Investigating the Application of Neural Architecture Search to Multi-Label, Medical Image Classification on Sparse Datasets

Tomas Slaven
slvtom001@myuct.ac.za
University of Cape Town
South Africa

ABSTRACT

This paper delves into Neural Architecture Search (NAS), a subset of Automatic Machine Learning (AutoML) focused on automating the creation of optimal neural network architectures. NAS presents a promising avenue for enhancing Convolutional Neural Networks (CNNs), particularly in demanding real-world applications like medical image classification, where sparse data poses challenges. Our research aims to address this issue and provide a comprehensive evaluation of NAS performance in this context. We conduct extensive experiments using state-of-the-art NAS models, including DeepMAD, ShapleyNAS, and ZenNAS, on pathological datasets of the neck and elbow. ZenNAS emerges as the top performer on Cervical Neck X-Rays, surpassing manually designed architectures. In contrast, the Elbow X-Ray dataset sees manually designed ResNet-50 outperform NAS models. Intriguingly, models trained on the elbow dataset exhibit potential for real-world application as first-readers in resource-constrained environments. Our study not only highlights the capabilities of NAS in medical image classification but also underscores the importance of dataset-specific model selection and optimization. These findings provide valuable insights into the practical applications of NAS, with implications for enhancing healthcare diagnostics and addressing data scarcity challenges.

1 INTRODUCTION

The analysis of pathological data poses significant challenges, including the need for expertise, resource intensiveness, and susceptibility to human errors. Even domain experts can encounter classification inaccuracies. Studies evaluating radiologists' performance in mammogram screening, for instance, revealed False Positive Rates of 8.9% and True Positive Rates of 83.8% in breast cancer detection, setting a benchmark for pathogen detection and underscoring room for improvement. Machine learning has generated substantial excitement for its potential to expedite data processing, uncover deeper insights, and accelerate decision-making in pathology. However, this field grapples with a critical issue: data availability. Medical contexts face a dual challenge of limited labeled datasets and severe class imbalances, notably an abundance of normal samples compared to rare medical abnormalities. This class imbalance is strikingly significant, as evident in our curated datasets (Table 7 and Table 8 in Appendix A).

Expanding a medical dataset with additional subjects is often impractical due to the costly and time-consuming nature of acquiring accurately labeled, task-specific data [1]. Data Augmentation (DA) emerges as a potential remedy, allowing the augmentation of existing data by generating new label-conforming samples [20]. However, research into DA reveals that it's not a universal solution to the challenges in medical image classification. Initial experiments indicate a modest improvement of approximately 2.5% in image classification performance on small datasets through DA [16]. While this improvement is valuable, it underscores that DA represents just one facet of enhancing image classification models and remains an evolving field of research. This suggests the existence of alternative, potentially more impactful enhancements and mechanisms for current models.

Effective image classification models that mitigate data availability limitations hinge on selecting the appropriate model architecture. Convolutional Neural Networks (CNNs) have long dominated computer vision [7]. However, the emergence of Vision Transformers (ViTs) has challenged this dominance. ViTs introduce a modern learning paradigm that outperforms CNNs, thanks to a self-attention mechanism and dedicated components [18][13]. Despite their strengths, ViTs come with their own set of challenges, including quadratic computational complexity in relation to token size [11], sub-optimal generalization on small datasets [10], and limited optimization and hardware integration in industry [13]. For medical image classification, the evidence suggests that CNNs remain the model of choice.

Deep learning's layered structure has created a vast search space for designing CNNs, offering significant room for improvement. However, the manual design process historically relied on experts' knowledge and time-consuming trial-and-error optimization. This limitation obstructs the exploration of the full potential of CNN design. To address this challenge, neuroevolution has emerged as a potential solution, employing automated methods to generate architectures for specific tasks. Early neuroevolution successes outperformed some top-performing fixed-topology models on challenging benchmarks, illustrating its theoretical and practical potential [15]. Advances in computing power further solidified neuroevolution algorithms as a competitive alternative that scales with the increase in computing resources, making them applicable to a wide range of neural network optimization problems [14].

Automatic Machine Learning (AutoML) represents the latest evolution of these programming concepts, advancing the entire machine learning pipeline automation. AutoML, broadly defined as automating the construction of an ML pipeline within a defined computing budget [4], encompasses Neural Architecture Search (NAS) and Hyperparameter Optimization (HO). For enhancing CNN performance on sparse pathological datasets, investigating NAS is pertinent because it addresses the optimization of network architectural design, a prerequisite for effective hyperparameter tuning.

The existing body of literature in the realm of Neural Architecture Search (NAS) within computer vision has showcased remarkable achievements, particularly in multi-class classification tasks

across well-established benchmarks like ImageNet [1], CIFAR-10 [2], and CIFAR-100[3]. However, a conspicuous research gap exists concerning the application of NAS in Medical Image Classification and multi-label classification, where each data sample can be assigned not just one but multiple labels, adding complexity to the task. The limited literature available on multi-label classification within pathological datasets for image classification is marred by misleading practices. These studies often emphasize overall accuracy, computed as the mean accuracy across individual labels, to present inflated performance figures, instead of considering the accuracy of all assigned labels for each instance — a more realistic but challenging metric. It's crucial to note that accuracy proves inadequate for evaluating machine learning model performance on medical datasets, as elaborated in Section Methodology, Sub-Section Evaluation Metrics.

This paper explores the efficacy of three cutting-edge Neural Architecture Search (NAS) methods, specifically DeepMAD, ShapleyNAS, and ZenNAS, for binary multi-label medical image classification [8, 13, 19]. Our investigation involves the application of these approaches to two sparse X-Ray datasets. This research endeavors to contribute to the field by shedding light on the capabilities of NAS as a potential solution for multi-label classification tasks within the constraints of sparse medical datasets.

1.1 Aim

This study aims to assess the potential of Neural Architecture Search (NAS) methods in enhancing the performance of Machine Learning for medical image classification, specifically focusing on sparse medical datasets. The task involves developing classifier models capable of analyzing medical X-Ray images and generating sets of predicted labels corresponding to potential medical characteristics present in each image. These predicted labels are represented as binary values, where one indicates label presence and zero indicates absence. Given the distinct nature of Cervical Neck and Elbow X-Rays, separate models will be constructed for each dataset. The research objectives are as follows:

(1) Demonstrate the superiority of Convolutional Neural Network (CNN) image classification model architectures discovered via Neural Architecture Search (NAS) over state-of-the-art manually designed architectures when applied to sparse pathological datasets.
(2) Evaluate the capability of NAS-derived architectures to deliver medical image classification models that achieve performance comparable to that of radiologists, thereby showcasing their potential for real-world applications.
(3) Present valuable empirical findings regarding the utilization of NAS in the context of multi-label medical image classification, offering insights into its practical applications.

2 BACKGROUND AND RELATED WORK

This section provides an overview of the literature pertaining to the chosen Neural Architecture Search (NAS) methods, along with the rationale behind the selection of each method to address the predefined research objectives.

2.1 DeepMAD

Mathematical Architecture Design for Deep CNNs (DeepMAD) is a systematic approach for designing high-performance convolutional neural network (CNN) models. DeepMAD leverages recent advancements in deep learning theories and employs a constrained Mathematical Programming (MP) problem to optimize CNN architecture parameters, including network width and depth [13].

Notably, the MP problem in DeepMAD has a low dimension, making it computationally efficient and solvable on CPUs without the need for custom MP solvers. This eliminates the requirement for GPUs or extensive memory, resulting in rapid execution, even on servers with limited resources. Once the MP problem is solved, it provides an optimized CNN architecture [13].

As highlighted in the introduction, while Vision Transformers (ViTs) have become the prevailing image classification technique, they come with substantial training overheads [5]. DeepMAD demonstrates that CNNs, despite extensive exploration, have untapped potential. DeepMAD achieves performance comparable to or surpassing state-of-the-art ViTs. The authors of DeepMAD showcase its effectiveness by optimizing architectures based on conventional convolutional layers. DeepMAD achieves results on par with or superior to ViT models of equivalent size and Floating-Point Operations Per Second (FLOPs). For instance, DeepMAD achieves an impressive top-1 accuracy of 82.8% on ImageNet-1k with 4.5G FLOPs and 29M parameters, outperforming models like ConvNeXt-Tiny and Swin-Tiny of comparable size. Furthermore, DeepMAD attains a top-1 accuracy of 77.7%, surpassing the original ResNet-18 by 8.9% and narrowly exceeding ResNet-50's performance while maintaining an equivalent CNN size [13].

Regarding its potential in medical image classification, DeepMAD demonstrates promising results, achieving accuracies of approximately 80% when trained on the CIFAR-100 dataset. Importantly, this experiment emphasizes the significance of the hyperparameter ρ rather than optimizing solely for CIFAR-100 benchmark performance. This showcases DeepMAD's strong model performance on a dataset similar in size to many medical datasets, which typically range from several hundred to tens of thousands of subjects [17]. Thus, DeepMAD presents a compelling case for its suitability in medical image classification [13].

2.2 ShapleyNAS

Shapley-NAS presents an innovative approach to neural architecture search (NAS), building upon the foundations laid by DARTS while addressing some of its limitations [9]. Shapley-NAS assesses the individual contributions of components within a supernet to a neural network's validation accuracy by leveraging the Shapley value. Unlike previous methods that primarily consider the magnitude of architectural parameters, Shapley-NAS also accounts for their practical impact on task performance. The Shapley value proves to be a powerful tool for handling complex relationships among individual elements, quantifying the average marginal contribution of all possible combinations of operations. To efficiently approximate the Shapley value, Monte Carlo sampling is employed,

[1] https://www.image-net.org/
[2] https://cifar.ca/
[3] https://cifar.ca/

with a momentum update mechanism to mitigate fluctuations introduced during the sampling process. Shapley-NAS surpasses previous methods, such as DARTS, by exhibiting a stronger correlation with task performance. This approach achieves superior results across various datasets and search spaces, including an impressive 2.43% error rate on CIFAR-10 and a top-1 accuracy of 23.9% on ImageNet under mobile settings [19].

However, a drawback of Shapley-NAS lies in the computational cost associated with computing the Shapley value, especially in common search spaces, where it demands numerous evaluations. This can result in substantial search costs that may limit its practicality for task-specific deployment. To address this challenge, Monte Carlo sampling is employed as a method for estimating the Shapley value when assessing the contribution of operations during architecture search. While Monte Carlo sampling offers computational efficiency, it may not provide the same level of precision as exact computation achieved by evaluating all possible subsets for the optimal Shapley value [19].

Shapley-NAS demonstrates remarkable performance across various benchmarks, including CIFAR-10, ImageNet, and NAS-Bench201, highlighting its efficacy in identifying optimal architectures [19]. Its particular success in a mobile setting underscores its potential as a lightweight NAS mechanism. Given its performance on CIFAR-10 and its ability to operate efficiently, it holds promise for the development of medical image classification models.

2.3 ZenNAS

Developing accurate predictors in the field of machine learning often incurs substantial computational costs. This is particularly true for brute-force and predictor-based methods, which necessitate the training of numerous networks. Even one-shot methods, which mitigate some of these expenses through parameter sharing, entail the training of a large supernet, a computationally intensive process that can suffer from model interference, leading to a degradation in predictor quality. Moreover, the requirement for sizable supernets presents challenges when searching for large target networks, especially when resources are limited [8].

To tackle these challenges, ZenNAS introduces an efficient proxy known as Zen-Score for Neural Architecture Search (NAS). Zen-Score quantifies a neural network's expressiveness, exhibiting a positive correlation with model accuracy. Importantly, it can be computed swiftly through forward inferences on randomly initialized networks using Gaussian inputs, rendering it both lightweight and independent of data. Zen-Score also adeptly addresses scale-sensitive issues arising from Batch Normalization [8].

Building upon Zen-Score, ZenNAS introduces a novel NAS algorithm, Zen-NAS. Zen-NAS operates by maximizing the Zen-Score of the target network within predefined inference budgets. Notably, Zen-NAS is a Zero-Shot method, which means it doesn't optimize network parameters during the search process. This innovative approach is employed to search for optimal networks across various inference budgets, resulting in state-of-the-art (SOTA) performance on datasets like CIFAR-10, CIFAR-100, and ImageNet. Particularly, ZenNets, designed using Zen-NAS, achieve competitive accuracy on ImageNet while maintaining faster inference speeds compared to EfficientNet-B5, marking a significant milestone as the first zero-shot method to outperform training-based methods on ImageNet [8].

This approach draws inspiration from deep learning advancements that highlight the superiority of deep models in expressiveness compared to shallow ones. The paper aligns with the bias-variance trade-off, indicating that more expressive networks lead to reduced bias error, especially with larger training datasets. Experimentation of ZenNAS on smaller training datasets in the region of 5000 images is scarce, however ZenNAS remains competitive with other SOTA NAS techniques - displaying a high potential for medical image classification tasks.[8].

2.4 ResNet

ResNet introduces a profound approach to deep learning by addressing the issue of vanishing gradients. Rather than having each layer aim to directly fit the underlying mapping, ResNet focuses on learning a residual mapping, denoted as $F(x) := H(x) - x$, where $H(x)$ represents the desired mapping. This transformation turns the original mapping into $F(x) + x$. ResNet implements this concept through feedforward neural networks with "shortcut connections," permitting identity mappings and facilitating more manageable optimization [3]. Notably, ResNet stands as one of the top-performing manually designed architectures and has consistently achieved state-of-the-art (SOTA) results in image classification benchmarks since its inception. It currently ranks seventh in performance for image classification on CIFAR-10 [4], a benchmark that closely mimics some characteristics of pathological datasets, such as dataset size and class count. We adopt ResNet as our baseline due to its widespread use and proven effectiveness in various computer vision tasks.

3 METHODOLOGY

This section provides the detailed approach in developing the DeepMAD, ShapleyNAS, and ZenNAS binary, multi-label, classification models for both Cervical Neck and Elbow X-Rays.

3.1 Data Exploration and Image Pre-Processing

Prior to establishing our training pipeline, we conducted an initial exploratory data analysis of our datasets, unearthing pivotal insights that guided our model development process. As demonstrated in Table 7 and Table 8 within Appendix A, we observed substantial label imbalances in our datasets. These significant imbalances, with certain labels appearing in as few as 0.5% of our training examples, carried the potential to profoundly impact our models, giving rise to issues such as overfitting, model instability, class bias, and distorted evaluation metrics. To address this challenge, we implemented specific amendments, which are elaborated upon in the subsequent sub-section titled "Optimizer and Loss Function Selection."

Furthermore, our in-depth exploratory data analysis revealed that, apart from the "Normal" labels common to both datasets, all other labels were independent of one another. Notably, the presence of a "Normal" label signified the absence of all other labels within a given example image.

[4] https://paperswithcode.com/sota/image-classification-on-cifar-10

Additionally, an examination of pixel value distributions across the training splits for both datasets underscored the wide range of values attributed to the high contrast and background noise inherent in X-Ray images. Consequently, we computed the mean and standard deviation for each training split to normalize our model input images. It's important to note that this normalization process solely considered the mean and standard deviation of the training splits, thus preventing any inadvertent information leakage and ensuring the unbiased evaluation of our models. This approach offered several advantages, including enhancing model generalization, ensuring numerical stability for mathematical operations, and promoting a more uniform convergence rate for our models.

The purpose of this research was to strictly investigate the base performance of NAS in this medical context and so the incorporation of Data Augmentation was voided.

3.2 Optimizer and Loss Function Selection

Our model training process employed the Adam optimizer [6], which proves particularly advantageous for architectures derived through NAS. This preference is attributable to Adam's adaptive learning rate mechanism and its adept handling of diverse gradients. Adam operates by individually adjusting learning rates for each parameter, drawing from historical gradient information. Such adaptability is invaluable when dealing with NAS-derived architectures, as these structures often exhibit intricate and varying sensitivities to gradient updates. The adaptive learning rates offered by Adam significantly enhance the optimization of these architectures.

To train our models effectively, we employed a Weighted Binary Cross-Entropy Loss Function [12]. This choice aligns seamlessly with the characteristics of our task, where each training instance may belong to multiple classes concurrently. The Binary Cross-Entropy Loss Function, by assuming label independence, aligns harmoniously with our dataset, a point discussed in the "Data Exploration and Image Pre-Processing" subsection. Additionally, it resonates with prevalent medical context evaluation metrics such as the F1-Score, as elaborated upon in the "Evaluation Metrics" subsection. Furthermore, to counteract the label imbalances identified earlier, we introduced weighting into our loss function. This step assumes paramount importance, given that most minority classes in our datasets correspond to critical or rare medical conditions requiring immediate attention. By assigning higher weights to these classes within the loss function, we effectively communicate to the model the heightened significance of correctly classifying instances from these critical classes.

3.3 Model Architecture Development

3.3.1 DeepMAD. The source code for DeepMAD is publicly available on GitHub [5].This source code encompasses diverse modules tailored for executing the DeepMAD neural architecture search, designed to operate within a predefined network size. In alignment with our initial hypothesis and the aim of applying NAS to deep learning for medical image classification, we opted for a network size comprising 50 layers. Once the architecture discovery phase was completed, seamlessly incorporating the identified architecture into our training pipeline became a straightforward endeavor, a process meticulously detailed in Section Methodology.

3.3.2 ShapleyNAS. The source code for ShapleyNAS is publicly available on GitHub [6]. In contrast to DeepMAD and ZenNAS, ShapleyNAS conducts architecture search concurrently with the training process. Consequently, our integration of ShapleyNAS necessitated the inclusion of methods from ShapleyNAS' train-and-search module into our established training pipeline. Furthermore, it's noteworthy that ShapleyNAS was explicitly designed to function in tandem with a stochastic gradient descent optimizer. Consequently, we made corresponding amendments to our ShapleyNAS models to ensure compatibility.

3.3.3 ZenNAS. The source code for ZenNAS is publicly available on GitHub [7]. Similarly, akin to DeepMAD, ZenNAS furnishes searcher modules for architecture search that are characterized by their relative simplicity of implementation. Once these modules were successfully incorporated and executed, the resultant architecture seamlessly fed into our training pipeline, as per the methodology elucidated earlier.

3.3.4 ResNet-50 Baseline. As elucidated in the sub-section Aim, we have adopted the ResNet-50 architecture as our baseline for comparative analysis against the results generated by NAS and traditional manually designed architectures. Acquiring the ResNet-50 architecture was facilitated by importing the PyTorch implementation of ResNet-50 [8]. Subsequently, this architecture was seamlessly integrated into the model loading phase of the training pipeline, as elaborated in our methodology. Furthermore, it's noteworthy that the PyTorch Implementation of ResNet offers the option to employ a pre-trained ResNet architecture. The incorporation of the pre-trained ResNet architecture serves as a potential indicator of how model performance might be further enhanced for the specific task under consideration.

3.4 Evaluation Metrics

The selection of appropriate evaluation metrics plays a pivotal role in comprehending a model's performance. Within the medical context, there exists a distinct emphasis on certain metrics, notably the False Positive Rate (FPR) (3), True Positive Rate (TPR) (4), and F1-Score (8), while giving relatively less weight to metrics like Mean Accuracy (9). This prioritization is attributed to the unique considerations in healthcare. FPR, for instance, meticulously gauges instances where a label is erroneously identified as positive (present), while TPR quantifies the model's proficiency in accurately identifying positive cases. In the context of the *Normal* label, FPR assumes paramount significance. This stems from the fact that the gravest scenario in medical practice is a misdiagnosis where a patient is incorrectly deemed *Normal* and consequently denied necessary medical attention. This erroneous diagnosis can have dire consequences, potentially exacerbating the patient's condition due to a lack of timely intervention. Conversely, it is equally undesirable for a model to indiscriminately classify every image as "not normal" solely to achieve a commendable FPR, as this approach renders

[5]https://github.com/alibaba/lightweight-neural-architecture-search

[6]https://github.com/euphoria16/shapley-nas

[7]https://github.com/idstcv/ZenNAS

[8]https://pytorch.org/hub/nvidia_deeplearningexamples_resnet50

the model functionally inept, providing uniform predictions for all scenarios. Striking a harmonious equilibrium between a high TPR and a low FPR emerges as the ultimate objective, reflecting a model's optimal performance in the medical domain - leading to the creation of our Weighted TPR & FPR metric (5).

3.4.1 Overall Model Metrics. To evaluate the overall performance of the model we use two calculations: Mean Accuracy and F1-Score.

Mean Accuracy is simply computed as the average of the total number of input images where all labels are correctly identified over the total number of input images.

$$Mean\ Accuracy = \frac{Total\ Correct\ Predictions}{Total\ Number\ of\ Images} \tag{1}$$

The F1-Score computed at the overall model level is referred to as a Macro F1-Score (2). This specific choice is rooted in our concurrent computation of F1-Scores for individual labels. By employing the Macro F1-Score, we achieve a comprehensive assessment of the model's performance across all classes, with each class receiving equitable consideration. This approach ensures an impartial representation of the models' performance across all classes, as it guarantees an equal contribution from each label to the overarching evaluation. Given our simultaneous calculation of label-specific metrics, this approach remains valid and, in fact, encourages a holistic appraisal of our models. The label-specific metrics offer intricate performance evaluations that enrich the overall assessment, thus fostering a balanced and thorough evaluation of our models.

$$Macro\ F1\ Score = \frac{Sum\ of\ per\ label\ F1\ Scores}{Number\ of\ Labels} \tag{2}$$

3.4.2 Label Specific Metrics. The prediction accuracy of the "Normal" label holds a central role in determining the efficacy of deep learning within this context. Its successful prediction is pivotal in demonstrating the practical applicability of our research. Consequently, we have meticulously recorded an array of label-specific metrics to conduct a comprehensive assessment of its performance. These metrics encompass:

$$FPR = 1 - Specificity \tag{3}$$

$$TPR\ (Sensitivity) = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4}$$

$$Weighted\ FPR\ \&\ TPR = \frac{TPR + (1 - FPR)}{2} \tag{5}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{6}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \tag{7}$$

$$F1\ Score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \tag{8}$$

Table 1: Table representing sample of hyperparamters used in initial Grid Search

Configuration	Batch Size	Learning Rate	Resize Resolution
1	32	0.01	64
2	32	0.001	128
3	64	0.01	128
4	64	0.001	64
5	128	0.01	64
6	128	0.001	128

$$Mean\ Accuracy = \frac{Total\ Correct\ Label\ Predictions}{Number\ of\ Examples} \tag{9}$$

3.4.3 Statistical Significance Tests. To ascertain the presence of statistically significant differences in the performance of our models, we subjected our overall model performance results to rigorous testing. We employed the Mann-Whitney U Test for this analysis, given our non-parametric, independent, and relatively small sample sizes.

3.5 Hyperparameter Optimization

Hyperparameters are pivotal determinants of machine learning model performance, influencing its effectiveness, efficiency, and generalization capability. In this study, we employ Grid Search to optimize key hyperparameters, including Batch Size, Learning Rate, and Image Resize Resolution, seeking the ideal configuration for our models. An example of our Grid Search implementation is displayed in Table 1.

Initial Grid Search revealed anticipated insights such as an increased batch size and Resize Resolution provided faster convergence times. Learning rate observed the opposite effect in certain instances: smaller learning rates with large batch sizes resulted in faster convergence times, whereas larger learning rates observed slower convergence times displaying symptoms of overfitting.

3.6 Experiment Design

3.6.1 Datasets. This study employs two datasets provided by the University of Cape Town. The first dataset comprises 933 Cervical Neck X-Rays, while the second contains 2740 Elbow X-rays. Each Cervical Neck X-Ray is captured from the same perspective, featuring eight labels corresponding to diagnosable medical characteristics potentially present in a single image. Similarly, the Elbow dataset includes a list of 14 labels for each image, with an approximate 50:50 distribution between X-Ray images taken with an AP view and those taken with a Lateral view. For the purpose of this research, both Elbow X-Ray views (AP and Lateral) are considered identical to address the challenge of a limited dataset size. For additional details on these datasets, including label information and class balances, please refer to Table 7 and Table 8 in Appendix A.

3.6.2 Dataset Split. Our datasets undergo a division into three distinct sets, each approximately comprising: 80% for Training, 10% for Validation, and 10% for Testing. The Training set is exclusively

utilized for model training, while the Validation set serves as the means to assess model performance after every training epoch. Following the training phase, we retain two model configurations for evaluation on the Test set. These configurations are selected based on their attainment of the highest Mean Accuracy and the highest Macro F1-Score. Subsequently, the Test set functions as a holdout dataset, enabling us to evaluate the generalization capabilities of our trained models on new, unseen data.

3.6.3 Setup. Every model is integrated into the uniform training pipeline, except for the ShapleyNAS models, which exhibit slight variations in their architecture search approach, as elaborated in the Model Architecture Development section, specifically in the ShapleyNAS sub-section. All models are executed using an Nvidia 40Gb A100 GPU. Subsequent to the conclusion of each experiment, the results are systematically recorded in CSV files for meticulous analysis, with the highest-performing models saved for future use. View Table 2 for a list of the experiments run.

4 RESULTS

In this section, we will present a comprehensive overview of our results, focusing on the comparative performance analysis of DeepMAD, ShapleyNAS, ZenNAS, and the baseline ResNet-50 models applied to both the Cervical Neck and Elbow X-Ray datasets. Additionally, we will incorporate statistical significance tests to highlight any notable differences between models across these datasets. Furthermore, we will provide specific insights into the models' performance concerning the *Normal* label, which is of utmost significance in each dataset, shedding light on their real-world effectiveness in this classification task. All results and code can be found at the linked GitHub repository [9].

4.1 Overall Performance on Cervical Neck Dataset

Table 5 compares the performance of DeepMAD, ShapleyNAS, ZenNAS and our ResNet-50 baseline across the Cervical Neck X-Ray dataset. The table displays the results for each NAS methods across two different model types: highest Mean Accuracy model, and highest Macro F1-Score model. The inclusion of these two models per dataset showcases that the model configuration that achieves the highest value of one metric does not translate to similarly achieving the highest value across the other metric. In fact, we observe that optimization of one metric results in a noticeably poor performance on the other, expect for DeepMAD which performs reliably across both metrics. This can be explained by the large class imbalance of the labels. Since the labels are heavily skewed, the models achieve the highest accuracy by simply predicting the most common occurrence of each label, however this approach leads to a poor F1-Score for each label leading to very low Macro F1-Scores. This also explains why DeepMAD and the baseline achieve the exact same Mean Accuracy. The importance of the metrics chosen to be maximised for this particular task is of paramount importance as a model that predicts the most common output recorded may be accurate, but useless at picking up abnormalities requiring medical treatment.

[9] https://github.com/tomslav/InvestigatingNAS

Table 2: This table summarizes the experiments conducted, categorizing them by the model and dataset used in conjunction with the specific research objectives tested.

Experiment	Model	Dataset	Objective
1	DeepMAD	Cervical Neck	1 & 2 & 3
2	DeepMAD	Elbow	1 & 2 & 3
3	ShapleyNAS	Cervical Neck	1 & 2 & 3
4	ShapleyNAS	Elbow	1 & 2 & 3
5	ZenNAS	Cervical Neck	1 & 2 & 3
6	ZenNAS	Elbow	1 & 2 & 3
7	ResNet	Cervical Neck	1 & 2 & 3
8	ResNet	Elbow	1 & 2 & 3
9	ResNet (Pre-trained)	Cervical Neck	2
10	ResNet (Pre-trained)	Elbow	2

The experimental results in Table 5 reveal that NAS methods generally outperform the baseline in both Mean Accuracy and Macro-F1 Score. Notably, ZenNAS excels with the highest Mean Accuracy (38.298%), while ShapleyNAS achieves the top Macro F1-Score (26.395%) compared to the baseline's results of 37.234% and 19.530%. However, ShapleyNAS is the sole NAS method performing worse than the baseline in Mean Accuracy. Statistically testing these results, as observed in Tables 9 and 10 in Appendix, demonstrates that while all NAS methods surpass the baseline in terms of Macro F1-Score, ZenNAS is the only method that displays a significant difference. This suggests that DeepMAD and ShapleyNAS perform comparably to the baseline for this metric. Regarding Mean Accuracy, all methods significantly outperform ShapleyNAS, while DeepMAD, ZenNAS, and the baseline perform equally well. Overall, these findings underscore NAS's superiority over the baseline in Macro F1-Score, while indicating a comparable performance in Mean Accuracy, except for ShapleyNAS. It is also worth noting that by pretraining our baseline the results for the baseline increased by 7.447% on Mean Accuracy and 9.276% on Macro F1-Score.

4.2 Overall Performance on Elbow Dataset

Table 6 compares the performance of DeepMAD, ShapleyNAS, ZenNAS and our ResNet-50 baseline across the Elbow X-Ray dataset. The table displays the results for each NAS methods across two different model types: highest Mean Accuracy model, and highest Macro F1-Score model. Unlike the results in Table 5, not all models experience a complete trade-off when optimizing for a particular metric. We observe for both ZenNAS and the baseline models that there is approximately a 2% difference in Macro F1-Scores between weight configurations of the same model type. This is not the same for Mean Accuracy with all models experiencing large disparities in Mean Accuracy depending on the weight configuration applied, this can be explained by the label imbalance as discussed in sub-section Overall Performance on Cervical Neck Dataset.

DeepMAD (26.036%) and ShapleyNAS (14.531%) represent the models that achieved the highest Mean Accuracy and Macro F1-Scores respectively. This is another display of NAS outperforming our baseline in terms of overall model performance, which recorded a Mean Accuracy of 24.260% and Macro F1-Score of 9.122%. Significance testing in Tables 12 and 13 of Appendix, reveal different

insights on NAS outperforming the baseline. The baseline significantly outperforms both ShapleyNAS and ZenNAS in Mean Accuracy, while performing similarly to DeepMAD. However, when it comes to Macro-F1 scores, no significant differences are observed among all models. This is likely due to the performance variability among NAS methods, as all NAS methods outperform the baseline in terms of largest Macro F1-Score recorded. It's worth noting that the NAS-discovered architectures exhibit higher complexity than the baseline, potentially leading to less reliable results as experienced in the significance test which takes in samples of the five highest recorded scores. This surprise in reliability of results is only thoroughly experienced with ShapleyNAS, suggesting that the ShapleyNAS architecture is prone to producing unreliable results. This aligns with the ShapleyNAS paradigm that conducts the architecture search during training; resulting in a changing complexity as the model learns on the dataset. Similarly to the Cervical Neck dataset, pretraining our baseline results in performance increases of 7.101% and 7.175% for Mean Accuracy and Macro F1-Score respectively, portraying that these initial experimental results do not define the full potential of the models used.

4.3 Normal Label of Neck Dataset

Table 3 displays the experimental results of the *Normal* label per each model. The most important metric for assessing real-world applicability.

The observed variations in Weighted True Positive Rate (TPR) and False Positive Rate (FPR) among different weight configurations within the same model type underline the influence of dataset imbalance on model performance. Notably, configurations optimizing F1-Score tend to yield the highest Weighted TPR and FPR. However, the largest F1-Score as obtained by DeepMAD (49.558%) did not correspond to the greatest Weighted TPR & FPR achieved by ShapleyNAS (60.157%), accentuating the importance of employing Weighted TPR and FPR metrics for real-world applicability assessment. Models with low FPR but equally low TPR are impractical, as they tend to classify all examples as normal regardless of their actual nature. Comparatively, Neural Architecture Search (NAS) methods consistently outperform the baseline model, with ShapleyNAS (60.157%) emerging as the top performer. Extensive testing as show in Table 11 reveals that only ZenNAS exhibits a significant improvement over the baseline, particularly in predicting the Normal label, showcasing its reliability compared to DeepMAD and ShapleyNAS, which, while surpassing the baseline, fail to reach statistical significance. While not tested for statistical significance, the pretrained baseline displays an improvement over the standard baseline of 17.083% across Weighted TPR & FPR.

4.4 Normal Label of Elbow Dataset

Table 4 displays the experimental results of the *Normal* label per each model. The most important prediction task for assessing real-world applicability.

Alike the models' performances on the Neck dataset, there is a limited correlation between their achievements in Weighted TPR and FPR, and their performance in other metrics. Attaining the highest scores in other metrics does not necessarily translate to the best ratings in Weighted TPR and FPR - reiterating the consequences of a weak dataset. The baseline model outperforms all NAS methods significantly in Weighted TPR and FPR, as observed in our statistical significance test in Table 14 of Appendix, with scores of 69.563%, compared to the next-best method, ShapleyNAS, at 64.203%. Following closely are ZenNAS at 59.114% and DeepMAD at 50.310%. This difference in performance and availability of many statistical significances is likely due to several factors: the balance of the normal class in the Elbow training split (as shown in Appendix Table 7), the larger size of the Elbow dataset compared to the Cervical Neck Dataset, and the relative simplicity of Elbow X-Rays. These elements collectively result in a more robust training dataset, enhancing result reliability. This underscores that balanced and larger datasets lead to increased result reliability and better baseline performance compared to other NAS methods. Pretraining of the baseline further improves its optimal performance over the NAS techniques as observed by a 10.778% increase in Weighted TPR and FPR.

5 DISCUSSION

There's a notable research deficit in the application of Neural Architecture Search (NAS) in the fields of Medical Image Classification and multi-label classification. Existing literature on multi-label classification within pathological datasets often suffers from misleading practices, such as calculating overall accuracy as the mean of individual label accuracies to inflate reported results. These gaps slow us from developing an automated AI system capable of accurately distinguishing normal and abnormal conditions that would immensely benefit in dealing with reporting backlogs and the scarcity of radiologists in low-resource settings leading to potentially significant health outcomes to lower income populations.

To address these weaknesses in the literature and further the development of an automated AI screening system, this research evaluated the application of DeepMAD [13], ShapleyNAS [19], and ZenNAS[8] in multi-label classification on sparse Cervical Neck and Elbow datasets, and provided an in-depth evaluation on their performance in distinguishing between normal and abnormal x-rays to assess their real-world application.

Regarding overall model performance on the Cervical Neck X-ray dataset, all NAS methods surpassed the baseline in terms of Macro F1 Score, with ShapleyNAS achieving the highest score of 26.395%. However, the only statistically significant improvement over the baseline was observed with ZenNAS. ShapleyNAS and DeepMAD showed no statistically significant differences compared to the baseline. In terms of Mean Accuracy, DeepMAD, ZenNAS, and the baseline demonstrated no statistical distinctions among themselves, while ShapleyNAS exhibited significantly lower performance than the others. These statistical tests highlight ZenNAS as the leading NAS method, as it exhibits a significant improvement over the baseline in terms of Macro F1-Score and matches the baseline's performance in Mean Accuracy. We attribute ZenNAS' success to its Zen-Score maximization, enabling the network to effectively model complex dependencies within sparse datasets like the Cervical Neck dataset [8]. These findings suggest that NAS methods perform at least as well as traditional manually designed methods in this multi-label classification task.

Table 3: Model Performance of the *Normal label* on Test Set of Neck Dataset. Note: (1) denotes weight configuration leading to highest Macro F1-Score, (2) denotes weight configuration leading to highest Overall Mean Accuracy

Model Type	Mean Accuracy (%)	F1-Score (%)	True Positive Rate (%)	False Positive Rate (%)	Weighted TPR & FPR (%)
DeepMAD (1)	39.362	**49.558**	**93.333**	85.938	53.698
DeepMAD (2)	50.000	44.706	63.333	56.250	53.542
ShapleyNAS (1)	63.830	46.875	50.000	29.687	**60.157**
ShapleyNAS (2)	68.085	0.000	0.000	**0.000**	50.000
ZenNAS (1)	65.957	45.714	53.333	37.500	57.917
ZenNAS (2)	**69.149**	6.451	3.333	**0.000**	51.667
ResNet (1)	64.894	23.256	16.667	12.500	52.084
ResNet (2)	68.085	0.000	0.000	**0.000**	50.000
Pretrained ResNet	*71.277*	*58.462*	*63.333*	*25.000*	*69.167*

Table 4: Model Performance of the *Normal label* on Test Set of Elbow Dataset. Note: (1) denotes weight configuration leading to highest Macro F1-Score, (2) denotes weight configuration leading to highest Overall Mean Accuracy

Model Type	Mean Accuracy (%)	F1-Score (%)	True Positive Rate (%)	False Positive Rate (%)	Weighted TPR & FPR (%)
DeepMAD (1)	85.799	7.692	5.882	5.263	50.310
DeepMAD (2)	88.757	0.000	0.000	**1.316**	49.342
ShapleyNAS (1)	34.911	21.429	**88.235**	71.053	58.591
ShapleyNAS (2)	49.704	24.779	82.352	53.947	64.203
ZenNAS (1)	82.840	25.641	29.412	11.184	59.114
ZenNAS (2)	85.799	25.000	23.529	7.237	58.146
ResNet (1)	82.840	38.298	52.941	13.816	**69.563**
ResNet (2)	**91.716**	**41.667**	29.411	**1.316**	64.048
Pretrained ResNet	*83.432*	*48.148*	*76.471*	*15.789*	*80.341*

Table 5: Model Performance on Test Set of Neck Dataset. Note: (1) denotes weight configuration leading to highest Macro F1-Score, (2) denotes weight configuration leading to highest Overall Mean Accuracy.

Model Type	Mean Accuracy (%)	Macro F1-Score (%)
DeepMAD (1)	27.660	21.788
DeepMAD (2)	37.234	20.017
ShapleyNAS (1)	0.000	**26.395**
ShapleyNAS (2)	8.511	5.768
ZenNAS (1)	12.766	25.788
ZenNAS (2)	**38.298**	9.772
ResNet (1)	26.596	19.530
ResNet (2)	37.234	7.987
Pretrained ResNet	*44.681*	*28.806*

Table 6: Model Performance on Test Set of Elbow Dataset. Note: (1) denotes weight configuration leading to highest Macro F1-Score, (2) denotes weight configuration leading to highest Overall Mean Accuracy.

Model Type	Mean Accuracy (%)	Macro F1-Score (%)
DeepMAD (1)	16.568	9.129
DeepMAD (2)	**26.036**	5.213
ShapleyNAS (1)	0.000	**14.531**
ShapleyNAS (2)	12.426	5.708
ZenNAS (1)	1.775	11.155
ZenNAS (2)	18.343	9.576
ResNet (1)	18.343	9.122
ResNet (2)	24.260	7.633
Pretrained ResNet	*31.361*	*16.297*

We further assessed the performance of our models on the Cervical Neck dataset, specifically focusing on the normal label to gauge their applicability in real-world scenarios. Our findings paralleled those described earlier: NAS methods consistently outperformed the baseline across all metrics, with particular emphasis on Weighted TPR & FPR. Statistical significance tests (see Table 14 of Appendix) underscore that only ZenNAS exhibits a significant performance enhancement over the baseline in terms of Weighted TPR & FPR. Conversely, ShapleyNAS and DeepMAD showed statistically comparable performance to the baseline. Notably, ShapleyNAS achieved the highest Weighted TPR & FPR scores and the highest Macro F1-Score for overall model performance but failed to achieve statistical significance over the baseline. This is likely attributed to ShapleyNAS's training paradigm, which involves evolving model complexity during training [19]. This dynamic nature introduces variability in ShapleyNAS results, rendering them less reliable, despite achieving the best Macro F1-Score and Weighted TPR & FPR. A study assessing radiologists' capacity to identify breast cancer reported a Weighted TPR & FPR of 87.45% [2]. This is considerably larger than ZenNAS' 57.917%, however it is worth noting that pretraining the baseline increased baseline performance by 17%. These findings imply that while our experimental outcomes may not presently endorse ZenNAS for real-world applications, the integration of techniques like pretraining and data augmentation could potentially enhance its suitability and performance.

In our evaluation of model performance on the Elbow dataset, NAS methods surpass the baseline in terms of Mean Accuracy and Macro F1-Score. However, statistical significance tests reveal a contrasting outcome, with the baseline demonstrating a significant improvement in Mean Accuracy over both ShapleyNAS and ZenNAS, as shown in Table 12, while performing equally well as DeepMAD. Concerning Macro F1-Score, all models exhibit statistically insignificant differences, even though NAS methods achieved superior individual scores. This underscores the baseline's ability to consistently produce competitive results. DeepMAD's competitiveness with the baseline may be attributed to its use of a constrained Mathematical Programming (MP) Problem for architecture optimization. In contrast to ZenNAS, which maximizes a ZenScore, and ShapleyNAS, whose architecture complexity changes during training, the constrained MP Problem seems to provide a more reliable architecture search mechanism, yielding consistent and competitive results. Overall, DeepMAD and the baseline emerge as the top-performing methods in terms of overall performance on the Elbow dataset, with DeepMAD potentially holding an edge due to its capacity to achieve higher individual scores for both Mean Accuracy and Macro F1-Score.

In our continued assessment of the Elbow dataset, we shift our focus from overall model performance to the normal label, a crucial aspect for real-world applicability. Here, the baseline stands out, achieving the highest individual Weighted TPR & FPR score and maintaining statistical significance over all NAS methods, as evident in the results of the statistical significance tests presented in Table 14 of Appendix. Notably, the highest Normal Label Accuracy and F1-Score do not align with the highest Weighted TPR & FPR score, underscoring the importance of selecting the appropriate metric for context dependent model evaluation. Further statistical analysis between the models reveals significant differences among each NAS method, with ShapleyNAS performing the best, followed by ZenNAS, and DeepMAD ranking last. The Elbow dataset's relative simplicity compared to the Cervical Neck dataset, coupled with its larger size, contributes to the increased stability in training, likely explaining the many significant differences observed among the models. This also underscores the significant divergence between NAS and manually designed methods, as this is the first instance where the baseline has distinctly outperformed NAS methods, suggesting that NAS methods hold a competitive advantage when trained on smaller, more complex datasets due to their complex architectures' ability to better capture the intricacies of such data. The pretraining of our baseline then reports an improvement of 10.811% on the Weighted TPR & FPR metric, boosting it to 80.341%.This finding is analogous to studies assessing radiologists' proficiency in discerning breast cancer presence in mammograms, where a median Weighted TPR & FPR score of 87.45% has been reported [2]. It serves as our initial illustration of a model achieving capabilities comparable with real-world professionals in this domain.

Our research focused on demonstrating the superiority of Convolutional Neural Network (CNN) image classification model architectures discovered via Neural Architecture Search (NAS) over state-of-the-art manually designed architectures when applied to sparse pathological datasets. Our findings confirm that NAS methods, particularly ZenNAS, provide a promising avenue for enhancing model performance, thereby surpassing traditional manually designed architectures in multi-label medical image classification tasks on more intricate datasets, namely the Cervical Neck dataset.

Next, we sought to evaluate the capability of NAS-derived architectures to deliver medical image classification models that approach the performance of radiologists, thereby showcasing their potential for real-world applications. While our NAS-derived models demonstrated significant improvements over the baseline in certain domains, we acknowledge that achieving classification accuracy on par with radiologists remains a formidable goal for NAS. Our experimental results display only initial applications of NAS in this context suggesting further developments are very much possible, however, we also realise several factors contribute to this discrepancy, including the limited dataset size, dataset complexity, and the challenges inherent in replicating the diagnostic expertise of human radiologists. Radiologists rely not only on image patterns but also on clinical context, patient history, and additional diagnostic modalities. Therefore, reaching radiologist-level performance with AI models necessitates not only robust image classification but possibly the integration of broader clinical knowledge and context — an intricate challenge that extends beyond the scope of this study.

Lastly, our research aimed to provide valuable empirical findings regarding the utilization of NAS in the context of multi-label medical image classification, offering insights into its practical applications. Our comprehensive evaluation of NAS methods on two distinct medical image datasets sheds light on their performance while also emphasizing the critical importance of metric selection for meaningful evaluation in this domain. The variability in the performance of NAS methods on different datasets underscores the necessity for dataset-specific model selection and optimization strategies.

6 CONCLUSIONS AND FUTURE WORK

In conclusion, this research addresses critical gaps in the field of medical image classification with Neural Architecture Search (NAS). We have demonstrated both successes and challenges in achieving our research objectives, which aimed to assess the potential of NAS in sparse pathological datasets.

Firstly, our findings confirm that NAS methods, particularly ZenNAS, offer a promising avenue for enhancing model performance, surpassing traditional manually designed architectures in one of our multi-label medical image classification tasks. This signifies the potential of NAS to advance the field.

Secondly, our research showcased the complexities of achieving radiologist-level performance with AI models. While our NAS models showed notable progress, our pre-trained baseline model achieved the best performance in distinguishing normal and abnormal x-rays. It performed comparably to radiologists in detecting breast cancer, with a Weighted TPR & FPR score of 80.341%, compared to radiologists' median score of 87.45%. Replicating the diagnostic expertise of human radiologists remains a formidable task. This emphasizes the need for further research in integrating clinical knowledge and context into AI-assisted diagnosis.

Thirdly, our research contributes meaningful empirical results that highlight the practical applications of NAS in multi-label medical image classification. By evaluating NAS methods on two distinct medical image datasets, we shed light on their performance and underscore the critical importance of metric selection for meaningful evaluation in this domain. The variability in the performance of NAS methods on different datasets underscores the necessity for dataset-specific model selection and optimization strategies. Looking ahead, future research should explore several avenues:

(1) Specialized Applications: Investigate NAS' suitability in specialized medical domains or rare diseases, where sparse data and unique challenges are prevalent. This can help tailor NAS approaches to specific clinical needs.
(2) Ethical Considerations: Address ethical considerations by focusing on model interpretability, fairness, and bias mitigation. Ensuring AI models are transparent and equitable is vital for their real-world applicability.
(3) Practical Applications: Explore practical applications of our insights within healthcare institutions, particularly those facing resource constraints. Develop strategies for integrating AI models into clinical workflows to enhance efficiency and accuracy.
(4) Advanced NAS Techniques: Investigate advanced NAS techniques and architectural modifications to further optimize model performance. Continuously evolve NAS methodologies to stay at the forefront of AI advancements.
(5) Collaborative Partnerships: Foster collaborations between AI researchers and medical experts or radiologists. These partnerships can lead to the development of AI-assisted diagnosis tools that harness the strengths of both AI and clinical expertise.

In summary, our research lays the foundation for NAS's potential in multi-label classification tasks on sparse pathological datasets. While we have made significant strides, the journey to fully harnessing the capabilities of NAS in the medical field has just begun. Through ongoing research and collaboration, we can bridge the gap between AI systems and human radiologists, ultimately enhancing patient outcomes and healthcare delivery while providing meaningful empirical results.

7 ACKNOWLEDGEMENTS

We extend our thanks to Prof. Geoff Nitschke, Bilal Aslan, Shaylin Chetty, and Bryan Kazaka for their valuable insights, guidance, and collaborative contributions to this research.

REFERENCES

[1] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. 2021. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology* 65, 5 (2021), 545–563.
[2] Joann G Elmore, Sara L Jackson, Linn Abraham, Diana L Miglioretti, Patricia A Carney, Berta M Geller, Bonnie C Yankaskas, Karla Kerlikowske, Tracy Onega, Robert D Rosenberg, et al. 2009. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology* 253, 3 (2009), 641–651.
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[4] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems* 212 (2021), 106622.
[5] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54, 10s (2022), 1–41.
[6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.
[8] Ming Lin, Pichao Wang, Zhenhong Sun, Hesen Chen, Xiuyu Sun, Qi Qian, Hao Li, and Rong Jin. 2021. Zen-nas: A zero-shot nas for high-performance image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 347–356.
[9] Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).
[10] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11976–11986.
[11] Sachin Mehta and Mohammad Rastegari. 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178* (2021).
[12] Usha Ruby and Vamsidhar Yendapalli. 2020. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng* 9, 10 (2020).
[13] Xuan Shen, Yaohua Wang, Ming Lin, Yilun Huang, Hao Tang, Xiuyu Sun, and Yanzhi Wang. 2023. DeepMAD: Mathematical Architecture Design for Deep Convolutional Neural Network. *arXiv preprint arXiv:2303.02165* (2023).
[14] Kenneth O Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. 2019. Designing neural networks through neuroevolution. *Nature Machine Intelligence* 1, 1 (2019), 24–35.
[15] Kenneth O Stanley and Risto Miikkulainen. 2002. Evolving neural networks through augmenting topologies. *Evolutionary computation* 10, 2 (2002), 99–127.
[16] Luke Taylor and Geoff Nitschke. 2018. Improving deep learning with generic data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 1542–1547.
[17] Gaël Varoquaux and Veronika Cheplygina. 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine* 5, 1 (2022), 48.
[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[19] Han Xiao, Ziwei Wang, Zheng Zhu, Jie Zhou, and Jiwen Lu. 2022. Shapley-NAS: Discovering Operation Contribution for Neural Architecture Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11892–11901.
[20] Larry Yaeger, Richard Lyon, and Brandyn Webb. 1996. Effective training of a neural network character classifier for word recognition. *Advances in neural information processing systems* 9 (1996).

A APPENDIX

In this appendix, we present essential information. It includes two tables illustrating class distribution in our datasets, along with details about the train, validation, and test splits. These tables are referred to in Section 3.6.1 (Datasets). Additionally, we provide a series of tables outlining significance tests employed to assess differences in model performance. These results substantiate the findings discussed in the Results and Discussion Sections of this paper.

Table 7: Elbow X-Ray Dataset

LABELS	**TRAIN**		**VALIDATION**		**TEST**	
	Present	Absent	Present	Absent	Present	Absent
Soft Tissue Swelling	202	2176	31	162	31	138
Joint Effusion	493	1885	62	131	61	108
Distal Humerus	64	2314	12	181	9	160
Supracondylar	691	1687	87	106	92	77
Medial Epicondyle Displaced	71	2307	14	179	8	161
Lateral Epicondyle Displaced	111	2267	12	181	18	151
Olecranon	52	2326	6	187	6	163
Elbow dislocation anterior	12	2366	2	191	2	167
Elbow dislocation posterior	47	2331	2	191	2	167
Proximal Radial	40	2338	2	191	6	163
Radial Head	14	2364	2	191	8	161
Radial Head Subluxation	18	2360	2	191	3	166
Proximal Ulnar Metaphysis	27	2351	1	192	5	164
Normal	1255	1123	57	136	17	152
Total Images	**2378**		**193**		**169**	

Table 8: Cervical Neck X-Ray Dataset

LABELS	**TRAIN**		**VALIDATION**		**TEST**	
	Present	Absent	Present	Absent	Present	Absent
Alignment	411	335	51	42	52	42
Soft Tissue Swelling	83	663	10	83	10	84
Listhesis	63	683	11	82	8	86
Fracture	99	647	9	84	15	79
Dislocation	26	720	2	91	3	91
Spinous	43	703	6	87	3	91
Other Pathogens	127	619	16	77	16	78
Normal	239	507	30	63	30	64
Total Images	**746**		**93**		**94**	

Table 9: Table depicting p-values for statistical significance test conducted on Mean Accuracy of Neck Dataset. Cells coloured in green represent significant values, while red cells represent insignificant p-values

MODELS	DeepMAD	ShapleyNAS	ZenNAS	ResNet
DeepMAD		0.0097	0.0807	0.6654
ShapleyNAS	0.0097		0.0112	0.0119
ZenNAS	0.0807	0.0112		0.5762
ResNet	0.6654	0.0119	0.5762	

Table 10: Table depicting p-values for statistical significance test conducted on Macro F1-Score of Neck Dataset. Cells coloured in green represent significant values, while red cells represent insignificant p-values

MODELS	DeepMAD	ShapleyNAS	ZenNAS	ResNet
DeepMAD		0.3095	0.1508	0.4206
ShapleyNAS	0.3095		0.1508	0.4206
ZenNAS	0.1508	0.1508		0.0159
ResNet	0.4206	0.4206	0.0159	

Table 11: Table depicting p-values for statistical significance test conducted on Weighted TPR & FPR of Neck Dataset. Cells coloured in green represent significant values, while red cells represent insignificant p-values

MODELS	DeepMAD	ShapleyNAS	ZenNAS	ResNet
DeepMAD		0.3457	0.0952	0.0556
ShapleyNAS	0.3457		0.8848	0.1160
ZenNAS	0.0952	0.8848		0.0158
ResNet	0.0556	0.1160	0.0158	

Table 12: Table depicting p-values for statistical significance test conducted on Mean Accuracy of Elbow Dataset. Cells coloured in green represent significant values, while red cells represent insignificant p-values

MODELS	DeepMAD	ShapleyNAS	ZenNAS	ResNet
DeepMAD		0.0156	0.1719	0.6752
ShapleyNAS	0.0156		0.0465	0.0079
ZenNAS	0.1719	0.0465		0.0465
ResNet	0.6752	0.0079	0.0465	

Table 13: Table depicting p-values for statistical significance test conducted on Macro F1-Score of Elbow Dataset. Cells coloured in green represent significant values, while red cells represent insignificant p-values

MODELS	DeepMAD	ShapleyNAS	ZenNAS	ResNet
DeepMAD		0.0571	0.3428	0.3112
ShapleyNAS	0.0571		0.2000	0.2363
ZenNAS	0.3428	0.2000		0.9861
ResNet	0.3112	0.2363	0.9861	

Table 14: Table depicting p-values for statistical significance test conducted on Weighted TPR & FPR of Elbow Dataset. Cells coloured in green represent significant values, while red cells represent insignificant p-values

MODELS	DeepMAD	ShapleyNAS	ZenNAS	ResNet
DeepMAD		0.0079	0.0159	0.0067
ShapleyNAS	0.0079		0.0317	0.0158
ZenNAS	0.0159	0.0317		0.0079
ResNet	0.0067	0.0158	0.0079	