Evolutionary Neural Architecture Search for 2D and 3D Medical Image Classification

Muhammad Junaid Ali[1] ⋆ †, Laurent Moalic[2] †, Mokhtar Essaid[3] †, and Lhassane Idoumghar[4] †
[1]muhammad-junaid.ali@uha.fr [2]laurent.moalic@uha.fr [3]mokhtar.essaid@uha.fr
[4]lhassane.idoumghar@uha.fr

† Université de Haute-Alsace, IRIMAS UR 7499, F-68093 Mulhouse, France

Abstract. Designing deep learning architectures is a challenging and time-consuming task. To address this problem, Neural Architecture Search (NAS) which automatically searches for a network topology is used. While existing NAS methods mainly focus on image classification tasks, particularly 2D medical images, this study presents an evolutionary NAS approach for 2D and 3D Medical image classification. We defined two different search spaces for 2D and 3D datasets and performed a comparative study of different meta-heuristics used in different NAS studies. Moreover, zero-cost proxies have been used to evaluate the performance of deep neural networks, which helps reduce the searching cost of the overall approach. Furthermore, recognizing the importance of Data Augmentation (DA) in model generalization, we propose a genetic algorithm based automatic DA strategy to find the optimal DA policy. Experiments on MedMNIST benchmark and BreakHIS dataset demonstrate the effectiveness of our approach, showcasing competitive results compared to existing AutoML approaches. The source code of our proposed approach is available at `https://github.com/Junaid199f/evo_nas_med_2d_3d`.

Keywords: Evolutionary Neural Architecture Search · Medical Image Classification · AutoML · AutoDL · Automatic Data Augmentation

1 Introduction

Deep Learning (DL) algorithms have been widely used for solving real-world tasks, but designing these architectures requires domain-expert knowledge. Multiple Neural Architecture Search (NAS) approaches have been proposed to automate DL architecture design for multiple tasks [17], but they require significant searching time due to the network evaluation phase.

Multiple performance estimation strategies have been proposed to reduce the search time. This would not only reduce the search time but also assist the search algorithm in the exploration of large search space. One such strategy is the Zero Cost (ZC) proxy, which evaluates a DL architecture on a small number of data

⋆ Corresponding Author

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

samples to quickly estimate individual performance [10]. This approach is particularly effective in medical image classification, where traditional NAS-based methods are computationally expensive due to the large number of samples.

Data Augmentation (DA) is crucial for performance enhancement in medical image analysis tasks, as it prevents overfitting and enhances the model's generalization ability to perform well on unseen data. By creating variations of existing data, DA can help prevent overfitting. However, due to the diverse nature of medical imaging datasets, a single DA strategy may perform differently on different datasets. To address this issue, multiple automatic DA strategies have been proposed, employing different optimization strategies to search for the best DA policy. The combination of architecture components and DA expands the search space, making it challenging to find an optimal set of DA policies and the best architecture simultaneously. To tackle this issue, we divided the proposed approach into two stages: (i) architecture search and (ii) automatic DA search. At first, an architecture is searched using the proposed NAS approach, and then the best-suited pair of DA techniques is searched using the proposed automatic DA approach. The main contributions of this study are as follows:

- We have proposed an evolutionary NAS approach for both 2D and 3D medical image classification.
- The proposition of a DA technique capable of searching the best augmentation topology for a given dataset.
- Experiments on both small-scale and large-scale datasets are conducted to demonstrate the effectiveness of the proposed approach.

2 Related Work

In recent years, numerous evolutionary NAS approaches have been proposed to solve different tasks, including image classification [6] and medical image classification [3]. These approaches have adopted different performance estimation strategies to reduce the searching time [17]. For experimentation, these approaches use standard benchmark datasets like CIFAR-10 and ImageNet for image classification and MedMNIST for medical image classification.

Multiple NAS studies have been proposed for medical image classification and used MedMNIST datasets for experiments [3] [2] [4] [6]. These studies have used different performance estimation strategies to reduce the searching time, such as surrogate models, ZC proxies, and One-Shot NAS approaches. The surrogate models, also known as performance predictors are machine learning models that predict the individual's fitness during evolution to reduce the search time. These machine learning models are trained on individual representations and their corresponding fitness values on the initialized population and retrained during evolution [2]. One-Shot NAS trains a supernet first, then samples sub-networks and uses a weight-sharing mechanism to save the time required for re-training. Additionally, ZC proxies use a mini-batch of data to quickly estimate the model's performance [6].

Moreover, studies have shown that DA plays a crucial role in enhancing the model generalization ability and the model performance on unseen data. These

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

studies proposed searching for both DA and network topology simultaneously. Zhang et al. proposed a unified approach for searching both DA policy and network topology [5]. They introduced an augmentation density matching algorithm that addresses the inefficiency of density matching caused by in-domain sampling bias. They first trained a Super-Net and then used an evolutionary algorithm to search for sub-networks with optimal augmentation policy for the given dataset.

Existing studies were proposed for either 2D or 3D architectures but not both. To this aim, this study addresses both 2D and 3D NAS architectures for medical image classification. Compared to 2D NAS approaches, 3D NAS approaches are computationally expensive due to the increasing model complexity and computational costs. Incorporating ZC proxies as a performance estimation strategy could reduce the search time of the overall NAS approach. To the best of our knowledge, this is the first study to use zero-cost proxies with 3D NAS for medical image classification.

Furthermore, different meta-heuristic algorithms have been used as a search strategy in different NAS studies. Some of the famous NAS approaches have used Genetic Algorithm (GA) [9], Particle Swarm Optimization (PSO) [8], Differential Evolution (DE) [7] and other different algorithms. Unfortunately, choosing an optimal metaheuristic for a given problem is a difficult task. In this study, we performed a comparative study of famous meta-heuristics to compare their performance to choose the best-performing one.

3 Proposed Methodology

The proposed methodology is mainly divided into two primary stages: (1) the architecture search stage and (2) the DA search stage, as illustrated in Figure 1. Initially, a population of individual neural networks is randomly generated, and then an evolutionary algorithm searches for the best-performing individual. The SynFlow ZC proxy is used for fitness evaluation, which assesses individual Neural Network (NN) performance on the validation set. The second stage involves the GA to search for the appropriate DA policy. The fitness is computed by training each model on the training set and evaluating it on the validation set. The best-performing DA strategy is used while training the final architecture. Then, the accuracy and Area Under the Curve (AUC) scores are computed on the test set. Further details regarding search space and the encoding scheme are provided in the following sections.

3.1 Search Space

In this study, we have used cell-based search space initially proposed in NASNet study by Zolph et al. [12] and DARTS study [11], which consists of small NN blocks called cells that can be repeated or connected in various ways to form a complete NN architecture. This search space has been widely adopted in different NAS-based studies due to its simplicity and flexibility in extending its components. These cells apply different convolution operations to get feature

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

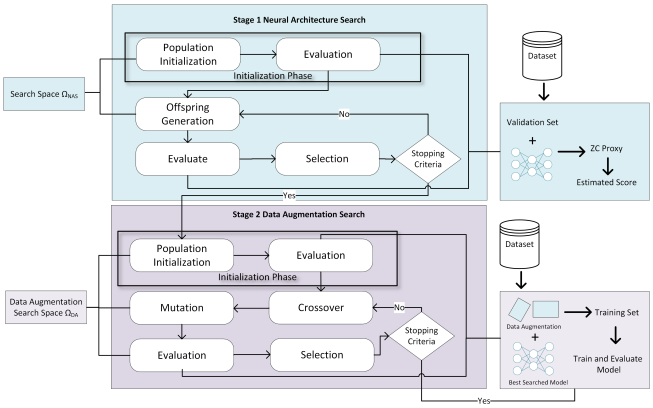

Fig. 1: Proposed methodology diagram consisting of two stages (i) neural architecture search using zero-cost proxies and (ii) data augmentation strategy using genetic algorithm

maps, which can be passed to other cells. It consists of two types of cells: normal and reduction cells. The normal cell computes the feature map of an input image, where convolution and pooling in the cell have a stride of 1 to keep the same resolution. In contrast, the reduction cell uses the stride of 2 to reduce the feature map dimension to down-sample the feature maps. The whole architecture is formed by stacking the cells one after another.

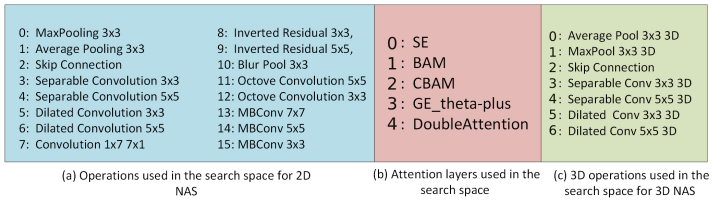

Fig. 2: Operations and attention layers for 2D architecture and 3D operations for 3D architecture used in the search space to design CNN architectures

Moreover, an attention-based search space is used which comprises 16 different convolution and pooling operations and 5 different attention layers as shown in Figure 2. These attention layers are adopted from the Attention-DARTS [13] study. The idea behind using attention layers is to focus more on salient regions. The attention layer is used after the candidate operations. For the 3D search space, we have used seven different operations, consisting of pooling layers, dilated convolution, and separable convolution with varying kernel sizes, as shown in Figure 2 [c].

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

In our 2D search space, we have used various lightweight and efficient CNN components from different architectures, including InceptionNets, OctoveNet, MobileNet, Invetable Residual Networks etc. These operations consist of mobile, dilated, separable, octove and inverted residual blocks with different kernel settings. Attention layers include Gather-Excite (GE) Attention, Squeeze and Excitation (SE), Convolution Bottleneck Attention Module (CBAM), Bottleneck Attention Module (BAM) and Double Attention (DA) blocks.

3.2 Encoding Scheme

For the representation of an individual, each candidate operation is expressed as a real value between 0 and 1. This real value is mapped to the corresponding operation by multiplying the number of operations and then applying floor operation to get the corresponding operation number. Besides, the associated attention layer is represented by a value between 1 and 5, where 1-5 means different attention types. Similarly, for 3D search space, each candidate operation in the individual is expressed by a real value between 0 and 1. An example representation of genotype and phenotype is shown in Figure 3.

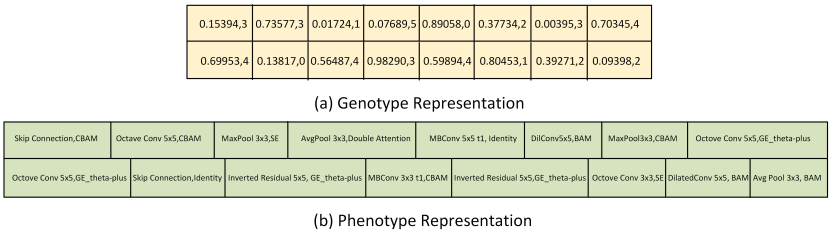

Fig. 3: Genotype and phenotype representations of a sample individual consisting of 16 genes

For instance, a genotype (0.56,4),(0.42,5),(0.29,4), (0.8926,1) which is decoded into (InvRes3x3,4), (DilatedConv5x5,5), (Separable Convolution 5x5,4), (MBConv 5x5,1) such as (0.56,4) mapped into (InvRes3x3,4) and 0.56x16 where each gene is multiplied with the number of candidate operations in the search space and the floor operation is applied and corresponding operation is fetched from the list consisting of different operations.

3.3 NAS Approach

As discussed above, multiple metaheuristic algorithms have been used as search strategies in NAS. Choosing the optimal meta-heuristic for a given problem is not straightforward. Thus, we have performed a comparative analysis of different meta-heuristics and chosen the optimal one in terms of final results. Evaluating the fitness of an individual is a time-consuming task. To address this challenge,

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

we have utilized zero-cost proxies that are based on recent pruning at initialization [27]. They use a single mini-batch of training data to compute a model score.

Abdelfattah et al. [10] proposed using existing proxies to estimate scoring in a DL architecture including SynapticFlow (SynFlow) and SNIP, which are used in this study. Recent works have shown that SynFlow assists the evolutionary algorithm in reducing the searching time while searching the optimal architecture [18]. It is noteworthy to mention that these proxies were used for both 2D and 3D datasets in this study.

3.4 Data Augmentation

DA is a series of transformations applied to the input data. It plays an essential role in DL-based medical image analysis. It increases the amount and diversity of the training data and reduces overfitting. However, manually developing a tailored DA strategy for each dataset is difficult because of the heterogeneity of medical imaging data and the different characteristics of each disease and modality. To overcome this issue, a GA-based approach is proposed to automatically search for a suitable DA policy for a given dataset.

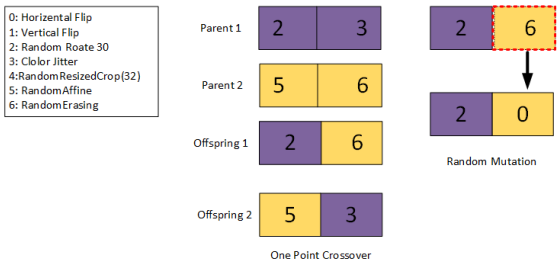

Fig. 4: Search space consisting of different data augmentation techniques and crossover and mutation operators

Figure 4 shows the search space, crossover and mutation operators used in the proposed approach. Seven DA strategies have been used in the search space, such as horizontal and vertical flipping, random rotation, resizing, cropping, random affine, erasing, and colour jitter. Each individual consists of two different kinds of DAs that have been used. The number of DA for each dataset is set to two, as experiments with more than two were not significant in terms of improvement rate. For reproduction and mutation, a one-point crossover is used, and a random mutation strategy is adopted, replacing a bit with a randomly selected DA from the search space.

To evaluate individual fitness, we have considered AUC and accuracy scores. These measures evaluate the model's performance in different aspects: accuracy measures the correctness of the model, and AUC measures the model's ability

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

to distinguish between positive and negative instances. The fitness function is given as:

$$Fitness = AUC + Accuracy \tag{1}$$

4 Experimental Settings

We have performed all the experiments with the same hardware configurations for a fair comparison. All the experiments were performed on a GPU cluster with a NVIDIA A100 GPU and 96 GB of RAM. The proposed framework is implemented in Python. PyTorch library is used for DL implementation and Mealpy for implementing metaheuristics. For NN training, we have used the Adam optimizer with a MultiStep Learning Rate scheduler and a learning rate of 0.0025, a gamma rate of 0.1, and a weight decay of 0.0003. The final architecture is trained on 300 epochs; for the GA-based DA strategy, each individual is trained on 25 epochs. The accuracy and AUC scores on validation sets are used as the fitness function.

Algorithm	Parameters	**Values**
GA (Data Augmentation)	**Crossover Probability**	0.95
	Mutation Probability	0.1
	Selection Technique	Tournament Selection
	Population Size	4
	Number of Generations	4
	Chromosome Length	2
PSO	**c1 and c2**	2
	Population Size	200
	Number of Generations	200
	Intertia weight	decreases from 0.95 to 0.4
DE	**F**	0.7
	Crossover Rate	0.9
	Population Size	200
	Number of Generations	200
	Mutation Strategy	DE/Current-to-best/2
LSHADE	μ_f	0.5
	Population Size	200
	Last Population Size	50
	Number of Generations	200
	μ_{CR}	0.5
ACO	**Z**	1.5
	q	0.7
	Population Size	200
	Number of Generations	200
	Sample Count	35

Table 1: Parameters' settings of different meta-heuristics

For MedMNIST implementation, we used the implementation provided by the authors and the same dataset split of train, val, and test. During the search phase, a validation set is used to evaluate the architecture and the test set is used to evaluate the final searched architecture with a given DA topology. For the Breast Cancer Histopathological Image Classification (BreakHIS) dataset, we divided the dataset into 80:10:10 ratios to create a train/validation and test set. We also used the same methodology for the MedMNIST dataset. The DA policy is only searched for MedMNIST 2D datasets, and the number of layers is

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

set from 8 to 12 for 2D datasets and 15 to 20 for 3D datasets. For the BreakHIS dataset, a layer size of 15 is used.

Multiple trials of experiments set up the parameters of GA for DA, and PSO, DE and ACO parameters are tuned using grid-search parameter tuning. The population size and number of generations are set to 200 for all the meta-heuristics. The parameter settings for the different algorithms are given in Table 1, where the best-reported parameters of each meta-heuristic algorithm are given.

4.1 Datasets Description

We have used the datasets from the MedMNIST benchmark and the BreakHIS (Breast Histopathology) dataset for experiments. MedMNIST is a large-scale MNIST-like standard dataset benchmark for biomedical imaging. It includes 12 2D and 6 3D datasets from different organs and modalities. All images of 2D and 3D datasets are pre-processed into 28x28 resolution with corresponding labels. These datasets belong to different organs and modalities, such as histopathology, abdominal computed tomography, breast mammograms, retinal fundus, microscopy, chest X-ray, and dermoscopy images.

Apart from MedMNIST, we have also performed experiments on the BreakHis dataset, which consists of 7,904 microscopic images of breast tumors collected from 82 patients with different magnification levels (40x, 100x, 200x and 400x) and 8 different classes. As these images are high resolution, we have resized them to 224x224 for ease of training. It contains 2,480 benign and 5,091 malignant samples. This dataset includes multiple samples. The idea behind using both small and large-scale datasets is to evaluate the performance of the proposed approach.

5 Experimental Results

We have performed experiments on 2D and 3D datasets from the MedMNIST benchmark and the BreakHIS dataset. The results of the proposed approach are compared with multiple variants of ResNet architectures with two different resolutions (28 and 224), AutoML approaches (AutoSKlearn, AutoKeras and Google AutoML Vision) and multiple NAS approaches proposed for natural images and medical images. The comparison in terms of accuracy is shown in Table 2, and the AUC is shown in Table 3.

Figure 5 shows the AUC scores of different NAS and deep learning approaches on MedMNIST 2D datasets. Each line represents a dataset, and dotted points represent the results of different methods. These lines show better performance on NAS-based approaches than other AutoML and deep learning approaches. Furthermore, the proposed approach also has better or equal performance than other NAS-based approaches. AutoSKLearn is an AutoML tool to automate the machine learning process which includes searching for optimal classifiers, features pre-processing and data processing. AutoKeras is also an AutoML approach that uses bayesian optimization to evolve the architecture for a given problem.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

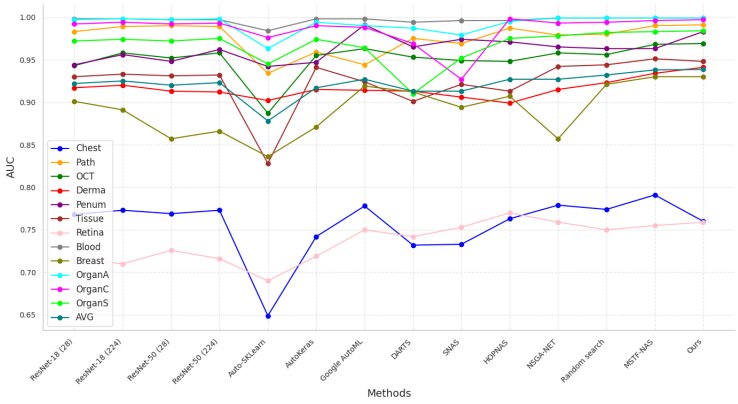

Fig. 5: Line graph representing the AUC scores of different NAS and deep learning approaches on MedMNIST 2D datasets

Google AutoML Vision is a cloud-based solution consisting of state-of-the-art NAS algorithms that design architecture automatically for the given dataset.

Methods	Chest	Path	OCT	Derma	Tissue	Retina	Blood	Breast	OrganA	OrganC	OrganS	Pneum	AVG.
ResNet-18 (28)	0.947	0.907	0.743	0.735	0.676	0.524	0.958	0.863	0.935	0.900	0.782	0.854	0.818
ResNet-18 (224)	0.947	0.909	0.763	0.754	0.681	0.493	0.963	0.833	0.951	0.920	0.778	0.864	0.821
ResNet-50 (28)	0.947	0.911	0.762	0.735	0.680	0.528	0.956	0.812	0.935	0.905	0.770	0.854	0.816
ResNet-50 (224)	0.948	0.892	0.776	0.731	0.680	0.511	0.950	0.842	0.947	0.911	0.785	0.884	0.821
Auto-SKlearn	0.779	0.716	0.601	0.719	0.532	0.515	0.878	0.803	0.762	0.829	0.672	0.855	0.721
AutoKeras	0.937	0.834	0.763	0.749	0.703	0.503	0.961	0.831	0.905	0.879	0.813	0.878	0.813
Google AutoML	0.948	0.728	0.771	0.768	0.673	0.531	0.966	0.861	0.886	0.877	0.749	**0.946**	0.809
DARTS [11]	0.934	0.872	0.712	0.749	0.648	0.510	0.953	0.832	0.926	0.791	0.808	0.874	0.800
SNAS [26]	0.938	0.850	0.708	0.737	0.708	0.515	0.946	0.811	0.918	0.891	0.778	0.871	0.805
HOPNAS [25]	0.947	0.912	0.761	0.759	0.698	0.523	0.958	0.853	0.937	0.911	0.803	0.852	0.826
NSGA-NET [24]	0.947	0.866	0.765	0.744	0.712	0.540	0.970	0.846	0.952	0.923	0.820	0.907	0.832
Random search	0.946	0.854	0.760	0.773	0.717	0.542	0.966	0.897	0.955	0.923	0.820	0.904	0.838
MSTF-NAS [23]	0.945	0.910	0.780	**0.774**	0.740	0.550	**0.976**	0.872	**0.962**	**0.936**	0.838	0.912	0.841
Ours	**0.949**	**0.920**	**0.800**	0.769	**0.760**	**0.561**	0.967	**0.911**	0.957	0.935	**0.864**	0.944	**0.861**

Table 2: Results comparison of the proposed approach with different NAS methods and deep learning approaches on the MedMNIST benchmark in terms of accuracy.

Existing NAS approaches to which the proposed approach is compared include Differentiable Architecture Search (DARTS), Stochastic Neural Architecture Search (SNAS), HOPNAS, Non-Dominated Sorting Genetic Algorithm (NSGA-Net), and Multi-Scale Training Free (MSTF-NAS). We have used the same results of these approaches given in the recent study named MSTF-NAS [23]. In MSTF-NAS, the authors have proposed a multi-scale, training-free NAS approach for medical image classification. They compared their proposed approach with existing NAS studies on 12 subsets from the MedMNIST benchmark.

These NAS approaches include DARTS, SNAS, and HOPNAS, which are one-shot NAS approaches based on SuperNet. Among these NAS approaches,

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

DARTS and NSGANet are specifically designed for natural images, and SNAS and HOPNAS are designed for medical images. The results of random search on 2D MedMNIST datasets by [23] are also included in the comparison. In random search, instead of using some guided search algorithm, an architecture is picked randomly till given iterations and the best architecture is selected at completion.

Methods	Chest	Path	OCT	Derma	Penum	Tissue	Retina	Blood	Breast	OrganA	OrganC	OrganS	AVG.
ResNet-18 (28)	0.768	0.983	0.943	0.917	0.944	0.930	0.717	0.998	0.901	0.997	0.992	0.972	0.922
ResNet-18 (224)	0.773	0.989	0.958	0.920	0.956	0.933	0.710	0.998	0.891	0.998	0.994	0.974	0.924
ResNet-50 (28)	0.769	0.990	0.952	0.913	0.948	0.931	0.726	0.997	0.857	0.997	0.992	0.972	0.920
ResNet-50 (224)	0.773	0.989	0.958	0.912	0.962	0.932	0.716	0.997	0.866	0.998	0.993	0.975	0.922
Auto-SKLearn	0.649	0.934	0.887	0.902	0.942	0.828	0.690	0.984	0.836	0.963	0.976	0.945	0.878
AutoKeras	0.742	0.959	0.955	0.915	0.947	0.941	0.719	0.998	0.871	0.994	0.990	0.974	0.917
Google AutoML	0.778	0.944	0.963	0.914	**0.991**	0.924	0.750	0.998	0.919	0.990	0.988	0.964	0.927
DARTS [11]	0.732	0.975	0.953	0.913	0.965	0.901	0.742	0.994	0.912	0.987	0.969	0.910	0.913
SNAS [26]	0.733	0.969	0.949	0.906	0.974	0.921	0.753	0.996	0.894	0.979	0.927	0.952	0.913
HOPNAS [25]	0.763	0.987	0.948	0.899	0.971	0.913	0.770	0.996	0.907	0.995	**0.998**	0.975	0.926
NSGA-NET [24]	0.779	0.979	0.958	0.915	0.965	0.942	**0.759**	**0.999**	0.857	**0.999**	0.993	0.978	0.926
Random search	0.774	0.980	0.956	0.923	0.963	0.944	0.750	**0.999**	0.921	**0.999**	0.994	0.982	0.932
MSTF-NAS [23]	**0.791**	0.990	0.968	0.934	0.963	0.951	0.755	**0.999**	**0.930**	**0.999**	0.996	0.983	0.938
Ours	0.760	**0.991**	**0.969**	**0.941**	0.983	**0.948**	**0.759**	**0.999**	**0.930**	**0.999**	0.997	**0.984**	**0.9383**

Table 3: Results comparison of the proposed approach with different NAS methods and deep learning approaches on the MedMNIST benchmark in terms of AUC

One of the reasons behind the good performance of random search in NAS is that, usually, NAS search spaces are high-dimensional, with a large number of possible configurations. Random sampling can effectively explore different regions of the search space in such spaces, thereby increasing the likelihood of discovering good solutions without the constraints of guided search methods.

Methods	Organ	Nodule	Fracture	Adrenal	Vessel	Synapse	Avg.
ResNet-18 + 2.5D	0.977	0.838	0.587	0.718	0.748	0.634	0.750
ResNet-18 + 3D	**0.996**	0.863	0.712	0.827	0.874	0.820	0.848
ResNet-18 + ACS	0.994	0.873	0.714	0.839	0.930	0.705	0.842
ResNet-50 + 2.5D	0.974	0.835	0.552	0.732	0.751	0.669	0.752
ResNet-50 + 3D	0.994	0.875	0.725	0.828	0.907	**0.851**	0.863
Auto-SKLearn	0.977	**0.914**	0.628	0.828	0.910	0.631	0.814
AutoKeras	0.979	0.844	0.642	0.804	0.773	0.538	0.763
Proposed	0.995	0.871	**0.728**	**0.857**	**0.940**	0.820	**0.868**

Table 4: Results comparison of the proposed approach with different NAS methods and deep learning approaches on the 3D datasets MedMNIST benchmark in terms of AUC score.

The reason behind the performance difference between NAS approaches addressing natural images and approaches addressing medical ones is mainly the domain gap. On the one hand, natural images depict landscapes, and objects that are captured by digital means. On the other hand, medical images visualize anatomical structures and pathological conditions within the body obtained

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

through different modalities. They also include small lesions and tumour regions. The attention mechanism highlights important regions to achieve an accurate feature extraction improving detection accuracy and interoperability. The results reveal that the proposed approach exhibits a better performance on average compared to the other studies. Regarding MSTF-NAS, the proposed approach has an advantage over DA as it helps improve model generalization ability on unseen data.

In Table 4 and Table 5, the comparison in terms of the performance metrics (AUC and Accuracy) of the proposed approach with different variants of ResNet and AutoML approaches on 3D datasets from the MedMNIST benchmark is given. The conclusion from the results is that our proposed approach yields better performance in comparison with existing approaches, as the average accuracy and AUC scores are better than existing DL and AutoML approaches.

The proposed approach exhibits an overall satisfactory performance on average and less searching time than existing approaches. Following the same context, the number of parameters of searched architectures is less compared to ResNet DL architectures. The proposed approach can achieve a better exploration and exploitation of individuals thanks to the ZC proxies, which allow reduced search costs to evaluate an individual.

Methods	Organ	Nodule	Fracture	Adrenal	Vessel	Synapse	Avg.
ResNet-18 + 2.5D	0.788	0.835	0.451	0.772	0.846	0.696	0.731
ResNet-18 + 3D	0.907	0.844	0.508	0.721	0.877	0.745	0.767
ResNet-18 + ACS	0.900	0.847	0.497	0.754	0.928	0.722	0.774
ResNet-50 + 2.5D	0.769	0.848	0.397	0.763	0.877	0.735	0.731
ResNet-50 + 3D	0.883	0.847	0.494	0.745	0.918	0.795	0.780
ResNet-50 + ACS	0.889	0.841	0.517	0.758	0.858	0.709	0.762
Auto-SKLearn	0.814	0.874	0.453	0.802	0.915	0.730	0.764
AutoKeras	0.804	0.834	0.458	0.705	0.894	0.724	0.736
Proposed	**0.908**	**0.877**	**0.690**	**0.805**	**0.940**	**0.846**	**0.844**

Table 5: Results comparison of the proposed approach with different NAS methods and deep learning approaches on the 3D datasets MedMNIST benchmark in terms of Accuracy score.

Similarly, experiments on the BreakHIS were conducted (the histopathology dataset, which consists of eight classes with a 200x magnification level). While prior NAS studies mainly focused on small-scale datasets, our research highlights the potential use of ZC proxies for efficiently exploring an architecture suited for large-scale datasets. A comparison of the best-performing architecture results with existing approaches is shown in Table 7. These results are based on the validation performance after the network training. The results reveal that the proposed approach performed better regarding multiple performance measures, taking only two hours to search an architecture with a ZC proxy.

5.1 Comparison of different Meta-heuristics

Although studies on evolutionary NAS have used different meta-heuristic algorithms, choosing a suitable meta-heuristic algorithm for a given problem remains

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

Metaheuristics		LSHADE	ACO	PSO	DE
Path	**Accuracy**	**0.89**	0.79	0.80	0.86
	AUC	**0.97**	0.96	0.96	0.93
OCT	**Accuracy**	0.74	0.73	0.74	**0.77**
	AUC	0.95	0.94	0.95	**0.96**
OrganA	**Accuracy**	**0.93**	0.92	0.92	0.91
	AUC	**0.99**	**0.99**	**0.99**	0.99
OrganC	**Accuracy**	0.89	0.85	0.89	**0.90**
	AUC	0.98	0.98	0.98	**0.99**
OrganS	**Accuracy**	0.75	**0.86**	0.78	0.77
	AUC	0.96	0.75	**0.97**	**0.97**
Breast	**Accuracy**	**0.90**	0.80	0.89	0.89
	AUC	0.89	0.85	**0.90**	0.88
Retina	**Accuracy**	**0.54**	0.52	**0.54**	0.51
	AUC	**0.75**	0.74	0.74	0.72
Pneumonia	**Accuracy**	**0.88**	0.83	0.84	0.82
	AUC	0.95	0.95	**0.96**	**0.96**
Derma	**Accuracy**	0.70	0.70	**0.74**	0.65
	AUC	0.89	0.89	**0.91**	0.85
Chest	**Accuracy**	**0.94**	**0.94**	**0.94**	**0.94**
	AUC	0.62	**0.66**	0.65	0.65
Tissue	**Accuracy**	**0.67**	0.65	0.64	0.57
	AUC	**0.92**	0.90	0.91	0.47
Blood	**Accuracy**	0.95	0.95	**0.96**	0.93
	AUC	**0.99**	0.99	**0.99**	**0.99**
Average Value Accuracy		**0.815**	0.80	0.80	0.79
Average Value AUC		**0.90**	0.88	**0.90**	0.86

Table 6: Comparison of Accuracy and AUC scores of different meta-heuristic algorithms without data augmentation strategy.

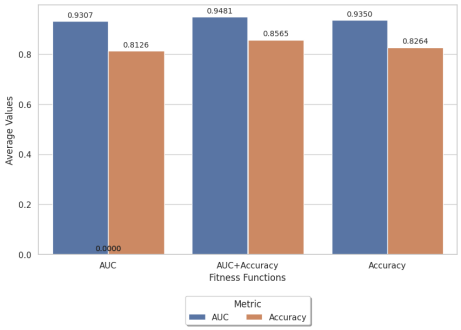

Fig. 6: Average fitness values on multiple 2D MedMNIST datasets of three different fitness functions of proposed Automatic Data Augmentation approach

challenging. In this study, we performed a comparative study of some of the well-known metaheuristics. It includes DE [19] alongside its variant LSHADE [22], ACO [21] and PSO algorithms [20].

The results of different meta-heuristics used to search architectures for 2D MedMNIST datasets are given in Table 6. It is worth noting that no meta-heuristic performed well on all the datasets, even if the PSO performed better on average than DE and ACO. As DE performance is sensitive towards its parameters, LSHADE performed well compared to DE, a variant of DE which is a successive history-based adoptive DE variant that keeps track of all best DE parameters with linear population size reduction. LSHADE algorithms have been widely used to solve large-scale optimization problems. It performed better despite the parameter tuning of DE due to several reasons, such as its adaptability, as it incorporates an adaptive mechanism that dynamically adjusts its parameters during the optimization process. This adaptability allows LSHADE to respond better to changes in the optimization landscape and maintain a balance between exploration and exploitation more effectively than DE.

Secondly, LSHADE explored the search space more efficiently than DE leading to the discovery of better solutions. Its methods explore diverse regions within the search space, which is crucial in finding near-optimal or optimal solutions to complex problems such as NAS.

6 Discussion and Conclusion

The study presents an efficient evolutionary neural architecture search method for 2D and 3D medical image classification by utilizing ZC proxies for fitness

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

	Accuracy	Precision	Recall	F1-Score
Bardou [14]	80.083	81.85	80.83	80.48
Yun Jaing [15]	92.270	90.71	92.24	91.42
Nouman et al [16]	94.710	91.42	91.63	91.76
Proposed	**95.05**	**93.76**	**93.59**	**93.66**

Table 7: Results comparison of the proposed approach with existing studies on BreakHIS dataset.

Dataset	Best data augmentation searched	
OCT	RandomAffine	RandomHorizentalFlip
OrganS	RandomVerticalFlip	RandomErasing
OrganC	RandomAffine	RandomErasing
OrganA	ColorJitter	RandomAffine
Path	ColorJitter	RandomVertical
Tissue	RandomVerticalFlip	RandomRotation
Pneumonia	RandomHorizental	RandomErasing
Chest	HorizentalFlip	VerticalFlip
Breast	ColorJitter	RandomRotate30
Blood	RandomAffine	RandomErasing

Table 8: Data Augmentation strategies searched by the proposed Automatic Data Augmentation approach

evaluation. It introduces an adaptive DA approach to address model generalization and overfitting issues and conducts a comparative study of metaheuristics to choose the optimal one for the problem at hand. The comparison of our proposed approach with existing NAS studies demonstrates its effectiveness. It outperforms existing approaches in terms of average performance. Furthermore, incorporating attention layers in the search space enables better feature extraction by prioritizing relevant regions of the image and capturing long-range dependencies.

This shows that the proposal is not only effective in the case of small-scale datasets such as MedMNIST but also in the case of large-scale datasets such as BreakHIS, requiring less searching time. DA is crucial for medical images, improving the robustness and generalization ability of DL models trained on limited datasets. Moreover, medical images often suffer from overfitting problems, which can be overcome by using DA. By searching for optimal DA sets for given datasets, an improvement in the performance is noticed. It reveals that it plays an essential role alongside NAS when searching architecture topology.

Data Augmentation	**Path**	**OCT**	**OrganA**	**OrganC**	**OrganS**	**Breast**	**Retina**	**Pneumonia**	**Derma**	**Chest**	**Tissue**	**Blood**
	0.900	0.780	0.930	0.900	0.841	0.900	0.540	0.926	0.970	0.740	0.940	0.670
✓	**0.920**	**0.800**	**0.962**	**0.935**	**0.864**	**0.911**	**0.561**	**0.944**	**0.983**	**0.769**	**0.949**	**0.760**

Table 9: Experimental Results on MedMNIST2D datasets on accuracy before and after data augmentation searched by the proposed Automatic Data Augmentation (ADA) approach

The best-reported set of DA searched by the proposed adaptive DA approach for different MedMNIST datasets is given in Table 8. This shows that the adaptive DA approach searches for various DAs. Besides, performance deterioration can emerge when using a single DA strategy. The accuracy and AUC scores before and after training the model with searched DAs are given in Table 9 and 10. The idea behind using the combined fitness value of AUC and accuracy is to find the data augmentation set that gives better performance on both AUC and accuracy measures. We also conducted an analysis study to compare the average accuracy and AUC values using different fitness metrics (AUC, accuracy, and AUC+accuracy), as shown in Figure 6. It clearly shows that combined

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

fitness function leads to better individual DAs in comparison with other fitness functions.

Data Augmentation	Path	OCT	OrganA	OrganC	OrganS	Breast	Retina	Pneumonia	Derma	Chest	Tissue	Blood
	0.970	0.960	0.990	0.990	0.970	0.925	0.750	0.970	0.910	0.941	0.670	0.920
✓	**0.991**	**0.969**	**0.999**	**0.997**	**0.984**	**0.930**	**0.759**	**0.983**	**0.944**	**0.941**	**0.769**	**0.967**

Table 10: Experimental Results on MedMNIST2D datasets on AUC before and after data augmentation searched by the proposed Automatic Data Augmentation (ADA) approach

As recent studies have been conducted on small-scale datasets, e.g., the MedMNIST benchmark, we have also performed experiments on the BreakHIS dataset and demonstrated the potential ability of ZC-NAS approaches to find the best-performing neural architectures for large-scale medical image datasets. Moreover, this study demonstrates that ZC is effective on both 2D and 3D datasets, as shown for multiple 3D MedMNIST datasets. Moreover, the comparison of meta-heuristics shows that the adaptive variant of DE and PSO algorithms performed better than other metaheuristics with satisfactory accuracy and AUC scores. In the near future, we aim to extend the 3D approach for large-scale 3D medical image datasets and propose a multi-objective approach.

Acknowledgment

This work was funded by ArtIC project “Artificial Intelligence for Care” (grant ANR-20-THIA-0006-01) and co-funded by IRIMAS Institute / Université de Haute Alsace. The authors would like to thank the Mesocentre of Strasbourg for providing access to the GPU cluster

References

1. Yang, Jiancheng, et al. "MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification." Scientific Data 10.1 (2023): 41.
2. Ali, M. J., et al. "Designing CNNs using Surrogate-assisted GA for Medical Image Classification." Proc. Companion Conf. on Genetic and Evolutionary Computation. 2023.
3. Ali, M. J., et al. "Designing Attention-Based CNNs for Medical Image Classification Using GA with Variable Length-Encoding." Intl Conf. on Artificial Evolution (EA). Springer, Cham, 2022.
4. Liao, P., Jin, Y., Du, W.: EMT-NAS: Transferring architectural knowledge between tasks from different datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023).
5. Zhang, J., Zhang, L., Li, D.: A Unified Search Framework for Data Augmentation and Neural Architecture on Small-scale Image Datasets. In: IEEE Transactions on Cognitive and Developmental Systems (2023).
6. Zhang, J., et al.: An Efficient Multi-Objective Evolutionary Zero-Shot Neural Architecture Search Framework for Image Classification. In: International Journal of Neural Systems, 33.05, pp. 2350016 (2023).
7. Huang, Junhao, et al. "EDE-NAS: An Eclectic Differential Evolution Approach to Single-Path Neural Architecture Search." Australasian Joint Conference on Artificial Intelligence. Cham: Springer International Publishing, 2022.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9

8. Niu, Ruicheng, et al. "Neural architecture search based on particle swarm optimization." 2019 3rd International Conference on Data Science and Business Analytics (ICDSBA). IEEE, 2019.
9. Deng, Shuchao, Yanan Sun, and Edgar Galvan. "Neural architecture search using genetic algorithm for facial expression recognition." Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2022.
10. Abdelfattah, Mohamed S., et al. "Zero-cost proxies for lightweight nas." arXiv preprint arXiv:2101.08134 (2021).
11. Liu, Hanxiao, Karen Simonyan, and Yiming Yang. "Darts: Differentiable architecture search." arXiv preprint arXiv:1806.09055 (2018).
12. Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
13. Nakai, Kohei, Takashi Matsubara, and Kuniaki Uehara. "Att-darts: Differentiable neural architecture search for attention." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.
14. Bardou, Dalal, Kun Zhang, and Sayed Mohammad Ahmad. "Classification of breast cancer based on histology images using convolutional neural networks." Ieee Access 6 (2018): 24680-24693.
15. Jiang, Yun, et al. "Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module." PloS one 14.3 (2019): e0214587.
16. Ahmad, N., Asghar, S., Gillani, S.A.: Transfer learning-assisted multi-resolution breast cancer histopathological images classification. In: The Visual Computer, 38.8, pp. 2751–2770 (2022).
17. Ren, Pengzhen, et al. "A comprehensive survey of neural architecture search: Challenges and solutions." ACM Computing Surveys (CSUR) 54.4 (2021): 1-34.
18. Vo, A., Pham, T.N., Luong, N.H.: Lightweight Multi-Objective and Many-Objective Problem Formulations for Evolutionary Neural Architecture Search with the Training-Free Performance Metric Synaptic Flow. In: Informatica, 47.3, (2023).
19. Feoktistov, Vitaliy. Differential evolution. Springer US, 2006.
20. Kennedy, J., Eberhart, R.: Particle swarm optimization (PSO). In: Proc. IEEE Intl Conf. on Neural Networks, Perth, Australia. Vol. 4. No. 1. pp. 1942–1948 (1995).
21. Dorigo, M., Birattari, M., Stutzle, T.: Ant colony optimization. In: IEEE Computational Intelligence Magazine, 1.4, pp. 28-39 (2006).
22. Tanabe, R., Fukunaga, A.S.: Improving the search performance of SHADE using linear population size reduction. In: 2014 IEEE Congress on Evolutionary Computation (CEC). IEEE, pp. 1658–1665 (2014).
23. Wang, Yan, et al. "MedNAS: Multi-Scale Training-Free Neural Architecture Search for Medical Image Analysis." IEEE Transactions on Evolutionary Computation (2024).
24. Lu, Zhichao, et al. "Nsga-net: A multi-objective genetic algorithm for neural architecture search." (2018).
25. Zhang, Jianwei, et al. "One-shot neural architecture search by dynamically pruning supernet in hierarchical order." International journal of neural systems 31.07 (2021): 2150029.
26. Xie, Sirui, et al. "SNAS: stochastic neural architecture search." arXiv preprint arXiv:1812.09926 (2018).
27. Wang, Huan, et al. "Recent Advances on Neural Network Pruning at Initialization."

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_9