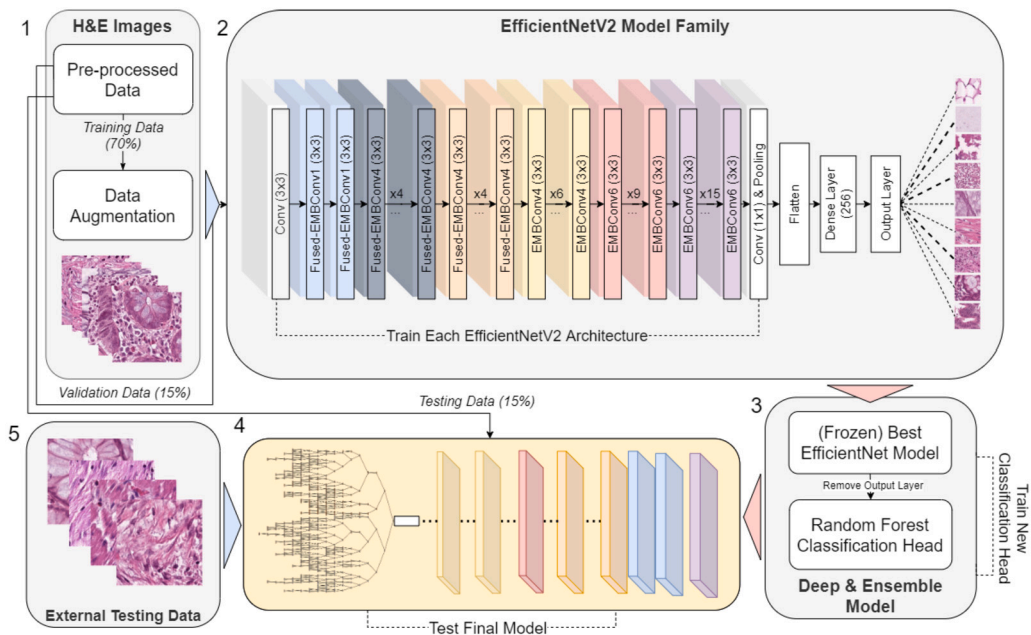


Fig. 1. Methodological Pipeline for Model Development. Blue arrows indicate data flow, red arrows represent model transfer, and numerical markers show the operational sequence.

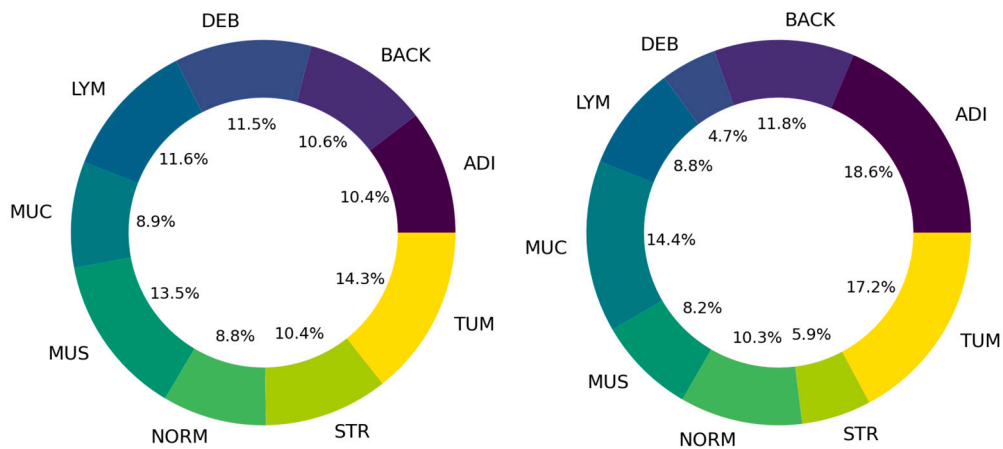


Fig. 2. Fig. 2 illustrates the distribution of data from the source [8]. On the left, the inner ring displays the percentage of data allocated for training, along with class labels on the outer ring. The right side shows a similar breakdown for the external-testing data.

2.2. Data augmentation

Data augmentation, a technique to artificially enhance the diversity of training images [49], is applied by randomly transforming some images before they are fed into the training. Simple examples include random image rotation or shifting. When using multiple augmentations, the methods are combined to increase the possible variations. In this study’s context, we employed six data augmentation methods bundled in the ‘advanced’ preset of the Deep Fast Vision Repository [50]. Each method and its specific configuration is detailed in Table 1. The same configuration was used for the training of the ViT transformer [46] from the ViT Keras [51] Library. Augmentation techniques were applied randomly to different images and in random combinations, in real-time as images were batched for training (with replacement).

2.3. Convolutional neural networks

One of the bedrocks of the recent surge in deep learning is the Convolutional Neural Network (CNN) [30]. A particular type of neural network, CNNs are often deployed for tasks in computer vision. They utilize the operation of convolution between an input and a filter-kernel. The kernels or filters are moved over the inputs to create feature maps, which represent highlighted features of the input. Different feature maps can be aggregated to form higher-level feature maps corresponding to more complex concepts. In a

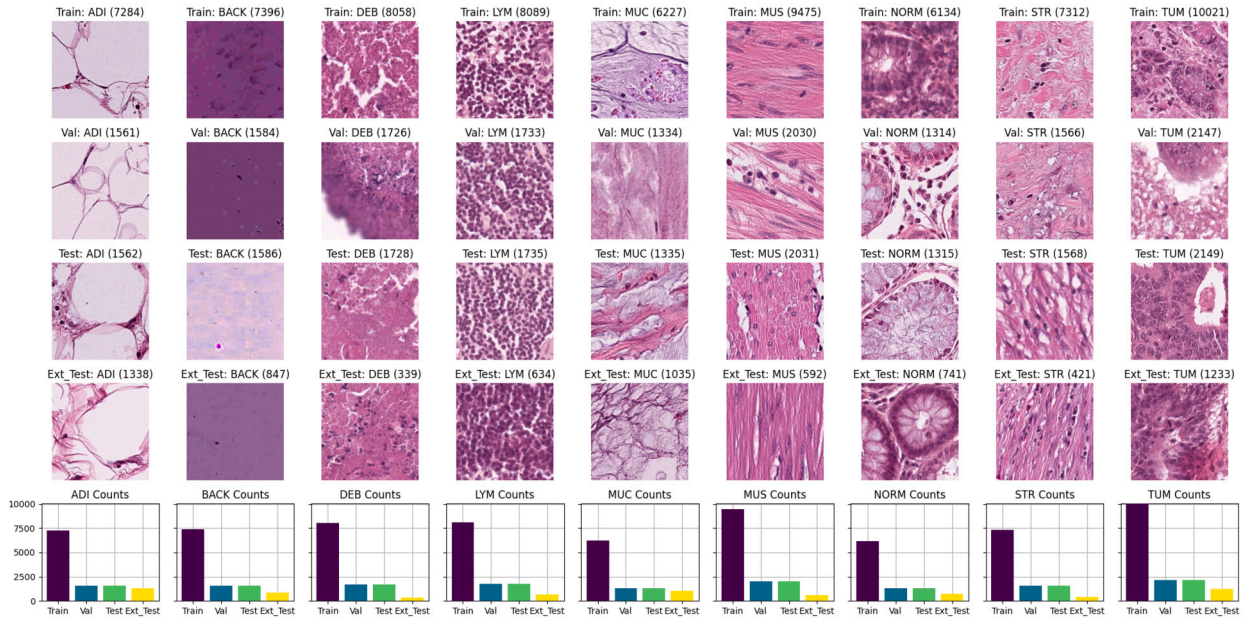


Fig. 3. Tissue sample tiles and distribution across dataset partitions for nine tissue types. Each row represents a different dataset partition (Training, Validation, Test, and External Test), denoted as Train, Val, Test, and Ext. Test, respectively. Below each tissue type column are bar graphs displaying the count of images in each dataset partition.

Table 1

A detailed description of the various image augmentation techniques utilized during the training of our neural networks.

Augmentation Technique	Technique Explanation	Training Settings
Rotation	Rotates the image in the plane.	Allows image rotation up to a maximum of 40 degrees.
Width Shift	Shifts the image horizontally.	Allows a shift of up to +- 45 pixels along the width.
Height Shift	Shifts the image vertically.	Allows a shift of up to +- 45 pixels along the height.
Shear	Distorts the image by “stretching” it either horizontally or vertically, providing a ‘skewed’ perspective.	Shearing of the image is allowed up to a maximum angle of 0.2 degrees.
Zoom	Changes the image’s apparent distance, making it seem closer or farther away.	Allows for a maximum zoom of up to 20%.
Horizontal Flip	Creates a mirrored version of the image, providing a reflection along the vertical axis.	Only horizontal flipping is supported.

formal context [52], given an image I with dimensions $m \times n$ and a filter-kernel K with dimensions $q \times r$, we generate the feature map F via convolution along the axes m, n with the kernel K as follows:

$$F(m, n) = \sum_q \sum_r I(m, n)K(m - q, n - r) \tag{1}$$

Typically, the values in the feature map undergo a transformation by an activation function. One common example is the rectified linear unit activation function [53] (ReLU) which transforms all negative values to zero.

2.4. EfficientNet architecture

EfficientNet [54] is a model that has gained traction in machine learning applications in recent years. The architecture of EfficientNet has been designed to systematically scale all dimensions of depth, width, and resolution.

The underlying idea of EfficientNet is based on the compound scaling method. This strategy aims to maintain a balance between the network depth (how many layers it has), width (how wide are the layers), and resolution (input image size). A set of fixed scaling coefficients guides this balance. Formally, for a baseline EfficientNet-B0, if α, β, γ are constants which maintain the balance, and ϕ is the user-specified coefficient, we can scale depth d , width w , and resolution r of the network according to:

$$d = \alpha^\phi d_0, \quad w = \beta^\phi w_0, \quad r = \gamma^\phi r_0 \tag{2}$$

Here, d_0, w_0, r_0 are the depth, width, and resolution of the base model, respectively.

A core component of EfficientNet is the MBConv block, inspired by the MobileNetV2 [55] architecture. This block applies a series of transformations: a 1×1 convolution (expansion), followed by a depth-wise convolution (represented by a depth-wise separable

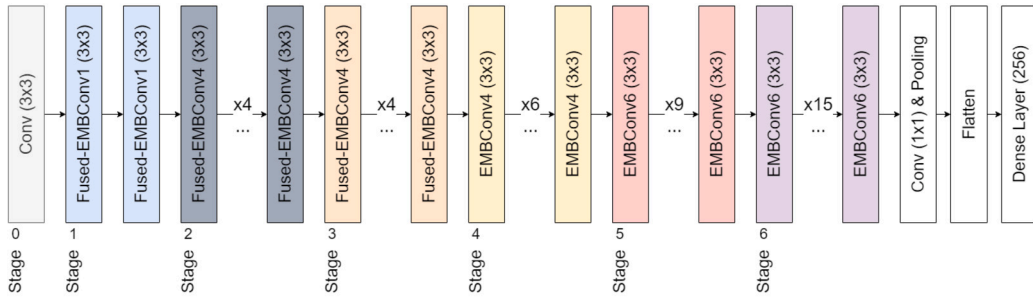


Fig. 4. Diagram illustrating the integrated architecture of the base EfficientNetV2 and our modifications after the pooling layer.

convolution kernel D), a Squeeze-and-Excitation (SE) operation [56], and another 1×1 convolution (projection). For an input image I , the transformation of the MBCConv block, T_{MB} , can be represented as:

$$T_{MB}(I) = K_2 * SE(D * (K_1 * I)) \tag{3}$$

In this equation, $*$ represents the convolution operation, K_1 and K_2 are the 1×1 convolutional filters, D represents the depth-wise convolutional filter, and $SE(\cdot)$ denotes the Squeeze-and-Excitation operation. Each convolution is followed by an activation function. This efficient use of computational resources within the MBCConv block, especially through the depth-wise convolution and the SE block, contributes significantly to the excellent performance of EfficientNet.

EfficientNetV2 [40] introduces several key changes to the original EfficientNet architecture. The depth, width, and resolution scaling remain the same as in the original EfficientNet. However, the use of the Fused-MBCConv block, which combines the initial 1×1 convolution and the depth-wise convolution into a single 3×3 convolution operation, and then applies the Squeeze-and-Excitation (SE) operation followed by a final 1×1 convolution, is a significant modification.

The transformation of the Fused-MBCConv block, T_{FMB} , can be expressed as:

$$T_{FMB}(I) = K_2 * (SE(K_f * I)) \tag{4}$$

In this equation, K_f is the 3×3 convolutional filter that combines the initial 1×1 convolution and the depth-wise convolution, K_2 is the final 1×1 convolutional filter and $SE(\cdot)$ denotes the Squeeze-and-Excitation operation. Each convolution, followed by an activation, is part of a block that may contain a skip connection.

Another key change in EfficientNetV2 is the progressive learning method, which adaptively adjusts the regularization and image size during training. EfficientNetV2 models to train faster and achieve better parameter efficiency than the original EfficientNet, even outperforming transformer models on key vision tasks, such as Image-Net Classification. Fig. 4 illustrates the EfficientNetV2's base architecture (B0).

In our experiment, we extensively explored the EfficientNetV2 family of models, ranging from the relatively compact EfficientNetV2-B0 to the more complex EfficientNetV2-M. We supplemented the architecture by incorporating a flattening layer, followed by a densely connected layer of 256 neurons. This configuration emerged from an early-stage grid search on training data (split into 60% train and 40% validation), which tested various setups between 64 to 512 neurons (in steps of 64 neurons) and between 1 to 3 layers of depth. A single layer with 256 neurons yielded the best early validation accuracy. The grid search was conducted up to 22 epochs with early stopping, and we observed that results plateaued after 13 to 15 epochs (and early stopping activated). The search was conducted using EfficientNetV2B0, S, and M architectures to encompass models of low, mid, and high capacities. The dense layer units utilized an Exponential Linear Unit (ELU) as the activation function, set before the final output layer. For the training process, we employed Categorical Cross-Entropy Loss to optimize the model's performance, and trained for 15 epochs each model family for the main experiment. The minimum validation loss was used as the early stopping criterion. All EfficientNet model training was performed with the Deep Fast Vision repository. The optimizer used was Adam [57], with a learning rate of 2×10^{-5} and a batch size of 32.

2.5. Vision transformers

Vision Transformers [46] (ViT) utilize the transformer architecture to perform computer vision tasks by processing images as a sequence of fixed-size patches. An image I of dimensions $H \times W \times C$ is segmented into patches P_i of size $P \times P \times C$, which are then linearly transformed into embeddings E_i using a projection matrix W_p :

$$X_0 = W_p \cdot \text{Flatten}(P_i) + E_{\text{pos}} \tag{5}$$

Here, E_{pos} are positional embeddings added to retain spatial information. The ViT encoder applies multi-headed self-attention and position-wise feed-forward networks to these embeddings, where the attention for a single head is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \tag{6}$$

This formula consolidates the image processing and self-attention steps into a single representation, encapsulating the core functionalities of ViTs. The design of ViT reflects a shift towards utilizing global context and attention-based mechanisms to interpret images, contrasting with the localized receptive fields typically employed in convolutional neural networks. This method has demonstrated effectiveness when trained on large-scale datasets, showcasing the potential of transformers to generalize and manage complex visual tasks without intensive domain-specific modifications. In this study we employed the ViT B32 architecture, which is a specific configuration of the Vision Transformer (ViT). In the ViT B32 model, “B” signifies the Base model size, which involves 12 transformer blocks (12 attention heads), connected to 256 hidden layer units, while “32” indicates that each image is segmented into 32x32 pixel patches. The ViT was connected to the exact classification head specification as the efficient net models (256 hidden layer units). The ViT trained maintaining the exact same batch size and augmentation approaches. Training lasted for 22 epochs, extending beyond the initial 15 to provide an overhead in effective capacity (training time). The Rectified Adam [58] optimizer was used, which may provide faster and improved convergence in terms of both time and accuracy.

2.6. Random forests

Random Forests [45] (RF) is an ensemble machine learning method that leverages multiple decision trees to make predictions. Its robustness stems from combining a multitude of weak learners, the individual trees, to form a strong learner, the Random Forest. The most typical form of decision tree uses binary decisions based on feature thresholds to make predictions. The prediction of a single decision tree, T_d , on an input vector I , can be mathematically expressed as:

$$T_d(I) = \sum_{i=1}^n v_i \delta(f_i(I) < t_i) \quad (7)$$

In this equation, n is the total number of nodes in the tree, v_i is the predicted value for the i -th node, $f_i(I)$ represents the i -th feature of the input vector I , t_i is the threshold for the i -th node, and δ is the indicator function.

A Random Forest aggregates the predictions from D decision trees to make a final prediction. For regression problems, the decision tree outputs are typically averaged, while for classification, they are aggregated through a majority vote (mode). Formally, the prediction of a Random Forest, F_D , on an input vector I , can be represented as:

$$F_D(I) = \frac{1}{D} \sum_{d=1}^D T_d(I) \quad (8)$$

In this equation, $T_d(I)$ is the output of the d -th decision tree for input vector I , and D is the total number of decision trees in the forest. The Random Forest model’s strength lies in its ability to reduce overfitting compared to a single decision tree by averaging the results over many trees. In our experiment, we employed 400 estimators and default parameters from the scikit-learn library [59]. Ultimately, the best-performing EfficientNet (Deep) model, determined by validation accuracy, had its classification head replaced by the Random Forest. After training and validating the RF, it was evaluated once on the test and external test sets.

2.7. t -distributed stochastic neighbor embedding

The t -Distributed Stochastic Neighbor Embedding (t-SNE) [60] is a tool for visualizing high-dimensional data in a lower-dimensional space, often two or three dimensions. It is particularly effective for visualizing complex data structures, such as those produced by deep learning models, in a way that preserves the relationships and structures within the data.

In our experiment for visualizing the external test data from the trained model, we used two dimensions with an automatic learning rate on the dense layer before the output layer of the best EfficientNetV2 model. Additionally, we sampled 160 examples per class from the lower-dimensional space and rasterized [61,62] the projected space. These two approaches can be seen in results Fig. 8, while Fig. 9 demonstrates the replacement of all data points with their corresponding whole-slide image patches.

2.7.1. Neural architecture search (NAS)

Neural Architecture Search (NAS) is an automated process aimed at discovering an approximately optimal neural network architecture for a specific task by systematically exploring a range of possible configurations. In the context of Autokeras NAS, NAS involves evaluating different combinations of network layers, their connections, and hyperparameters to identify the architecture that yields improved performance on a given dataset.

2.7.2. Autokeras NAS configuration

Autokeras NAS [63] is an open-source Auto-ML library for deep learning, built on top of Keras and offers functionalities like automatic model selection and hyperparameter tuning. The library can handle various data types, including images, text, and structured data. For our experiment, we initiated a search incorporating instance normalization block, image-augmentation, ResNet V2 blocks pre-trained with ImageNet, all followed by flatten and dense layer blocks. We trained for 15 epochs and ran a maximum of 6 trials; early stopping (with restore best weights) was employed, and the search focused on validation loss. The NAS process ran for 52 hours with multiple parallel instances in a P100 GPU cluster.

Table 2

Comparison of accuracy and parameter count across different EfficientNet models, the Autokeras NAS and Transformer-ViT32 baseline models. The 95% CI refers to the 95% confidence interval from bootstrapping.

Model Name	Training Accuracy	Validation Accuracy	Testing Accuracy	External Testing Accuracy	Parameter Count (M)
EfficientNetV2B0	99.5%	99.57%	-	-	21.98
EfficientNetV2B1	99.5%	99.73%	-	-	22.99
EfficientNetV2B2	99.51%	99.73%	-	-	26.43
EfficientNetV2B3	99.59%	99.75%	-	-	32.20
EfficientNetV2S	99.68%	99.79%	-	-	36.39
EfficientNetV2M	99.72%	99.8%	-	-	69.21
EfficientNetV2M + RF	100%	99.91%	99.89%	96.74% (95% CI:96.3%-97.1%)	69.21
Autokeras NAS Model (Baseline)	100%	98.47%	98.52%	94.21%	26.77
Transformer-ViT32 (Baseline 2)	99.45%	99.32%	99.32%	92.55%	87.65

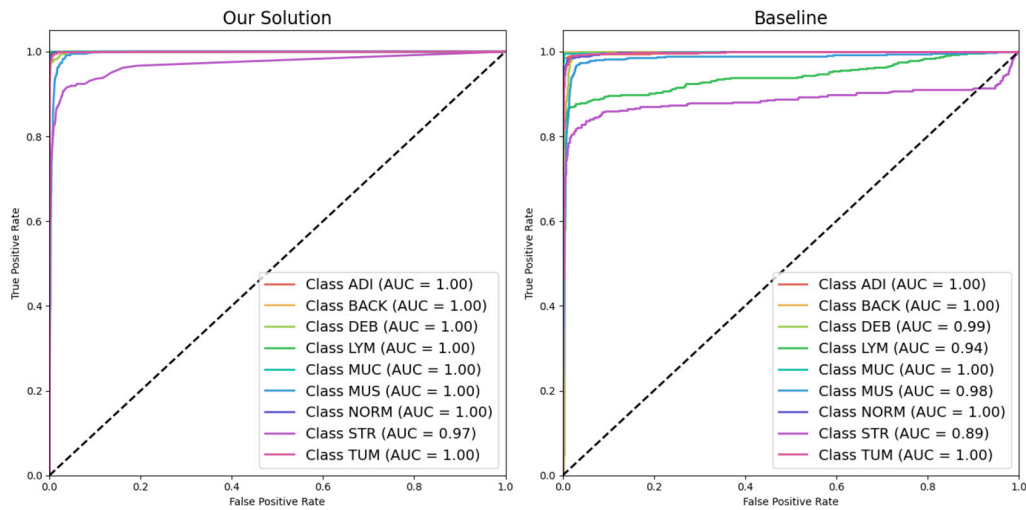


Fig. 5. Benchmarking EfficientNetV2M + RF hybrid vs. Autokeras NAS using one-vs-all ROC curves.

3. Results

3.1. Classification

As presented in Table 2, we evaluated the performance of a series of models, including several EfficientNet architectures (without testing), the final hybrid model, and an Autokeras NAS baseline. The table provides a comprehensive summary of the training, validation, testing accuracy scores, and parameter counts for each model. Notably, the EfficientNet models consistently showcased a rise in accuracy scores corresponding to an increase in model complexity, emphasizing the positive association between the number of parameters and model performance. Among the investigated models, EfficientNetV2M emerged with the highest validation accuracy. This deep model subsequently served as the basis for our hybrid model, complemented by a Random Forest (RF) classification head. Our hybrid model underwent further evaluations on internal and external test sets to ascertain its robustness and applicability. The model accuracy scores of 99.89% on the internal test set and 96.74% on the external test set. The 95% confidence interval was obtained with 1000 iterations of bootstrapping [64]. In comparison, the Autokeras NAS and Transformer baseline models did not match the accuracy levels exhibited by the hybrid model. Table 3 offers a detailed comparative analysis, juxtaposing the performance metrics of our model with those from other studies. Notably, our approach achieved improved results in both internal and external testing datasets.

As shown in Fig. 5, using the one-vs-all scheme, our hybrid model and the Autokeras NAS baseline model’s performance on the external testing set were assessed using the Area Under the Receiver Operating Characteristic (ROC) Curves. Our hybrid model displayed improved performance across all classes. It achieved the maximum AUC score of 1.00 for ADI, BACK, DEB, LYM, MUC, MUS, NORM, and TUM classes. For the STR class, the AUC slightly dropped, and our model registered a score of 0.97. On the contrary, the Autokeras NAS baseline model, while demonstrating high AUC scores for ADI, BACK, MUC, NORM, and TUM, showed diminished performance for the remaining classes. The most notable dip was observed for class STR, registering an AUC of 0.89. Notably, the ROC curve for the Autokeras NAS baseline model intersected with the baseline after the 0.8 threshold. This intersection was not observed in the ROC curves of the hybrid model.

Table 3

Comparative analysis of accuracy metrics and validation techniques across research studies. The 95% CI refers to the 95% confidence interval from bootstrapping.

Research Work	Training Set (TR)	Validation Set (V)	Test Set (T)	Validation Approach	External Test Set (T2)
This study	100%	99.91%	99.89%	TR-V-T-T2	96.74% (95% CI:96.3%-97.1%)
Prezja, et al. [39]	99.4%	99.3%	99.5%	TR-V-T-T2	95.6%
Kather et al. [8]	-	-	98.7%	TR-V-T-T2, TR-T2	94.3%
Peng et al. [31]	-	-	-	TR-V-T-T2	95%
Qi et al. [32]	-	99%	-	TR-V-T2	95%
Shen et al. [33]	-	-	-	TR-V-T2	94.8%
Wang et al. [34]	-	-	-	TR-V-T2	94.8%
Yang et al. [35]	-	-	-	TR-V-T2	91.1%
Yang et al. [65]	-	-	-	TR-V-T2	86.4%

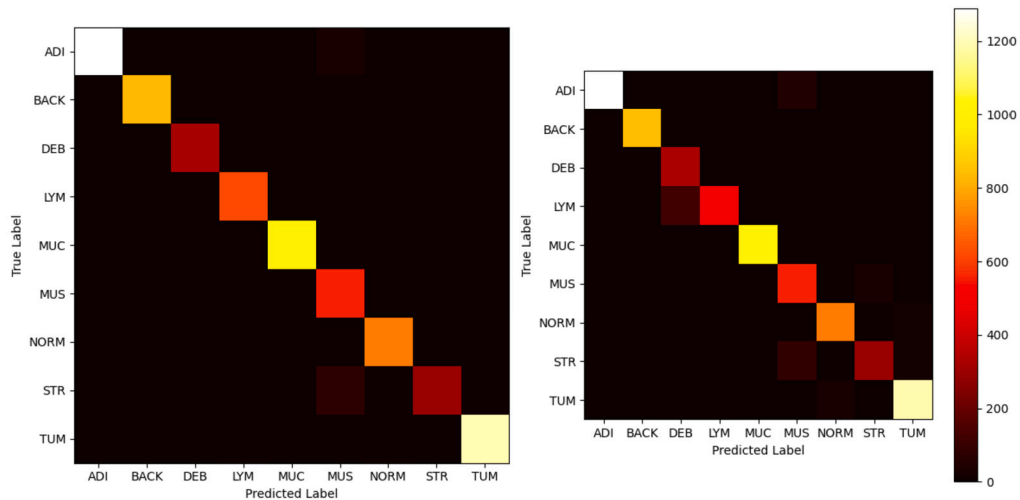


Fig. 6. Confusion matrices for external testing: Hybrid Model (left), Autokeras NAS baseline (right). Identical color format as Kather et al. [8] for direct comparison.

Table 3 comprehensively compares accuracy scores across all relevant studies that employed the original training and external testing data. As shown, our model demonstrated improved performance, improving upon previous studies.

Furthermore, our study stands out as the only one providing a complete and transparent report of errors across all stages - from training through to external testing, and a confidence interval for the external testing benchmark. In evaluating prior studies, some [66,67] did not use conventional validation approaches, evident from the absence of validation and testing data to detect overfitting, including the search for parameters and hyperparameters on external testing data, which further complicate comparisons. Moreover, instances that employed techniques like few-shot learning and testing [36], shuffling of external testing data within training data [37], and using external testing as validation and testing with their own testing data [68], also make it challenging to conduct a comparative analysis.

Fig. 6 demonstrates exceptional performance across most classes, with the vast majority of predictions falling on the diagonal, signifying correct classifications. Class BACK, LYM, and NORM are particularly noteworthy, which were nearly perfectly classified. However, the model exhibits some confusion between classes MUC-MUS, and STR-MUS, suggesting shared features or characteristics that led to these misclassifications.

The confusion matrix from the original study conducted by Kather et al. revealed certain areas of confusion between particular classes, specifically LYM-DEB, MUS-MUC, NORM-TUM, TUM-STR, and STR-MUSC-MUC. When we compare this with the confusion matrix generated by our hybrid model, we see substantial improvements in classification accuracy, especially in the problematic areas identified in Kather’s study. Fig. 6 (on the left) shows that our model successfully eliminated most of these previously reported confusions. For instance, our model has resolved the confusion between LYM-DEB, indicating its improved ability to distinguish between these two classes. Similarly, the confusion between MUS-MUC and NORM-TUM seen in Kather’s study has also been resolved. One of the most substantial improvements was observed in the classification of LYM, where our model demonstrated almost perfect classification accuracy. In Fig. 6 (on the right), the Autokeras NAS model, still achieving overall solid performance, underperformed in several areas relative to the hybrid model. Specifically, there is apparent confusion between classes DEB-LYM, and STR-MUS. It is also worth noting that, unlike the hybrid model, the Autokeras NAS model made several misclassifications between classes TUM and NORM.

The hybrid model’s improved performance is most noticeable when dealing with classes ADI, MUS, LYM, NORM, and STR, suggesting that its architecture may be more adept at recognizing the nuanced features that distinguish these classes. These features are relevant to the Deep Stroma score calculation proposed by Kather.

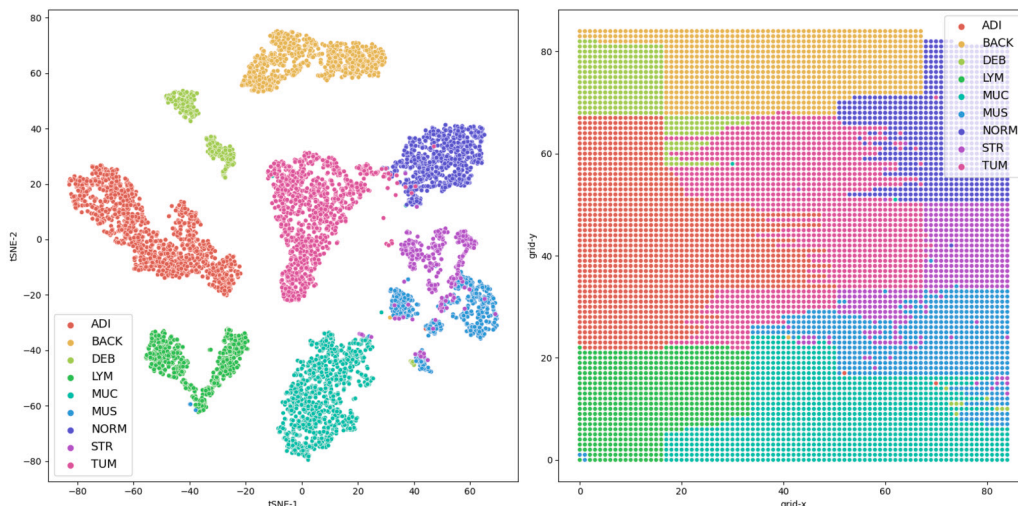


Fig. 7. Left: t-SNE projection illustrating near-optimal class separation with TUM class at center. Right: Rasterized t-SNE map preserving relative inter-class distances and eliminating inter-point distance.

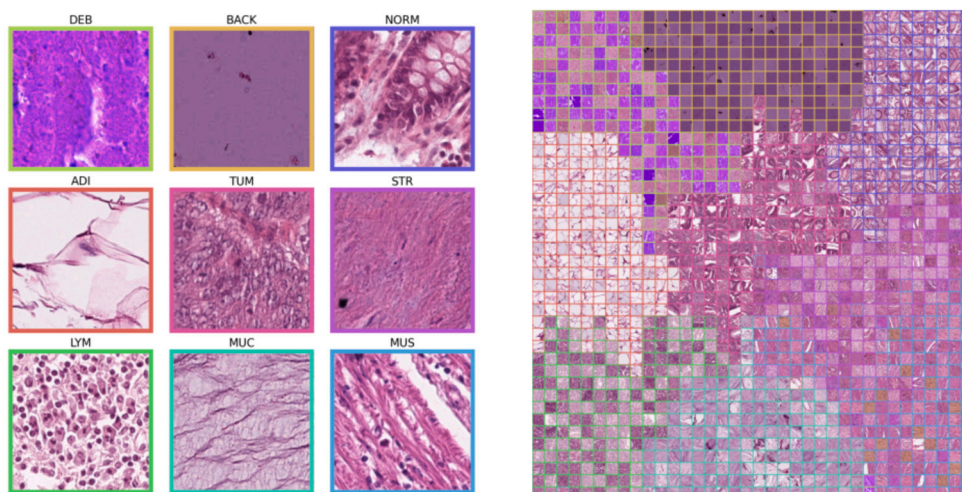


Fig. 8. Left: Nearest intra-class image to class mean in the t-SNE projection. Right: Rasterized t-SNE space with 160 samples per class, data points replaced with images. Boundary colors indicate class origin.

In the t-SNE visualization of the high-dimensional feature space (Fig. 7), the classes align almost seamlessly around the tumor (TUM) class at the approximate center. This near-optimal separation of classes demonstrates the effectiveness of the hybrid model in distinguishing between various histopathological types. Notably, there is no observed fragmentation within the classes, underlining the consistency of the learned features even as a two-dimensional projection. Moreover, the relative positioning of the classes aligns with histopathological anticipations. The TUM and normal (NORM) tissues, displaying histological similarities, are situated adjacently on the t-SNE map. Analogously, the proximity of the stroma (STR) and muscle (MUS) classes suggests vector space congruence. Transitioning the t-SNE map into a rasterized depiction preserves relative inter-class distances and condenses the visualization by eliminating inter-point space, enabling a more compact visualization of class distributions. Fig. 8 illustrates a single instance from each class closest to each class mean as projected by t-SNE. Concurrently, we subsample the t-SNE space with 160 samples per class to build a rasterized representation, substituting data-points with corresponding images. Fig. 9 follows the same pattern but without sub-sampling, using all external test data.

Overall, these results demonstrate the robustness of the hybrid model’s feature extraction and highlight its potential to provide clinically relevant insights.

4. Discussion

The present study was attempted to advance the decomposition classification of colorectal cancer by developing a novel hybrid model. Our methodology integrated EfficientNetV2M and Random Forest (RF) techniques, with the results demonstrating the model’s

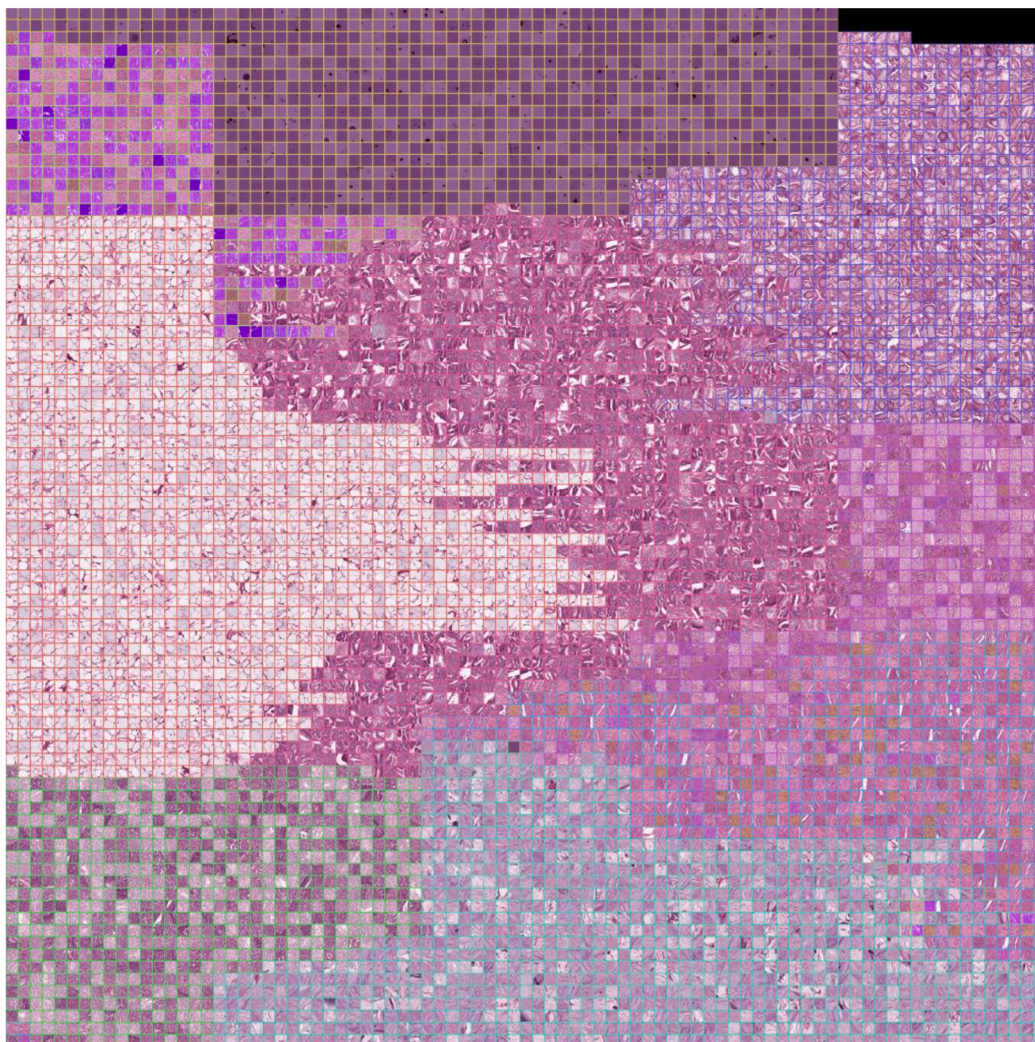


Fig. 9. Rasterized t-SNE visualization of all external test data showing class distributions. Black indicates blank grid space, boundary colors denote class origin. The full resolution is made available under data availability.

improved performance over previously proposed solutions (Table 3). This was achieved by combining the strengths of both deep learning and traditional ensemble machine learning. In addition to achieving high accuracy rates, our hybrid model demonstrated improved performance across all classes when assessed using the Area Under the Receiver Operating Characteristic (ROC) Curves in a one-vs-all scheme (Fig. 5). The hybrid model achieved the maximum AUC score of 1.00 for classes ADI, BACK, DEB, LYM, MUC, MUS, NORM, and TUM. While the model performance slightly dropped for the STR class, it still registered an AUC of 0.97. The confusion matrices demonstrated the hybrid model's improved ability to correctly classify various histopathological types, even those that presented difficulties in previous studies, such as the original study by Kather et al. [8].

Combining Random Forests with Convolutional Neural Networks (CNNs) represents a relatively unexplored area within medical imaging. Prior to this study, this approach was applied using a two-layer CNN trained from scratch with a binary target, specifically with MRI images [69]. Although promising, these initial applications did not fully explore the potential efficacy of combining more complex neural network architectures and leveraging transfer learning. Our study extended this methodology into a new domain by employing advanced CNN architectures such as EfficientNetV2 and incorporating transfer learning techniques along with a multi-class target. The results demonstrated improved task performance with these new parameters and more complex neural networks. This underscored the potential of combining traditional machine learning algorithms with state-of-the-art deep learning models for enhanced predictive accuracy in histopathological image analysis.

Progress in the task of colorectal cancer histology decomposition appears necessarily incremental, given that less than 5% improvement remains to reach a maximum score in available testing data. However, this seemingly modest threshold is crucial due to the scalability potential of the system. Small enhancements in performance may lead to substantial benefits when applied across large datasets, where tens of thousands of histology slides are analyzed. This phenomenon mirrors trends observed in the ImageNet Challenge, where similar incremental improvements still drive researchers to choose newer architectures over older ones due to the

competitive edge they provide. Likewise, in our study, we chose the EfficientNetV2 architecture for its proven incremental improvements over previous models in the ImageNet Challenge.

One significant limitation encountered in our study is the inability to apply Gradient-weighted Class Activation Mapping (Grad-CAM) [70] for visualizing class activation regions in input images. Grad-CAM relies on the gradients flowing through the final convolutional layers to produce heat maps that highlight the areas of the image most relevant to the classification decision. However, when employing a Random Forest (RF) as the final classification head, the gradients necessary for generating these heat maps are not computable. Consequently, the integration of RF with CNNs, while effective in enhancing classification performance, inherently precluded the use of gradient-based visualization techniques such as Grad-CAM. This limitation suggests the need for alternative methods to visualize class-related activations in hybrid models that combine deep learning and traditional machine learning algorithms.

The t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization of the high-dimensional feature space demonstrated the effectiveness of the hybrid model in distinguishing between various histopathological types. Additionally, t-SNE revealed the relationships among class instances in the projected vector space, though the two dimensions of this vector space do not yet have a higher-level interpretation. To this end, we are providing the rasterized t-SNE plot in high definition to assist future medical experts in identifying the projected features. This effort aims to enhance the interpretability of the projected space, enabling medical professionals to gain better insights into the relationships and structures within the data.

In terms of data augmentation and validation, our findings suggest that robust data augmentation is a powerful tool for mitigating overfitting when paired with an appropriate validation technique. This conclusion is consistent with established practices, as data augmentation is a commonly used method in numerous classification studies. We believe that more advanced forms of augmentation, particularly those that utilize synthetic data from Generative Neural Networks [71,72,13,73], could also prove beneficial and may warrant a dedicated future study. As illustrated in Table 3, there is considerable variation in validation methods, which poses a challenge when validation, testing, and external testing datasets are missing. The practice of validation in this context could be significantly improved by adopting a standardized approach. From a limitation perspective, the training of such systems necessitates a substantial volume of annotated data. For potential advancements in these systems, the annotation of even more data by medical professionals may be required. Notably, several studies [33,35,67] are still pending peer review.

In reflecting upon the prevailing CRC scientific literature, we have noted considerable advancements achieved over a comparably short time. However, this progress has not been devoid of limitations. Firstly, our survey identified a distinct lack of uniformity in the evaluation methodologies employed across different systems. Emphasizing consistency and adherence to best practices during evaluation is crucial for reducing biases and enabling direct system comparisons. Moreover, estimates for metrics such as label noise are absent. Ascertaining estimates for label noise could be pivotal in establishing a benchmark for future comparative analyses.

In our study, the hybrid model's deep section was trained using our new Deep Fast Vision open-source library [50], whereas the baseline employed Autokeras NAS [63]. Both these libraries uniquely leverage AutoML in vision-related contexts with transfer learning capabilities. The hybrid model demonstrated improvement in scores compared to Autokeras NAS. Although Autokeras NAS employs ImageNet pre-trained blocks, such blocks function independently, and only low-level features are loaded. On the contrary, Deep Fast Vision utilizes the entire hierarchy of learned features from already validated architectures. This methodology gives Deep Fast Vision an expected early advantage over Autokeras NAS, allowing it to harness the transfer learning capability of previously validated deep learning models.

The combination of EfficientNet and Random Forests introduces a novel approach by integrating a cutting-edge deep learning architecture with a traditional machine learning algorithm, a specific pairing that has not been explored previously. In addition, to the best of our knowledge, this is the first instance where the EfficientNetV2 architecture is utilized in colorectal cancer (CRC) histopathology imaging. Intriguingly, the hybrid model outperformed a vision transformer baseline, which possessed notably greater parameter counts. These findings are both unique and promising. However, it is important to note a significant limitation: we were unable to utilize transformers as the classification head. The primary reason is that the input features to tabular transformers [74] would neither be categorical nor a mix of categorical and numeric. As a result, these features bypass the advantages of transformers' attention mechanisms and proceed directly to the MLP phase. Additionally, it is crucial to highlight that Random Forest cannot receive gradients directly from the CNN or its flattening layer, necessitating the application of the Random Forests asynchronously.

On Transformer models and Autokeras NAS baselines, it is essential to address the variations in performance between different architectural approaches and training strategies within this task. Specifically, the difference in performance between Transformers and Autokeras NAS was not surprising when considering the depth and scope of the neural architecture search (NAS) implemented by Autokeras. This approach extended the search duration significantly, optimizing for the most effective model configurations over a prolonged period. The NAS process ran for 52 hours with multiple parallel instances in a P100 GPU cluster. In contrast, EfficientNet, which was included in our hybrid model, had previously demonstrated improved performance over Vision Transformer (ViT) models in benchmarks like ImageNet. This result could explain the increased performance of EfficientNet over ViTs. The performance of larger Transformer models remains untested in this context, suggesting that this topic warrants further investigation through a future study. Additionally, utilizing a transformer as a second baseline is crucial to ground and contrast our results, moving away from solely relying on automated machine learning (AutoML) solutions. Moreover, Autokeras NAS was not constrained by a capacity threshold, and trained long enough to also gain more effective capacity (training time). This implies that the Autokeras NAS was capable of searching long enough to potentially outperform both Transformer and EfficientNet models but did not surpass EfficientNet, indicating that factors other than model capacity were at play, which could warrant further investigation in future work.

In our study, we utilized Autokeras NAS for automated neural architecture search (NAS), specifically incorporating the ResNet block due to its availability of pre-trained weights from ImageNet. This decision was critical for maintaining a fair comparison across different models. ResNet blocks, being pre-trained, allow for the use of pre-learned features, which significantly accelerate the

convergence process during the training phase. This is similar to the advantage that EfficientNet gains from its pre-training. Without utilizing blocks that come with pre-trained weights, comparisons with EfficientNet would naturally be skewed. EfficientNet benefits from a pre-training phase that enhances its initial performance and learning speed, attributes that would be lacking in a NAS-generated model built from scratch. By using the ResNet block within Autokeras NAS, we aimed to level the playing field, ensuring that all models being compared had similar advantages in terms of initial feature learning and convergence speed.

Concerning the quality of microscope images, factors such as the slide's quality and the presence of artifacts or pixel noise could significantly contribute to misclassifications. Normalization and contrast enhancement processes performed on tiles before their entry into the classifier might accentuate pixel noise or other non-tissue artifacts, such as dust or hair, thereby potentially skewing results. The 'Picasso' effect [75] may compound these distortions in Convolutional Neural Networks (CNNs). Furthermore, instances, where a tile contains minimal or no relevant tissue and isn't labeled as background could also instigate misclassifications. Incorporating a background class can somewhat mitigate these issues, but this effect is only partial, and similar errors would be anticipated. To address this more comprehensively, employing adversarial augmentations [76] could help uncover additional vulnerabilities. Future systems could benefit from introducing pixel noise during the augmentation phase, ideally in a randomized manner and with replacement. Randomization and replacement are critical to prevent the introduction of biases and potential overfitting by the classifier. Lastly, the focus factor ('blur') can also influence results, particularly when it coexists with pixel noise. As a preventative measure, further augmentation involving a variety of blur intensities might be beneficial. Such recommendations are particularly crucial given the variability in focus and quality across different patient slides. Additionally, it is important to note that staining normalization [77–81,15] could also affect results and might warrant an independent future study to thoroughly assess its impact.

In our previous study [39], we highlighted the need for a deeper analysis of classifier probability profiles [82]. While Kather's approach [8] pinpointed one classifier, it did not address the influence of model probability calibration on the deep stroma score. Given that different architectures can produce varying probability profiles, even with similar performance metrics, there's a gap in understanding these profiles' effects on deep stroma and outcome prediction. Addressing this could open new avenues for research and refine our approach to model architectures and outcome predictions. Before advancing to bio-marker extraction and patient outcome assessment, it's imperative that these profiles undergo further validation.

In Kather, et al. [8] from which we sourced our data, there was a stringent manual review process for all slides. Slides with pronounced artifacts, be it tissue folds or tears, were set aside. This exclusion underscores an inherent gap: our models, along with others, have not been trained with these challenges. In many real-world scenarios where such artifacts may exist, the desired performance might decline, emphasizing the importance of training datasets that mirror extensive real-world challenges. While remedies like affine augmentations might provide some mitigation, they fall short of fully addressing these challenges. Additionally, it should be noted that to our knowledge no further annotated data is available for additional training, refinement or validation.

5. Conclusion

Our study successfully developed a hybrid model that combines the EfficientNetV2 and Random Forest algorithms, notably advancing the task scores of automatic colorectal cancer tissue decomposition. Our model demonstrated high average accuracy and excelled in performance across all classes, as indicated by the AUC scores. It improved upon the results of previous research in this domain, including the seminal work by Kather et al. [8].

It is imperative that this work is subjected to rigorous clinical validation before any deployment into routine clinical practice. Our research presents considerable potential in the classification of CRC slides and the prospect of improving CNN-based biomarkers that rely on classification accuracy. This improvement, in turn, could lead to improved predictions for CRC patient outcomes.

Lastly, we have made a commitment to transparency and replicability by providing unrestricted access to our models.

CRedit authorship contribution statement

Fabi Prezja: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Leevi Annala:** Data curation, Formal analysis, Writing – review & editing. **Sampsa Kiiskinen:** Data curation, Formal analysis, Writing – review & editing. **Suvi Lahtinen:** Data curation, Formal analysis, Writing – review & editing. **Timo Ojala:** Data curation, Formal analysis, Writing – review & editing. **Pekka Ruusuvoori:** Data curation, Formal analysis, Writing – review & editing. **Teijo Kuopio:** Conceptualization, Data curation, Formal analysis, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Materials and best models from the current study are accessible in the Google Drive repository: https://drive.google.com/drive/folders/1ypFyU2V6ifRkLB6hRkLqB6C2D_fvL_5Y?usp=sharing

Acknowledgements

The authors extend their gratitude to Kimmo Riihiahio, Rodion Enkel and Leevi Lind.

References

- [1] C.-N. Qian, Y. Mei, J. Zhang, Cancer metastasis: issues and challenges, *Chin. J. Cancer* 36 (1) (2017) 1–4.
- [2] WHO, Cancer (2022).
- [3] Colorectal Cancer Alliance, Colorectal Cancer Information (2022).
- [4] J. Malik, S. Kiranyaz, S. Kunhoth, T. Ince, S. Al-Maadeed, R. Hamila, M. Gabbouj, Colorectal cancer diagnosis from histology images: a comparative study, arXiv preprint, arXiv:1903.11210, 2019, <http://arxiv.org/abs/1903.11210>.
- [5] R. Parveen, S.S. Rahman, S.A. Sultana, Z.H. Habib, Cancer types and treatment modalities in patients attending at delta medical college hospital, *Delta Med. Coll. J.* 3 (2) (2015) 57–62, <https://doi.org/10.3329/dmcj.v3i2.24423>.
- [6] J.D. Schiffman, P.G. Fisher, P. Gibbs, Early detection of cancer: past, present, and future, *Am. Soc. Clin. Oncol. Educ. Book* 35 (1) (2015) 57–65.
- [7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
- [8] J.N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Hergel, C.-A. Weis, T. Gaiser, A. Marx, N.A. Valous, D. Ferber, et al., Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study, *PLoS Med.* 16 (1) (2019) e1002730.
- [9] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P.E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, J. Lundin, Deep learning based tissue analysis predicts outcome in colorectal cancer, *Sci. Rep.* 8 (1) (2018) 1–11.
- [10] O.-J. Skrede, S. De Raedt, A. Kleppe, T.S. Hveem, K. Liestøl, J. Maddison, H.A. Askautrud, M. Pradhan, J.A. Nesheim, F. Albrechtsen, et al., Deep learning for prediction of colorectal cancer outcome: a discovery and validation study, *Lancet* 395 (10221) (2020) 350–360.
- [11] F. Calimeri, A. Marzullo, C. Stamile, G. Terracina, Biomedical data augmentation using generative adversarial neural networks, in: *International Conference on Artificial Neural Networks*, Springer, 2017, pp. 626–634.
- [12] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, *Neurocomputing* 321 (2018) 321–331.
- [13] F. Prezja, J. Paloneva, I. Pölonen, E. Niinimäki, S. Äyrämö, DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification, *Sci. Rep.* 12 (1) (2022) 1–16.
- [14] F. Prezja, L. Annala, S. Kiiskinen, S. Lahtinen, T. Ojala, Adaptive variance thresholding: a novel approach to improve existing deep transfer vision models and advance automatic knee-joint osteoarthritis classification, arXiv preprint, arXiv:2311.05799, 2023.
- [15] F. Prezja, I. Pölonen, S. Äyrämö, P. Ruusuvoori, T. Kuopio, H&E multi-laboratory staining variance exploration with machine learning, *Appl. Sci.* 12 (15) (2022) 7511.
- [16] U. Khan, S. Koivukoski, M. Valkonen, L. Latonen, P. Ruusuvoori, The effect of neural network architecture on virtual H&E staining: systematic assessment of histological feasibility, *Patterns* 4 (5) (2023).
- [17] B.F. Kurland, E.R. Gerstner, J.M. Mountz, L.H. Schwartz, C.W. Ryan, M.M. Graham, J.M. Buatti, F.M. Fennessy, E.A. Eikman, V. Kumar, et al., Promise and pitfalls of quantitative imaging in oncology clinical trials, *Magn. Reson. Imaging* 30 (9) (2012) 1301–1312.
- [18] J.L. Sprattin, N.J. Serkova, S.G. Eckhardt, Clinical applications of metabolomics in oncology: a review, *Clin. Cancer Res.* 15 (2) (2009) 431–440.
- [19] J.P.B. O'Connor, A. Jackson, M.-C. Asselin, D.L. Buckley, G.J.M. Parker, G.C. Jayson, Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives, *Lancet Oncol.* 9 (8) (2008) 766–776.
- [20] A.D. Waldman, A. Jackson, S.J. Price, C.A. Clark, T.C. Booth, D.P. Auer, P.S. Tofts, D.J. Collins, M.O. Leach, J.H. Rees, Quantitative imaging biomarkers in neuro-oncology, *Nat. Rev. Clin. Oncol.* 6 (8) (2009) 445–454.
- [21] H.E. Danielsen, T.S. Hveem, E. Domingo, M. Pradhan, A. Kleppe, R.A. Syvertsen, I. Kostolomov, J.A. Nesheim, H.A. Askautrud, A. Nesbakken, et al., Prognostic markers for colorectal cancer: estimating ploidy and stroma, *Ann. Oncol.* 29 (3) (2018) 616–623.
- [22] J.N. Kather, A.T. Pearson, N. Halama, D. Jäger, J. Krause, S.H. Loosen, A. Marx, P. Boor, F. Tacke, U.P. Neumann, et al., Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer, *Nat. Med.* 25 (7) (2019) 1054–1056.
- [23] K. Sirinukunwattana, E. Domingo, S.D. Richman, K.L. Redmond, A. Blake, C. Verrill, S.J. Leedham, A. Chatziplis, C. Hardy, C.M. Whalley, et al., Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning, *Gut* 70 (3) (2021) 544–554.
- [24] A. Echle, N.G. Laleh, P.L. Schrammen, N.P. West, C. Trautwein, T.J. Brinker, S.B. Gruber, R.D. Buelow, P. Boor, H.I. Grabsch, et al., Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review, *Immunoinformatics* (2021) 100008.
- [25] L.H. Sobin, M.K. Gospodarowicz, C. Wittekind, *TNM Classification of Malignant Tumours*, John Wiley & Sons, 2011.
- [26] C. Isella, A. Terrasi, S.E. Bellomo, C. Petti, G. Galatola, A. Muratore, A. Mellano, R. Senetta, A. Cassenti, C. Sonetto, et al., Stromal contribution to the colorectal cancer transcriptome, *Nat. Genet.* 47 (4) (2015) 312–319.
- [27] S.J. Wagner, D. Reisenbüchler, N.P. West, J.M. Niehues, J. Zhu, S. Foersch, G.P. Veldhuizen, P. Quirke, H.I. Grabsch, P.A. van den Brandt, et al., Transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study, *Cancer Cell* 41 (9) (2023) 1650–1661.
- [28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [29] P. Ruusuvoori, M. Valkonen, L. Latonen, Deep learning transforms colorectal cancer biomarker prediction from histopathology images, *Cancer Cell* 41 (9) (2023) 1543–1545.
- [30] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, *Handb. Brain Theory Neural Netw.* 3361 (10) (1995) 1995.
- [31] T. Peng, M. Boxberg, W. Weichert, N. Navab, C. Marr, Multi-task learning of a deep k-nearest neighbour network for histopathological image classification and retrieval, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 676–684.
- [32] L. Qi, J. Ke, Z. Yu, Y. Cao, Y. Lai, Y. Chen, F. Gao, X. Wang, Identification of prognostic spatial organization features in colorectal cancer microenvironment using deep learning on histopathology images, *Med. Omics* 2 (2021) 100008.
- [33] Y. Shen, Y. Luo, D. Shen, J. Ke, RandStainNA: learning stain-agnostic features from histology slides by bridging stain augmentation and normalization, arXiv preprint, arXiv:2206.12694, 2022.
- [34] K.-S. Wang, G. Yu, C. Xu, X.-H. Meng, J. Zhou, C. Zheng, Z. Deng, L. Shang, R. Liu, S. Su, et al., Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence, *BMC Med.* 19 (1) (2021) 1–12.
- [35] J. Yang, R. Shi, B. Ni, Medmnist classification decathlon: a lightweight autml benchmark for medical image analysis, in: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2021, pp. 191–195.
- [36] W. Shuai, J. Li, Few-shot learning with collateral location coding and single-key global spatial attention for medical image classification, *Electronics* 11 (9) (2022) 1510.
- [37] S. Ghosh, A. Bandyopadhyay, S. Sahay, R. Ghosh, I. Kundu, K.C. Santosh, Colorectal histology tumor detection using ensemble deep neural network, *Eng. Appl. Artif. Intell.* 100 (2021) 104202.
- [38] D. Schuhmacher, S. Schörner, C. Küpper, F. Großeschenkamp, C. Sternemann, C. Lugnier, A.-L. Kraeft, H. Jütte, A. Tannapfel, A. Reinacher-Schick, et al., A framework for falsifiable explanations of machine learning models with an application in computational pathology, *Med. Image Anal.* 82 (2022) 102594.

- [39] F. Prezja, S. Äyrämö, I. Pölönen, T. Ojala, S. Lahtinen, P. Ruusuvoori, T. Kuopio, Improved accuracy in colorectal cancer tissue decomposition through refinement of established deep learning solutions, *Sci. Rep.* 13 (1) (2023) 15879.
- [40] M. Tan, Q. Le, Efficientnetv2: smaller models and faster training, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 10096–10106.
- [41] R. Raza, F. Zulfiqar, M.O. Khan, M. Arif, A. Alvi, M.A. Iftikhar, T. Alam, Lung-effnet: lung cancer classification using efficientnet from ct-scan images, *Eng. Appl. Artif. Intell.* 126 (2023) 106902.
- [42] S.-j. Byeon, J. Park, Y.A. Cho, B.-J. Cho, Automated histological classification for digital pathology images of colonoscopy specimen via deep learning, *Sci. Rep.* 12 (1) (2022) 12804.
- [43] A. Kallipolitis, K. Revelos, I. Maglogiannis, Ensembling efficientnets for the classification and interpretation of histopathology images, *Algorithms* 14 (10) (2021) 278.
- [44] C. Munien, S. Viriri, Classification of hematoxylin and eosin-stained breast cancer histology microscopy images using transfer learning with efficientnets, *Comput. Intell. Neurosci.* (2021) 2021.
- [45] L. Breiman, Random forests, *Mach. Learn.* 45 (2001) 5–32.
- [46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, *arXiv preprint*, arXiv:2010.11929, 2020.
- [47] J.N. Kather, N. Halama, A. Marx, 100,000 histological images of human colorectal cancer and healthy tissue (2018), <https://doi.org/10.5281/zenodo.1214456>, 2018.
- [48] M. Macenko, M. Niethammer, J.S. Marron, D. Borland, J.T. Woosley, X. Guan, C. Schmitt, N.E. Thomas, A method for normalizing histology slides for quantitative analysis, in: *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, ISBI 2009, IEEE, 2009, pp. 1107–1110.
- [49] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data* 6 (1) (2019) 1–48.
- [50] F. Prezja, Deep fast vision: a python library for accelerated deep transfer learning vision prototyping, *arXiv preprint*, arXiv:2311.06169, 2023.
- [51] F. Morales, ViT-keras: implementation of vision transformer, a simple way to achieve sota in vision models, using keras, <https://github.com/faustomorales/vit-keras>. (Accessed 10 May 2024), 2024.
- [52] I.J. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA, 2016.
- [53] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: *Icml*, 2010.
- [54] M. Tan, Q.V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2020.
- [55] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [56] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [57] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, *arXiv preprint*, arXiv:1412.6980, 2014.
- [58] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the variance of the adaptive learning rate and beyond, *arXiv preprint*, arXiv:1908.03265, 2019.
- [59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830.
- [60] L. der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (11) (2008).
- [61] G. Kogan, ofxTSNE, <https://github.com/genekogan/ofxTSNE>, 2016.
- [62] M. Klingemann, RasterFairy-Py3, <https://github.com/Quasimondo/RasterFairy>, 2016.
- [63] H. Jin, F. Chollet, Q. Song, X. Hu, AutoKeras: an AutoML library for deep learning, *J. Mach. Learn. Res.* 24 (6) (2023) 1–6, <http://jmlr.org/papers/v24/20-1355.html>.
- [64] B. Efron, Bootstrap methods: another look at the jackknife, in: *Breakthroughs in Statistics: Methodology and Distribution*, Springer, 1992, pp. 569–593.
- [65] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2: a large-scale lightweight benchmark for 2d and 3d biomedical image classification, *arXiv preprint*, arXiv:2110.14795, 2021.
- [66] M.-J. Tsai, Y.-H. Tao, Deep learning techniques for the classification of colorectal cancer tissue, *Electronics* 10 (14) (2021) 1662.
- [67] R.A. Shawesh, Y.X. Chen, Enhancing Histopathological Colorectal Cancer Image Classification by using Convolutional Neural Network, *medRxiv* (2021).
- [68] D. Schuchmacher, S. Schoerner, C. Kuepper, F. Grosserueschkamp, C. Sternemann, C. Lugnier, A.-L. Kraeft, H. Juette, A. Tannapfel, A. Reinacher-Schick, et al., A Framework for Falsifiable Explanations of Machine Learning Models with an Application in Computational Pathology, *medRxiv* (2021).
- [69] F. Khozeimeh, D. Sharifrazi, N.H. Izadi, J.H. Joloudari, A. Shoeibi, R. Alizadehsani, M. Tartibi, S. Hussain, Z.A. Sani, M. Khodatars, et al., Rf-cnn-f: random forest with convolutional neural network features for coronary artery disease diagnosis based on cardiac magnetic resonance, *Sci. Rep.* 12 (1) (2022) 11178.
- [70] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [71] A. Mauricio, J. López, R. Huaaya, J. Diaz, High-resolution generative adversarial neural networks applied to histological images generation, in: *Artificial Neural Networks and Machine Learning—ICANN 2018: 27th International Conference on Artificial Neural Networks*, Proceedings, Part II 27, Rhodes, Greece, October 4-7, 2018, Springer, 2018, pp. 195–202.
- [72] L. Jose, S. Liu, C. Russo, A. Nadort, A. Di Ieva, Generative adversarial networks in digital pathology and histopathological image processing: a review, *J. Pathol. Inform.* 12 (1) (2021) 43.
- [73] F. Prezja, L. Annala, S. Kiiskinen, S. Lahtinen, T. Ojala, Synthesizing bidirectional temporal states of knee osteoarthritis radiographs with cycle-consistent generative adversarial neural networks, *arXiv preprint*, arXiv:2311.05798, 2023.
- [74] X. Huang, A. Khetan, M. Cvitkovic, Z. Karnin, Tabtransformer: tabular data modeling using contextual embeddings, *arXiv preprint*, arXiv:2012.06678, 2020.
- [75] V. Gliozzi, G.L. Pozzato, A. Vales, Combining neural and symbolic approaches to solve the Picasso problem: a first step, *Displays* 74 (2022) 102203.
- [76] F. Prezja, L. Annala, S. Kiiskinen, T. Ojala, Exploring the efficacy of base data augmentation methods in deep learning-based radiograph classification of knee joint osteoarthritis, *Algorithms* 17 (1) (2023) 8.
- [77] E.L. Clarke, D. Treanor, Colour in digital pathology: a review, *Histopathology* 70 (2) (2017) 153–163.
- [78] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A.M. Schlitter, I. Esposito, N. Navab, Structure-preserving color normalization and sparse stain separation for histological images, *IEEE Trans. Med. Imaging* 35 (8) (2016) 1962–1971.
- [79] S. Roy, A. Kumar Jain, S. Lal, J. Kini, A study about color normalization methods for histopathology images, *Micron* 114 (2018) 42–61.
- [80] A.M. Khan, N. Rajpoot, D. Treanor, D. Magee, A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution, *IEEE Trans. Biomed. Eng.* 61 (6) (2014) 1729–1738.
- [81] P. Salehi, A. Chalechale, Pix2pix-based stain-to-stain translation: a solution for robust stain normalization in histopathology images analysis, in: *2020 International Conference on Machine Vision and Image Processing (MVIP)*, IEEE, 2020, pp. 1–7.
- [82] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: *International Conference on Machine Learning*, PMLR, 2017, pp. 1321–1330.