Medical Neural Architecture Search: Survey and Taxonomy

Hadjer Benmeziane [1] , **Imane Hamzaoui** [2] , **Zayneb Cherif** [3] , **Kaoutar El Maghraoui** [4]
[1]IBM Research Europe, 8803 Rüschlikon, Switzerland.
[2]École nationale Supérieure d'Informatique, Algiers, Algeria.
[3] Yorktown High school, Yorktown Heights, 10598, USA .
[4] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA.
Hadjer.Benmeziane1@ibm.com, ji_hamzaoui@esi.dz, kelmaghr@us.ibm.com

Abstract

This paper presents a comprehensive survey of Medical Neural Architecture Search (MedNAS), a burgeoning field at the confluence of deep learning and medical imaging. With the increasing prevalence of FDA-approved medical deep learning models, MedNAS emerges as a key area in leveraging computational innovations for healthcare advancements. Our survey examines the paradigm shift introduced by Neural Architecture Search (NAS), which automates neural network design, replacing traditional, manual designs. We explore the unique search spaces tailored for medical tasks on different types of data from images to EEG, the methodologies of MedNAS, and their impact on medical applications.

1 Introduction

The rapidly evolving field of medical deep learning (DL) applications has garnered significant interest, evidenced by numerous models achieving FDA approval [Benjamens *et al.*, 2020]. This surge in development and regulatory endorsement underscores the critical role of advanced computational techniques in transforming healthcare. Among these, Medical Neural Architecture Search (MedNAS) emerges as a pivotal area, bridging the gap between state-of-the-art machine learning methodologies and the intricate demands of medical data analysis. Figure 1 shows the increasing number of papers targeting automatic deep learning architecture design for medical tasks.

Neural Architecture Search (NAS) [Elsken *et al.*, 2019], a cornerstone in this evolution, signifies a transformative shift in the realm of DL. It marks the transition from the traditional, expertise-driven, and often heuristic approach to the design of neural network architectures to a more systematic and algorithm-driven process. NAS harnesses sophisticated algorithms to autonomously conceive potential network architectures, with an emphasis on optimizing performance metrics and computational efficiency. Several optimization algorithms are used to explore different architecture search spaces, among them evolutionary algorithms and gradient-based approaches are predominant. These strategies involve approximating the time-expensive performance measurements, using surrogate models and weight sharing. For medical tasks, applying NAS involves: 1 defining a specific search space. The search space depends on the targeted tasks and input modality. For segmentation tasks, it is generally inspired by the U-Net [Weng *et al.*, 2019a] architecture. 2 carefully choosing the objectives. While in conventional NAS, objectives such as fairness, interpretability, and certainty are overlooked, in medical settings those become critical. 3 implementing a search strategy and an evaluation methodology. The evaluation methodology is the bottleneck component of conventional NAS. Given that MedNAS is multi-objective by nature, this component is even more crucial. Finding a fast yet accurate methodology to approximate the different objectives is the main differentiation between the multiple MedNAS frameworks. The search strategy usually depends on the evaluation methodology. If the search space is defined as a supernetwork to allow the use of weight sharing, a gradient-optimization strategy is used. If instead a surrogate model is trained, or a zero-cost metric is used a proxy, evolutionary algorithm or bayesian optimization are used. A large search space may require a combination of gradient-optimization and evolutionary to speed up the exploration.

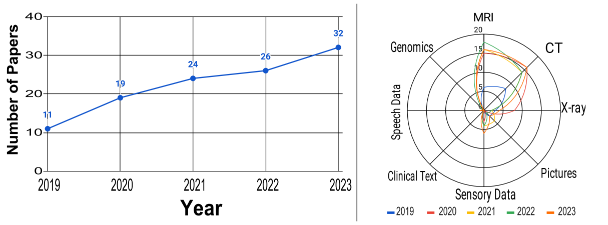

Figure 1: MedNAS statistics and growth

In this survey, we explain the details of each step to build a MedNAS framework. We provide a comprehensive analysis of how the unique characteristics of medical data and the stringent requirements of healthcare applications influence the design of neural architectures. This includes a deep dive into the nuances of defining search spaces tailored to specific medical tasks, such as diagnostic imaging, genomics, or patient data analysis. We highlight how each MedNAS objective can be measured, and how it can be approximated. Moreover, we explore the intricacies of implementing effective search strategies and evaluation methodologies in Med-

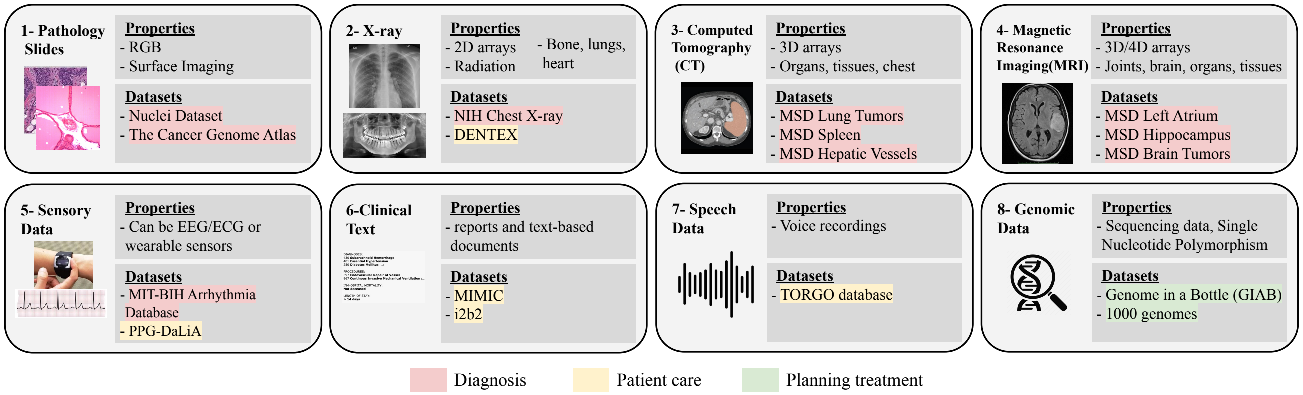

Figure 2: Common medical data modalities.

NAS. Given the multi-objective nature of medical tasks and the computational constraints inherent in healthcare applications, we examine how different frameworks approach the challenge of balancing speed, accuracy, and complexity. Finally, this survey identifies gaps in current research and proposes potential future directions for MedNAS. We discuss emerging trends, such as the integration of federated learning for privacy-preserving MedNAS, and the exploration of more efficient and scalable NAS techniques suited for real-world medical applications.

2 Existing Surveys

Numerous surveys have thoroughly explored the realms of deep learning in medical applications [Kumar *et al.*, 2023] and Neural Architecture Search (NAS) [Elsken *et al.*, 2019] separately, each contributing valuable insights into their respective fields.

However, the specialized intersection of NAS within medical applications, known as Medical Neural Architecture Search (MedNAS), represents a nuanced and emerging field that has not been as comprehensively surveyed. This book's chapter [Vo-Ho *et al.*, 2023] highlight some of the work done on NAS for medical image segmentation. Our survey aims to fill this gap by providing a detailed exploration of MedNAS. We endeavor to connect the dots between the general principles of NAS and their tailored application in the complex landscape of medical tasks. By offering a comprehensive overview of the current state of MedNAS, discussing the unique challenges it faces, and projecting future directions, our work seeks to serve as a cornerstone for researchers and practitioners in this intersection of technology and healthcare.

3 MedNAS Problem Formulation

NAS is conceptualized as a bi-level optimization problem, defined in equation 1. At its core, NAS operates on two interrelated levels: the upper level focuses on the architecture search space, where optimal neural network architectures are identified, and the lower level deals with the training of these architectures to minimize a predefined loss function.

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & L_{val}(w^*(\alpha), \alpha) \\ \text{subject to} \quad & w^*(\alpha) = \underset{w}{\text{argmin}}\, L_{train}(w, \alpha), \end{aligned} \tag{1}$$

In the context of MedNAS, this formulation takes on additional layers of complexity. The search space extends beyond a mere assembly of architectural elements, as it is deeply intertwined with the particularities of the the medical datasets. These datasets often exhibit high variability and distinctive features, adding layers of complexity to the search space. This necessitates a search strategy that can navigate a space rich in diverse architectural possibilities while maintaining a focus on medical-specific performance metrics such as diagnostic accuracy and interpretability.

Furthermore, the training process at the lower level is confronted with challenges inherent to medical data, including limited sample sizes, imbalance in class distributions, and the critical need for model robustness and generalization. The optimization process in MedNAS must, therefore, be adept at handling these nuances, often requiring bespoke strategies that go beyond conventional NAS methodologies.

4 Medical Data Modalities

Medical datasets are characterized by a range of data modalities, each presenting distinct features and requiring specialized handling. These modalities play critical roles in various aspects of healthcare, such as diagnosis, patient care, and treatment planning. Figure 2 show the different modalities and datasets.

Diagnostic modalities are pivotal in identifying and understanding medical conditions, providing clear images of internal body structures. Pathology slides, or microscopic images, are used to diagnose various types of cancer and other tissue abnormalities, with datasets like The Cancer Genome Atlas (TCGA) [Tomczak *et al.*, 2015] offering extensive histopathological data. The Nuclei dataset [Caicedo *et al.*, 2019] , extracted from TCGA, is commonly used for breast cancer detection. The Medical Segmentation Decathlon (MSD) [Antonelli *et al.*, 2022] is a ten datasets benchmark with CTs and MRIs to enable brain tumor, lung tumor, and spleen segmentation. The NIH Chest X-rays [Wang *et al.*,

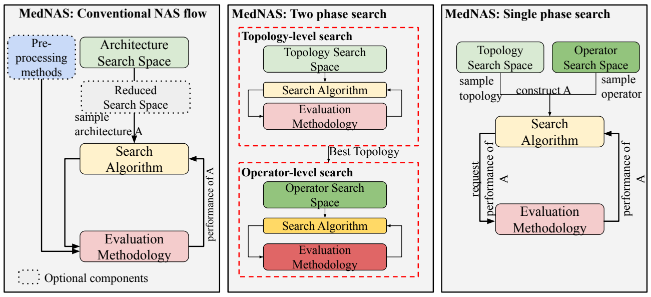

Figure 3: Taxonomy of MedNAS strategies

2017] , used for pneumonia detection on chest X-rays. Additionally, Ultrasound images, used in datasets such as DUS Dataset [Cordts *et al.*, 2016] , are vital for prenatal diagnostics and examining internal organs. Electroencephalogram (EEG) and electrocardiogram (ECG) are non-invasive tests that record electrical activity datasets such as MIT-BIH Arrhythmia Database [Moody and Mark, 2001].

For ongoing patient care, data from wearable sensors and electronic health records (EHRs) are indispensable, offering continuous monitoring and comprehensive patient history, respectively. Wearable sensors, as in the PPG-DaLiA dataset [Reiss *et al.*, 2019] , track health metrics like heart rate and physical activity, which are crucial for chronic disease management. EHRs, exemplified by datasets like MIMIC-III [Johnson *et al.*, 2016] and i2b2, provide a holistic view of patient history, treatments, and progress, facilitating ongoing care and management in a text format. These datasets are used to train language models that generate reports.

In treatment planning, genomics data is increasingly crucial, enabling personalized medicine approaches based on individual genetic profiles. Datasets like the 1000 Genomes Projectand Genome in a Bottle (GIAB) [Zook *et al.*, 2019] offer insights into genetic variations, aiding in the development of targeted therapies. Notably, MRI not only serves a key role in diagnosis, particularly for soft tissue conditions but also aids significantly in planning treatments, especially for surgeries and cancer therapies.

Given the plurality of datasets and modalities, there is a necessity for an automated system that can design the best-performing architecture. Selecting an appropriate architecture for a given medical task is another layer of complexity. The range of potential architectures is vast, from conventional convolutional networks to more recent, intricate designs like capsule networks or attention-based models. Each architecture has its strengths and weaknesses, making the selection process non-trivial. Additionally, pre-processing of medical data plays a crucial role in model performance. Techniques like normalization, augmentation, and feature extraction can significantly impact the effectiveness of the chosen architecture, necessitating careful consideration in the NAS process.

4.1 Taxonomy & Trends

The field of MedNAS has witnessed a significant rise in popularity, which can be attributed to the growing adoption of DL models in medical tasks. In figure 3, we provide a general taxonomy of the methodology used in MedNAS.

First, multiple MedNAS works [Song *et al.*, 2021; Wang *et al.*, 2024] use the same NAS flow, specifically those that target tiny machine learning tasks such as arrhythmia detection [Moody and Mark, 2001]. The NAS flow, involves sampling an architecture from the search space, evaluating that architecture and based on the performance, efficiently sampling the next architecture. For these tasks, real-time training is usually applied to obtain the evaluation metrics. These tasks defines a large search space of small networks, which are fast to evaluate. For large search spaces, methods to reduce the search space are used. The goal is to prune architectures which are considered inefficient or less likely to yield optimal performance. Techniques such as network morphism, constraint-based search, or heuristic pruning are often employed. These methods enable a more focused search by eliminating architectures that do not meet predefined criteria. For example, BiX-NAS [Wang *et al.*,] includes a differentiable NAS to narrow down the search space, followed by a novel progressive evolutionary search.

The most complex medical tasks are segmentation and detection. These tasks are hindered by the large size of their datasets and networks which makes training during the search completely impractical. To mitigate this, MedNAS uses two-level search strategies: the topology-level and the operator-level. An operator, in this paper, refers to a block of layers, such as a residual block or a VGG block. In optimizing the topology-level, the focus is on determining the optimal number of operators and their interconnections while keeping the individual operator configurations constant. On the other hand, optimizing the operator-level involves fine-tuning the configuration of each operator, assuming the overall number of operators and their connections are predefined. Strategies can then choose to optimize each level independently, which gives rise to the two-phase search class, or jointly optimizing the topology and operator levels by sampling from each search space at the same time to build the architecture.

C2FNAS [Yu *et al.*, 2020] searches for both the topology- and operator- levels through a two-phase optimization problem for different segmentation tasks. Similarly, Thrifty NAS [Chen *et al.*, 2022] involves a two-phase optimization where they re-engineered a U-Net-like backbone architecture with dense connections, enabling feature map reuse and consequently lowering the parameter count. At the operator level, their focus shifts to crafting a operator structure search space that selectively retains feature maps, thereby diminishing GPU memory usage. On the same tasks, EMONAS-NET [Baldeon-Calisto and Lai-Yuen, 2021] proposed a simultaneous approach to solve the two optimization problems to speed up the search process. Other works, only focus on one search space. While fixing the topology, ENAS U-Net [Gessert and Schlaefer, 2019] searches new operator blocks. Contrary, Resource Optimized NAS [Bae *et al.*, 2019] focuses on the topology level only and uses a standard U-Net block as an operator.

An optional but that appears critical in Medical settings, is to explore and search for the best pre-processing strategy along side the architecture. NN-Unet [Isensee *et al.*, 2020] offers a streamlined data pre-processing approach for medical imaging. It standardizes dataset resolution through isotropic

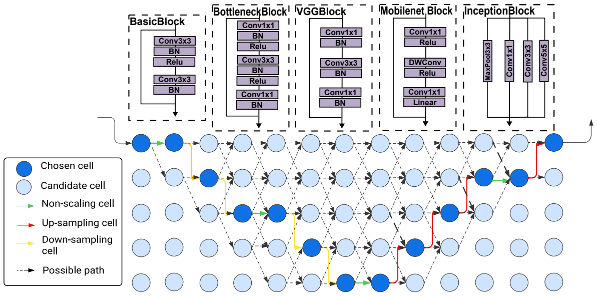

Figure 4: Medical imaging search space based on U-Net-like architecture. Each cell can be one of the blocks defined on top.

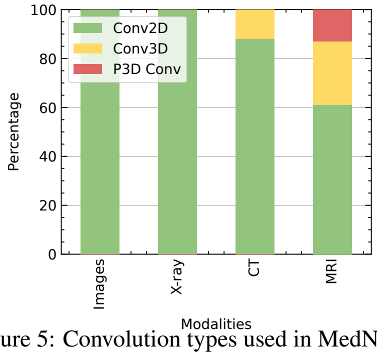

Figure 5: Convolution types used in MedNAS.

resampling to address varying voxel sizes in medical scans. Each image is then normalized, generally by adjusting the intensity values to a standard scale. They search for the adequate data augmentation as well. NAS-PPG [Song *et al.*, 2021] includes a specific heuristic for data pre-processing according to the sensory data types. RO-NAS [Bae *et al.*, 2019] includes the input images resolution in their search space.

5 Search Space

In this section, we detail the commonly used search space and design techniques. As images modalities are the most common and most complex, we start with search spaces that target segmentation and detection tasks on CT, MRI or X-rays. Most of these search spaces are based on the U-Net architecture. We then provide a detailed description of different search spaces for other medical tasks including those dealing with sensory data.

5.1 Medical Imaging Search Spaces

While recent works [Yu *et al.*, 2020; Weng *et al.*, 2019b; Isensee *et al.*, 2020] adhere to the U-Net-like architecture, it is worth mentioning that some the early works such as MM-NAS [Peng *et al.*, 2020], uses a search space inspired from NASNet [Zoph *et al.*, 2018] with a sequence of normal and reduction convolutional blocks. Enhanced MRI Reconstruction NAS (EMR-NAS) [Huang *et al.*, 2020] creates a unique search space, featuring eight operators sequentially tailored for MRI reconstruction. However, these search spaces result in suboptimal performance [Kim *et al.*, 2019].
When considering a U-Net-like architecture, we define two search spaces: the topology and operator levels.

Topology-level

Figure 4 illustrates the possible paths in a U-Net-like search space. This architecture has a downsampling path, i.e., the left side of the U-shape, and an upsampling path. The downsampling path reduces the feature map size to extract small features. The upsampling enlarges back the feature maps. This strategy is essential to detect tiny tumors and larger ones in medical images. The search space implementation can be either sequential [Yu *et al.*, 2020; Baldeon-Calisto and Lai-Yuen, 2021] or recursive [Weng *et al.*, 2019b]. Recursive implementation forces the upsampling path to have the same number of downsampling operators. While it restricts the search space, it makes up for efficient search.

The same search space is used by several methodologies including, V-NAS [Zhu *et al.*, 2019], Resource optimized NAS [Bae *et al.*, 2019], Efficient NAS U-Net (ENAS-U-Net) [Gessert and Schlaefer, 2019], Quantum inspired NAS (segQNAS) [Carlos *et al.*, 2023] and Mixed Block NAS [Bosma *et al.*,], NG-NAS [Qin *et al.*, 2023]. However, they can differ in the allowed maximum depth and additional nodes in both the encoder and decoder. For instance, NG-NAS [Qin *et al.*, 2023] uses a carefully designed skip connections to avoid additional computational costs while. Different types of skip connections have been exploited as they improve accuracy, however, they increase the computational resources needed for the search space and many works tried to overcome this issue [Gessert and Schlaefer, 2019; Bae *et al.*, 2019]. The types include either applying an element-wise sum or a concatenation to merge the feature maps while maintaining a basic U-shape backbone. This method consciously avoids adding superfluous skip connections between nodes of varying resolutions, which, while minimally impacting model performance, significantly escalates computational costs and latency. Thrifty-NAS reimagines the backbone architecture by integrating dense connections for effective feature map reuse and incorporates both downsampling and upsampling operators within a densely connected framework. Bix-NAS [Wang *et al.*,] stands out with a modified multiscale bi-directional NAS on the backbone of Bio-Net.

Some works used unconventional backbone networks, while still targeting the same tasks. NAS-DBN [Qiang *et al.*, 2020] design a search space of Deep Belief Networks (DBN). They vary the number of layers and the neuron counts in the fully-connected architecture. The NAS framework for adversarial medical image segmentation, as presented in [Dong *et al.*,], marks a significant innovation by incorporating Generative Adversarial Networks (GANs) as its backbone. This framework automates the design of discriminator architectures, essential to GANs, using NAS.

Operation-level

The operators used in the topology are searched at this level. Generally, the same operator is used in the whole architecture, modifying only the output channel number to extract more features.

Due to volumetric nature of medical imaging modalities, We distinguish different types of convolutions such as: dilated, depth, depthwise separable, 2D, 3D, and P3D convolutions. 2D convolutions are suitable for processing 2D images, while 3D convolutions are used for volumetric data, such as the slices of different organ regions in MRIs. Pseudo-3D (P3D) convolutions, a variant of 3D convolutions, can also be applied to such volumetric data for efficient and effective analysis. Figure 5 illustrates the diversity of operators in the search spaces over the years and modalities.

Figure 5 shows an increasing number of search spaces using 3D and pseudo 3D convolutions for CT and MRI datasets. Although 3D convolutions achieve higher performance with MRI and CT datasets, their training cost is expensive, which hinders the search.

Figure 6 shows the types of operators used in MedNAS over the years. A combination of multiple common blocks are used including VGGBlock, Basic residual blocks, Bottleneck blocks, Inception and Mobilenet blocks. The bottleneck block are the most used type. This block is the originally used one in U-Net architecture, which makes it practical. Recently, the use of attention-based blocks such as SwinBlock in Swin U-Net [Cao *et al.*,] have succeded in outperforming end-to-end convolution models. However, there are no medNAS methodology yet with an attention-based search space.

MB-NAS [Bosma *et al.*,] stands out as it defines the search space with a variety of different pre-defined blocks, including VGG, Residual, Dense, and Inception blocks, to streamline the architecture design process.

In addition to the block type, other hyperparameters are also tuned such as kernel size, stride, and padding.

5.2 Search Space for Sensory Data

Sensory datasets include EEG, ECG, and wearable device sensors such as Photoplethysmogram (PPG). These data are usually time series and require different architectural backbones for the search spaces.

For EEG and ECG data, MedNAS search spaces are represented with small Convnets. [Li *et al.*, 2023a]'s search space is a sequence of reduction and normal convolutions, in which different hyperparameters, including kernel size, are searched. TNAS [Li *et al.*, 2023b] considered incorporating transformers and proposed a multi-objective NAS framework that finds the optimal number of heads and the number of hidden layers to maximize accuracy and minimize the number of parameters.

[Wang *et al.*, 2022] have built two search spaces for a CNN baseline: spectral and temporal. Both of them are defined with a convolutional neural network for which convolution and its hyperparameters are searched.

AutoEER [Wu *et al.*, 2023] extends the definition of the search space with a wider set of operators including convolution, transformer, 2D convolution, Local-Global-Graph Network (LGGNet), Channel Wise Attention (CWA) in addition to a skip connection and a zero operation.

Wearable devices have also seen significant advancements in the application of MedNAS. PPG sensors are generally used for pulse rate estimation. These tasks are fast and practical for NAS. NAS-PPG [Song *et al.*, 2021] defines a search space based on convolution, long-short-term memory (LSTM), and fully-connected layers.

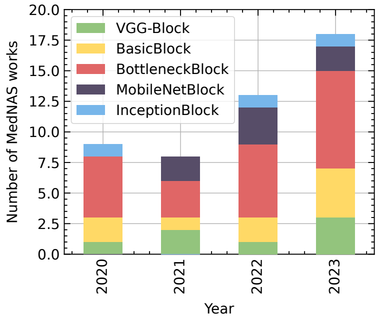

Figure 6: Operators used in MedNAS

6 Search Algorithms and Strategies

Diverse search algorithms have been used for the different MedNas techniques. In this section, we briefly describe these methodologies.

Evolutionary Algorithm mimic the process of natural selection by iteratively generating, evaluating, and selecting candidate solutions based on their performance or 'fitness'. These algorithms typically start with a randomly generated population of solutions, and over successive generations, they apply operations akin to genetic crossover and mutation to evolve increasingly effective solutions to a given problem. C2FNAS [Yu *et al.*, 2020], for example, uses a two phase evolutionary algorithm with the dice score as a fitness for segmentation tasks.

Gradient-based optimization uses the supernetwork methodogy inspired by DARTS [Liu *et al.*, 2018]. This approach involves training a large network that encompasses many different sub-network architectures, allowing the optimization process to explore a vast search space. By backpropagating gradients through this supernetwork, the method efficiently identifies the most effective architecture, balancing performance and computational efficiency. BiX-NAS [Wang *et al.*,] and MixSearch [Liu *et al.*, 2021] use these supernetworks for both topology and operator-level exploration. Although training these supernetworks can be resource-intensive, it is typically a one-time process, after which the optimized network can be deployed for various applications without the need for further extensive training. However, this is true for each targeted task.

Reinforcement Learning is used as a generator of efficient architectures by framing the architecture design process as a sequential decision-making problem. In this approach, an agent, typically a controller network, iteratively proposes architectures, which are then evaluated for their performance on a given task. The feedback from this evaluation, often in the form of accuracy or other performance metrics, is used as a reward signal to train the controller using reinforcement learning algorithms. RO-NAS [Bae *et al.*, 2019] uses this strategy and defines a large RNN controller trained to predict the activation type, pooling type, convolution delated rate and other hyperparameters.

Bayesian Optimization This methodology leverages Bayesian inferential statistics to efficiently explore neural network architectures for medical use, integrating prior

Category	Example tasks	Common metrics
Classification	Disease diagnosis from images Skin lesion classification Cell type classification	Accuracy Precision Recall F1 Score AUC-ROC
Detection	Tumor detection in radiology images Polyp detection in endoscopy Lesion detection in dermatology images	IoU mAP Recall Precision
Segmentation	Organ segmentation in CT/MRI Lesion segmentation in radiology images Cell segmentation in microscopy images	Dice Coefficient Jaccard Index (IoU) Pixel Accuracy Sensitivity Specificity
Sequence Prediction	ECG signal classification EEG signal analysis Time-series prediction in medical data	Accuracy Precision Recall F1 Score PRC
Regression	Estimating organ/tumor volume Predicting patient's length of stay Age estimation from medical images	MSE MAE RMSE
Anomaly Detection	Identifying abnormal radiology images Detecting irregular heartbeats in ECG Unusual patterns in medical time-series data	Sensitivity Specificity AUC-ROC F1 Score
NLP	Information extraction from clinical notes Automated report generation Sentiment analysis in patient feedback	Accuracy Precision Recall F1 Score BLEU score

Table 1: Medical deep learning tasks and their task-performance metrics

knowledge and empirical data for iterative refinement. This approach is advantageous in medical contexts with limited or sensitive data. [Odema *et al.*, 2021] employs this strategy for multi-objective hyperparameter search in binary convolutions, aiming at energy-efficient myocardial infarction detection on wearable devices.

7 Multi-Objective MedNAS

MedNAS is inherently multi-faceted, aiming not just to optimize a single objective but multiple, often conflicting, objectives. This section delves into the various objectives of MedNAS, emphasizing their definitions and significance in the medical domain.

Task-specific Performance: The foremost objective in MedNAS is performance, defined as the model's ability to correctly interpret and classify medical data. In clinical settings, high accuracy is vital, as misdiagnosis or incorrect predictions can have serious implications. MedNAS strives to develop architectures that yield the highest possible accuracy. The task-specific performance may differ from one task to another. Table 1 summarizes the performance metrics used in common medical deep learning applications. Most MedNAS strategies, fully train a supernetwork [Yu *et al.*, 2020; Isensee *et al.*, 2020; Wang *et al.*,] or generally the sampled networks [Song *et al.*, 2021; Li *et al.*, 2023b]. Currently, this makes MedNAS strategies extremely time-consuming. In NAS, performance estimators are used to overcome this challenge. However, these methods compromise optimality. A MedNAS method [Wang *et al.*, 2024] uses the FLOPs as a metric for performance, assuming larger models always yield better result. This significantly speed up the search process at the expense of performance. Note that fairness, in this context, because of its high importance, is included with the task-specific performance.

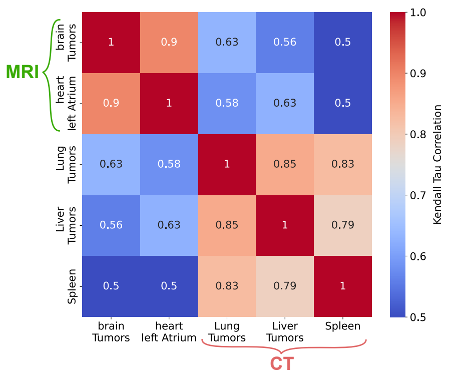

Figure 7: Cross-datasets architecture ranking correlation.

Multi-tasking: The ability to use a single architecture for multiple datasets and task is especially appealing for medical settings. In a single modality, the input data are highly similar. Figure 7 shows a small experiments in which we extract 1000 architectures and compare their ranking on different tasks using the kendal tau correlation. Datasets are extracted from MSD [Antonelli *et al.*, 2022] and the search space is represented with U-Net-like architecture akin to C2FNAS [Yu *et al.*, 2020]. We see the same architecture can be used for multiple datasets, which can significantly enhance and speed MedNAS frameworks. Besides, given the low-memory devices used by radiologist all over the world, deploying a single architecture, would greatly improve the medical tools.

Interpretability: is the degree to which a human can understand the cause of a decision made by the model. In medical settings, interpretability is vital for gaining clinicians' trust and for validating the model's decisions. It is essential that these AI systems provide insights that are comprehensible to medical professionals. While qualitative, it can be approached through feature importance scores and visualization techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and shapely values (SHAP). [Zhang *et al.*, 2023] uses SHAP values to find interpretable convolution networks to understand enzymatic reactions.

Robustness: Robustness in MedNAS refers to the model's ability to maintain performance despite variations in input data, such as noise or artifacts typical in medical images. Ensuring robustness is crucial, as real-world medical data often come with such imperfections. [Dong *et al.*,] uses an adversarial training during the search to find robust architecture, however, this strategy is extremely time-consuming and more efficient methods are interesting future works.

Uncertainty: Another important metric in medical applications is uncertainty. Particularly in medical diagnostics where decisions must be made with confidence. Metrics to measure uncertainty include: confidence intervals for predictions, entropy-based measures for classification uncertainty, and bayesian approaches for quantifying model uncertainty. NAS-DBN [Qiang *et al.*, 2020] focuses on bayesian neural networks to find an architecture that is performant and confident.

Another important objective includes hardware efficiency. Specifically, for on-going patient care using wearable resource constrained devices. The following section is dedi-

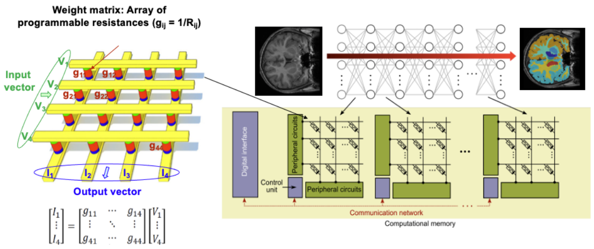
Weight matrix: Array of programmable resistances ($g_{ij} = 1/R_{ij}$)
Input vector
V_1
V_2
V_3
V_4
g_{11}
g_{12}
g_{21}
g_{22}
g_{44}
I_1
I_2
I_3
I_4
Output vector
$\begin{bmatrix} I_1 \\ \vdots \\ I_4 \end{bmatrix} = \begin{bmatrix} g_{11} & \cdots & g_{14} \\ \vdots & \ddots & \vdots \\ g_{41} & \cdots & g_{44} \end{bmatrix} \begin{bmatrix} V_1 \\ \vdots \\ V_4 \end{bmatrix}$
Digital interface
Peripheral circuits
Control unit
Peripheral circuits
Communication network
Computational memory

Figure 8: Analog In-Memory Computing Concept

cated on hardware-aware MedNAS.

8 Hardware-Aware MedNAS

In the medical domain, where real-time processing and portability can be crucial, designing DL architectures need to consider hardware efficiency. This includes considerations for memory footprint, computational power, energy consumption, and latency.

8.1 Edge Computing Potential

The reality in many healthcare environments, particularly in hospitals with limited budgets, is that the availability of computationally intensive resources is often constrained. This limitation necessitates the development of DL models that are not overly reliant on high-resource environments. Besides, wearable devices with constrained resources have become an integral part of the on-going patient care. This is where Hardware-aware Medical Neural Architecture Search (MedNAS) plays a crucial role. It aims to tailor neural network architectures to achieve not just high performance in medical tasks but also optimal compatibility with diverse hardware configurations.

RO-NAS [Zeng *et al.*, 2020] searched for a real-time 3D cardiac cine MRI segmentation on a supernetwork search space. They included the latency in their loss function as a regularizer. These task are extremely important for cardiac intervention assistance.

In contrast, [Odema *et al.*, 2021] searches for an energy-efficient binary convolutional network using bayesian optimization. They target myocardial infarction detection on low-power wearable devices. Each sampled architecture is deployed on a SmartCardia INYU and the energy consumption is measured.

8.2 In-Memory Computing Potential

Analog In-memory computing (AIMC) [Sebastian *et al.*, 2020], a paradigm shift in computational architecture, holds substantial promise for enhancing the capabilities of MedNAS. AIMC addresses the shortcomings of traditional Von-Neumann computing in handling the increasing volume of medical data and the demand for real-time analysis. As illustrated in Figure 8, an AIMC chip consists of crossbar arrays, each representing a neural network layer. Neural network synaptic weights are stored as charge or conductance states in memory devices at crosspoints, handling both positive and negative weights. Data is processed through these layers in a single step, with input on the rows and output from the columns, followed by a nonlinear neuron function at the crossbar edge. In feed-forward networks like MLPs or CNNs, each array interfaces with the subsequent layer's array, while in RNNs, the output feeds back into its own input. However, the inherent noise and susceptibility to conductance drift in these chips pose significant challenges to AIMC's effectiveness, particularly in maintaining model accuracy.

A study by [Hamzaoui *et al.*, 2024], focused on AIMC's role in medical AI, specifically in brain tumor detection, spleen segmentation, and nuclei identification, showed that introducing noise could be a strategic advantage in AIMC, showcasing its potential benefits in providing robust model training, enhancing noise resilience, and improving prediction certainty. Another key finding was that transformer models have greater noise tolerance compared to pyramidal alternatives, ultimately contributing to more robust and certain predictions in healthcare settings. A novel closed-loop, continuous-time AIMC-based resistive memory circuit has shown significant promise for compressed sensing (CS) recovery [Wang *et al.*, 2023], which shows the interest in these types of hardware for medical settings.

9 Conclusion and Future Directions

NAS is crucial for advancing medical imaging through automated neural network design, yet it faces multiple challenges. The **computational expense** is substantial due to the need for training and evaluating numerous architectures on complex datasets. Issues with the **generalizability** of MedNAS-designed architectures across varied medical imaging data and patient populations also persist. The inherent black-box nature of NAS hinders the **incorporation of essential domain-specific knowledge**, leading to a **lack of interpretability**—a critical obstacle for clinical adoption. Additionally, the **scarcity and diversity of medical imaging data** and **data privacy concerns** complicate NAS model training and deployment. Effectively addressing these concerns requires a tailored approach that considers the unique needs and limitations of the medical imaging domain. Looking ahead, future research in MedNAS should concentrate on developing more efficient algorithms, enhancing the generalizability of NAS architectures, and integrating domain-specific knowledge to improve interpretability. Key focuses also include addressing deployment challenges by creating NAS frameworks compliant with healthcare regulations and seamlessly integrating the validation reporting required by FDA.

Acknowledgements

This research work was supported by Abu Dhabi National Oil Company (ADNOC), Emirates NBD, Sharjah Electricity Water & Gas Authority (SEWA), Technology Innovation Institute (TII) and GSK as the sponsors of the 4th Forum for Women in Research (QUWA): Sustaining Women's Empowerment in Research & Innovation at University of Sharjah. We acknowledge support from the IBM Research AI Hardware Center and the Center for Computational Innovation (CCI) at Rensselaer Polytechnic Institute for computational resources on the AiMOS Supercomputer.

References

[Antonelli *et al.*, 2022] Michela Antonelli, Annika Reinke, and al. The Medical Segmentation Decathlon. *Nature Communications*, 2022.

[Bae *et al.*, 2019] Woong Bae, Seungho Lee, Yeha Lee, Beomhee Park, Minki Chung, and Kyung-Ho Jung. *Resource Optimized Neural Architecture search for 3D medical image segmentation*. 2019.

[Baldeon-Calisto and Lai-Yuen, 2021] María Baldeon-Calisto and Susana K. Lai-Yuen. EMONAS-Net: Efficient multiobjective neural architecture search using surrogate-assisted evolutionary algorithm for 3D medical image segmentation. *Artificial Intelligence in Medicine*, 119:102154, 9 2021.

[Benjamens *et al.*, 2020] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 2020.

[Bosma *et al.*,] Martijn Bosma, Arkadiy Dushatskiy, Monika Grewal, Tanja Alderliesten, and Peter A. N. Bosman. Mixed-block neural architecture search for medical image segmentation. *Medical Imaging 2022: Image Processing*.

[Caicedo *et al.*, 2019] Juan C Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, Cherkeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Hossein Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods*, pages 1247–1253, 2019.

[Cao *et al.*,] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV 2022 Workshops*, pages 205–218.

[Carlos *et al.*, 2023] Guilherme Carlos, Karla Figueiredo, Abir Hussain, and Marley Vellasco. Segqnas: Quantum-inspired neural architecture search applied to medical image semantic segmentation. In *IJCNN*, pages 1–8, 2023.

[Chen *et al.*, 2022] Ruibin Chen, Miao Zhang, Xin Zheng, and Shirui Pan. Thrifty neural architecture search for medical image segmentation. *AAAI*, 36:12925–12926, 2022.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[Dong *et al.*,] Nanqing Dong, Min Xu, Xiaodan Liang, Yiliang Jiang, Wei Dai, and Eric Xing. Neural architecture search for adversarial medical image segmentation. In *MICCAI 2019*.

[Elsken *et al.*, 2019] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, pages 1–21, 2019.

[Gessert and Schlaefer, 2019] Nils Gessert and A. Schlaefer. Efficient neural architecture search on low-dimensional data for oct image segmentation. *ArXiv*, abs/1905.02590, 2019.

[Hamzaoui *et al.*, 2024] Imane Hamzaoui, Hadjer Benmeziane, Zayneb Cherif, and Kaoutar El Maghraoui. Are analog in-memory computing the future of medical imaging segmentation? *ArXiv*, abs/1905.01392, 2024.

[Huang *et al.*, 2020] Qiaoying Huang, Dong yang, Yikun Xian, Pengxiang Wu, Jingru Yi, Hui Qu, and Dimitris Metaxas. Enhanced mri reconstruction network using neural architecture search. In Mingxia Liu, Pingkun Yan, Chunfeng Lian, and Xiaohuan Cao, editors, *Machine Learning in Medical Imaging*, pages 634–643, 2020.

[Isensee *et al.*, 2020] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, pages 203–211, 2020.

[Johnson *et al.*, 2016] Alistair E. W. Johnson, Tom Pollard, Lu Shen, Li Wei H. Lehman, Mengling Feng, Mohammad M. Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 2016.

[Kim *et al.*, 2019] Sungwoong Kim, Ildoo Kim, Sungbin Lim, Woonhyuk Baek, Chiheon Kim, Hyungjoo Cho, Boogeon Yoon, and Taesup Kim. Scalable neural architecture search for 3d medical image segmentation. In *MICCAI*, 2019.

[Kumar *et al.*, 2023] Rakesh Kumar, Pooja Kumbharkar, Sandeep Vanam, and Sanjeev Sharma. Medical images classification using deep learning: a survey. *Multimedia Tools and Applications*, 2023.

[Li *et al.*, 2023a] Chang Li, Zhongzhen Zhang, Rencheng Song, Juan Cheng, Yu Liu, and Xun Chen. Eeg-based emotion recognition via neural architecture search. *IEEE Transactions on Affective Computing*, pages 957–968, 2023.

[Li *et al.*, 2023b] Chang Li, Zhongzhen Zhang, Xiaodong Zhang, Guoning Huang, Yu Liu, and Xun Chen. Eeg-based emotion recognition via transformer neural architecture search. *IEEE Transactions on Industrial Informatics*, pages 6016–6025, 2023.

[Liu *et al.*, 2018] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv:1806.09055*, 2018.

[Liu *et al.*, 2021] Luyan Liu, Zhiwei Wen, Songwei Liu, Hong-Yu Zhou, Hongwei Zhu, Weicheng Xie, Linlin Shen, Kai Ma, and Yefeng Zheng. Mixsearch: Searching for domain generalized medical image segmentation architectures. abs/2102.13280, 2021.

[Moody and Mark, 2001] GB Moody and RG Mark. The impact of the mitbih arrhythmia database. ieee engineering in medicine and biology magazine. *IEEE Engineering in Medicine and Biology Magazine, 2001*, 2001.

[Odema *et al.*, 2021] Mohanad Odema, Nafiul Rashid, and Mohammad Abdullah Al Faruque. Energy-aware design methodology for myocardial infarction detection on low-power wearable devices. In *Asia and South Pacific DAC*, 2021.

[Peng *et al.*, 2020] Yige Peng, Lei Bi, Michael Fulham, Dagan Feng, and Jinman Kim. Multi-modality information fusion for radiomics-based neural architecture search. In *MICCAI*, 2020.

[Qiang *et al.*, 2020] Ning Qiang, Qinglin Dong, Wei Zhang, Bao Ge, Fangfei Ge, Hongtao Liang, Yifei Sun, Jie Gao, and Tianming Liu. Modeling task-based fmri data via deep belief network with neural architecture search. *Computerized Medical Imaging and Graphics*, 2020.

[Qin *et al.*, 2023] Shixi Qin, Zixun Zhang, Yuncheng Jiang, Shuguang Cui, Shenghui Cheng, and Zhen Li. Ng-nas: Node growth neural architecture search for 3d medical image segmentation. *Computerized Medical Imaging and Graphics*, 108:102268, 2023.

[Reiss *et al.*, 2019] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors*, 19(14):3079, 7 2019.

[Sebastian *et al.*, 2020] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou. Memory devices and applications for in-memory computing. *Nature nanotechnology*, 15(7):529–544, 2020.

[Song *et al.*, 2021] Seok Bin Song, Jung Woo Nam, and Jin Heon Kim. Nas-ppg: Ppg-based heart rate estimation using neural architecture search. *IEEE Sensors Journal*, 21(13):14941–14949, 2021.

[Tomczak *et al.*, 2015] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Wspolczesna Onkologia-Contemporary Oncology*, 2015.

[Vo-Ho *et al.*, 2023] Viet-Khoa Vo-Ho, Kashu Yamazaki, Hieu Hoang, Minh-Triet Tran, and Ngan Le. Chapter 19 - neural architecture search for medical image applications. In *Meta Learning With Medical Imaging and Health Informatics Applications*, pages 369–384. 2023.

[Wang *et al.*,] Xinyi Wang, Tiange Xiang, Chaoyi Zhang, Yang Song, Dongnan Liu, Heng Huang, and Weidong Cai. Bix-nas: Searching efficient bi-directional architecture for medical image segmentation. In *MICCAI 2021*.

[Wang *et al.*, 2017] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 3462–3471, 2017.

[Wang *et al.*, 2022] He Wang, Xinshan Zhu, Peiyin Chen, Yuxuan Yang, Chao Ma, and Zhongke Gao. A gradient-based automatic optimization cnn framework for eeg state recognition. *Journal of Neural Engineering*, 19(1):016009, 2022.

[Wang *et al.*, 2023] Shiqing Wang, Yubiao Luo, Pushen Zuo, Lunshuai Pan, Yongxiang Li, and Zhong Sun. In-memory analog solution of compressed sensing recovery in one step. *Science Advances*, 2023.

[Wang *et al.*, 2024] Yan Wang, Liangli Zhen, Jianwei Zhang, Miqing Li, Lei Zhang, Zizhou Wang, Yangqin Feng, Yu Xue, Xiao Wang, Zheng Chen, Tao Luo, Rich Siow Mong Goh, and Yong Liu. Mednas: Multi-scale training-free neural architecture search for medical image analysis. *IEEE Transactions on Evolutionary Computation*, 2024.

[Weng *et al.*, 2019a] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019.

[Weng *et al.*, 2019b] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, pages 44247–44257, 2019.

[Wu *et al.*, 2023] Yixiao Wu, Huan Liu, Dalin Zhang, Yuzhe Zhang, Tianzheng Lou, and Qinghua Zheng. AutoEER: automatic EEG-based emotion recognition with neural architecture search. *Journal of Neural Engineering*, 20(4):046029, 2023.

[Yu *et al.*, 2020] Qihang Yu, Dong Yang, Holger R. Roth, Yutong Bai, Yixiao Zhang, Alan Loddon Yuille, and Daguang Xu. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. *CVPR*, pages 4125–4134, 2020.

[Zeng *et al.*, 2020] Dewen Zeng, Weiwen Jiang, Tianchen Wang, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang, Jingtong Hu, and Yiyu Shi. Towards cardiac intervention assistance: hardware-aware neural architecture exploration for real-time 3d cardiac cine mri segmentation. In *ICCAD*, 2020.

[Zhang *et al.*, 2023] Zijun Zhang, Adam R Lamson, Michael Shelley, and Olga Troyanskaya. Interpretable neural architecture search and transfer learning for understanding CRISPR-Cas9 off-target enzymatic reactions. *Nat Comput Sci*, pages 1056–1066, 2023.

[Zhu *et al.*, 2019] Zhuotun Zhu, Chenxi Liu, Dong Yang, Alan Loddon Yuille, and Daguang Xu. V-nas: Neural architecture search for volumetric medical image segmentation. *International Conference on 3D Vision (3DV)*, pages 240–248, 2019.

[Zook *et al.*, 2019] Justin M. Zook, Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, Len Trigg, Rebecca Truty, Cory Y. McLean, Francisco Vega, Chunlin Xiao, Stephen T. Sherry, and Marc Salit. An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, pages 561–566, 2019.

[Zoph *et al.*, 2018] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, pages 8697–8710, 2018.