

Neural Architecture Search for biomedical image classification: A comparative study across data modalities

Zeki Kuş^a,*, Musa Aydin^a, Berna Kiraz^b, Alper Kiraz^c

^a Fatih Sultan Mehmet Vakıf University, Department of Computer Engineering, Istanbul, Türkiye

^b Fatih Sultan Mehmet Vakıf University, Department of Artificial Intelligence and Data Engineering, Istanbul, Türkiye

^c Koç University, Department of Physics, Electrical and Electronics Engineering, Istanbul, Türkiye

ARTICLE INFO

Keywords:

Neural Architecture Search
Biomedical image classification
Opposition-based differential evolution
MedMNIST

ABSTRACT

Deep neural networks have significantly advanced medical image classification across various modalities and tasks. However, manually designing these networks is often time-consuming and suboptimal. Neural Architecture Search (NAS) automates this process, potentially finding more efficient and effective models. This study provides a comprehensive comparative analysis of our two NAS methods, PBC-NAS and BioNAS, across multiple biomedical image classification tasks using the MedMNIST dataset. Our experiments evaluate these methods based on classification performance (Accuracy (ACC) and Area Under the Curve (AUC)) and computational complexity (Floating Point Operation Counts). Results demonstrate that BioNAS models slightly outperform PBC-NAS models in accuracy, with BioNAS-2 achieving the highest average accuracy of 0.848. However, PBC-NAS models exhibit superior computational efficiency, with PBC-NAS-2 achieving the lowest average FLOPs of 0.82 GB. Both methods outperform state-of-the-art architectures like ResNet-18 and ResNet-50 and AutoML frameworks such as auto-sklearn, AutoKeras, and Google AutoML. Additionally, PBC-NAS and BioNAS outperform other NAS studies in average ACC results (except MSTF-NAS), and show highly competitive results in average AUC. We conduct extensive ablation studies to investigate the impact of architectural parameters, the effectiveness of fine-tuning, search space efficiency, and the discriminative performance of generated architectures. These studies reveal that larger filter sizes and specific numbers of stacks or modules enhance performance. Fine-tuning existing architectures can achieve nearly optimal results without separating NAS for each dataset. Furthermore, we analyze search space efficiency, uncovering patterns in frequently selected operations and architectural choices. This study highlights the strengths and efficiencies of PBC-NAS and BioNAS, providing valuable insights and guidance for future research and practical applications in biomedical image classification.

1. Introduction

Deep neural networks have significantly advanced medical image classification, demonstrating remarkable performance across various modalities and diagnostic tasks [1,2]. The success of deep learning in medical image analysis has led to significant improvements in diagnostic accuracy, efficiency, and consistency. Convolutional neural networks (CNNs) have been successfully applied to diverse medical imaging problems, including pathology, radiology, and dermatology [3,4]. These networks have shown the ability to extract complex features from medical images, often surpassing human-level performance in specific tasks. However, CNNs are manually designed networks possessing various architectural structures and many hyper-parameters that need to be tuned. Tuning is time-consuming, and networks may need to

be redesigned for other problems. Therefore, automating the design of neural networks for different issues is crucial.

Neural Architecture Search (NAS) has become a powerful technique for automating the design of neural networks, achieving state-of-the-art performance across various tasks and domains [5–7]. The importance of NAS is that it can automatically explore many candidate architectures, finding network structures that human experts might miss. This automated approach not only saves time and resources but also has the potential to create more efficient and effective models for specific tasks. There are many NAS approaches categorized by search methods: Reinforcement-learning based [8], gradient-based [9], evolutionary based [10,11]. Most of these studies are focused on natural image classification tasks. Sun et al. [12] have introduced a new evolutionary algorithm called EvoCNN, which aims to automatically

* Corresponding author.

E-mail addresses: zkus@fsm.edu.tr (Z. Kuş), maydin@fsm.edu.tr (M. Aydin), bkiraz@fsm.edu.tr (B. Kiraz), akiraz@ku.edu.tr (A. Kiraz).

<https://doi.org/10.1016/j.artmed.2024.103064>

Received 11 July 2024; Received in revised form 26 December 2024; Accepted 30 December 2024

Available online 4 January 2025

0933-3657/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

design CNN architectures and initialize connection weights for image classification tasks. They use a flexible gene encoding scheme, unique genetic operators, and an efficient fitness evaluation approach to evolve CNN architectures and weight initialization values without requiring extensive computational resources. Experimental results on 9 benchmark datasets demonstrate that EvoCNN significantly outperforms 22 existing algorithms, achieving better classification accuracy with fewer parameters in most cases. Xue et al. [13] present a novel approach to NAS called Progressive Evaluation and Sub-Population Preservation (PEPNAS). It aims to optimize CNN designs by reducing computational costs and enhancing search efficiency. They employ a progressive evaluation strategy to incrementally increase training epochs for network candidates and a sub-population preservation strategy to maintain diversity by retaining medium and large architectures. In another study, Xue et al. [14] propose a novel hybrid approach called GENAS that combines gradient-based and evolutionary methods elements. GENAS uses a supernet to define the search space and evolves a population of candidate architectures, incorporating global evolutionary search and local gradient-based refinement. Han et al. [15] propose a novel SaDENAS algorithm to address premature convergence issues in evolutionary NAS. SaDENAS incorporates local and global search strategies balanced by an adaptive scaling factor. The method uses continuous encodings for architectures and evolves them using modified differential evolution operators. These algorithms outperform many state-of-the-art algorithms in terms of accuracy and computational efficiency on CIFAR10, CIFAR100, and ImageNet. Existing NAS methods have primarily focused on natural image classification. However, medical images have smaller regions of interest than natural images, making medical image classification problems highly challenging and leaving a critical gap in their application to biomedical image analysis. Therefore, recent NAS studies have focused on medical image analysis problems. In the context of medical image classification, several NAS studies have shown promising results [16–19]. Kwasigroch et al. [16] apply the NAS framework to malignant melanoma detection, addressing the challenge of manual architecture selection. Their approach uses hill-climbing search and network morphism operations, enabling efficient search space exploration. The method produces architectures comparable to manually designed ones but with 20 times fewer parameters for real-world medical image classification tasks. Wang et al. [17] propose MSTF-NAS, a novel evolutionary NAS method for medical image analysis. MSTF-NAS introduces a new reduction cell search space for multi-scale feature extraction and uses training-free indicators for efficient architecture evaluation. Experiments on medical image datasets show that MSTF-NAS outperforms human-designed architectures and existing NAS algorithms, demonstrating its effectiveness in this domain. Ali et al. [18] propose an evolutionary NAS approach to develop robust architectures for medical image classification. Their method uses Differential Evolution with an attention-based search space including five distinct attention mechanisms with sixteen convolution and pooling operations. Experiments on MedMNIST datasets show that this approach outperforms traditional deep learning architectures and robust NAS methods. Ali et al. [19] propose an evolutionary NAS approach for 2D and 3D medical image classification. Their method includes separate search spaces for 2D and 3D datasets, uses zero-cost proxies for efficient evaluation, and incorporates a genetic algorithm-based automatic Data Augmentation strategy. The proposed method shows competitive results on MedMNIST and BreakHis datasets compared to existing AutoML approaches. When examining these studies, it is observed that NAS for multiple data modalities in biomedical image classification is a hot topic, and designing search space and determining the search algorithm is critical for NAS approaches. Furthermore, while some NAS studies have explored medical imaging, they often lack comprehensive evaluations across diverse data modalities, analysis of search space efficiency and generated architectures, or failure to address computational efficiency—a crucial factor in real-world clinical settings.

This study aims to fill existing gaps by offering experimental insights into our Neural Architecture Search (NAS) techniques, namely PBC-NAS [20] and BioNAS [21]. The primary objective of this study is to evaluate the performance of two NAS methods, PBC-NAS and BioNAS, in biomedical image classification across multiple data modalities. We study multiple data modalities in biomedical image classification, analyzing the classification performance, computational efficiency, and generalization capabilities of our proposed search space and cell structure strategies in detail. Additionally, ablation studies are conducted to analyze the impact of key architectural parameters, such as the number of stacks, modules, and initial filters, as well as the effectiveness of fine-tuning for cross-modality datasets. The study also investigates the efficiency of the search space and the discriminative performance of the generated architectures. Furthermore, the proposed NAS methods are compared with state-of-the-art architectures, AutoML frameworks, and other NAS studies to highlight their strengths and limitations. Our previous investigations in [20,21] demonstrate the effectiveness of NAS techniques in biomedical image analysis. In this study, we extend these previous studies and perform a comprehensive comparative analysis of these methods across different data modalities, evaluating their performance, computational complexity, and generalization capabilities. We have expanded the number of evaluated datasets to 12 and added performance metrics that differ from those in other studies: F1-Score, Precision, Recall, and Specificity, to assess the robustness of the methods. We have also included an analysis of computational complexity using Floating Point Operation Counts (FLOPs) which is crucial for fast inference in the medical diagnosis. This study differs from the literature in the following aspects and is important for this field. PBC-NAS and BioNAS include a dynamic architectural structure that automatically determines the number of stacks and modules. This helps to control complexity and enhance classification performance, distinguishing our work from existing methods [22,23]. Additionally, two different encoding strategies that can work in continuous space are proposed in PBC-NAS and BioNAS for two different cell structure representations that are opposite to each other. Furthermore, novel search spaces are presented, which include a more diverse set of operations that can be selected for each node or edge compared to other studies [13–15]. Moreover, unlike other studies [15,18], the opposition-based differential evolution algorithm, known to converge faster than the classical differential evolution algorithm, is used. Unlike other studies in this field, we have analyzed two different cell-based search space designs (PBC-NAS and BioNAS) and search space diversity and evaluated their efficiency compared to NAS-Bench-101 [15,18,19]. We have deeply analyzed generated architectures and examined the frequency of selected decision parameters such as operations, initial filters, stacks, and modules. Moreover, in contrast to other studies [19], we have evaluated NAS methods based on their computational complexity alongside their classification effectiveness. These analyses underscore the importance of designing efficient and diverse search spaces that balance performance and complexity. It also provides important information about search space and cell-based architecture design for other researchers working in NAS. We utilize the MedMNIST dataset, a collection of pre-processed medical image datasets designed to benchmark machine learning algorithms, to conduct our experiments. MedMNIST includes several subsets, each focusing on a specific type of medical imaging and diagnostic task. The dataset contains diverse imaging modalities such as pathology, X-rays, dermatoscopes, optical coherence tomography (OCT), and computed tomography (CT). These subsets address various diagnostic tasks, including multi-class and binary classification, as well as ordinal regression. For this study, we focus on the 2D datasets within MedMNIST, which provide a rich and diverse set of medical images representative of real-world clinical scenarios.

Our experimental studies involve a thorough evaluation of both PBC-NAS and BioNAS methods. We compare their classification performance using metrics such as Accuracy (ACC) and Area Under the Curve (AUC), as well as their computational complexity measured in

Floating Point Operation Counts (FLOPs). The results demonstrate the effectiveness of both PBC-NAS and BioNAS methods across various biomedical image classification tasks. In a direct comparison between the two methods, BioNAS models slightly outperform PBC-NAS models on average, with BioNAS-2 achieving the highest average accuracy of 0.848. However, PBC-NAS models demonstrate superior computational efficiency, with PBC-NAS-2 achieving the lowest average FLOPs of 0.82 GB. Compared to state-of-the-art methods such as ResNet-18 and ResNet-50, both PBC-NAS and BioNAS models show superior performance. The best PBC-NAS and BioNAS models achieve average accuracies of 0.843 and 0.848, respectively, outperforming ResNet-18 (0.821) and ResNet-50 (0.821) across the 12 datasets. When compared to AutoML frameworks, our methods demonstrate robust performance. All PBC-NAS and BioNAS models reveal a range of average accuracies between 0.838 and 0.848, outperforming auto-sklearn (0.722), AutoKeras (0.813), and Google AutoML (0.809). While Google AutoML achieves a slightly higher average AUC (0.927), our methods remained highly competitive, with BioNAS-3 achieving an average AUC of 0.925. Finally, when compared to other recent NAS studies, our methods have shown highly competitive results. While MSTF-NAS achieves the highest average accuracy (0.850), BioNAS-2 (0.848) and PBC-NAS models (0.841–0.843) followed closely behind. Our methods outperform or match the performance of other NAS approaches such as DARTS, SNAS, HOPNAS, and NSGA-NET across various datasets. These findings highlight the strengths and efficiencies of each method across various datasets.

In addition to the primary experiments, we conduct several ablation studies to further understand the impact of different architectural parameters and the generalization performance of the NAS methods and to analyze the search space efficiency and best network patterns. These studies include evaluating the effect of hyper-parameter settings, evaluating performance on cross-modality datasets with fine-tuning, and analyzing the discriminative performance of the generated architectures. The ablation studies reveal that larger filter sizes and a specific number of stacks or modules lead to better performance while fine-tuning existing architectures can yield nearly optimal results without the need for separate NAS for each dataset. Additionally, the top three networks generated by both methods tend to select certain operations, such as applying activation before convolution operations, more frequently and prefer simpler architectures with fewer stacks (1). Both methods also favor larger initial filter sizes (64 and 128) and a moderate number of modules (2). Furthermore, visualization of the embeddings using t-SNE demonstrates that BioNAS produces more compact and better-separated clusters, indicating better discriminative performance. In summary, this study provides valuable insights into how effective NAS methods are for biomedical image classification. It provides guidance for future research and practical applications in this field.

The primary contributions of this paper are outlined as follows:

- This study compares two NAS methods, PBC-NAS and BioNAS, using 12 diverse biomedical image datasets from MedMNIST with recent state-of-the-art models, the AutoML framework, and NAS methods. Unlike many existing studies, this analysis evaluates classification performance (accuracy, AUC, F1-score, precision, recall, and specificity), computational complexity (FLOPs), and generalization capabilities. This comprehensive evaluation highlights the strengths and weaknesses of both methods, offering valuable insights for researchers and moving beyond simply proposing a new method to rigorously testing existing ones.
- We provide insights into the impact of architectural parameters, the effectiveness of fine-tuning for cross-modality datasets, the generalization performance of generated networks, and search space efficiency with detailed ablation studies. It provides valuable insights into why certain architectures perform better, which can inform future NAS design.

- We analyze the most frequently selected operations and architectural choices for optimal performance in biomedical image classification tasks. It reveals uncovering patterns and preferences that balance performance and complexity. This is valuable for researchers who want to design more efficient search spaces and understand the underlying principles of effective architectures. The study also identifies specific operations (like applying activation before convolution) and architectural choices (larger initial filter sizes, moderate number of modules) frequently selected for optimal performance.
- The study helps researchers choose the most appropriate NAS method based on their specific tasks and computational constraints. By sharing the source code and best models, we allow other researchers to make fair comparisons and advance their own research. This promotes reproducibility and accelerates progress in the field.

The structure of the paper is as follows: The next section provides the details of the NAS approaches used, including the search space, encoding scheme, and search algorithm. The experimental design and empirical analysis of NAS approaches on a set of biomedical image classification problems are presented in Sections 3 and 4, respectively. Furthermore, Section 5 introduces the ablation studies. Finally, Section 6 provides the conclusion.

2. Neural architecture search methods

In this section, we describe two different Neural Architecture Search (NAS) methods that are introduced in PBC-NAS [20] and BioNAS [21]. Both methods share the same architectural structure based on the cell-based network architecture developed in [22]. The architecture consists of three sequential stacks, each with three modules. These modules have a cell-based network structure represented by a directed acyclic graph (DAG). The stacks extract features from input images, halving the feature dimensions after each stack using max-pooling, except for the final pooling layer. The extracted features are then passed into the classifier to produce the classification results. Unlike the fixed number of stacks and modules in the original network structure [22], these methods dynamically determine the number of stacks and modules.

As previously mentioned, PBC-NAS and BioNAS have the same architectural structure, but differ in their search space and encoding steps. The search space used in both methods is illustrated in Fig. 1. Here, PBC-NAS represents 2D operations as nodes, while BioNAS represents them as edges. However, the encoding process shown in Fig. 2 consists of the following steps: (1) Determine the links between nodes; (2) Select 2D operations; (3) Set the values for the relevant parameters; (4) Generate the cell structures within each module; (5) Construct the modules within each stack, place them in the stacks and build the overall architecture. Detailed information specific to each method, as well as the meta-heuristic used as the search algorithm, is provided in the relevant subsections. The general structures of PBC-NAS and BioNAS are presented in Sections 2.1 and 2.2, respectively. Fig. 1 visually illustrates the overall architecture, including the stack, module, and cell structures. We have clarified how these components interact and how the dynamic architectural structure of PBC-NAS and BioNAS differs from other NAS methods (see Sections 2.1.1 and 2.2.1). The architecture encoding process for both PBC-NAS and BioNAS has been described in detail in Sections 2.1.2 and 2.2.2. Fig. 2 visually represents the encoding steps, including the mapping of continuous values to discrete architectural parameters. This ensures that the encoding process is transparent and reproducible. The steps of the Opposition-based Differential Evolution (ODE) algorithm, used as the search method, have been explained in detail in Section 2.3. Fig. 3 provides a visual representation of the search algorithm, including the opposition-based initialization, mutation, crossover, and generation jumping steps.

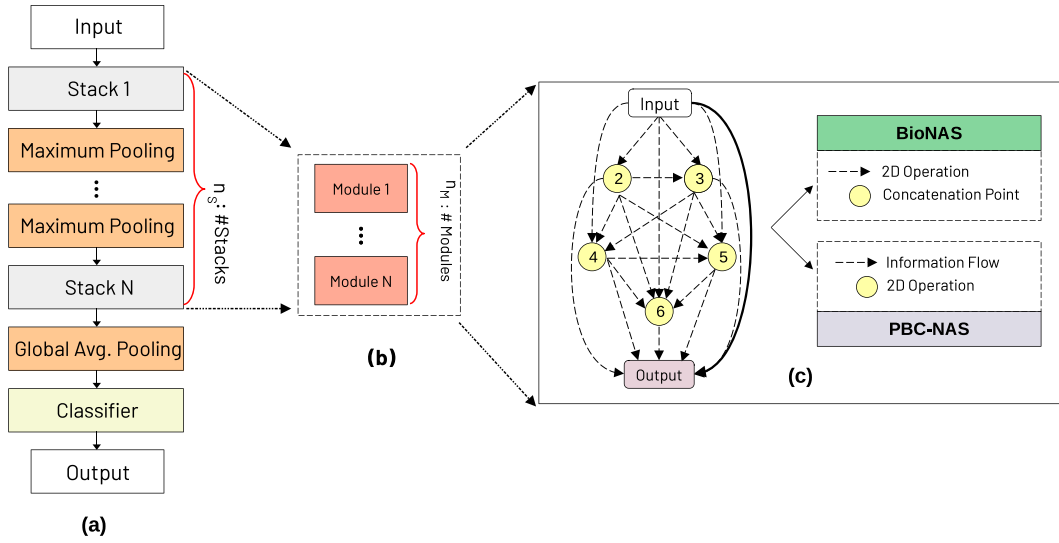


Fig. 1. (a) General structure of PBC-NAS and BioNAS architectures. (b) Structure of the modules in each stack. (c) Structure of the cells in each module. In PBC-NAS, the nodes represent 2D operations, whereas in BioNAS, the edges represent 2D operations.

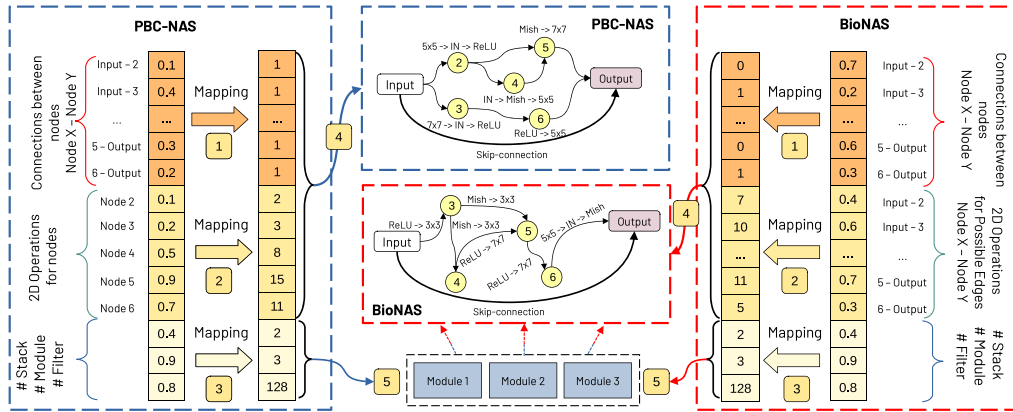


Fig. 2. Summary of the architecture encoding process: (1) Identify the connections between nodes, (2) Choose 2D operations, (3) Set the values for n_s , n_m , and n_f^{init} , (4) Create the cell structure in each module, (5) Construct modules in each stack, place them into the stacks (see Fig. 1) and construct the architecture. Blue and red arrows and dashed lines indicate mapping steps for PBC-NAS and BioNAS, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.1. PBC-NAS

2.1.1. Search space

In PBC-NAS [20], cell structures within the modules responsible for feature extraction are represented using DAGs. Each node within the DAG represents a 2D convolution operation that can be selected, while the edges indicate the information flow between these nodes. To control the rapid expansion of the solution space and to minimize the computational complexity of the generated network architectures, the number of nodes and edges within the cell is restricted. The nodes in the DAG are denoted by set N , and the edges by set E . The sizes of sets N and E are limited to 7 (including input, output, and 5 intermediate nodes) and 9, respectively. Based on superior findings on multiple NAS benchmark studies, we have identified search space parameters and restrictions relying on the NAS-Bench-101 study [22]. Moreover, the selectable 2D convolution operations for the nodes in set N are represented by set O , which consists of 15 elements. The operations in set O , chosen based on their reported success in Wei et al. [24], provide a range of combinations to enhance diversity. When choosing this collection, we selected operations that improve classification performance while reducing complexity. The search algorithm can adapt the architecture to specific biomedical image data using various kernel sizes and activation functions. Smaller kernels might

be more effective for capturing fine details, while larger kernels can capture more global patterns (Operations 1 to 6). The use of Instance Normalization helps in normalizing the feature maps, which can be beneficial in handling variations in medical images. The choice between ReLU and Mish activations provides flexibility in the non-linear transformations applied to the data, with Mish potentially offering smoother gradients compared to ReLU. Operations 7 to 9 apply Instance Normalization followed by Mish activation and then a convolution. This sequence can be particularly useful for stabilizing the training process and improving convergence, especially in deep networks. Operations 10 to 15 reverse the typical order by applying an activation function (ReLU or Mish) before the convolutional operation. This can help in scenarios where pre-activation might lead to better feature extraction or gradient flow [24]. The operations included in set O are detailed in Table 1.

The cell structure within stacks and modules, as well as the overall architecture of the network, is shown in Fig. 1. The modified cell structure ensures that input information is always relayed to the output layer, with the goal of reducing information loss that typically occurs after pooling layers [25]. In addition to the cell structure, three additional decision variables are integrated into the solution space: n_s : the number of stacks, n_m : the number of modules, and n_f^{init} : the number of initial feature maps. These variables are included in the search space

Table 1
Operations included in the set O .

#	Operation	#	Operation
1	$3 \times 3 \rightarrow IN \rightarrow ReLU$	9	$IN \rightarrow Mish \rightarrow 7 \times 7$
2	$5 \times 5 \rightarrow IN \rightarrow ReLU$	10	$ReLU \rightarrow 3 \times 3$
3	$7 \times 7 \rightarrow IN \rightarrow ReLU$	11	$ReLU \rightarrow 5 \times 5$
4	$3 \times 3 \rightarrow IN \rightarrow Mish$	12	$ReLU \rightarrow 7 \times 7$
5	$5 \times 5 \rightarrow IN \rightarrow Mish$	13	$Mish \rightarrow 3 \times 3$
6	$7 \times 7 \rightarrow IN \rightarrow Mish$	14	$Mish \rightarrow 5 \times 5$
7	$IN \rightarrow Mish \rightarrow 3 \times 3$	15	$Mish \rightarrow 7 \times 7$
8	$IN \rightarrow Mish \rightarrow 5 \times 5$		

IN: Instance Normalization
Mish, ReLU: Activation Function

to enable the creation of dynamic network structures and to manage complexity.

2.1.2. Encoding

Each solution in PBC-NAS is represented with 29 real-valued vectors. The first 26 elements are employed in constructing the cell structure, while the last 3 elements represent decision variables outside the cell structure. Each element in the solution vector can take continuous values within the range of $[0, 1]$. The initial 21 elements indicate the connections between nodes within the cell, where a value of 1 denotes a connection, and 0 denotes no connection. Accordingly, the range $[0, 1]$ is divided into discrete bins (2 for the first 21 elements), with values in $[0, 0.5]$ mapped to 1, and values in $[0.5, 1]$ mapped to 0. The next 5 elements denote the chosen 2D convolution operations assigned to nodes from set O . Table 1 shows that set O contains 15 elements, so the range $[0, 1]$ is divided into 15 parts. For example, values in the range $[0, 0.06]$ select the $3 \times 3 \rightarrow IN \rightarrow ReLU$, and values in $[0.06, 0.13]$ select the $5 \times 5 \rightarrow IN \rightarrow ReLU$. The last 3 elements in the vector specify the decision variables $n_S \in \{1, 2, 3\}$, $n_M \in \{1, 2, 3\}$ and $n_f^{init} \in \{32, 64, 128\}$. As there are 3 possible values for each of these decision variables, the range $[0, 1]$ is split into 3 parts, and continuous values are converted into discrete values using similar mapping processes. The network architecture is constructed for evaluation after applying the mapping process for all 29 elements in the solution vector. The encoding steps, indicated with blue arrows, are illustrated in Fig. 2.

2.2. BioNAS

2.2.1. Search space

Similar to PBC-NAS, DAGs in BioNAS are utilized to illustrate the cell structure within each module used for feature extraction (Fig. 1). In PBC-NAS, each DAG includes a specified number of nodes and edges, with edges representing the flow of information and nodes signifying the 2D operations employed for feature extraction. In contrast, BioNAS modifies this representation by assigning 2D operations to edges, while nodes indicate the points where feature maps from different nodes are combined (see Fig. 1(c)). Consequently, more features are extracted by assigning 2D operations to edges. Similar to PBC-NAS, in BioNAS, the size of the E and N is limited to 7 and 9, respectively. Additionally, 15 different 2D operations (see Table 1) can be chosen for each edge in DAG. Moreover, the decision variables explained in Section 2.1 are also included in the NAS and the same rules apply for determining the search space here.

2.2.2. Encoding

In BioNAS, each network architecture/solution is represented by a 45-dimensional real-valued vector. This vector's first 42 elements define the cell structure; the decision variables, n_S , n_M , and n_f^{init} , are represented with its last three values. A continuous value in $[0, 1]$ can be assigned to each entry in the solution vector. Potential connections between nodes are represented by the first 21 elements, where 1 signifies a connection, and 0 signifies no connection. Therefore, a

mapping method is required to transform these continuous values into discrete ones. The proposed mapping method splits the $[0, 1]$ range into bins, with the number of bins corresponding to the number of possible discrete values. For instance, for the first 21 elements, which have two possible values, the range $[0, 1]$ is split into two bins: if the continuous value falls within $[0, 0.5]$, it is set to 1; otherwise, if it falls within $[0.5, 1]$, it is set to 0. This mapping scheme determines the node connections. The next 21 values, which denote the chosen 2D operations for these connections, are then mapped using the same method. For these connections, there are fifteen distinct selectable operations ($|O|$), so the range is split into fifteen bins, each of which corresponds to a particular operation in $|O|$: for values between $[0.13, 0.20]$, choose $7 \times 7 \rightarrow IN \rightarrow ReLU$; for values between $[0.20, 0.26]$, choose $3 \times 3 \rightarrow IN \rightarrow Mish$. The mapping technique is also used for the next 21 elements, which determines the 2D operations for possible edges. The vector's last three values, n_S , n_M and n_f^{init} , have three distinct values for each decision variable. As a result, the range is split into three bins, and the associated discrete value is chosen accordingly. This mapping scheme is employed to map the 45 continuous values into discrete ones, leading to the construction of the network architecture. The steps of the architecture encoding (mapping scheme), indicated with red arrows, are illustrated in Fig. 2.

2.3. Search algorithm

In this work, we employed the Opposition-based Differential Evolution (ODE) algorithm as the search method. ODE [26] is the enhanced version of the Differential Evolution (DE) algorithm, integrating the concept of opposition-based learning (OBL) into the population initialization and generation steps. OBL considers the candidate solution itself as well as their opposites during the search process. The opposite solutions provide additional insights into the search space, potentially leading to improved exploration and faster convergence.

We considered the ODE used in NAS studies [27]. This ODE operates through the following steps: (1) opposition-based initialization; (2) the operators in DE, i.e. mutation, crossover, and selection; (3) the opposition-based generation jumping step with a probability. The search process starts by randomly initializing the population as in the traditional DE, and then computing the opposites of initial solutions. The candidate solutions and their opposites are then evaluated and merged. The algorithm selects the superior solutions from the merged population to form the initial population. All these steps correspond to the opposition-based initialization in ODE. The algorithm then proceeds with the operators in the classical variant of DE (DE/rand/1/bin) [28]. Moreover, it performs opposition-based generation jumping step with a probability. In this step, the opposites of candidate solutions in the current population are computed and evaluated. The current population and its opposites are merged and better solutions are selected to replace the current population.

The overall steps of ODE are illustrated in Fig. 3. The candidate solution is represented by a real-valued vector. Then, it is converted into the corresponding network architecture described in Sections 2.1.2 and 2.2.2. The fitness of each candidate solution is defined as the validation accuracy. The implementation details are given in Section 3.2.

3. Experimental design

This section outlines the experimental design for evaluating the NAS methods, PBC-NAS and BioNAS, on biomedical image classification tasks. Section 3.1 describes the dataset properties, including imaging modalities, class distributions, and train/validation/test splits, as well as the preprocessing steps applied. Section 3.2 details the training procedures, divided into Approximate Evaluation and Final Evaluation phases, with strategies like early stopping and reduced training data for computational efficiency. Section 3.3 defines the performance metrics, including Accuracy (ACC), Area Under the Curve (AUC), F1-Score, Precision, Recall, Specificity, and computational complexity measured in Floating Point Operation Counts (FLOPs).

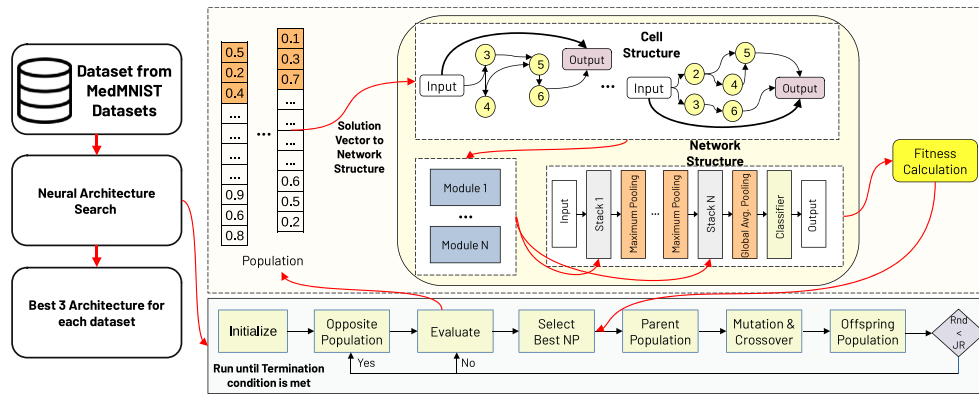


Fig. 3. Overall steps of the opposition-based differential evolution algorithm. The steps begin with initializing a population of neural network solutions, constructing network structures from these solutions, and evaluating their fitness on selected datasets. ODE performs the following steps: Generate the opposite population, evaluate fitness, choose the best solutions, perform mutation and crossover, and update the population until the termination condition is met.

3.1. Dataset

We have used the MedMNIST dataset [2] to evaluate and compare the performance of various methods used in the experiments. MedMNIST is a set of pre-processed medical image datasets designed to benchmark machine learning algorithms. The dataset includes several subsets, each focusing on a specific type of medical imaging and diagnostic task. Table 2 gives an overview of the different datasets included in MedMNIST with their corresponding properties. Each subset in MedMNIST is chosen to provide a diverse set of medical images representative of real-world clinical scenarios. The datasets contain various imaging modalities such as pathology, X-rays, dermatoscopes, optical coherence tomography (OCT), and computed tomography (CT), and address a variety of diagnostic tasks from multi-class and binary classification to ordinal regression. These datasets include PathMNIST, which focuses on colorectal cancer histology images with 9 classes; ChestMNIST, featuring chest X-ray images for multi-label classification of thoracic diseases across 14 classes; and DermaMNIST, which contains dermatoscope images of skin lesions categorized into 7 classes. OCTMNIST presents optical coherence tomography images of retinal diseases with 4 classes, while PneumoniaMNIST offers chest X-ray images for binary classification of pneumonia. RetinaMNIST provides fundus images for diabetic retinopathy grading with 5 classes in an ordinal regression task, and BreastMNIST features breast ultrasound images for binary classification of tumors. The dataset also includes BloodMNIST, containing blood cell microscopy images across 8 classes, and TissueMNIST, which focuses on kidney cortex microscopy images with 8 classes. Lastly, OrganAMNIST, OrganCMNIST, and OrganSMNIST all contain abdominal CT images for multi-organ segmentation, each with 11 classes. MedMNIST contains both 2D and 3D datasets, but in this study, we have focused on the 2D datasets. We have used publicly shared train/validation/test splits in MedMNIST for our experimental studies and have not applied any augmentation and preprocessing steps.

3.2. Implementation details

This section details the training and evaluation steps in the NAS methodology. The training and evaluation steps are split into Approximate and Final Evaluation. During the Approximate Evaluation phase, we aimed to quickly evaluate a large number of potential solutions. For datasets containing fewer than 15,000 samples, we used 50% of the training data. For larger datasets, we utilized 10% of the total training samples. We have also used the same split ratios for the validation set to evaluate the approximate performance of the candidate solution (fitness value). The EarlyStopping mechanism is employed to reduce training time and evaluate more solutions. Each candidate solution is trained

for a maximum of 120 epochs, and training is stopped if there is no improvement in the validation loss value for 30 consecutive epochs. This significantly reduced training times, allowing us to evaluate various neural architectures without exhausting computational resources and time. Following the Approximate evaluation (end 500 solutions are evaluated on GPU), we conducted a Final Evaluation for the best three architectures obtained from the Approximate evaluation phase. In the final evaluation, the models are trained with entire training and validation sets during 120 epochs, and their actual performance is evaluated on the whole test set. We have used the following hyperparameters: learning rate: $1e-2$, batch size: 128, and optimizer: Adam. We have also used the Pytorch library for all implementations and have not employed any data augmentation techniques. The source code and best models are shared on [GitHub](#).

3.3. Performance measures

This study evaluates classification performance using Accuracy (ACC), area under the Curve (AUC), F1-Score, Precision, Recall, and Specificity metrics. These metrics are widely used to measure and compare the model's performance. Accuracy is one of the most commonly used evaluation metrics in classification problems. The accuracy of a model is defined as the ratio of correctly predicted samples to total samples. Accuracy gives a general idea of the model's overall performance and provides a simple and understandable metric to measure classification success. On the other hand, AUC represents the area under the ROC curve and summarizes the model's classification performance. AUC evaluates the model's ability to distinguish between positive and negative classes, with a value between 0.5 and 1. AUC is especially useful for unbalanced datasets as it measures overall performance without dependence on a threshold to evaluate class discrimination. F1-Score, Precision, and Recall are included to address the challenges of imbalanced datasets and to evaluate the trade-offs between false positives and false negatives, which are critical in biomedical applications. Specificity is added to evaluate the models' ability to correctly identify negative instances, which is particularly important in reducing over-diagnosis and unnecessary treatments. We have also measured the complexity of the optimal networks obtained from NAS in terms of the Floating Point Operation Counts (FLOPs). It evaluates the computational complexity of a model; higher FLOPs indicate the more complex models that potentially run slower [27].

4. Results and discussion

This study provides a comprehensive evaluation of two NAS methods, namely PBC-NAS and BioNAS. The experiments are conducted in four parts. First, we compare the two methods in terms of classification

Table 2

Overview of the 2D MedMNIST datasets used in this study.

Dataset	Modality	# Classes/Labels	# Samples	# Training/Validation/Test
PathMNIST	Colon Pathology	Multi-Class (9)	107,180	89,996/10,004/7180
ChestMNIST	Chest X-ray	Multi-Label (14) Binary-Class (2)	112,120	78,468/11,219/22,433
DermaMNIST	Dermatoscope	Multi-Class (7)	10,015	7007/1003/2005
OCTMNIST	Retinal OCT	Multi-Class (4)	109,309	97,477/10,832/1000
PneumoniaMNIST	Chest X-ray	Binary-Class (2)	5856	4,708/524/624
RetinaMNIST	Fundus Camera	Ordinal Regression (5)	1600	1,080/120/400
BreastMNIST	Breast Ultrasound	Binary-Class (2)	780	546/78/156
BloodMNIST	Blood Cell Microscope	Multi-Class (8)	17,092	11,959/1712/3421
TissueMNIST	Kidney Cortex Microscope	Multi-Class (8)	236,386	165,466/23,640/47,280
OrganAMNIST	Abdominal CT	Multi-Class (11)	58,830	34,561/6491/17,778
OrganCMNIST	Abdominal CT	Multi-Class (11)	23,583	12,975/2392/8216
OrganSMNIST	Abdominal CT	Multi-Class (11)	25,211	13,932/2452/8827

Table 3

Results of different methods on the MedMNIST benchmark in terms of accuracy. AVG. denotes the average performance over the 12 subsets.

Methods	Path	Chest	Derma	OCT	Pneum	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	AVG
ResNet-18 (28)	0.907	0.947	0.735	0.743	0.854	0.524	0.863	0.958	0.676	0.935	0.900	0.782	0.819
ResNet-18 (224)	0.909	0.947	0.754	0.763	0.864	0.493	0.833	0.963	0.681	0.951	0.920	0.778	0.821
ResNet-50 (28)	0.911	0.947	0.735	0.762	0.854	0.528	0.812	0.956	0.680	0.935	0.905	0.770	0.816
ResNet-50 (224)	0.892	0.948	0.731	0.776	0.884	0.511	0.842	0.950	0.680	0.947	0.911	0.785	0.821
auto-sklearn	0.716	0.779	0.719	0.601	0.855	0.515	0.803	0.878	0.532	0.762	0.829	0.672	0.722
AutoKeras	0.834	0.937	0.749	0.763	0.878	0.503	0.831	0.961	0.703	0.905	0.879	0.813	0.813
Google AutoML	0.728	0.948	0.768	0.771	0.946	0.531	0.861	0.966	0.673	0.886	0.877	0.749	0.809
DARTS	0.872	0.934	0.749	0.712	0.874	0.510	0.832	0.953	0.648	0.926	0.791	0.808	0.801
SNAS	0.850	0.938	0.737	0.708	0.871	0.515	0.811	0.946	0.708	0.918	0.891	0.778	0.806
HOPNAS	0.912	0.947	0.759	0.761	0.852	0.523	0.853	0.958	0.698	0.937	0.911	0.803	0.826
NSGA-NET	0.866	0.947	0.744	0.765	0.907	0.540	0.846	0.970	0.712	0.952	0.923	0.820	0.833
Random Search	0.854	0.946	0.773	0.760	0.904	0.542	0.897	0.966	0.717	0.955	0.923	0.820	0.838
MSTF-NAS	0.910	0.945	0.774	0.780	0.912	0.550	0.872	0.976	0.740	0.962	0.936	0.838	0.850
PBC-NAS-1	0.908	0.947	0.784	0.783	0.875	0.560	0.891	0.972	0.687	0.943	0.928	0.812	0.841
PBC-NAS-2	0.914	0.948	0.785	0.801	0.875	0.545	0.917	0.975	0.669	0.943	0.929	0.812	0.843
PBC-NAS-3	0.913	0.947	0.787	0.784	0.877	0.540	0.897	0.974	0.695	0.946	0.927	0.812	0.842
BioNAS-1	0.920	0.947	0.780	0.796	0.881	0.545	0.891	0.973	0.678	0.943	0.931	0.805	0.841
BioNAS-2	0.926	0.947	0.798	0.819	0.901	0.545	0.891	0.974	0.688	0.944	0.923	0.820	0.848
BioNAS-3	0.912	0.947	0.776	0.793	0.869	0.553	0.865	0.976	0.688	0.944	0.934	0.798	0.838

performance (ACC and AUC) and time complexity (FLOPs). Second, we benchmark them against ResNet models to evaluate their performance with state-of-the-art architectures. Third, we compare PBC-NAS and BioNAS with various AutoML frameworks to evaluate their effectiveness compared to automated machine learning solutions. Finally, we expand our experiments by comparing them with other NAS studies. All these experimental results, together with the results of the compared algorithms reported in [17], are presented in Tables 3 and 4. In these tables, PBC-NAS-1, 2, 3 and BioNAS-1, 2, 3 indicate the top three models obtained from PBC-NAS and BioNAS for each dataset. Further results and discussion specific to each of the experiments are given in the relevant subsections.

4.1. Comparison between the PBC-NAS and the BioNAS

In this part of the experiment, we compare the performance of PBC-NAS and BioNAS models in terms of accuracy, AUC, and computational complexity.

The results show that both the PBC-NAS and BioNAS models are highly competitive in terms of accuracy metrics (see Table 3). On average, BioNAS models slightly outperform the PBC-NAS models, with BioNAS-2 achieving the highest average accuracy of 0.848, closely followed by PBC-NAS-2 and PBC-NAS-3. For individual datasets, the BioNAS models achieve top results for 8 out of the 12 datasets, with BioNAS-2 performing best in the Path, Derma, OCT, Pneum, and OrganS datasets. Additionally, PBC-NAS models perform strongly in the Chest and Breast datasets, with PBC-NAS-2 achieving top results. Furthermore, there is a slight difference between the best results obtained by PBC-NAS and BioNAS models in Retina, Blood, and Organs (OrganA,

OrganC, OrganS) datasets, highlighting their comparable effectiveness across various datasets.

On the other hand, the average AUC scores for all datasets show that the PBC-NAS and BioNAS models perform similarly well. Across the 12 datasets, BioNAS models achieve the best results in 5 datasets, have equivalent results in 4, and PBC-NAS models outperform in the remaining 3. The BioNAS-3 and PBC-NAS-3 models have the highest average AUC scores. In individual dataset analysis, BioNAS-2 and BioNAS-3 perform the best in the Path dataset, with AUC scores of 0.993 and 0.991. However, both PBC-NAS and BioNAS models demonstrate lower performance in the Chest dataset, indicating a potential area for improvement. In the Derma, PBC-NAS-1 achieves the highest AUC score of 0.941, demonstrating its strength in dermatological data analysis. For the OCT, BioNAS-1 and BioNAS-2 give the top two results. In the Breast, BioNAS-1 leads with an AUC of 0.916, followed closely by the PBC-NAS-3 model (0.915). All PBC-NAS and BioNAS models achieve perfect AUC scores of 0.999 in the Blood dataset, demonstrating their robustness and reliability in blood cell classification. They also achieve the same results for OrganA, with only slight differences observed in the OrganC and OrganS datasets.

The comparison of the two methods in terms of computational complexity is provided in Table 5 and shown in Fig. 4. Overall, the PBC-NAS models demonstrate superior computational efficiency compared to the BioNAS models. Specifically, PBC-NAS-2 achieves the lowest average FLOPs of 0.82, closely followed by PBC-NAS-3 with 0.85. In contrast, the BioNAS models have higher average FLOPs, with BioNAS-2 and BioNAS-3 averaging 1.54 and 1.63, respectively. These results suggest that the PBC-NAS models generally require fewer computational resources, making them more efficient. PBC-NAS-2 is often the most efficient model, notably in the Chest (0.16), OCT (0.28),

Table 4

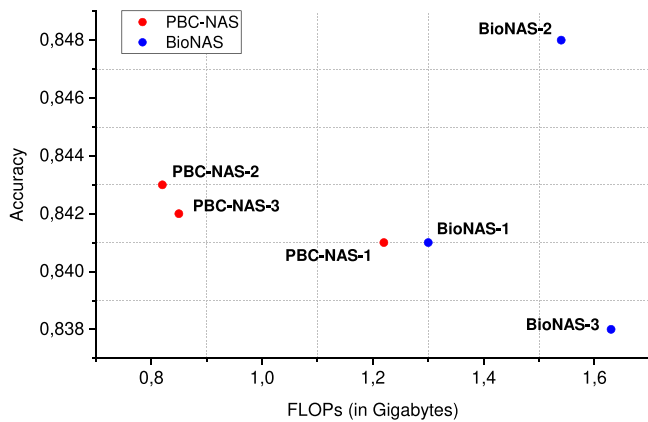
Results of different methods on the MedMNIST benchmark in terms of AUC. AVG. denotes the average performance over the 12 subsets.

Methods	Path	Chest	Derma	OCT	Pneum	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	AVG
ResNet-18 (28)	0.983	0.768	0.917	0.943	0.944	0.717	0.901	0.998	0.930	0.997	0.992	0.972	0.922
ResNet-18 (224)	0.989	0.773	0.920	0.958	0.956	0.710	0.891	0.998	0.933	0.998	0.994	0.974	0.925
ResNet-50 (28)	0.990	0.769	0.913	0.952	0.948	0.726	0.857	0.997	0.931	0.997	0.992	0.972	0.920
ResNet-50 (224)	0.989	0.773	0.912	0.958	0.962	0.716	0.866	0.997	0.932	0.998	0.993	0.975	0.923
auto-sklearn	0.934	0.649	0.902	0.887	0.942	0.690	0.836	0.984	0.828	0.963	0.976	0.945	0.878
AutoKeras	0.959	0.742	0.915	0.955	0.947	0.719	0.871	0.998	0.941	0.994	0.990	0.974	0.917
Google AutoML	0.944	0.778	0.914	0.963	0.991	0.750	0.919	0.998	0.924	0.990	0.988	0.964	0.927
DARTS	0.975	0.732	0.913	0.953	0.965	0.742	0.912	0.994	0.901	0.987	0.969	0.910	0.913
SNAS	0.969	0.733	0.906	0.949	0.974	0.753	0.894	0.996	0.921	0.979	0.927	0.952	0.913
HOPNAS	0.987	0.763	0.899	0.948	0.971	0.770	0.907	0.996	0.913	0.995	0.998	0.975	0.927
NSGA-NET	0.979	0.779	0.915	0.958	0.965	0.759	0.857	0.999	0.942	0.999	0.993	0.978	0.927
Random Search	0.980	0.774	0.923	0.956	0.963	0.750	0.921	0.999	0.944	0.999	0.994	0.982	0.932
MSTF-NAS	0.990	0.791	0.934	0.968	0.963	0.755	0.930	0.999	0.951	0.999	0.996	0.983	0.938
PBC-NAS-1	0.976	0.693	0.941	0.956	0.975	0.728	0.895	0.999	0.931	0.998	0.996	0.978	0.922
PBC-NAS-2	0.976	0.689	0.937	0.965	0.971	0.724	0.898	0.999	0.924	0.998	0.995	0.981	0.921
PBC-NAS-3	0.977	0.672	0.935	0.962	0.975	0.738	0.915	0.999	0.934	0.998	0.995	0.978	0.923
BioNAS-1	0.978	0.667	0.926	0.966	0.972	0.732	0.916	0.999	0.927	0.998	0.996	0.976	0.921
BioNAS-2	0.993	0.656	0.941	0.962	0.969	0.735	0.873	0.999	0.930	0.998	0.995	0.979	0.919
BioNAS-3	0.991	0.685	0.935	0.965	0.977	0.749	0.898	0.999	0.931	0.998	0.996	0.979	0.925

Table 5

Complexity comparison of the PBC-NAS and BioNAS models in terms of Floating Point Operations in Gigabytes.

Methods	Path	Chest	Derma	OCT	Pneum	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	AVG
PBC-NAS-1	1.11	0.16	0.24	0.96	0.64	0.06	4.02	0.24	1.27	2.59	0.75	2.60	1.22
PBC-NAS-2	0.26	0.20	0.25	0.28	0.79	0.55	0.32	0.47	0.04	1.31	2.23	3.17	0.82
PBC-NAS-3	1.31	0.55	0.24	0.53	0.13	0.70	0.48	0.35	0.87	1.87	0.28	2.90	0.85
BioNAS-1	1.11	1.27	0.22	1.98	0.73	2.37	0.81	2.78	0.13	1.04	1.29	1.90	1.30
BioNAS-2	3.63	0.48	1.62	2.78	2.24	1.49	0.49	0.38	0.58	0.75	2.19	1.90	1.54
BioNAS-3	1.31	1.29	0.44	0.93	0.66	0.16	4.02	1.74	0.61	0.75	3.89	3.71	1.63

**Fig. 4.** Comparison of PBC-NAS and BioNAS models in terms of accuracy and computational cost (FLOPs in gigabytes). The average accuracy and FLOPs results reported in Tables 3 and 5 are used.

Pneum (0.13), Retina (0.06), and Tissue (0.04) datasets. On the other hand, while the BioNAS models are competitive in some datasets, they generally achieve higher computational costs, particularly in datasets such as OCT (BioNAS-2 at 2.78) and OrganS (BioNAS-3 at 3.71). For the Chest dataset, PBC-NAS-1 and PBC-NAS-2 are the most computationally efficient models, significantly outperforming the BioNAS models. In the Derma dataset, PBC-NAS-2 and PBC-NAS-3 show competitive results with 0.25 and 0.24 FLOPs, respectively, while BioNAS-1 leads slightly in efficiency with 0.22 FLOPs. PBC-NAS models give less complex models for 9 of the 12 datasets and the best two results for 7 of the 12 datasets.

Overall, the PBC-NAS models consistently show lower FLOPs across most datasets, highlighting their computational efficiency. Therefore,

the PBC-NAS models, especially PBC-NAS-2, are preferable for applications requiring computational efficiency. Conversely, the BioNAS models, while competitive in certain datasets, generally require more computational resources. These findings highlight the PBC-NAS models, especially PBC-NAS-2, as preferable choices for applications where computational efficiency is crucial.

We have compared PBC-NAS and BioNAS methods in terms of average ACC and AUC obtained from five different runs (using 5 different seed values) on 12 datasets to highlight the head-to-head performance of the two methods. The results are presented in Fig. 5, and they show slight differences in terms of ACC and AUC. Both methods perform well for the Blood dataset, with BioNAS slightly edging out PBC-NAS in accuracy (0.974 vs. 0.973) while both achieve a perfect AUC of 0.999. This trend of BioNAS slightly outperforming PBC-NAS continues across most datasets. For instance, in the OrganA dataset, BioNAS achieves an accuracy of 0.942 compared to BioNAS's 0.938, with both methods maintaining an AUC of 0.998. In OCT, BioNAS obtains slightly better accuracy than PBC-NAS (0.779 vs. 0.771) with similar AUC (0.967 vs. 0.966). In the Chest dataset, PBC-NAS shows higher accuracy (0.947 vs. 0.945) but slightly worse AUC (0.694 vs. 0.697). The Path dataset highlights a more noticeable difference in AUC, with BioNAS achieving 0.992 compared to PBC-NAS's 0.977, alongside a higher accuracy (0.917 vs. 0.905). This pattern persists in the Pneum dataset, where BioNAS outperforms with 0.878 accuracy compared to PBC-NAS's 0.861. For the OrganA and OrganC datasets, BioNAS consistently outperforms PBC-NAS in accuracy and AUC with smaller margins. Notably, in the OrganS dataset, both methods achieve the same accuracy of 0.818, but PBC-NAS has a slightly higher AUC (0.981 vs. 0.979). In the Retina and Tissue datasets, PBC-NAS demonstrates better performance with higher accuracy. Specifically, in the Tissue dataset, PBC-NAS achieves an accuracy of 0.691 and an AUC of 0.934, surpassing BioNAS's 0.686 accuracy and 0.931 AUC. In the Retina dataset, PBC-NAS records an accuracy of 0.546, outperforming BioNAS's 0.536 accuracy. We have also reported standard deviation values obtained from five different runs using error bars in Fig. 5.

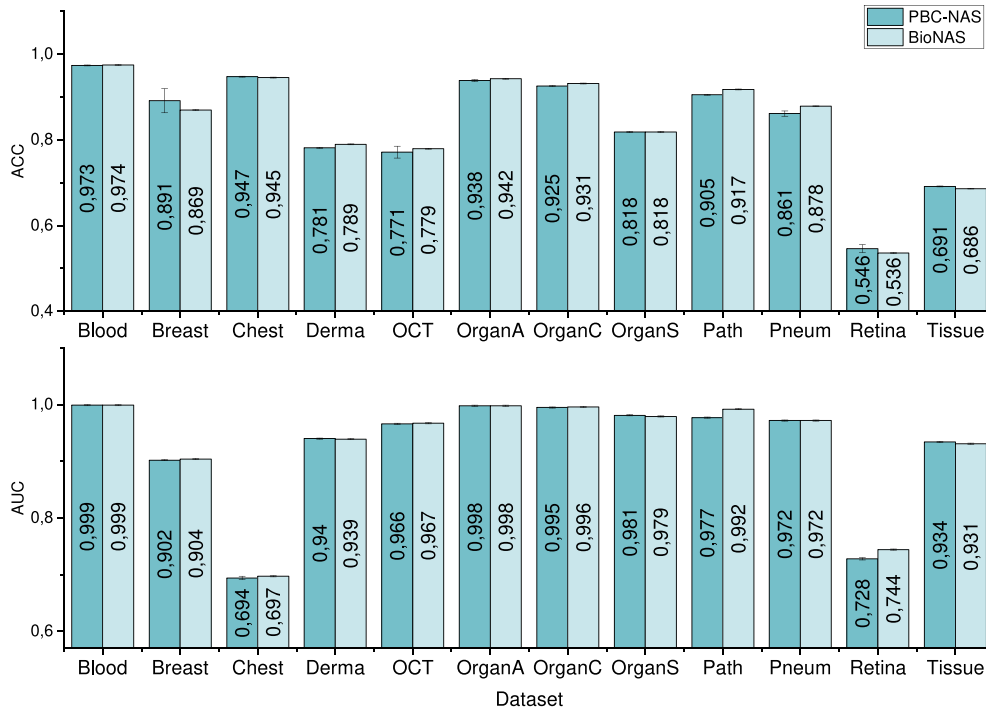


Fig. 5. Comparison of PBC-NAS and BioNAS performance across 12 medical imaging datasets. The figure displays average Accuracy (ACC) and Area Under the Curve (AUC) scores from five independent runs, with error bars representing standard deviations.

As shown in the figure, the standard deviation values are very low, indicating that the methods give stable results. In summary, BioNAS outperforms PBC-NAS in 7 of the 11 datasets in terms of ACC, achieves better performance for 6 in terms of AUC, and gives the same results for 3 datasets.

Table 6 compares the performance of PBC-NAS and BioNAS models across various datasets using key metrics: F1-Score, Precision, Recall, and Specificity. Each metric is reported as an average from five different runs, along with the standard deviation, providing insights into the consistency and reliability of each model's performance. F1-Score, which balances precision and recall, is a critical metric for evaluating the overall effectiveness of a model in handling both false positives and false negatives. BioNAS models perform well across most datasets, with the highest F1-Score for the Path, Derma, Pneum, OCT, Blood, OrganA and OrganS datasets. On the other hand, although PBC-NAS models perform well in certain datasets, such as the Breast, Retina, Tissue, and OrganS datasets, they generally do not exhibit the consistently high performance of BioNAS models across a broader range of datasets. This suggests that BioNAS models are more versatile and effective in achieving a balanced performance, making them a more reliable choice for tasks where both precision and recall are crucial. In terms of precision and recall, BioNAS models generally outperform PBC-NAS models across most datasets, indicating a more balanced and effective performance. Precision measures the accuracy of positive predictions, and BioNAS models consistently achieve higher precision scores in seven of the datasets. This suggests that BioNAS models are more adept at minimizing false positives, thereby making more accurate positive predictions. However, while PBC-NAS models show competitive precision in certain datasets, they often achieve lower precision levels than those achieved by BioNAS models. Regarding recall, which evaluates the model's ability to capture all relevant instances, BioNAS models again demonstrate superior performance. BioNAS-2 and BioNAS-3 frequently achieve the highest recall scores, indicating their effectiveness in identifying true positives and minimizing false negatives. PBC-NAS models, although effective in specific scenarios like the Breast dataset, generally do not match the recall performance of BioNAS models. This consistent performance in both precision and recall underscores

the robustness and reliability of BioNAS models in achieving accurate and comprehensive predictions. The BioNAS models generally perform better than PBC-NAS models in terms of specificity across most datasets. This means they are better at correctly identifying negative instances and reducing false positives. In particular, BioNAS-2 and BioNAS-3 achieve higher specificity scores, suggesting that BioNAS models are more effective at distinguishing between relevant and irrelevant instances. While PBC-NAS models perform adequately, they often do not match the specificity levels of BioNAS models, indicating a higher tendency for false positives. Overall, the superior specificity of BioNAS models highlights their robustness and reliability in applications where minimizing false positives is important.

As a result, we have comprehensively compared the proposed methods using various metrics and analyses in this subsection. Firstly, we have compared the proposed methods based on ACC and AUC, referencing studies in the literature (see Tables 3 and 4). Subsequently, we have assessed model complexity in terms of FLOPs as shown in Table 5. Additionally, Fig. 5 illustrates the head-to-head comparison for both methods using the mean and standard deviation values derived from five different runs across all datasets. In Table 6, we have further compared the proposed methods using F1-score, precision, recall, and specificity, discussing these metrics throughout the subsection. We have also performed the Wilcoxon signed rank test (with p -value = 0.05) using the ACC and AUC values obtained by running the best-reported models (see Tables 3 and 4 for each dataset) for 5 different seed values (180 data points = Top three models $\times$ Five different runs $\times$ Twelve different datasets). It is a non-parametric statistical test used to compare paired samples, which does not assume a normal distribution of the data. The test evaluates whether there is a statistically significant difference between the paired results of two methods, providing a reliable basis for our comparative analysis. The Wilcoxon signed-rank test for the results in Table 6 is also conducted. The following p -values are obtained for ACC, AUC, F1-Score, Recall, Precision, and Specificity: 0.2726, 0.6548, 0.3875, 0.3408, 0.8541, and 0.6952. The results indicate no statistically significant difference between the two methods in terms of ACC, AUC, F1-Score, Recall, Precision, and Specificity, as p -values are greater than the significance

Table 6

Comparison of the PBC-NAS and BioNAS models in terms of performance metrics for various datasets. The best value for each metric within each dataset is highlighted in bold.

Dataset	F1-Score					
	PBC-NAS-1	PBC-NAS-2	PBC-NAS-3	BioNAS-1	BioNAS-2	BioNAS-3
Path	0.902 ± 0.001	0.904 ± 0.003	0.905 ± 0.005	0.913 ± 0.002	0.917 ± 0.003	0.91 ± 0.001
Derma	0.778 ± 0.002	0.781 ± 0.001	0.779 ± 0.002	0.776 ± 0.001	0.789 ± 0.007	0.773 ± 0.001
Pneum	0.856 ± 0.039	0.861 ± 0.006	0.859 ± 0.017	0.868 ± 0.047	0.878 ± 0.035	0.857 ± 0.01
OCT	0.770 ± 0.014	0.769 ± 0.029	0.771 ± 0.02	0.778 ± 0.036	0.779 ± 0.042	0.77 ± 0.018
Retina	0.546 ± 0.01	0.531 ± 0.012	0.52 ± 0.021	0.533 ± 0.009	0.536 ± 0.018	0.535 ± 0.017
Breast	0.874 ± 0.019	0.891 ± 0.028	0.971 ± 0.001	0.869 ± 0.012	0.863 ± 0.034	0.859 ± 0.003
Blood	0.968 ± 0.001	0.973 ± 0.001	0.971 ± 0.001	0.971 ± 0.001	0.970 ± 0.001	0.974 ± 0.001
Tissue	0.686 ± 0.001	0.668 ± 0.001	0.691 ± 0.001	0.676 ± 0.001	0.685 ± 0.001	0.686 ± 0.001
OrganA	0.936 ± 0.005	0.938 ± 0.002	0.937 ± 0.009	0.934 ± 0.009	0.94 ± 0.001	0.942 ± 0.001
OrganC	0.927 ± 0.001	0.928 ± 0.001	0.927 ± 0.001	0.929 ± 0.001	0.92 ± 0.001	0.931 ± 0.001
OrganS	0.803 ± 0.002	0.818 ± 0.001	0.802 ± 0.003	0.801 ± 0.001	0.818 ± 0.001	0.794 ± 0.003
Dataset	Precision					
	PBC-NAS-1	PBC-NAS-2	PBC-NAS-3	BioNAS-1	BioNAS-2	BioNAS-3
Path	0.870 ± 0.002	0.878 ± 0.001	0.878 ± 0.002	0.89 ± 0.004	0.891 ± 0.005	0.886 ± 0.0
Derma	0.624 ± 0.019	0.647 ± 0.017	0.610 ± 0.026	0.614 ± 0.16	0.643 ± 0.026	0.641 ± 0.178
Pneum	0.903 ± 0.009	0.904 ± 0.002	0.903 ± 0.004	0.909 ± 0.009	0.913 ± 0.016	0.905 ± 0.005
OCT	0.820 ± 0.007	0.820 ± 0.008	0.823 ± 0.007	0.828 ± 0.004	0.821 ± 0.01	0.817 ± 0.011
Retina	0.359 ± 0.143	0.367 ± 0.092	0.336 ± 0.069	0.346 ± 0.28	0.343 ± 0.139	0.313 ± 0.093
Breast	0.866 ± 0.018	0.865 ± 0.043	0.828 ± 0.156	0.843 ± 0.018	0.833 ± 0.05	0.84 ± 0.029
Blood	0.966 ± 0.001	0.972 ± 0.001	0.970 ± 0.001	0.969 ± 0.001	0.968 ± 0.001	0.973 ± 0.001
Tissue	0.620 ± 0.002	0.602 ± 0.002	0.633 ± 0.018	0.612 ± 0.005	0.625 ± 0.002	0.627 ± 0.01
OrganA	0.937 ± 0.009	0.939 ± 0.002	0.935 ± 0.011	0.932 ± 0.01	0.941 ± 0.002	0.942 ± 0.001
OrganC	0.922 ± 0.001	0.922 ± 0.002	0.922 ± 0.001	0.924 ± 0.001	0.914 ± 0.001	0.925 ± 0.001
OrganS	0.766 ± 0.002	0.784 ± 0.002	0.768 ± 0.002	0.769 ± 0.002	0.787 ± 0.001	0.758 ± 0.002
Dataset	Recall					
	PBC-NAS-1	PBC-NAS-2	PBC-NAS-3	BioNAS-1	BioNAS-2	BioNAS-3
Path	0.869 ± 0.001	0.876 ± 0.003	0.876 ± 0.004	0.887 ± 0.001	0.888 ± 0.005	0.881 ± 0.001
Derma	0.552 ± 0.005	0.566 ± 0.009	0.544 ± 0.115	0.528 ± 0.038	0.554 ± 0.081	0.479 ± 0.005
Pneum	0.809 ± 0.072	0.817 ± 0.011	0.813 ± 0.032	0.825 ± 0.087	0.838 ± 0.06	0.811 ± 0.017
OCT	0.77 ± 0.014	0.769 ± 0.029	0.771 ± 0.020	0.778 ± 0.036	0.779 ± 0.042	0.77 ± 0.018
Retina	0.367 ± 0.04	0.366 ± 0.019	0.358 ± 0.002	0.363 ± 0.036	0.372 ± 0.005	0.361 ± 0.018
Breast	0.803 ± 0.070	0.856 ± 0.072	0.827 ± 0.058	0.817 ± 0.072	0.808 ± 0.106	0.789 ± 0.038
Blood	0.965 ± 0.001	0.972 ± 0.001	0.969 ± 0.001	0.968 ± 0.001	0.967 ± 0.002	0.971 ± 0.001
Tissue	0.585 ± 0.005	0.561 ± 0.001	0.585 ± 0.005	0.566 ± 0.004	0.577 ± 0.005	0.578 ± 0.008
OrganA	0.93 ± 0.004	0.933 ± 0.003	0.933 ± 0.008	0.929 ± 0.015	0.934 ± 0.002	0.935 ± 0.001
OrganC	0.918 ± 0.002	0.919 ± 0.001	0.915 ± 0.002	0.918 ± 0.001	0.907 ± 0.001	0.923 ± 0.001
OrganS	0.759 ± 0.004	0.776 ± 0.002	0.76 ± 0.002	0.756 ± 0.001	0.775 ± 0.001	0.746 ± 0.009
Dataset	Specificity					
	PBC-NAS-1	PBC-NAS-2	PBC-NAS-3	BioNAS-1	BioNAS-2	BioNAS-3
Path	0.988 ± 0.001	0.988 ± 0.001	0.988 ± 0.001	0.989 ± 0.001	0.99 ± 0.001	0.989 ± 0.001
Derma	0.939 ± 0.001	0.939 ± 0.001	0.939 ± 0.001	0.94 ± 0.001	0.943 ± 0.001	0.934 ± 0.001
Pneum	0.809 ± 0.072	0.817 ± 0.011	0.813 ± 0.032	0.825 ± 0.087	0.838 ± 0.06	0.811 ± 0.017
OCT	0.923 ± 0.002	0.923 ± 0.003	0.924 ± 0.002	0.926 ± 0.004	0.926 ± 0.005	0.923 ± 0.002
Retina	0.875 ± 0.003	0.874 ± 0.002	0.871 ± 0.001	0.873 ± 0.001	0.875 ± 0.001	0.873 ± 0.002
Breast	0.803 ± 0.070	0.856 ± 0.072	0.827 ± 0.058	0.817 ± 0.072	0.808 ± 0.106	0.789 ± 0.038
Blood	0.995 ± 0.001	0.996 ± 0.001	0.996 ± 0.001	0.996 ± 0.001	0.996 ± 0.001	0.996 ± 0.001
Tissue	0.951 ± 0.001	0.948 ± 0.001	0.952 ± 0.001	0.949 ± 0.001	0.951 ± 0.001	0.951 ± 0.001
OrganA	0.994 ± 0.001	0.994 ± 0.001	0.994 ± 0.001	0.993 ± 0.001	0.994 ± 0.001	0.994 ± 0.001
OrganC	0.993 ± 0.001	0.993 ± 0.001	0.993 ± 0.001	0.993 ± 0.001	0.992 ± 0.001	0.993 ± 0.001
OrganS	0.98 ± 0.001	0.982 ± 0.001	0.98 ± 0.001	0.98 ± 0.001	0.981 ± 0.001	0.979 ± 0.001

level of 0.05. However, as presented in Table 5, the networks generated with PBC-NAS exhibited significantly lower complexity. This finding suggests that the PBC-NAS method can achieve highly competitive results without a statistically significant difference while maintaining much lower network complexity.

4.2. Comparisons with the state-of-the-art methods

The comparison between PBC-NAS and BioNAS methods with state-of-the-art methods on the MedMNIST benchmark provides valuable insights into their performance across 12 medical image datasets. In this part, we compare our models to four different ResNet models [29] (see Table 3). The PBC-NAS methods show superior accuracy when evaluating the average performance (AVG) across all datasets. PBC-NAS-1, PBC-NAS-2, and PBC-NAS-3 achieve average accuracies of 0.841, 0.843, and 0.842, respectively, outperforming both ResNet-18 (0.819

for 28×28 input size and 0.821 for 224×224 input size) and ResNet-50 (0.816 for 28×28 input size and 0.821 for 224×224 input size). Similarly, the BioNAS methods also demonstrate robust performance, with BioNAS-2 leading with an average accuracy of 0.848, followed by BioNAS-1 at 0.841 and BioNAS-3 at 0.838. These results indicate that the proposed methods provide better overall performance compared to the state-of-the-art architectures in most of the 12 datasets. The proposed methods consistently outperform ResNet models in dataset-specific evaluations for most datasets. For example, in the Path dataset, BioNAS-2 achieves an accuracy of 0.926, higher than both ResNet-18 and ResNet-50, which achieve 0.907 and 0.911, respectively, for the 28×28 input size. Similarly, in the OCT dataset, BioNAS-2 achieves the highest accuracy of 0.819, whereas ResNet models perform in the range of 0.743 to 0.776. The PBC-NAS methods also give the top two models in challenging datasets like Breast. Furthermore, PBC-NAS and BioNAS methods perform best in Retina (PBC-NAS-1 and

BioNAS-3), Derma (PBC-NAS-2 and BioNAS-3) and Blood (PBC-NAS-2 and BioNAS-3) datasets compared to ResNet models.

We have also compared PBC-NAS and BioNAS methods in terms of AUC, as shown in Table 4. The average AUC values indicate that the proposed PBC-NAS and BioNAS methods show competitive results with the ResNet models. The PBC-NAS methods achieve average AUC scores of 0.922 for PBC-NAS-1, 0.921 for PBC-NAS-2, and 0.923 for PBC-NAS-3, which are competitive with ResNet-18 and ResNet-50. The BioNAS methods also perform well, with BioNAS-1 at 0.921, BioNAS-2 at 0.925, and BioNAS-3 leading at 0.925. Across individual datasets, the proposed methods achieve superior or comparable results with the ResNet models on most datasets. For example, in the Path dataset, BioNAS methods perform the best two results compared to ResNet models. Additionally, in the Blood dataset, all PBC-NAS and BioNAS models reach the highest possible AUC of 0.999. In challenging datasets such as Retina and Breast, the proposed methods also show strong performance. For the Retina dataset, all PBC-NAS and BioNAS models (except PBC-NAS-2) achieve better AUC scores than ResNet models. For the Breast dataset, BioNAS-1 matches the highest AUC compared to ResNets, demonstrating its effectiveness in identifying cancerous tissues. Furthermore, they also perform the best two AUC scores in Derma, OrganA, OrganC, and OrganS datasets compared to ResNets.

Overall, all PBC-NAS and BioNAS models surpass the ResNet-18 and ResNet-50 models in terms of average accuracy and AUC. They also achieve superior performance than state-of-the-art methods in dataset-specific evaluations. This suggests that the PBC-NAS and BioNAS methods more effectively capture and learn diverse features in medical images, leading to significant improvements over traditional state-of-the-art architectures in classification tasks.

4.3. Comparisons with automl frameworks

We have compared the PBC-NAS and BioNAS methods with AutoML frameworks (auto-sklearn [30], AutoKeras [31], and Google AutoML) to evaluate their performance on the MedMNIST benchmark. Our evaluation of accuracy and AUC across 12 medical image datasets demonstrates their effectiveness and robustness in diverse classification tasks and data modalities.

First, we have compared PBC-NAS and BioNAS models in terms of accuracy, as shown in Table 3. The average accuracy values indicate that all PBC-NAS and BioNAS models outperform the AutoML frameworks. These results clearly show that PBC-NAS and BioNAS methods provide higher accuracy across various datasets. When examining individual datasets, the proposed methods often achieve superior accuracy compared to the AutoML frameworks. In more challenging datasets such as Retina, Blood, and Organ datasets, the PBC-NAS and BioNAS models consistently outperform AutoML frameworks. For the Retina dataset, all PBC-NAS and BioNAS models achieve higher accuracy than any of the AutoML frameworks, highlighting their robustness in complex tasks. The PBC-NAS and BioNAS methods also achieve top scores for the Blood, indicating their ability to handle complex classification tasks effectively. Furthermore, all PBC-NAS and BioNAS models perform better than all AutoML frameworks for all datasets (except Chest, Pneum, Tissue, and OrganS). This shows their superior generalization capabilities across diverse medical image modalities.

We have also performed a comparison in terms of AUC, as given in Table 4. Based on the average AUC scores, the PBC-NAS and BioNAS models perform competitively with AutoML frameworks. All PBC-NAS and BioNAS models (except BioNAS-2) achieve an average AUC score above 0.92, outperforming Auto-sklearn (0.878) and AutoKeras (0.917). Although Google AutoML achieves a higher average AUC score of 0.927, the PBC-NAS and BioNAS models still outperform or match the AutoML frameworks on most individual datasets. Both PBC-NAS and BioNAS models perform better in all Organ datasets than AutoML frameworks and achieve top results on more challenging datasets such as Retina and Blood. Furthermore, for the Path dataset, all PBC-NAS

and BioNAS models show significantly better performance than AutoML frameworks.

Overall, all PBC-NAS and BioNAS models generally give better or highly competitive performance in accuracy and AUC for most of the datasets compared to AutoML frameworks. They demonstrate significantly superior average accuracy while maintaining highly competitive average AUC scores.

4.4. Comparisons with other NAS studies

We have compared PBC-NAS and BioNAS models with other recent NAS studies in the literature in this part. As seen in Table 3, the proposed PBC-NAS and BioNAS models achieve superior performance in average accuracy compared to all NAS methods (except MSTF-NAS). They perform well performance across various datasets, often outperforming or matching NAS methods such as DARTS [9], SNAS [32], HOPNAS [33], NSGA-NET [34], Random Search, and MSTF-NAS [17]. In terms of average accuracy, MSTF-NAS leads with a score of 0.850, slightly ahead of BioNAS-2 (0.848) and the PBC-NAS models, which range from 0.841 to 0.843. This indicates that while MSTF-NAS is slightly better than others, the PBC-NAS and BioNAS models are highly competitive. In individual dataset analysis, the PBC-NAS and BioNAS models frequently achieve top performance. For example, BioNAS-2 achieves the highest accuracy in the Path dataset (0.926) and the OCT dataset (0.819), while PBC-NAS-2 performs the best in the Chest dataset (0.948) and the second-best in the OCT dataset (0.801). In the Blood dataset, BioNAS-3 and MSTF-NAS both achieve the highest accuracy (0.976). Despite some NAS methods like HOPNAS, NSGA-NET, and Random Search performing well in specific datasets, they do not consistently outperform the PBC-NAS and BioNAS models across all datasets. The PBC-NAS models mainly give better results in the Derma and Breast datasets, with PBC-NAS-2 achieving the second-best accuracy in Derma (0.785) and the best in Breast (up to 0.917). In the Retina dataset, PBC-NAS-1 and BioNAS-3 achieve the top two results, while PBC-NAS-2 and PBC-NAS-3 lead in the Breast dataset. Additionally, all PBC-NAS and BioNAS models outperform all other NAS studies in the Derma and OCT datasets and achieve better or at least equal results compared to all NAS studies (except MSTF-NAS) for Derma, OCT, Blood, and OrganC.

In Table 4, we have also compared PBC-NAS and BioNAS models with other NAS studies based on their AUC results. The MSTF-NAS and Random Search methods give top results in average AUC. The PBC-NAS and BioNAS models achieve top and highly competitive performance when examining individual datasets. For example, BioNAS-2 achieves the highest AUC for the Path dataset (0.993), while PBC-NAS-1 and BioNAS-2 outperform other NAS methods in the Derma dataset with AUCs of 0.941. In the Blood dataset, all PBC-NAS and BioNAS models achieve perfect AUC scores (0.999). Moreover, all BioNAS models demonstrate superior performance compared to other NAS studies (except MSTF-NAS) in the OCT dataset. In the OrganA dataset, PBC-NAS and BioNAS models provide the second-best results, and PBC-NAS-1, BioNAS-1, and BioNAS-3 achieve second-best in OrganC. Additionally, BioNAS-3 achieves the second-best performance in Pneum. Furthermore, in the OCT, OrganA, OrganC, and OrganS datasets, PBC-NAS and BioNAS models remain highly competitive in terms of AUC, closely trailing the top performer.

In summary, while MSTF-NAS holds a slight edge in average accuracy, the PBC-NAS and BioNAS models exhibit superior and more consistent performance across a variety of individual datasets. This robustness and adaptability across different data types highlight the effectiveness and competitiveness of the PBC-NAS and BioNAS approaches compared to other NAS methods.

4.5. Discussion

The comprehensive evaluation of PBC-NAS and BioNAS methods across multiple biomedical image classification tasks reveals their strengths and efficiencies in achieving high classification performance while maintaining computational efficiency. The comparison between PBC-NAS and BioNAS (Section 4.1) models in terms of accuracy, AUC, and computational complexity highlights their competitive performance. On average, BioNAS models slightly outperform PBC-NAS models in accuracy, with BioNAS-2 achieving the highest average accuracy of 0.848. However, PBC-NAS models demonstrate superior computational efficiency, with PBC-NAS-2 achieving the lowest average FLOPs of 0.82 GB. BioNAS is more effective at identifying architectures with strong discriminative power, while PBC-NAS is known for its computational efficiency, consistently producing models with lower FLOPs. This creates a trade-off: BioNAS emphasizes accuracy, potentially increasing computational demands, whereas PBC-NAS prioritizes efficiency without significantly sacrificing performance. These findings highlight the need to consider computational constraints when selecting a NAS method for clinical applications.

In comparison to state-of-the-art architectures (Section 4.2) such as ResNet-18 and ResNet-50, both PBC-NAS and BioNAS methods show superior performance in terms of average accuracy and AUC. Notably, BioNAS-2 achieved the highest average accuracy of 0.848, while PBC-NAS-2 demonstrated the best computational efficiency with the lowest average FLOPs of 0.82 GB. This suggests that the automatically discovered architectures by PBC-NAS and BioNAS are better suited for the specific characteristics of medical images within this benchmark than manually designed architectures like ResNet. This demonstrates that NAS can uncover more tailored and effective architectures for specialized fields like medical imaging.

The comparison with AutoML frameworks (Section 4.3) like auto-sklearn, AutoKeras, and Google AutoML highlights the robustness and effectiveness of PBC-NAS and BioNAS methods in handling diverse medical imaging tasks. Both PBC-NAS and BioNAS methods outperform AutoML frameworks in terms of average accuracy and AUC, demonstrating their ability to adapt to different data modalities and tasks. This indicates that while AutoML frameworks provide a general approach to model selection and hyperparameter tuning, NAS methods like PBC-NAS and BioNAS, which focus specifically on architectural design, can achieve better results for particular tasks, such as biomedical image classification. This highlights the value of specialized NAS approaches for achieving optimal performance in particular domains.

Compared to other recent NAS studies (Section 4.4), PBC-NAS and BioNAS demonstrated highly competitive results. While MSTF-NAS achieved the highest average accuracy in this comparison, BioNAS and PBC-NAS closely followed, outperforming or matching the performance of other notable NAS approaches like DARTS, SNAS, HOPNAS, and NSGA-NET across various datasets. This highlights the adaptability and robustness of our methods within the broader landscape of NAS research for medical image analysis.

Analyzing these comparisons provides valuable insights into the architectural strengths of PBC-NAS and BioNAS. The ability of both methods to outperform ResNet models suggests that their dynamic architectural structures play a significant role in their success. These structures automatically determine the number of stacks and modules, which contributes to improved feature extraction and adaptation to the specific details of medical image data. The different encoding strategies employed by PBC-NAS (node-based operation selection) and BioNAS (edge-based operation selection) likely influence the types of architectures discovered and contribute to the observed performance differences. BioNAS's slightly higher accuracy may result from its broader exploration of the operational space at the edge level, which might lead to more complex and expressive cell structures. Conversely, PBC-NAS's focus on node-level operation selection might result in more efficient and computationally effective architectures. The consistent

outperformance of both methods compared to AutoML frameworks further supports the idea that a dedicated search for optimal network architecture, as implemented in PBC-NAS and BioNAS. It is the core motivation behind NAS research and crucial for achieving high performance. The competitive performance against other recent NAS studies indicates that the search algorithm and search spaces employed by PBC-NAS and BioNAS effectively navigate the complex architectural landscape for medical imaging tasks. In particular, continuous search space, and opposition-based differential evolution algorithms contribute to generating more diverse solutions.

5. Ablation studies

Ablation studies are essential in NAS studies to determine the contribution of each component within a network architecture. In this section, we present the ablation studies, focusing on the following aspects: (1) impact of hyper-parameter settings; (2) performance on the cross-modality dataset with fine-tuning; (3) generalization performance; (4) discrimination performance; (5) search space efficiency; (6) experimental analysis of architectures.

5.1. The effect of the hyper-parameter settings

This ablation study investigates the impact of various architectural parameters on the performance and computational complexity of PBC-NAS and BioNAS. By evaluating different configurations of initial filters, stacks, and modules (n_M , n_S , n_F^{init}), we aim to understand how these factors affect classification performance and computational complexity. This ablation study aims to find the optimal balance between model complexity and performance, providing results for designing NAS-based models for various biomedical image classification tasks. The corresponding results are shown in Fig. 6. We have reported average ACC and AUC results obtained for 12 different datasets and two NAS methods (PBC-NAS and BioNAS), however, we have also discussed results in terms of FLOPs for each dataset in this subsection.

These results provide significant insights into the impact of the number of init filters on classification performance and computational complexity. The ACC results show that, on average, 64 and 128 filters lead to the highest accuracy for the PBC-NAS and BioNAS methods. For individual datasets, 128 filters are the most effective for 6 out of 12 datasets in BioNAS and for 7 out of 12 datasets in PBC-NAS. While 128 filters achieve the highest average accuracy, the difference is not significantly greater than with lower filter counts. The results for average AUC are similar, with 64 and 128 filters performing best for PBC-NAS, and 64 filters being optimal for BioNAS. This suggests that 64 filters provide a good balance between achieving high performance and managing complexity. The analysis of FLOPs highlights the significant increase in computational complexity with more filters. For example, for the Blood dataset in BioNAS, the complexity rises from 0.007G FLOPs with 8 filters to 1.738G FLOPs with 128 filters. This pattern is consistent across all datasets for the two NAS methods.

The number of stacks has a minimal effect on performance for both methods. The ACC and AUC values remain relatively consistent across different stack configurations. There is a slight increase in ACC with 1 stack and the highest AUC with 2 stacks for both PBC-NAS and BioNAS. However, the complexity increases FLOPs by up to 3 times on average with more stacks. Similarly, the number of modules has a less significant impact. The best results are achieved with 2 modules in ACC for both NAS methods, but there is no direct relation for AUC (3 modules for PBC-NAS; 1 and 2 modules for BioNAS).

In summary, 128 filters generally offer the best performance, but at a high cost in terms of computation. Many datasets achieve comparable performance with fewer filters, particularly 32 or 64, which provide a good balance between ACC/AUC and computational efficiency. Therefore, when choosing the number of filters, it is important

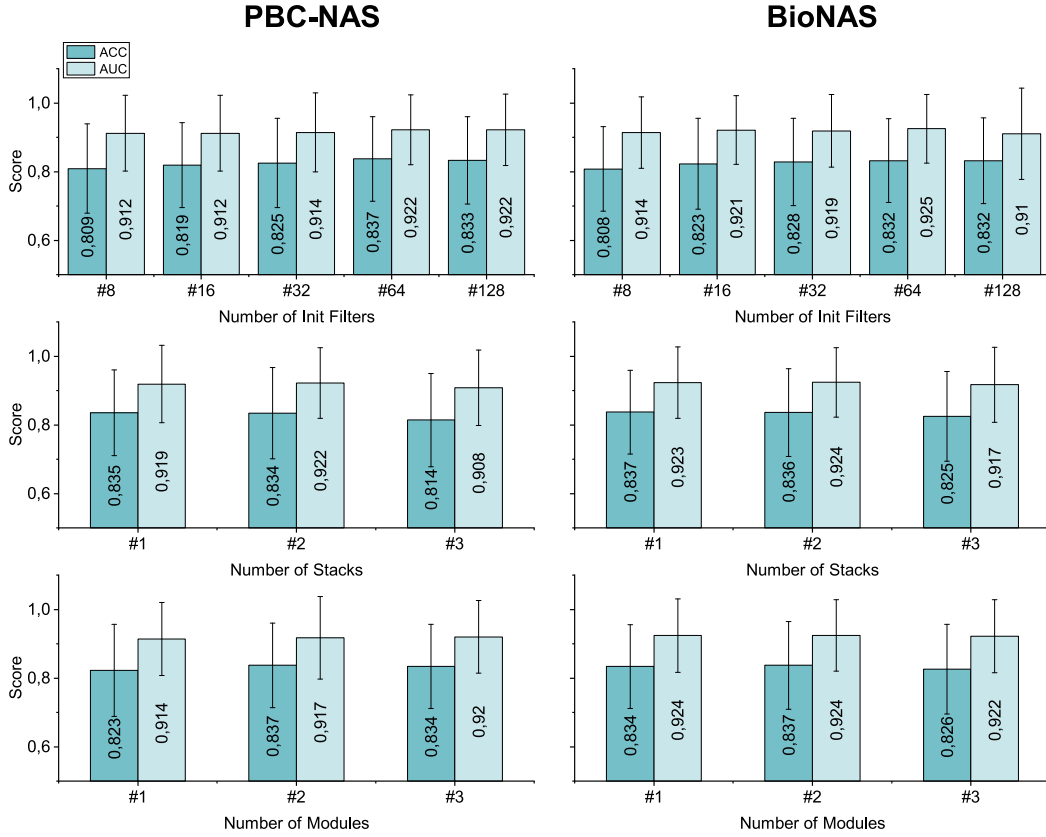


Fig. 6. Comparative analysis of the effect of the hyper-parameters. The top row shows the effect of the number of initial filters (8, 16, 32, 64, 128) on ACC and AUC. The middle row examines the impact of the number of stacks (1, 2, 3) on ACC and AUC. The bottom row illustrates the influence of the number of modules (1, 2, 3) on ACC and AUC. Average ACC and AUC results are reported and error bars represent the standard deviation, highlighting the variability in performance across 12 different datasets.

to consider both the desired performance and the available computational resources. For environments with limited computational power, using 16 or 32 filters could be a good option, maintaining acceptable performance while managing complexity. In high-performance environments where accuracy is critical, using 64 or 128 filters may be chosen despite the increased computational demands. Additionally, there is no direct relationship between number of stacks/modules and complexity. Moreover, as the number of stacks/modules increases, classification performance does not improve, and complexity rises. Therefore, choosing 1 or 2 stacks/modules can balance classification performance and computational complexity.

5.2. Evaluation on the cross-modality datasets with fine-tuning

In recent years, Neural Architecture Search (NAS) has become a powerful technique for automating the design of neural networks, achieving state-of-the-art performance across various tasks and domains. Despite its success, NAS is computationally intensive and often requires significant resources and time. Also, datasets in similar modalities show similar characteristics. This challenge raises a critical question: Can we achieve comparable performance by fine-tuning existing architectures rather than conducting separate NAS for each dataset, especially within similar data modalities?

To address this question, we have conducted a comprehensive ablation study comparing two NAS methods across 12 medical image datasets from different data modalities. We have evaluated two approaches in terms of ACC and AUC. The datasets are grouped into two categories based on their data modalities: Group 1 includes Path, Retina, Derma, and Blood datasets, while Group 2 comprises Pneum, Tissue, Breast, OrganA, OrganC, and OrganS datasets. For Group 1, the best model for the Path dataset is selected as the baseline model (PBC-NAS-2 and BioNAS-2 in Table 3), and the best model for the Tissue

dataset is chosen as a baseline model for Group 2 (PBC-NAS-3 and BioNAS-2 in Table 3). The datasets with the most samples are selected for both groups, and the best model for these datasets is chosen as the baseline (see Table 3). Subsequently, the best models are fine-tuned on other datasets in the same group, and these results are reported in Fig. 7. For example, the best model for Path (PBC-NAS-2) is retrained on Retina, Derma, and Blood datasets separately, and the results are reported for Group 1. Similarly, BioNAS-2 is retrained on these datasets in Group 1. The same steps are applied to Group 2 similarly.

The experimental results are shown in Fig. 7. The comparative analysis of PBC-NAS and BioNAS shows that their ACC and AUC scores are very close to their best results, even in ablation studies. This indicates that fine-tuning existing architectures can yield nearly optimal performance without the need for separate NAS for each dataset. There are slight decreases in ACC and AUC in Group 1 for PBC-NAS when comparing best results to the ablation study results. For example, in the Blood dataset, the ACC decreased slightly from 0.975 (best) to 0.970 (ablation), while the AUC remained constant at 0.999. Similarly, for the Derma dataset, there are minor differences between ablation (0.788) and best (0.787) results in terms of ACC and AUC. On the other hand, in the Retina dataset, the ACC significantly decreased from 0.562 (best) to 0.522 (ablation). All datasets in Group 2 (except Breast) also show minor decreases. These minor differences suggest that the performance of the ablation studies is very close to the best results, indicating that the ablation approach captures most of the performance potential of the entire NAS process.

BioNAS demonstrates a similar trend, where the performance metrics from ablation studies are very close to the best results. In Group 1, the Retina dataset shows a slight decrease in ACC from 0.541 (ablation) to 0.553 (best) and a minor drop in AUC from 0.732 to 0.749. The Derma and Blood datasets show stable performance, with

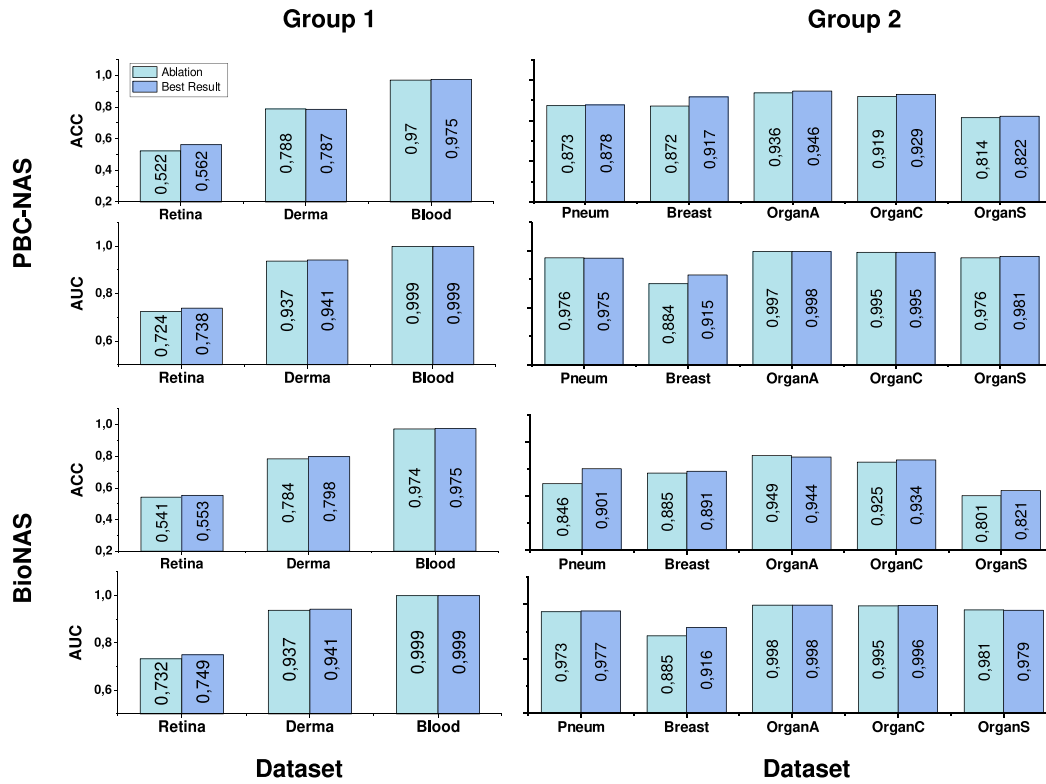


Fig. 7. This figure illustrates the performance of two neural architecture search methods, PBC-NAS and BioNAS, evaluated on multiple datasets categorized into Group 1 (Retina, Derma, Blood) and Group 2 (Pneum, Breast, OrganA, OrganC, OrganS) based on their data modalities. The bar charts represent the ACC and AUC scores for each dataset. Light blue bars denote the results obtained from ablation studies, while dark blue bars represent the best results achieved. The comparison aims to address whether similar results can be achieved through fine-tuning without the necessity of conducting separate NAS for datasets within similar data modalities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

only minor differences. In Group 2, a significant difference is observed in the Pneum dataset, with ACC decreasing from 0.901 (best) to 0.846 (ablation). In the OrganC dataset, ACC decreased from 0.934 to 0.925 and AUC from 0.996 to 0.995. The other datasets show only slight differences in performance metrics.

In summary, the results suggest that the performance obtained from ablation studies is very close to the best possible results, demonstrating that there is no significant need to run separate NAS for each dataset. Fine-tuning existing architectures proves to be an effective and resource-efficient strategy for enhancing model performance.

5.3. Generalization performance

Traditional NAS methods often require individual searches for each dataset, which can be computationally intensive and time-consuming. The efficiency and effectiveness of these methods are of significant interest, especially in biomedical domains where datasets can be highly diverse and complex. We have observed that PBC-NAS and BioNAS methods give good results for different datasets in similar data modalities without requiring NAS separately (see Section 5.2). Therefore, we have performed a new ablation study to evaluate the generalization performance of PBC-NAS and BioNAS methods. In this ablation study, we investigate whether we could achieve results comparable to the best result (in Tables 3 and 4) without performing NAS independently for each dataset. We have evaluated PBC-NAS and BioNAS in terms of ACC and AUC. We have chosen the best models in terms of ACC for the Tissue dataset (PBC-NAS-3 and BioNAS-3) as a baseline because they have the most image samples. Then, we fine-tune these baseline models on the remaining 11 datasets, and the test results obtained are reported in Fig. 8.

For PBC-NAS, the ACC results indicate that datasets such as Blood, Chest, and Derma have almost identical ablation results compared to

the best results. This suggests that similar performance can be achieved without a separate NAS. Additionally, for Pneum, OrganA, and OrganS, the ablation results are slightly lower but still comparable to the best results, with differences ranging from 0.004 to 0.01, indicating only minor performance degradation. However, Breast and Retina show more notable differences in ACC (0.039 and 0.038, respectively), highlighting significant performance drops without dataset-specific NAS. Furthermore, minor differences are observed for other datasets. In terms of AUC, PBC-NAS shows that for Blood, OCT, and OrganA, the ablation results are very close to the best results, with slight differences (0 to 0.002), demonstrating that a general NAS approach can be highly effective. However, most other datasets show minor differences (0.001 to 0.005), while Breast and Derma display significant differences (0.031 and 0.013).

For BioNAS, the ACC values for Breast, Chest, OrganA, and Retina are nearly the same as the best results, with only minimal differences (0.002 to 0.006). Additionally, for Derma and OrganC, slight differences (0.013 to 0.009) suggest a minor decrease in performance, but the results remain comparable. Blood shows a minimal difference (0.007), while more significant performance declines are observed for OCT, Path, and Pneum (0.042, 0.05, and 0.055), indicating varying levels of dependence on separate NAS. In terms of AUC, BioNAS results for Blood, OrganA, and OrganC are almost identical to the best results (0 to 0.001), demonstrating effective general NAS performance. However, Breast displays a significant difference (0.031), while most other datasets show minor differences (0.001 to 0.007).

The comparative analysis of PBC-NAS and BioNAS results shows that achieving similar performance to the best results is possible without conducting NAS separately for each dataset. Some datasets like Blood, Chest, and OrganA showed nearly identical ablation results to the best ones, indicating that a generalized NAS approach can be effective. However, the performance drops significantly for other

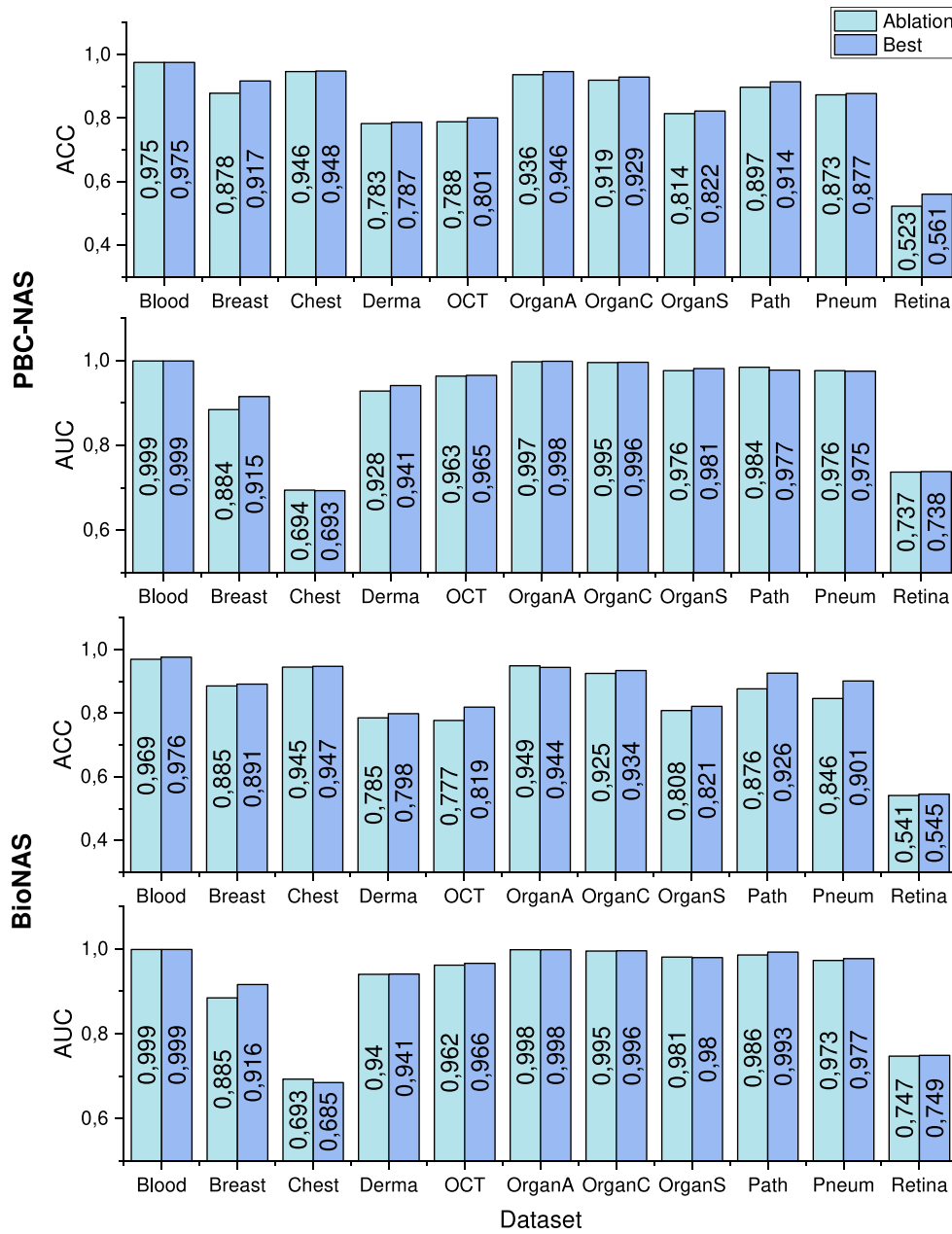


Fig. 8. Comparison of generalization performance of BioNAS and PBC-NAS methods on 12 datasets. Each bar plot compares the ablation results (applying a single architecture across multiple datasets) with the best results (obtained from separate NAS for each dataset). The light blue bars represent the ablation results, and the blue bars represent the best results. This comparison aims to evaluate whether similar performance can be achieved without conducting NAS independently for each dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

datasets, mainly Breast and Path, highlighting the value of dataset-specific NAS. This suggests that while generalized NAS can save computational resources and time, there is a trade-off in performance for specific datasets that may require tailored NAS to achieve the best accuracy and AUC. In conclusion, this ablation suggests that a balanced approach that combines generalized NAS for some datasets with specific NAS for others could efficiently optimize overall performance.

5.4. Discrimination performance

Visualization of embeddings obtained from different methods is crucial for evaluating the discriminative performance of a model, especially in high-dimensional data. The most commonly used method for visualizing high-dimensional data is the t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE visually examines how well different

classes are separated in the embedding space by reducing data to two or three dimensions. Effective discriminative performance is indicated by clear, distinct clusters corresponding to different classes, with minimal overlap and outliers. This visualization technique is valuable for comparing the performance of different models or methods.

In Fig. 9, we have visually compared PBC-NAS and BioNAS methods on the BloodMNIST and PathMNIST datasets. We have visualized the embeddings obtained from test samples in 2D space. For the BloodMNIST dataset, the clusters generated by PBC-NAS are relatively well-separated. However, there is a noticeable small isolated cluster (red circle) from the main clusters, indicating potential issues with discrimination. In contrast, the BioNAS method produces more compact and better-separated clusters. Both methods show an overlap between classes four and six. Similarly, for the PathMNIST dataset, the clusters formed by PBC-NAS are distinct, but multiple small clusters are

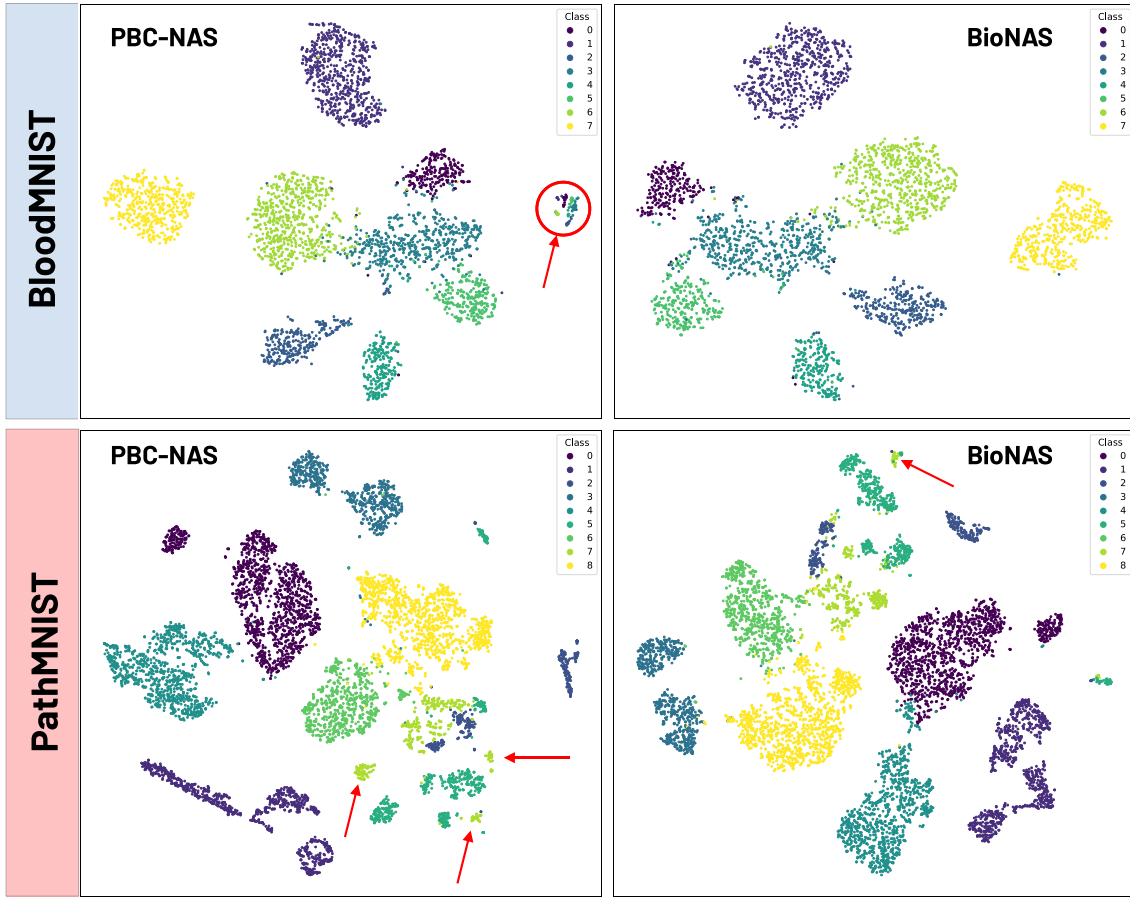


Fig. 9. t-SNE visualizations of embeddings generated by PBC-NAS and BioNAS methods for BloodMNIST and PathMNIST datasets. The plots indicate the embeddings from test dataset samples, highlighting the discriminative performance of the two methods. The circles and arrows indicate the potential issues with discrimination.

scattered around, indicating less effective discrimination. Conversely, BioNAS shows more coherent and better-separated clusters. Similar results are observed for the other 10 datasets; both PBC-NAS and BioNAS methods generate similar compact and well-separated clusters for each class.

Overall, BioNAS demonstrates better discriminative performance compared to PBC-NAS for both datasets. The clusters in BioNAS are more distinct and compact, indicating better separation between classes and higher within-class similarity. This suggests that BioNAS is more effective at generating discriminative embeddings, handling outliers, and maintaining cluster compactness, making it a more reliable method for these datasets (see in Table 3).

5.5. Analysis of search space efficiency

In this ablation study, we have compared the performance of three different search spaces (PBC-NAS, BioNAS, and NAS-Bench-101) in terms of average ACC, AUC and FLOPs. We have selected NAS-Bench-101 as the third search space because it is the baseline for cell-structured NAS studies, and PBC-NAS and BioNAS are proposed based on it. We have generated 100 solutions in each search space using the random search method and repeated these steps across 12 different datasets. In this ablation study, we aim to evaluate the efficiency of the different search spaces when the search method is fixed. The results are shown in Table 7.

The results show that all three search spaces perform well, with only slight differences in average ACC and AUC. NAS-Bench-101 has achieved slightly better performance among the three search spaces, with an average ACC (0.837) and AUC (0.925). The BioNAS follows closely with an average ACC of 0.836 and an AUC of 0.924. While still

performing well, the PBC-NAS performs poorly, with an average ACC of 0.831 and an AUC of 0.923. When considering the FLOPs, we observe differences among the three search spaces. PBC-NAS demonstrates the lowest computational cost (0.607), while BioNAS has the highest computational cost (1.264). NAS-Bench-101 falls in between with 0.781 Gigabytes of FLOPs. It is important to note that the differences between the search spaces are minimal, with only a 0.006 difference in ACC and 0.002 in AUC between the highest and lowest-performing spaces. Despite PBC-NAS having the lowest performance, it also shows the lowest computational complexity. The difference between NAS-Bench-101 and PBC-NAS is minor in terms of AUC (0.002) and ACC (0.006), but significant in terms of complexity (1.28 times). This suggests that while PBC-NAS is more computationally efficient, it may show slightly lower performance in terms of accuracy and the ability to distinguish between classes. These slight differences in performance metrics compared to NAS-Bench-101 show that PBC-NAS and especially BioNAS are highly effective search spaces.

These results are expected because of the limited search space size of the NAS-Bench-101. In the NAS-Bench-101 study, the number of filters (n_f^{init}), number of stacks (n_s), and modules (n_m) are fixed and not included in the search space. Additionally, only three 2D operations (3×3 convolution, 1×1 convolution, and 3×3 max-pooling) can be selected in NAS-Bench-101. However, in PBC-NAS and BioNAS, fifteen different 2D operations can be chosen, and also n_f^{init} , n_s , and n_m parameters are included in the search space. Therefore, the search space size grows enormously, leading to increased diversity of candidate architectures. The search space requires more than 100 generated solutions to find optimal, making the random search insufficient.

In conclusion, while all three search spaces are effective, NAS-Bench-101 shows a slight difference in terms of ACC and AUC. BioNAS

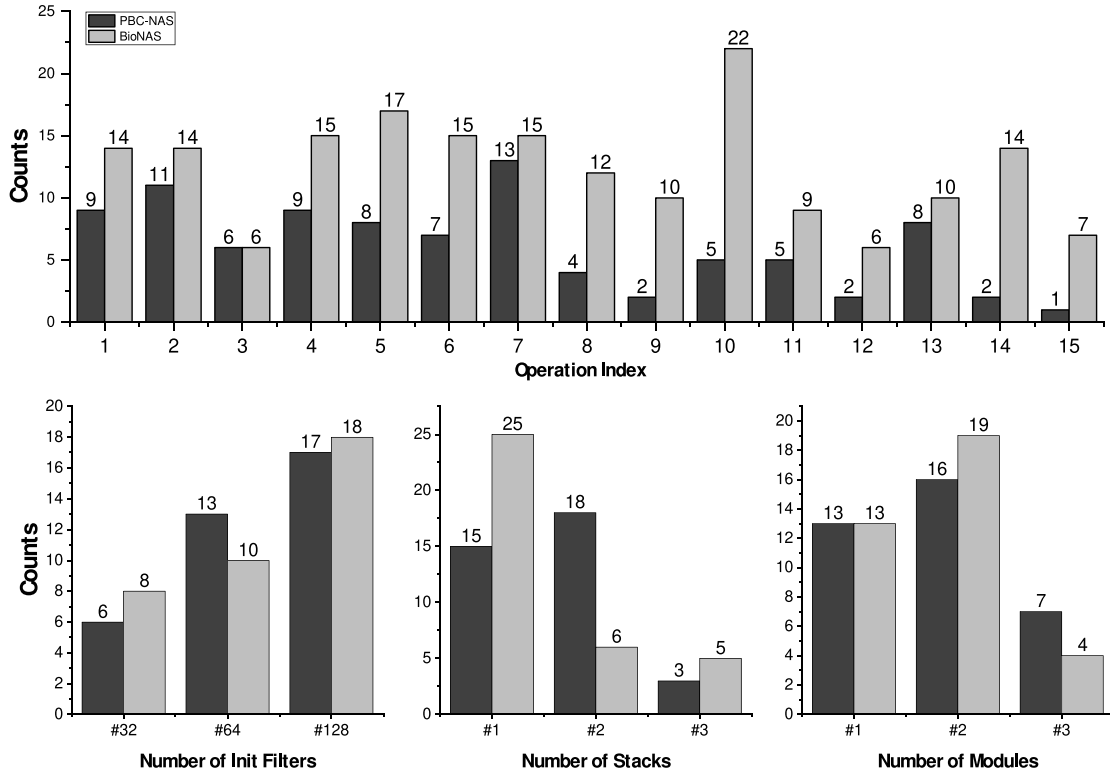


Fig. 10. Comparative analysis of the best three architectures obtained by PBC-NAS and BioNAS. The top plot shows the frequency of the most selected operations for the best networks. Similarly, the bottom row of plots illustrates the most frequent selections for n_f^{init} , n_s and n_m , respectively.

Table 7

Performance and complexity comparison of different search spaces on 12 different datasets. Average results from 5 different runs are reported.

Search space	ACC	AUC	FLOPs (GB)
PBC-NAS [20]	0.831	0.923	0.607
BioNAS [21]	0.836	0.924	1.264
NAS-Bench-101 [22]	0.837	0.925	0.781

is a close second, and PBC-NAS, although slightly behind, still offers competitive performance with the lowest complexity. These results can help guide future research in selecting the most suitable search space for neural architecture search problems. NAS-Bench-101 can be good alternative when the search time and capacity are limited. However, due to the fixed number of filters, stacks, and modules, it may generate more complex solutions for different problems. Conversely, PBC-NAS and BioNAS search spaces support the dynamic structures that balance the classification performance and complexity. Additionally, the results highlight the importance of the design of the search space and the relation between the search space and the search algorithm.

5.6. Experimental analysis of generated architectures

In this ablation study, we have conducted experimental analysis on the top three networks from 12 datasets. The most frequently selected operations, the number of initial filters, stack, modules, and the number of edges and operations in each cell are investigated in detail. We aim to offer guidance to fellow researchers in designing suitable search spaces that balance performance and complexity and identify some patterns. The results are illustrated in Fig. 10. We also compare our findings to previous ablation studies to better understand the optimization processes used by PBC-NAS and BioNAS.

The figure compares the PBC-NAS and BioNAS methods in terms of the most frequently selected operations, initial filters, stacks, and modules for the best three architectures generated across 12 datasets. The

top plot shows that BioNAS selects certain operations (e.g., operation 10) significantly more frequently than PBC-NAS, indicating differing preferences in the architecture encoding strategies of the two methods. Operations 2, 7, and 13 are selected at the same frequency for both methods. BioNAS tends to choose a significantly higher number of operations, possibly due to its encoding strategy of selecting operations for each edge, whereas PBC-NAS selects operations for each node (see Sections 2.1.2 and 2.2.2). When we analyze the number of edges and nodes in the cell, we find that the most frequent number of edges selected is 7 for BioNAS, while the most frequent number of nodes is 3 for PBC-NAS. This explains the difference in the number of operations used for the two methods. Additionally, for both methods, applying activation before convolution operations is more selected, indicating good performance. Furthermore, when we examine the most frequent operations, we observe that Mish activation functions are chosen more often for good architectures than ReLU.

In the bottom left plot, both methods strongly prefer larger initial filter sizes, particularly 128 filters, chosen 17 times by PBC-NAS and 18 times by BioNAS. In addition, for the two methods, it is seen that the number of filters smaller than 32 is not selected for the best models. This suggests that larger filter sizes are associated with better model performance but also increase the complexity (see Section 5.1). The bottom middle plot indicates the structural complexity of models generated by the two methods. BioNAS significantly favors a simpler architecture with 1 stack (25 times) compared to PBC-NAS (15 times), while PBC-NAS chooses 2 stacks (18 times) more than BioNAS (6 times). Both methods rarely select 3 stacks. The bottom right plot shows that both methods often choose a moderate number of modules (2), with BioNAS selecting it 19 times and PBC-NAS 16 times. A number of modules (1) are chosen equally for both methods and (3) are selected rarely. These architectural choices indicate that both methods optimize their architectures based on performance-driven decisions. These results are relative to the performed ablation study in Section 5.1. For instance, Section 5.1 showed that larger filter sizes and a specific

number of stacks or modules (1 and 2) lead to better performance; these preferences for these configurations in the results shown are consistent.

6. Conclusion

In this study, we provide a comprehensive evaluation of two NAS methods, PBC-NAS and BioNAS, for biomedical image classification across multiple data modalities. Our experiments, conducted on the MedMNIST dataset, demonstrate the effectiveness of these methods in achieving high classification performance while maintaining computational efficiency. Both PBC-NAS and BioNAS methods show superior performance compared to state-of-the-art architectures such as ResNet-18 and ResNet-50. Notably, BioNAS-2 achieved the highest average accuracy of 0.848, while PBC-NAS-2 demonstrated the best computational efficiency with the lowest average FLOPs of 0.82 GB. Additionally, the proposed NAS methods outperform AutoML frameworks like auto-sklearn, AutoKeras, and Google AutoML in terms of average accuracy and AUC, highlighting their robustness and effectiveness in handling diverse medical imaging tasks. Our comparative analysis with other NAS studies shows that while MSTF-NAS slightly outperforms in average accuracy, PBC-NAS and BioNAS models exhibit highly competitive performance across various datasets, often achieving top results. This highlights the adaptability and robustness of our methods.

We have also performed ablation studies to evaluate NAS methods in detail. Our ablation studies reveal that larger filter sizes and a specific number of stacks or modules lead to better performance. Importantly, we found that fine-tuning existing architectures can yield nearly optimal results without the need for separate NAS for each dataset, especially within similar data modalities. Additionally, BioNAS demonstrates better discriminative performance, producing more compact and well-separated clusters in t-SNE visualizations. The analysis of search space efficiency shows that while NAS-Bench-101 achieves slightly better performance, PBC-NAS and BioNAS offer highly competitive results with lower computational complexity (PBC-NAS). This highlights the importance of designing efficient search spaces that balance performance and complexity.

Our study provides a thorough experimental evaluation of PBC-NAS and BioNAS methods in terms of classification performance, computational complexity, and generalization capability. Detailed ablation studies provide insights into the impact of architectural parameters, the effectiveness of fine-tuning, and the discriminative performance of the generated architectures. Furthermore, our analysis uncovers patterns and preferences in the search space, guiding future research in designing efficient NAS methods for biomedical image classification. In summary, this study demonstrates the potential of PBC-NAS and BioNAS methods in biomedical image classification and provides valuable guidance for researchers in selecting and designing appropriate NAS methods for their specific tasks and computational constraints. The choice between BioNAS and PBC-NAS in real-world clinical settings depends on several factors:

- **Available computational resources:**

- If computational resources are limited (e.g., in smaller clinics or for edge computing applications), PBC-NAS would be the preferred choice due to its significantly lower computational cost.
- For well-resourced hospitals or research centers with access to high-performance computing infrastructure, BioNAS could be used to achieve slightly higher accuracy.

- **Specific clinical task requirements:**

- For critical diagnostic tasks, such as early cancer detection, the slightly higher accuracy of BioNAS may justify its increased computational cost.

- For routine screening or less critical applications, the computational efficiency of PBC-NAS may be more valuable, allowing for faster processing of large volumes of images.

- **Time constraints:**

- In time-sensitive situations, such as emergency radiology, PBC-NAS's lower computational cost leads to faster results, which could be crucial.
- For non-urgent cases where processing time is less critical, BioNAS could be used to achieve higher accuracy.

- **Model updating and retraining needs:**

- In scenarios that require frequent model updates or retraining (e.g., to account for new data or equipment changes), the lower computational cost of PBC-NAS could be advantageous. This allows for more agile and frequent model iterations.

While BioNAS and PBC-NAS demonstrate strong performance in biomedical image classification tasks, they do have some limitations. BioNAS, in particular, shows higher computational costs in terms of FLOPs compared to PBC-NAS, which may limit its applicability in resource-constrained environments. The study also indicates that both methods struggle with certain challenging datasets, such as the Chest dataset, where they showed lower performance in AUC scores. In addition, fine-tuning is required to generalize both methods to other datasets. In future work, researchers can investigate integrating more advanced search algorithms and focusing on the mentioned disadvantages. Additionally, NAS approaches that reduce the evaluation cost, such as surrogate-assisted NAS and training-free NAS, can be applied to medical imaging tasks. Furthermore, compact search spaces that balance prediction performance and computational complexity can be proposed.

CRediT authorship contribution statement

Zeki Kuş: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Musa Aydın:** Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Berna Kiraz:** Writing – original draft, Methodology, Formal analysis, Conceptualization. **Alper Kiraz:** Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alper Kiraz reports financial support was provided by TÜBA (Turkish Academy of Sciences). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

A. Kiraz acknowledges partial support from the Turkish Academy of Sciences (TÜBA).

References

- [1] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88.
- [2] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, Pfister H, Ni B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Sci Data* 2023;10(1):41.

- [3] Kumar R, Kumbharkar P, Vanam S, Sharma S. Medical images classification using deep learning: a survey. *Multimedia Tools Appl* 2023;83(7):19683–728.
- [4] Sharma D, Mishra I, Parepalli R, S J. Biomedical image classification using convolutional neural networks. In: 2023 international conference on intelligent and innovative technologies in computing, electrical and electronics. IITCEE, 2023, p. 840–5. <http://dx.doi.org/10.1109/IITCEE57236.2023.10091043>.
- [5] Elsken T, Metzen JH, Hutter F. Neural architecture search: A survey. *J Mach Learn Res* 2019;20(55):1–21, URL <http://jmlr.org/papers/v20/18-598.html>.
- [6] Baldeon Calisto M, Lai-Yuen SK. Emonas-net: Efficient multiobjective neural architecture search using surrogate-assisted evolutionary algorithm for 3d medical image segmentation. *Artif Intell Med* 2021;119:102154.
- [7] Benmeziane H, Hamzaoui I, Cherif Z, El Maghraoui K. Medical neural architecture search: Survey and taxonomy. In: International joint conference on artificial intelligence. 2024.
- [8] Zoph B, Le QV. Neural architecture search with reinforcement learning. 2016.
- [9] Liu H, Simonyan K, Yang Y. DARTS: differentiable architecture search. 2018, CoRR abs/1806.09055. [arXiv:1806.09055](https://arxiv.org/abs/1806.09055).
- [10] Lu Z, Whalen I, Boddeti V, Dhebar Y, Deb K, Goodman E, Banzhaf W. Nsganet: neural architecture search using multi-objective genetic algorithm. In: Proceedings of the genetic and evolutionary computation conference. 2019, p. 419–27.
- [11] Ma L, Li N, Yu G, Geng X, Cheng S, Wang X, Huang M, Jin Y. Pareto-wise ranking classifier for multiobjective evolutionary neural architecture search. *IEEE Trans Evol Comput* 2024;28(3):570–81.
- [12] Sun Y, Xue B, Zhang M, Yen GG. Evolving deep convolutional neural networks for image classification. *IEEE Trans Evol Comput* 2020;24(2):394–407. <http://dx.doi.org/10.1109/TEVC.2019.2916183>.
- [13] Xue Y, Zha J, Pelusi D, Chen P, Luo T, Zhen L, Wang Y, Wahib M. Neural architecture search with progressive evaluation and sub-population preservation. *IEEE Trans Evol Comput* 2024. <http://dx.doi.org/10.1109/TEVC.2024.3393304>, 1–1.
- [14] Xue Y, Han X, Neri F, Qin J, Pelusi D. A gradient-guided evolutionary neural architecture search. *IEEE Trans Neural Netw Learn Syst* 2024;1–13. <http://dx.doi.org/10.1109/TNNLS.2024.3371432>.
- [15] Han X, Xue Y, Wang Z, Zhang Y, Muravev A, Gabbouj M. Sadenas: A self-adaptive differential evolution algorithm for neural architecture search. *Swarm Evol Comput* 2024;91:101736. <http://dx.doi.org/10.1016/j.swevo.2024.101736>.
- [16] Kwasigroch A, Grochowski M, Mikołajczyk A. Neural architecture search for skin lesion classification. *IEEE Access* 2020;8:9061–71.
- [17] Wang Y, Zhen L, Zhang J, Li M, Zhang L, Wang Z, Feng Y, Xue Y, Wang X, Chen Z, Luo T, Goh RSM, Liu Y. Mednas: Multiscale training-free neural architecture search for medical image analysis. *IEEE Trans Evol Comput* 2024;28(3):668–81.
- [18] Ali MJ, Moalic L, Essaid M, Idoumghar L. Robust neural architecture search using differential evolution for medical images. In: applications of evolutionary computation. Springer Nature Switzerland; 2024, p. 163–79.
- [19] Ali MJ, Moalic L, Essaid M, Idoumghar L. Evolutionary neural architecture search for 2d and 3d medical image classification. In: Computational science – ICCS 2024. Cham: Springer Nature Switzerland; 2024, p. 131–46.
- [20] Kuş Z, Kiraz B, Aydın M, Kiraz A. PBC-NAS: Neural architecture search for peripheral blood cells classification. In: 32nd signal processing and communications applications conference. 2024, p. 1–4.
- [21] Kuş Z, Kiraz B, Aydın M, Kiraz A. BioNAS: Neural architecture search for multiple data modalities in biomedical image classification. In: recent trends and advances in artificial intelligence, 2024. Cham: Springer Nature Switzerland; 2024, p. 539–50.
- [22] Ying C, et al. Nas-bench-101: Towards reproducible neural architecture search. 2019.
- [23] Gülcü A, Özkan İTS, Kuş Z, Karakuş OF. Triplet maml for few-shot classification problems. In: advanced engineering, technology and applications. Cham: Springer Nature Switzerland; 2024, p. 437–49.
- [24] Wei J, et al. Genetic u-net: Automatically designed deep networks for retinal vessel segmentation using a genetic algorithm. *IEEE Trans Med Imaging* 2022;41(2):292–307.
- [25] Koçak T, et al. Empirical comparison of deep neural networks for brain vessel segmentation. In: UBMK. 2021, p. 751–6.
- [26] Rahnamayan S, Tizhoosh HR, Salama MMA. Opposition-based differential evolution. *IEEE Trans Evol Comput* 2008;12(1):64–79. <http://dx.doi.org/10.1109/TEVC.2007.894200>.
- [27] Kuş Z, Kiraz B, Göksu TK, Aydın M, Özkan E, Vural A, Kiraz A, Can B. Differential evolution-based neural architecture search for brain vessel segmentation. *Eng Sci Technol Int J* 2023;46:101502.
- [28] Storn R, Price K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim* 1997;11(4):341.
- [29] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR, 2016.
- [30] Feurer M, et al. Efficient and robust automated machine learning. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in neural information processing systems. Vol. 28, Curran Associates, Inc.; 2015.
- [31] Jin H, et al. Auto-keras: An efficient neural architecture search system. In: Proceedings of the 25th ACM SIGKDD. Association for Computing Machinery; 2019, p. 1946–56.
- [32] Xie S, Zheng H, Liu C, Lin L. SNAS: stochastic neural architecture search. 2018, CoRR abs/1812.09926. [arXiv:1812.09926](https://arxiv.org/abs/1812.09926).
- [33] Zhang J, Li D, Wang L, Zhang L. One-shot neural architecture search by dynamically pruning supernet in hierarchical order. *Int J Neural Syst* 2021;31(07):2150029. <http://dx.doi.org/10.1142/s0129065721500295>.
- [34] Lu Z, Whalen I, Boddeti V, Dhebar Y, Deb K, Goodman E, Banzhaf W. Nsganet: neural architecture search using multi-objective genetic algorithm. In: Proceedings of the genetic and evolutionary computation conference. GECCO '19, ACM; 2019.