UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Project Final Paper 2023

Title: Exploring Supervised Learning for Small Data Pathology Classifiers

Author: Winner Bryan Kazaka

Project Abbreviation: DEEPPC

Supervisor(s): Geoff Nitschke, Deshen Moodley

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	
Theoretical Analysis	*0*	*25*	
Experiment Design and Execution	*0*	*20*	20
System Development and Implementation	*0*	*20*	5
Results, Findings and Conclusions	*10*	*20*	20
Aim Formulation and Background Work	*10*	*15*	15
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	
Total marks		**80**	

Exploring Supervised Learning for Small Data Pathology Classifiers

Winner Bryan Kazaka
KZKWINN001@myuct.ac.za
University of Cape Town
Cape Town, South Africa

Abstract

In the field of Orthopedic Pathology, expert diagnoses can be prone to errors due to human limitations, and having specialized teams is not always feasible. Conversely, machine learning (ML) algorithms excel in image recognition tasks, akin to those performed by medical specialists. Nonetheless, the scarce availability of extensive datasets in the medical field can hamper the optimal performance of these algorithms. This research assesses the efficacy of state-of-the-art (SOTA) Supervised Learning models and strategies, namely Data Augmentation (DA) and Transfer Learning (TL), in enhancing classification accuracy on two medical datasets. This paper establishes that ConvNeXts are the most proficient image classifiers among the tested models and underscores the effectiveness of pre-training and fine-tuning methodologies in optimizing Deep Learning (DL) models. While DA shows marginal benefits, particularly through Neural Augment, its impact on model accuracy is not substantial. Importantly, the introduction of a new validation technique, Default Validation, notably improves model accuracy by a margin of 1.6% to 6.9%.

Keywords

Pathology, Supervised Learning, Multi-label Classification, ConvNeXt, Transfer Learning, Neural Augment.

1 Introduction

Deep Learning harnesses advanced computer architectures and rich datasets to discern patterns in computer vision tasks. One notable technique within DL is Supervised Learning, where neural networks learn from labeled data, subsequently using these trained models to classify unseen data [36]. At its core, neural networks consist of interconnected units with specific values, determined by activation functions and parameters [20].

Recent advancements have seen the integration of DL tools in medical image analysis, yielding promising outcomes [2]. A prevalent task in this realm is multi-label classification image recognition, where an image may be associated with multiple labels, and the goal is to identify them all [9, 39]. However, the ongoing success in image processing largely hinges on the availability of labeled images, a luxury often absent in medical imaging. Open-source medical imaging datasets typically contain significantly fewer data points, ranging from 267 to 65,000 subjects. Such small datasets are plagued by various issues, including low statistical confidence, high error rates, class imbalances, and overfitting [37]. Overfitting, where a model narrowly adapts to its training data rather than understanding its broader traits, can be reduced through regularization techniques, such as DA [40]. DA enhances datasets by generating new, yet consistent data points, even when the dataset is minute in size, thereby boosting classification accuracy [7]. While some DA techniques involve simple transformations, others, like adversarial training, generative methods, and neural augmentation, are more intricate [32]. Another regularization strategy is TL, where insights from one model are transferred to another [10].

Parallel efforts to counteract overfitting have driven innovations in network design. Notable outcomes include architectures such as ConvNeXt [23], DenseNet [16], EffNet [27], and ResNet [14], designed for a spectrum of visual recognition tasks. However, manually calibrating these architectures demands deep expertise and considerable time. This challenge has fueled interest in automatic machine learning (Auto-ML), wherein models autonomously refine their topology and parameters through techniques like stochastic optimization [1, 31]. Prominent examples include ADAM [18] and Neuro-evolution of Augmenting Topologies (NEAT) [24]. Despite this, our study narrows its focus to DA and TL.

1.1 Research Questions

The contemporary literature in the field is more inclined towards augmenting dataset volumes instead of pinpointing the optimal model biases for small-scale datasets. It is observed that a significant focus is being placed on general computer vision tasks as opposed to specialized tasks such as image recognition. Furthermore, despite the implementation of renowned DL architectures on medical datasets, the performance outcomes have varied considerably. Often, the existing research on applying DL techniques in medical contexts accentuate metrics which, albeit academically intriguing, may not translate to practical utility for medical practitioners.

Another glaring trend in existing literature is the concentration on binary or multi-class classification issues, somewhat neglecting the intricate challenges associated with multi-label classification. Simultaneously, there appears to be a scarcity of explorations involving the application of SOTA Supervised Learning classifiers in medical datasets. This study is motivated by the necessity to address these existing gaps. Consequently, the following research questions have been formulated:

(1) What is the most adept DL model architecture for conducting multi-label classification tasks on small medical datasets?
(2) Is DA effective in enhancing the performance of small medical datasets?
(3) Can cross-domain TL facilitate efficient learning during the deployment of DL models on medical datasets?

In response to these questions, we undertake an evaluation of contemporary Supervised Learning models such as ConvNeXt,

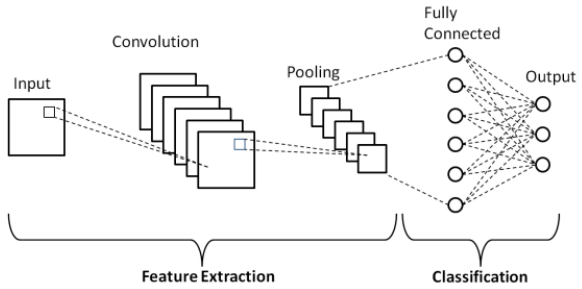


Figure 1: The overall general structure of a Convolutional Neural Network [30].

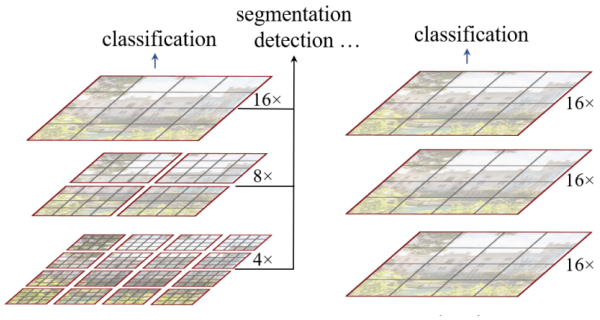


Figure 2: Comparison of Swin-T (left) with standard ViT feature maps (right). Swin-T, leveraging shifted windows, achieves linear complexity relative to image size by confining self-attention to its local vicinity, as illustrated by the red lines[22].

SwinNetv2 [22], EfficientNetv2, ResNet, and DenseNet, specifically in the context of multi-label classification tasks on restricted medical datasets. Additionally, we scrutinize the relevance of incorporating TL and field leading DA methodologies in these settings. Notably, our findings are reported employing the Exact Match metric [35], a criterion potentially more meaningful to the medical community. Complementing this, we propose a fresh training methodology, advocate for refined metric reporting conventions, and advise potential directions for ensuing research in this field.

2 Background & Related Work

This section delves into the latest Supervised Learning models, techniques optimizing their efficiency, and the evolution of DL in pathology.

2.1 Supervised Learning Models

Since the groundbreaking success of AlexNet [19] in the Image-1K challenge of 2012, Convolutional Neural Networks (CNNs) have dominated as the primary backbone for image classification tasks. As illustrated in Figure 1 above, CNNs encompass feature extractors and a classifier. While feature extractors discern the intrinsic patterns in data, the classifier predicts the corresponding output class. However, a recent shift from CNNs to Vision Transformers (ViTs) for visual tasks has been observed, spearheaded by Liu et al.'s [22] introduction of SwinTransformer. Soon after, the second version of the SwinTransformer emerged [21], aiming to bridge the gap between vision and Natural Language Processing (NLP) tasks. Yet, challenging this shift, Mao et al. [23] reintroduced the merits of CNNs for visual tasks, showcasing the ConNeXt, an evolved ResNet-50 model.

2.1.1 SwinTransformer. In their seminal 2021 paper, Lui et al. [22] introduced the Swin Hierarchical Vision Transformer (Swin-T). It quickly outperformed its CNN predecessors in various computer vision tasks. The prowess of Swin-T is attributed to its capability in handling long-range dependencies via the attention mechanism. This mechanism signifies how an input's representation is framed by evaluating its interplay with other sequence elements [21]. The original Swin-T's drawback was its quadratic complexity concerning image resolution. Nonetheless, this was rectified in its subsequent version [21] by employing a shifted window strategy, as visualized in Figure 2 above. This method partitions an image into non-overlapping windows, optimizing self-attention within each. However, the enhanced Swin-T still demands considerable GPU memory and extended training durations.

2.1.2 ConvNeXt. In 2022, Mao et al. [23] introduced the ConvNeXt, a revised ResNet-50 model that draws inspiration from ViTs but remains purely a CNN. They demonstrated that ConvNeXt outperforms the premier ViT model, Swin-T, across key computer vision tasks like image classification, object detection, and segmentation. These findings were consistent on both ImageNet-1K and ImageNet-22K datasets. Notably, despite its superior classification accuracy, ConvNeXt records a higher FLOP average and is optimized mainly for large datasets, not necessarily for subverting overfitting.

2.1.3 EfficientNet. Tan and Le [34] unveiled EffNetv2 with an aim to cut down model training times. The rationale is to facilitate frequent model modifications and testing. Employing progressive training, EffNet dynamically tweaks its settings. Earlier models like EffNetv1 [24] employed a strategy of incrementally enlarging image size. Conversely, EffNetv2 [27] couples this with DA to procure novel data points. As training progresses, image size and regularization strength escalate, enhancing classification accuracy. A limiting factor, however, is the need for sizable datasets to replicate these results.

2.1.4 ResNet. ResNets have been a consistently well performing deep model in the computer vision realm with numerous variations. A ResNet's salient feature is the uninterrupted data flow from one layer to the next, devoid of intermediaries. Bello et al. [3] train a canonical ResNet with advanced training and scaling techniques, finding faster training times than EffNets but similar accuracy levels. Their research underscores the pivotal role of training methods in model performance and replication.

2.1.5 DenseNet. In 2018, Huang et al. [16] introduced DenseNets to Supervised Learning. Unique to DenseNet is an architecture where every layer connects to every other in a sequential manner. As shown in Figure 3 below, each layer uses the feature maps of all prior layers as inputs and feeds into all subsequent layers. This design minimizes redundancy, leveraging identity transformations

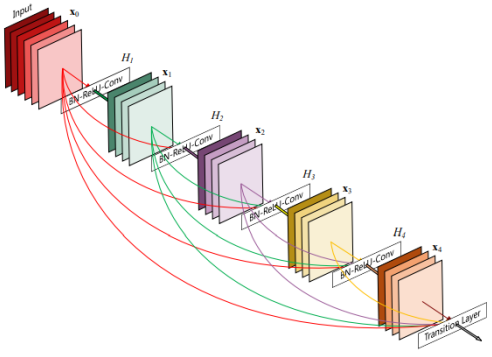


Figure 3: A depiction of a DenseNet's convolution layers, focusing on a 5-layer dense block with a growth rate of k = 4. Each layer integrates all preceding feature maps [16].

between layers. Mathematically, given $x_0, \ldots, x_{\text{‘}-1}$ as input, the relationship is captured as:

$$x = H([\boldsymbol{x_0}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_{\text{‘}-1}}]) \tag{1}$$

where in equation (1), $[\boldsymbol{x_0}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_{-1}}]$ refers to the concatenation feature maps produced in layers 0, ..., $l-1$. This architecture substantially trims the parameter count, ensuring efficient training as layers reuse gradients from the loss function. DenseNets also have inherent regularization effects, aiding datasets susceptible to overfitting. Regrettably, the original DenseNet paper [16] omits ImageNet benchmarks, restricting direct comparisons with other models.

2.2 Regularization and Optimization

The performance of Supervised Learning models is directly proportional to the size and quality of their training data [32]. Sparse datasets tend to lead models into overfitting, causing them to fail in generalizing to unseen data. Generalization refers to the capability of a model to perform well on new, unseen data after being trained on a particular dataset. Overfitting is effectively countered by employing regularization strategies, which can be categorized into data enhancement methods known as DA and techniques that adjust the model's training process. This research paper emphasizes the significance and efficiency of DA and TL among the wide array of regularization techniques such as Dropout [25], Batch normalization [17], TL [38], Pre-training and Fine-tuning [10], DA [8] and finally One-shot and Zero-shot learning [26].

2.2.1 Transfer Learning. TL and its subset, Pre-training, both revolve around utilizing knowledge from a broad domain to improve a narrower one. TL entails transferring both the network's architecture and weights, whereas Pre-training involves only the latter [10, 32, 38]. Homogeneous TL, where source and target models share similar features, is widely used in medical image analysis. Typically, in Pre-training, a model is initially exposed to extensive supervised or unsupervised data similar to the training dataset, allowing it to grasp general data representations. This process establishes the model's weights, decreasing the training data required. Subsequently, Fine-tuning takes the initialized weights and transfers them to a target model, refining either some or all of its weights. This research applies the Pre-training and Fine-tuning methodologies in TL.

2.2.2 Data Augmentation. DA bolsters DL training by increasing both the volume and diversity of training datasets. Broadly, DA techniques fall into Generic and Smart Augmentations. The former involves straightforward image alterations such as geometric and photometric transformations, including flipping, color adjustments, cropping, rotations, and noise additions. This can effectively double the size of a given dataset. Taylor and Nitschke determined that for smaller datasets, cropping proved most beneficial [36].

Conversely, Smart DAs harness advanced DL techniques for image transformations. They include Automated Augmentation, Adversarial Training, Neural Augmentation (Neural Aug), and Meta-learning DAs. While Automated Augmentation seeks the best transformation policy within a set of geometric transformations, Generative Adversarial Networks (GANs) create entirely new synthetic images that retain the distribution characteristics of the original data. Neural Augmentation similarly generates synthetic images but relies on the convolution of randomly selected images from a dataset. Among state-of-the-art Automated Augmentation algorithms, RandAugment (Rand Aug) was identified as being exceptionally efficient and effective [5]. A comprehensive study by Wang and Perez contrasted traditional augmentations, GANs, and various neural augmentation methods. Their analysis unveiled that Neural Augmentation, especially when content loss is omitted, offers compelling results, second only to generic augmentations in certain applications [8].

2.3 Deep Learning in Pathology

The utilization of computerized image analysis in pathology started in the 1970s, evolving significantly with the integration of Supervised Learning techniques in the 1990s and gaining momentum with the success of AlexNet [19] in image classification challenges.

2.3.1 Medical Datasets. In pathological studies, tasks typically involve single or multiple images. When referring to multiple images, the term "exam" or "subject" is commonly used. Presently, the primary applications of DL techniques in medical image analysis encompass segmentation, classification, abnormality detection, and computer-aided detection. The magnitude of images in medical datasets are often much smaller than those present in computer vision, but this is not the only cause of sub-optimal classification accuracies. A common phenomenon is class imbalance, where data of rare cases may be difficult to find or simply do not exist [2].

2.3.2 Caveats in DL for Pathology. When applying supervised learning to medical datasets, it is essential to consider factors such as the necessity of multi-label classification for certain tasks and leveraging domain-specific knowledge for data processing and augmentation. Often, the quality of training data suffers due to dataset bias and inadequate reporting of data demographics, affecting the generalizability of results.

Drawing from existing literature, we hypothesize that ConvNeXts demonstrate superior image classification abilities for the tasks proposed, with DenseNets being more apt for medical image analysis. Moreover, optimization of DL models is more effective when incorporating TL and utilizing constrained Rand Aug DA, ensuring the preservation of medical image label integrity.

3 Methods

In this section, we elucidate our experimental approach and the underlying rationale. Our aim is to offer a comprehensive understanding of the methodologies employed. By doing so, we aim to ensure that other researchers can replicate our results.

3.1 Research Aims

To recap, in this study, we are driven by the motivation to offer a nuanced understanding of how SOTA DL models, coupled with DA and TL techniques, perform in multi-label classification tasks on two medical image datasets. We recognize that the realm of medical datasets presents unique challenges, and hence, demands a specialized approach. Our overarching aim is to navigate the landscape of DL models focusing especially on CNNs and ViTs, and to evaluate their performance dynamics when applied to medical datasets. This deep dive promises to be an invaluable resource for researchers and practitioners alike, guiding them towards informed choices about model and training methodologies that align with their specific needs.

Our exploration will involve a mix of CNNs and ViTs as DL models. These will be tested under two experimental conditions: DA and TL. Moreover, an assessment across models of varying architectural complexities will be conducted, aiming to gauge the influence of model size on the designated evaluation metrics and increase variation in the results. The desired evaluation metrics are detailed in Section 3.4. Our ultimate vision is two-fold: First, we aspire to pinpoint the models or techniques that demonstrates the highest efficacy in the given context. Second, we aim to present a holistic view of the identified solution's advantages and constraints. We are committed to arming medical professionals with comprehensive insights, empowering them to choose techniques and strategies that are seamlessly integrated into their clinical workflows, ensuring precision and efficiency.

3.2 Data Collection and Ethics

The image data pertaining to patients, furnished by the Department, has undergone thorough anonymization and is devoid of any sensitive or personally identifiable particulars. Dr. Kruger from Groote Schuur Hospital has granted the necessary ethical clearance for this data. It is imperative to note that the University of Cape Town retains proprietary rights over all developed models and their corresponding codebase - these are made public on Github[1].

[1]https://github.com/bryankazaka/DEEPPC-Supervised-Learning/

3.3 Tools and Software

The following tools were essential for the effective execution of the experimental design and were consistently employed throughout the project's duration.

(1) The models used in the experiments were made available by and imported from Pytorch[2]
(2) This paper used Torch Vision for Data Augmentation[3]
(3) This paper used models pre-trained on the publicly available ImageNet1K-V1[4]
(4) Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team[5]

3.4 Evaluation Methods

In the realm of neural networks, traditional benchmarks often employ classification accuracy for performance assessment. However, when evaluating medical image systems, the focus shifts to more clinically-relevant metrics such as Precision, Recall, F1 score, AUC and Hamming Loss. The formulas for these metrics are as follows:

$$F1\text{score} = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{2}$$

Where Precision and Recall are defined by:

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

In multi-label classification scenarios, macro-average accuracy can be represented as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{5}$$

And Hamming Loss as:

$$Hamming\text{Loss} = \frac{FP + FN}{TP + FP + TN + FN} \tag{6}$$

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively, and make up the confusion matrix [2, 6]. To clarify terminologies: TP implies an accurate prediction of label presence, TN indicates a correct prediction of label absence, FP refers to a mistaken prediction of label presence, and FN denotes an erroneous prediction of label absence. Equation (3) evaluates a model's Precision, indicating the trustworthiness of a positive prediction. Equation (4) measures Recall, highlighting the model's adeptness in identifying false negatives. The synergy of these metrics allows for the construction of the Precision-Recall curve (AUC), balancing the two. Equation (2) formulates the F1 score, which offers a comprehensive view of the model's balance between Precision and Recall, particularly useful for imbalanced datasets. Equation (5) quantifies accuracy, reflecting the proportion of samples classified correctly. It's noteworthy that for F1 score,

[2]https://pytorch.org/vision/stable/models.html
[3]https://pytorch.org/vision/stable/index.html
[4]https://www.image-net.org/download.php
[5]hpc.uct.ac.za

Precision, Recall, Accuracy, and AUC, a higher value implies better performance. Equation (6), Hamming Loss gauges the average discrepancy between true and predicted labels. Crucially, a lower Hamming Loss indicates improved model efficacy, contrasting with previously discussed metrics [41].

As Singh et al. [33] highlight, not all errors carry the same implications; for instance, a false negative in the context of a life-threatening disease poses a significantly greater threat than a false positive. Given the gravity of medical decisions and the differences among errors, this paper emphasizes model practicality and task sensitivity. Thus, we adopt the Exact Match [35] as our primary accuracy metric, along with micro-averaged Precision and Recall. The implementation of the Exact Match metric is described in Section 3.5. The rationale behind using Exact Match and micro averaging lies in their emphasis on individual predictions, in contrast to macro averaging methods which may oversimplify class disparities. F1, AUC, and Hamming Loss are reported as secondary results. For a comprehensive perspective, we have also included results in a macro-averaged format, for overall and per label accuracy, following the form from literature such as those on the ChestX-ray8 dataset by Kalantari et al. [29]. These are available in Appendix D and are referenced as supplementary findings.

3.5 Problem Formulation

In our study, we address a multi-label image classification challenge within the realm of computer vision. Specifically, models are trained on two distinct datasets: Neck and Elbow X-ray images. The details regarding the class distribution for each dataset are provided in Appendix A. Each image in these datasets can contain one or multiple class instances, with the "normal" class signifying the absence of any instances. For processing, image labels are transformed into one-hot encoded arrays, with binary representation, indicating the presence or absence of a particular category within an image. Adhering to a conventional split, our dataset is divided into training, validation, and testing subsets at an 80/10/10 ratio, a decision aimed at optimizing computational efficiency.

We evaluate five architectural paradigms: ConvNeXt, SwinTransformerv2, DenseNet, EfficientNetv2, and ResNet, across three escalating architecture complexities. These architectures are tested against four distinct DA techniques: No DA, Random Cropping, Rand Aug, and Neural Aug. Furthermore, evaluations are conducted both with and without TL on each of the two datasets. With all the combined permutations, our assessment comprehensively covers 240 models. For a visual grasp, Appendix B illustrates sample images from our dataset, while Appendix C provides a glimpse of the outcomes from different DA techniques. A more detailed explanation of the Experiment Design can be found in section 4.

The models receive as input, medical images, courtesy of Dr. Kruger, converted into tensors. The output is comprised of predicted label arrays, each element of which undergoes a softmax function. A defined threshold of 0.5 is set: values below this threshold translate to 0, while those above result in a score of 1. Successful classification is achieved only if each array element is predicted accurately, aligning with the Exact Match criterion. Given the high stakes of medical image analysis, this stringent criterion is deliberately employed.

4 Experiment Overview

This section provides a comprehensive breakdown of the experimental procedures followed in this study, summarized in Table 1, the experimental design which are formulated to support or refute the research questions defined in Section 1.1. It is pivotal to note that, to facilitate fair comparisons across architectures, the training methods remained consistent for each model as advised by Bello et al. [3], unless otherwise stated.

Experiment	Methods	Variables	Datasets	Research Question
DL model architecture selection	ConvNeXt, Swin-Tv2, DenseNet, EffNetv2, ResNet	Parameter count, TL method, DA method	Elbow and Neck	Q1
DA method selection	Random Crop, Rand Aug, Neural Aug, No DA	Model architecture, TL method	Elbow and Neck	Q2
TL method selection	ImageNet-1K pre-training, No TL	Model architecture, No DA	Elbow and Neck	Q3

Table 1: Experimental Design. See Section 2.1 for model architecures, Section 2.2 for DA and TL methods, and Section 1.1 for research questions.

4.1 Data Processing

The provided datasets consist of two distinct directories: Neck and Elbow images, each labelled by an image ID. Accompanying .csv files detail the image number, ID, and corresponding label configuration. We partitioned these datasets into training, validation, and test subsets, adhering to an 80/10/10 split. This manual division ensures each label appears proportionally across subsets as in the original dataset. The data was then cleansed for inconsistencies. To prepare the data for training, images were converted into tensors after undergoing a GrayScale transformation and resizing. Likewise, the images' labels were converted to tensors. Our approach of shuffling the training image dataset draws inspiration from a prevalent ResNet-50 implementation[6]. We save the class labels for each training set to later output per label accuracy.

4.2 Training

In summary, our experiment encompasses the training and evaluation of 240 distinct models. Depending on the dataset (Elbow or Neck), we employ three different architecture complexities for five model architectures: ConvNeXt, SwinTransformerv2, DenseNet, EfficientNetv2, or ResNet. These models can either leverage pre-trained weights or not and undergo one of four image augmentation types: No DA, Random Cropping, Rand Augm, or Neural Aug. All models were trained on the UCT HPC a100 partition with a configuration of 4g:20gb:1 with 32 task nodes.

[6]https://www.kaggle.com/code/pmigdal/transfer-learning-with-resnet-50-in-pytorch/notebook

4.2.1 Model Sizes. Each architecture has three variants based on complexity.

For ConvNeXt, proposed by Mao et al. [23], we use:

- ConvNeXt-T: 29 million parameters
- ConvNeXt-S: 50 million parameters
- CovNeXt-B: 89 million parameters

For Swin-T, following Liu et al. [21], we use:

- Swinv2-T: 28 million parameters
- Swinv2-S: 49 million parameters
- Swinv2-B: 88 million parameters

DenseNet configurations, Huang et al. [16] inspired, we train:

- DenseNet-121: 8 million parameters
- DenseNet-169: 14 million parameters
- DenseNet-201: 20 million parameters

EfficientNetv2, by Tan and Le [34], we use:

- EfficientNetv2-S: 22 million parameters
- EfficientNetv2-M: 54 million parameters
- EfficientNetv2-L: 120 million parameters

Lastly, for Resnet, based on He et al. [14], we employ:

- Resnet18: 12 million parameters
- ResNet50: 26 million parameters
- ResNet152: 60 million parameters

Hyperparameter	**Value**
Activation Function	Sigmoid, ReLU
Cost Function	Binary Cross Entropy
Learning Rate	1×10^{-3}
Weight Decay	1×10^{-4}/10 steps
Optimizer	Adam
Epochs	15-TL, 30-No TL
Batch Size	32
Image Size	224x224
Training Callbacks	Early Stopping, Normalization

Table 2: Hyperparameters values

4.2.2 Hyperparameter Tuning. In this section, we detail the hyperparameter settings employed during model training, as outlined in Table 2. Our choices for these hyperparameters draw inspiration from approaches in the domain, with further refinement achieved through manual tuning. For activation, our custom classifier uses the ReLU function, while the model outputs employ the Sigmoid function. Binary Cross Entropy serves as our loss function.

In setting the learning rate, we experimented with values 0.05 and 0.001, finding the latter to be more favorable. The weight decay is fixed at $0.1x$ the learning rate. Amongst the optimizers, we compared Adam and Stochastic Gradient Descent, with Adam demonstrating superior performance. To set the number of epochs, we monitored validation metrics, including accuracy and loss. Our experiments with batch sizes: 16, 32, and 48, whereby 32 was highlighted as the optimal choice. In terms of image dimensions, we tested 144, 244, 360, and 480, settling on 244 in alignment with widely-accepted practices, as emphasized by Sabottke et al. [28].

To avoid overfitting and to streamline our training process, we employed early stopping, saving the model iteration with the peak validation accuracy. For image normalization, the dataset's mean and standard deviation for each RGB channel were used. Lastly, for model evaluation, we relied on the scikit-learn library[7] to compute metrics such as Hamming Loss, F1 score, AUC, Precision, and Recall. The numpy library[8] was used to compute attributes of the confusion matrix for both the overarching model and individual labels.

4.2.3 Transfer Learning Implementation. Models not pre-trained are set to have all their parameters tunable. Pre-trained model have their weights initialized to being fit on the ImageNet-1K dataset and have their parameters frozen. After which all models have their classification layer replaced with a linear layer with dimensions: number of features x 128, with a ReLU activation function and a final linear layer of dimensions 128 x number of output class labels for that dataset. This configures the models to be able to predict for their specific dataset. These 2 additional layers have their parameters tunable. This follows a well established training consideration of a TL ResNet50 model[9]. The epoch consideration for the experiments were largely chosen manually by observation accuracy trends over time to a point of convergence, but the values used by He et al. [13] we used as a starting point.

4.2.4 Data Augmentation Implementation. As of the DA implementation, for models without DA, the images were loaded into their respective data loaders, were set to grayscale with 3 output channels (RGB), resized to 224x224 and then finally tensor-converted.

For the Random Cropping augmentation method, Torch Vision's RandomResizedCrop was used. The image dimensions of the resize was kept at 224x224 and the same grayscale and tensor conversion were applied to the images.

For the RandAugment implementation method, The image was first grayscaled in order to nullify an colour augmentations which would not keep the integrity of medical images, we then used Torch Vision's Rand Aug, with a number of operations = 3, and magnitude = 25. These values were inspired by values produced by Tan and Le's [34] experiments with RandAugment for considerations for our batch size of 32 and image size of 224x224.

The previous three augmentations are made "on the fly", meaning as the images are loaded into the data loader. In contrast, the implementation of Neural Aug followed the Offline Augmentation method due to its computational intensity. For each dataset, each image is trained through a VGG model for 2000 epochs, with the style image, among the available tested by Gatys et al. [11], most similar to those in medical image analysis. This image is included in Appendix C. The augmented image is added back into the dataset,

[7] https://scikit-learn.org/stable/
[8] https://numpy.org/
[9] https://www.kaggle.com/code/pmigdal/transfer-learning-with-resnet-50-in-pytorch

	Models						DA Methods					TL Methods	
	ConvNeXt	Swin-T	DenseNet	EffNet	ResNet		No DA	Cropping	Rand Aug	Neural Aug		No TL	TL
ConvNeXT	x		✓0.091			No DA	x		✓0.013		No TL	x	✓$1.41e-15$
Swin-T		x	✓0.0131			Cropping		x	✓0.096		TL	✓$1.41e-15$	x
DenseNet	✓0.091	✓0.131	x	✓0.003	✓0.172	Rand Aug	✓0.013	✓0.096	x	✓0.015			
EffNet			✓0.003	x	✓0.044	Neural Aug			✓0.015	x			
ResNet			✓0.172	✓0.044	x								

Table 3: Elbow Dataset results of the Mann-Whitney U Test. Intersections marked by checks indicate where there was a statistically significant difference in test accuracy from the two groups ($p <= 0.20$). Refer to Section 2.1 for a recap of the model architectures and Section 2.2 for an explanation of the DA and TL methods.

	Models						DA Methods					TL Methods	
	ConvNeXt	Swin-T	DenseNet	EffNet	ResNet		No DA	Cropping	RandAug	Neural Aug		No TL	TL
ConvNeXT	x	✓0.047			✓0.196	No DA	x				No TL	x	✓0.041
Swin-T	✓0.047	x				Cropping		x			TL	✓0.041	x
DenseNet			x			RandAug			x				
EffNet				x		Neural Aug				x			
ResNet	✓0.196				x								

Table 4: Neck Dataset results of the Mann-Whitney U Test. Intersections marked by checks indicate where there was a statistically significant difference in test accuracy from the two groups ($p <= 0.20$). Refer to Section 2.1 for a recap of the model architectures and Section 2.2 for an explanation of the DA and TL methods.

and the corresponding .csv file appended with the new image and the label of its sample image. Effectively each dataset is augmented to have its magnitude doubled. All modifiable variables, including the amount of epochs to train for was kept consistent with the Neural Augment implementation proposed by Gatys et al. [11]. The code implementation for the method used in this paper is made publicly available[10] and was modified to the use case of this paper's experiment. Synonomous to the other DA methods in this paper, the images were set to grayscale and resized to 224x224 before being added to the data loader.

5 Results

In this section, we present the outcomes of the experiments. Following the Shapiro-Wilk test for normality, with a threshold set at $p = 0.05$, it was ascertained that the data is not normally distributed. Consequently, the non-parametric Mann–Whitney U test was employed at $p = 0.20$ to investigate the discrepancies in test accuracies across different methodologies. Significant distinctions between model architectures, DA and TL methods are denoted with a check mark and the associated p-value in Table 3 and 4. The comparisons between the models are displayed in Figure 4 on the following page. A comprehensive breakdown of the performance metrics of the top models is available in Table 10, alongside macro-averaged accuracies depicted in Table 8 and Table 9.

5.1 Elbow Dataset

5.1.1 Model Architectures. The data denoted in Table 3 and visualized in Figure 4(a) signifies that the DenseNet model notably surpasses other architectures in terms of test accuracy, although it faces close competition from ConvNeXt, particularly in the realms of Precision and Recall. Furthermore, EffNets are identified as the least effective model, lagging behind ResNets, a trend corroborated by the macro-averaged scores expressed in Table 8

5.1.2 DA Methods. An analysis of the data represented in Table 3 and Figure 4(c) reveals the subpar performance of Rand Aug relative to other DA techniques. Despite the marginal statistical differences between DA methods, both Figure 4(c) and Table 8 underline that Neural Aug tends to produce superior results, with No DA also demonstrating competitive performance in several instances.

5.1.3 TL Methods. Table 3 and Figure 4(e) emphasize the enhanced effectiveness of incorporating TL techniques compared to approaches devoid of it, a pattern echoed across all evaluated metrics.

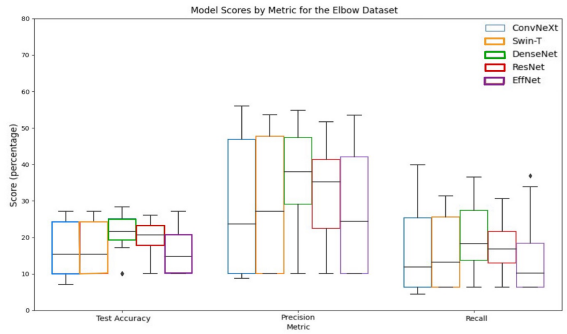
5.2 Neck Dataset

5.2.1 Model Architectures. Upon examining Table 4 and Figure 4(b), it is evident that the Swin-T and ResNet architectures exhibit superior performance in comparison to the ConvNeXt model for the Neck dataset, albeit based on mean outcomes. Notably, ConvNeXt excels in the top quartile and maximum performance cases, a trend substantiated by its dominance in the highest ranks for the Neck dataset experiments showcased in Table 10. Consistently, EffNets register the lowest peak test accuracy results.

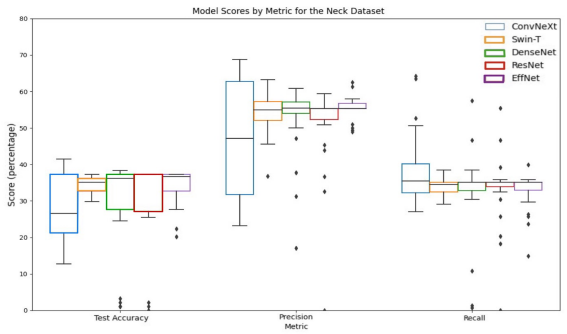
5.2.2 DA Methods. Figure 4(d) highlights No DA as the optimal choice concerning test accuracy, trailed closely by Neural Aug, which prevails in Precision and Recall metrics. These observations are in harmony with the macro-averaged outcomes presented in the supplementary tables 8 and 9.

5.2.3 TL Methods. As corroborated by Table 4 and Figure 4(f), TL methods substantially enhance Test Accuracy and Precision. However, Figure 4(f) indicates a noticeable dip in performance for Recall metrics, wherein models without TL showcased better results.

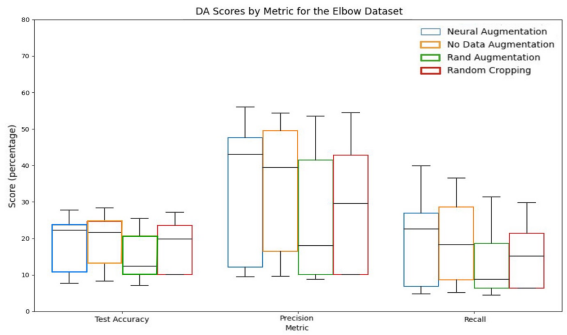
[10]https://github.com/aladdinpersson/Machine-Learning-Collection/tree/master/ML/Pytorch/more_advanced/neuralstyle



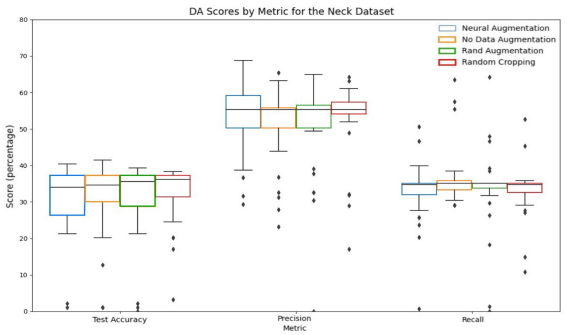
Model Scores by Metric for the Elbow Dataset. The experimental variables: (a) transfer learning and data augmentation are aggregated ($N = 24$ models per architecture).



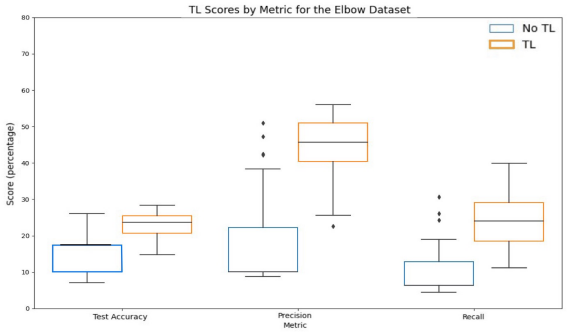
Model Scores by Metric for the Neck Dataset. The experimental variables: (b) transfer learning and data augmentation are aggregated ($N = 24$ models per architecture).



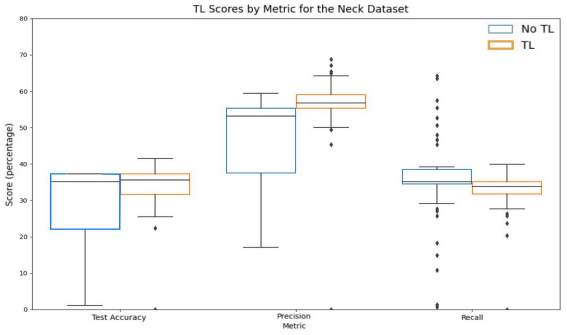
(c) DA Scores by Metric for the Elbow Dataset. The experimental variables: architecture and transfer learning are aggregated ($N = 30$ models per DA method).



(d) DA Scores by Metric for the Neck Dataset. The experimental variables: architecture and transfer learning are aggregated ($N = 30$ models per DA method).



(e) TL Scores by Metric for the Elbow Dataset. The experimental variables: architecture and DA method are aggregated ($N = 60$ models per TL method).



(f) TL Scores by Metric for the Neck Dataset. The experimental variables: architecture and DA method are aggregated ($N = 60$ models per TL method).

Figure 4: Experiment Results for model architectures, DA and TL by metric. The independent axes show the metric scores normalized into a percentage. The dependent axes shows the different regularization methods and their results for each metric. The methods are colour coded by the legend displayed on the top right of each graph.

6 Discussion

In this section, we scrutinize the outcomes obtained from our experiments, juxtaposing them with the research queries delineated in Section 1.1 to ascertain their alignment or contradiction with our initial hypotheses. Furthermore, we contextualize these findings within the broader corpus of existing literature, critically appraising their significance and potential contributions to the domain. We also propose prospective directions for future research, stimulated by the insights gleaned from our data analysis. As illustrated in Table 10, there is a consistent synergy among the top model architectures, TL methods, and DA methods, collaboratively enhancing the performance metrics of the resulting models.

6.1 Models

In our study, EffNets emerged as the least effective models across both datasets, recording the lowest test accuracy scores amongst top models. This performance gap can be attributed to EffNets' characteristic approach of scaling up the image size [27]. Our experimentation indicated an optimal image dimension of 224x224 during the parameter tuning phase, where any increase beyond this threshold led to diminished accuracy, corroborating Sabottke et al's findings [28] on the plateau effect of image resolution in DL pathology studies. Specifically, this research reinforces that while larger image dimensions can enhance DL for pathology, the effectiveness peaks at approximately 224x224. Beyond this, increased parameter space counteract the benefits of added image features.

Contrarily, DenseNet exhibited superior performance on the Elbow dataset, closely trailed by ConvNeXt models. Moreover, the ConvNeXt architecture displayed optimal results for the Neck dataset, with DenseNet following closely. This alignment with the existing literature supports our hypothesis that DenseNet's high regularization effect due to dense connections makes it well-suited for small datasets. Meanwhile, the complexity and small size of the Neck dataset necessitate a more robust and deeper model, hence better performance with the novel ConvNeXt architecture which boasts the top performing metric results in the field [23].

Responding to Research Question (1) outlined in Section 1.1 which begs to which SOTA DL model is optimal for small medical datasets, we deduce that the efficacy of DL model architectures for multi-label classification tasks on compact medical datasets is contingent on the specific dataset and problem use case. However, in a generalized scenario considering all criterion, in which despite Test Accuracy, the ConvNeXt architecture dominates; we consider the ConvNeXt architecture to demonstrate the superior performance.

6.2 Data Augmentations

In our evaluation, the Neural Aug method demonstrated a noticeable underperformance in the Recall tests for the Neck dataset as illustrated in Figure 4(d), despite achieving comparable results in Precision to other methods. This discrepancy indicates a tendency of the Neural Aug method to yield higher false negative rates, manifesting a conservative predictive model. This phenomenon is further supported by analyzing the dataset image pixel values dispersion, both pre and post Neural Aug application. As depicted in the table

Table 5: Mean and Standard Deviation for Offline Datasets. Refer to Section 2.2.1 for an explanation of DA and Neural Aug.

	No DA		Neural Aug	
	mean	std	mean	std
Elbow	0.1256	0.2226	0.1746	0.1922
Neck	0.5468	0.2548	0.5414	0.2522

above, the application of Neural Aug leads to a reduction in the standard deviation of pixel values across both datasets, implying an increased homogeneity and smoothing of the datasets. This aspect can potentially hinder the discernment of finer features pivotal in medical image analysis, making the dataset more uniform and less diverse in features.

On the other side, Neural Aug seems to excel in Test Accuracy compared to other DA methods, albeit without statistical significance, predominantly driven by its high Precision scores. This suggests a higher probability of correct predictions when the model identifies positive cases, indicating a robustness against class imbalance issues compared to other DA techniques. This resilience can be attributed to the preservation of pixel integrity in Neural Aug, mitigating the risks of pathogen exclusion or distortion that might occur with other methods such as Random Cropping and Rand Aug. These issues are particularly accentuated in medical datasets where distinguishing between intricate pathogen patterns is vital.

Despite the observed differences, statistical tests reveal no significant disparities between the DA methods at a $p = 0.20$ significance level, possibly due to the sensitive nature of medical imagery and the nuanced manner of their presentation. Yet, a review of Table 8, Table 9, and the top-ranking models in Table 10 signifies a marginal advantage in employing DA techniques in this domain as opposed to not.

In response to Research Question (2) delineated in Section 1.1, which queries the efficacy of DA in optimizing the utility of compact medical datasets, we affirm the positive role of DA in enhancing the performance metrics in supervised learning for medical image analysis, given an appropriate method is employed. Hence, we advocate the consideration of Neural Aug as a viable strategy [11].

6.3 Transfer Learning

In the experiments conducted on both datasets, we notice an improvement in performance, affirming the observations made by Yadav and Jadhav in their study [39]. TL has emerged as inarguably useful in CNN training. It serves to elevate accuracy levels and to foster quicker convergence when retraining particular features on novel datasets. Moreover, it mitigates the issue of overfitting by capitalizing on the knowledge acquired from analyzing larger capacity datasets compared to those available in pathology.

Conversely, during the sessions without the use of pretrained weights, we noted a surge in training and validation accuracy until reaching a peak, followed by a dip to zero percent before the

completion of ten epochs. This decline is attributable to the shrinking weight values for all labels, except for the 'normal' label. This phenomenon is analogous to the vanishing gradients problem, a prevalent issue where repeated multiplications during backpropagation render the gradients exceedingly small. This scenario also hints at the model's inability to adequately generalize over the training set in the absence of TL, resulting in uniform predictions across different models, which subsequently explains the plateaued upper sections of the box plots in Figure 4.

Moreover, TL proves instrumental, especially in the fine-tuning phase, by imparting knowledge of general features to models grappling with feature extraction in their existing domain. This strategy also augments weight initialization, preventing the derivative functions of model weights from being initialized too near to zero, thereby averting the implications of vanishing gradients [12, 15].

Addressing Research Question (3) introduced in Section 1.1 on the efficacy of TL in this domain, we affirm that TL manifests noticeable advantages in both datasets under investigation.

6.4 Dataset

Medical images harbor a substantial number of label classes possible per image, which amplifies the potential for incorrect model predictions. This complexity, we theorize, contributes to the diminished performance observed in the Elbow dataset, despite its larger image repository when juxtaposed with the Neck dataset. Moreover, the X-ray images in the datasets present significant variations in aspects like shades, zoom levels, angles, and poses, adding layers of complexity to the data. For instance, elbow images display stark differences with regards to arm positioning, which can either be straight or bent, while neck images introduce complexity due to the intricate spinal patterns. These variations, consequently, influence how pathogens appear across individuals, further complicating the pattern recognition process. Table 5 reports the image pixel spread for the Neck dataset at $mean = [0.5468, 0.5468, 0.5468]$, $std = [0.2548, 0.2548, 0.2548]$ which is noticeably higher than ImageNet's $mean = [0.485, 0.456, 0.406]$, $std = [0.229, 0.224, 0.225]$, despite all the Neck dataset being homogeneous.

A prominent inverse relationship is observed between the quantity of examples available for a particular pathogen and its corresponding per-label accuracy, with an R^2 value of 0.94 and 0.95 for the Neck and Elbow datasets, respectively, as illustrated in Figure 15. This correlation underscores that the prevailing issue in medical datasets is not necessarily data scarcity, but rather class imbalance. This imbalance suggests that models are yet to grasp the fundamental structures within the data fully, given that adding marginal examples of a class tends to diminish the accuracy for that respective class. This imbalance manifests prominently in the disparity observed between Precision and Recall metrics, signifying a higher rate of false negative predictions. Essentially, the models exhibit a propensity to predict negative outcomes for niche classes, capitalizing on the higher likelihood of being correct, synonomous with the phenomenon of overfitting.

This scenario compellingly argues for further research aimed at evaluating the potential benefits of integrating oversampling techniques with DA to mitigate the class imbalance prevalent in medical datasets. Implementing strategies such as the Synthetic Minority Oversampling Technique (SMOTE) [4] may present a viable avenue to address this issue effectively.

6.5 Default Validation

The depicted data in the preceding diagram notably highlights an elevated error rate for normal images, despite their prevalence in the dataset. This contradiction instigated further investigation into the standard practices of medical image labelling. Subsequently, we introduce a novel testing methodology dubbed "Default Validation", envisioned to rectify logical impossiblilities encountered during label predictions in multi-label classifications of medical images.

Distinct from other applications, the multi-label classification in medical imaging is somewhat nuanced, encompassing not only various pathogen classes but also incorporating a default normal class. Generally, the normal class, symbolizing the absence of pathogenic labels, gets treated as a pathogen label during the annotation process. This discrepancy in labelling practices is not uncommon, as observed in prominent datasets like ChestX-ray14[11], which harbors over 112K images. Default Validation aims to align the labelling process with logical consistency by imposing two primary constraints: 1) The prediction should indicate 'normal' if no pathogens are identified in an image, and 2) The 'normal' label should be excluded if one or more pathogens are identified in an image. This logic is encapsulated in Algorithm 1, where 'prediction' refers to the output tensor encapsulating label predictions for an individual image.

Algorithm 1 Default Validation

```
for prediction in predictions do
    if all(prediction[:-1] == 0) then
        prediction[-1] ⟵ 1
    else
        prediction[-1] ⟵ 0
    end if
end for
```

By integrating this method, we enhance the predictive accuracy of the model, fostering informed decisions grounded in a comprehensive understanding of the problem domain. Implemented within the testing pipeline post the application of activation and softmax functions, Default Validation serves to streamline the prediction process. To gauge the effectiveness of this pioneering approach, we retested all $N = 240$ models, now incorporating Default Validation, and analyzed the outcomes using the Mann-Whitney U Test (at a significance level of $p = 0.05$). This analysis revealed a statistically significant improvement in the Exact Match results, as documented in Table 12.

Further analysis, reflected in Table 11, showcases the augmented performance of top-tier models in the Neck and Elbow datasets

[11] https://paperswithcode.com/dataset/chestx-ray14

when Default Validation is implemented, indicating an enhancement in accuracy ranging between 1.6% and 6.9%. Although, it does this by improving Recall, sacrificing performance on Precision. Despite this, it makes for a more practical model for medical diagnoses. We notice a more pronounced effect noted in the Neck dataset and therefore we conclude that this can be attributed to the higher accuracy levels associated with the normal class in the Neck dataset, implying a heightened efficacy in datasets with more discernible pathogens.

This notable improvement not only advocates for the integration of Default Validation in DL applications on medical datasets but also fosters avenues for additional research focused on developing specialized medical classifiers. A two-step classification process, distinguishing between pathogenic and non-pathogenic images initially using a binary classifier, followed by a more nuanced multi-label classification for pathogenic instances, presents a promising direction for future endeavors.

7 Conclusions

In culmination, this paper finds without statistical but with good reason that ConvNeXt is the superior DL model architecture for multi-label image classification. We find the EffNet architecture to underperform all other architectures due to aggressive image scaling and therefore further deduce 224x224 to be the optimal image resolution size for medical images. The results also point to the belief, albeit without statistical significance, that DA improves the performance of DL models trained on medical datasets, yet importantly, with augmentations that preserve the integrity of each pixel in the image as much as possible, such as Neural Aug. On experiments testing the efficacy of TL, we find with substantial statistical significance that TL increases the performance of models and mitigate overfitting phenomenon pertinent with small and unbalanced datasets. We find with substantial evidence that in medical datasets, class imbalance is a more detrimental issue than the shear magnitude of the datasets. Conclusively, this paper proposes Default validation, a validation algorithm which eliminates logical impossibilities in label predictions in multi-label classification implementations for medical datasets. We find this technique, with substantial statistical significance to improve the Exact Match accuracy by 1.6%-6.9% across both datasets.

8 Future Work

The current study identifies potential areas for extended research, including exploring additional DA methods optimized for multi-label classification and investigating the benefits of oversampling to address class imbalance. While TL has proven beneficial, further studies could explore the effectiveness of transferring knowledge from models specifically trained on medical datasets. There's also room to expand upon the Default Validation approach, perhaps by developing a more sophisticated, hierarchical medical classifier that integrates both multi-label and binary classifiers, ensuring a focused analysis of medical images rather than a generalized, less optimal solution.

9 Acknowledgements

We extend our gratitude to Associate Prof. Nitschke for guidance and Dr. Kruger for providing the medical datasets. This research was facilitated by the support from the University of Cape Town ICTS High Performance Cluster team and Mr. Lewis, who assisted in navigating the technical environment.

References

[1] S. Acton, S. Ambramowitz, L. Toledo, and G. Nitschke. Efficiently coevolving deep neural networks and data augmentations. In *IEEE Symposium Symposium Series on Computational Intelligence*, Cape Town, CT, SA, 2020. IEEE.

[2] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan. Medical image analysis using convolutional neural networks: A review. *J Med Syst*, 42:226–239, 2014.

[3] I. Bello, W. Fedus, X. Du, E. D. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph. Revisiting resnets: Improved training and scaling strategies. 2021.

[4] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 06 2002.

[5] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020.

[6] L. Deininger, B. Stimpel, A. Yucel, S. Abbasi-Sureshjani, S. Schönenberger, P. Ocampo, K. Korski, and F. Gaire. A comparative study between vision transformers and cnns in digital pathology. 2022.

[7] S. Dolgikh. Analysis and augmentation of small datasets with unsupervised machine learning. *medRxiv*, pages 2021–04, 2021.

[8] M. Elgendi, M. U. Nasir, Q. Tang, D. Smith, J. P. Grenier, C. Batte, B. Spieler, W. D. Leslie, C. Menon, R. R. Fletcher, N. Howard, R. Ward, W. Parker, and S. Nicolaou. The effectiveness of image augmentation in deep learning networks for detecting covid-19: A geometric transformation perspective. *Frontiers in Medicine*, 8:629134, 2021.

[9] M. J. Er, R. Venkatesan, and N. Wang. An online universal classifier for binary, multi-class and multi-label classification. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 003701–003706, 2016.

[10] D. Erhan, Y. Bengio, A. Courville, P. Manzagol, P. Vincent, and S. Bengio. Why does unsuervised pre-training help deep learning? journal of machine learning research. 2010.

[11] L. Gatys, A. Ecker, and M. Bethge. A neural algorithm of artistic style. *Journal of Vision*, 16:326, September 2016. Vision Sciences Society Annual Meeting Abstract.

[12] B. Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[15] K. Hee, A. Cosa, N. Santhanam, M. Jannesari, M. Maros, and T. Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22, 04 2022.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. pages 4700–4708, 2017.

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 2015. ICLR 2015.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[20] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[21] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, pages 11976–11986, 2022.

[24] R. Miikkulainen, J. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzyan, N. Duffy, et al. Evolving deep neural networks. In *Artificial intelligence in the age of neural networks and brain computing*, pages 293–312. Elsevier, 2019.

[25] S. Nitish, H. Geoffrey, K. Alex, S. Ilya, S. Ruslan, et al. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[26] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[27] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár. Designing network design spaces. pages 10428–10436, 2020.

[28] C. F. Sabottke and B. M. Spieler. The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, 2(1):e190015, 2020.

[29] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[30] P. Sharma. Basic introduction to convolutional neural network in deep learning. *Data Science Blogathon*, March 2022. Last Modified On August 24th, 2023.

[31] N. Shazeer and M. Stern. Adafactor: Adaptive learning rates with sublinear memory cost. pages 4596–4604, 2018.

[32] C. Shorten and T. Khoshgoftaa. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(60):1–48, Apr. 2019.

[33] R. K. Singh, R. Pandey, and R. N. Babu. Covidscreen: explainable deep learning framework for differential diagnosis of covid-19 using chest x-rays. *Neural Computing and Applications*, 33(14):8871–8892, 2021.

[34] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. pages 10096–10106, 2021.

[35] L. Tang, S. Rajan, and V. K. Narayanan. Large scale multi-label classification via metalabeler. In *Proceedings of the 18th international conference on World wide web*, pages 211–220, 2009.

[36] L. Taylor and G. Nitschke. Improving deep learning with generic data augmentation. In *IEEE Symposium Symposium Series on Computational Intelligence*, SSCI 2018, pages 1542–1546, Cape Town, CT, SA, 2018. IEEE.

[37] G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1):48, 2022.

[38] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(9):1–40, Apr.

[39] S. Yadav and S. Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6, 12 2019.

[40] X. Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2), Feb.

[41] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 26:1819–1837, 08 2014.

Appendix A: Dataset Analysis

In this appendix section, we delve into the distribution of the Elbow and Neck datasets examined in the study. Analyzing these datasets offers a glimpse into the nuances of small and imbalanced pathology datasets and their prevalent labelling conventions. While this data fortifies our understanding of model behavior and supports our results, it remains supplementary and is therefore not integral to the core discussion of the paper. By examining this data, professionals in pathology can discern the nature of the training data, helping them ascertain the potential applicability of the models. Even though this dataset analysis augments our discussion and bolsters our results, it isn't pivotal for comprehending the primary methodologies, outcomes, or the paper's conclusions.

Table 6: Pathogen distribution for the Elbow dataset

LABELS	Train	Validation	Test
Number of images	2378	193	169
Soft tissue swelling	202	31	31
Joint effusion	493	62	61
Distal humerus	64	12	9
Supracondylar	691	87	92
Medial epicondyle displaced	71	14	8
Lateral epicondyle displaced	111	12	18
Olecranon	52	6	6
Elbow dislocation anterior	12	2	2
Elbow dislocation posterior	47	2	2
Proximal radial	40	2	6
Radial head	14	2	8
Radial head subluxation	18	2	3
Proximal ulnar metaphysis	27	1	5
Normal	1255	57	17

Table 7: Pathogen distribution for the Neck dataset

LABELS	Train	Validation	Test
Number of images	746	93	94
Alignment	411	51	52
Soft tissue swelling	83	10	10
Listhesis	63	11	8
Fracture	99	9	15
Dislocation	26	2	3
Spinous	43	6	3
Other pathogens	127	16	16
Normal	239	30	30

Appendix B: Dataset Image Examples

In this appendix, we showcase select images from both the Elbow and Neck datasets, serving as a supplementary visual tool to enhance the readers' comprehension of the dataset's inherent complexity and distinct features. This visual aid is instrumental in elucidating the pixel distribution values identified during our analysis. Although these images are not central to understanding the primary results or the subsequent discussion, they provide a nuanced insight into the dataset's characteristics, thereby enriching the context around our research questions. Furthermore, while the main text offers a macro analysis of the dataset, these visuals offer a more granular perspective, allowing readers to appreciate the finer details and complexities that are woven into the data, potentially sparking further scholarly discourse or exploration.

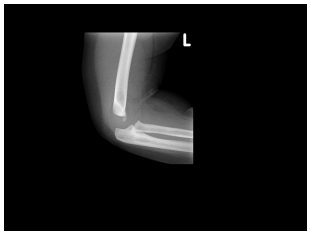


Figure 5: An image from the Elbow dataset, classified: Normal.

Figure 6: An image from the Neck dataset, classified: Normal.

Figure 7: An image from the Elbow dataset, classified: Soft tissue swelling and Dislocation (posterior).

Figure 8: An image from the Neck dataset, classified: Alignment, Soft tissue swelling, Listhesis, Dislocation, Spinous, and Other pathogens.

Appendix C: Data Augmentation Examples

In this appendix, we present a series of visual representations illustrating the effects of various data augmentation techniques on a sample image from the dataset. This initiative is designed to enhance the reader's comprehension of the different data augmentation methodologies employed, offering a tangible glimpse into the transformations applied during the experiment, albeit more for informational enrichment and curiosity satisfaction. While these visuals are not crucial to grasping the theoretical or practical aspects of the augmentations or interpreting the core findings and results, they provide an enriching supplementary perspective, offering an intuitive understanding that complements the theoretical discussions addressed in the main body of the paper.

Figure 9: A sample elbow image with No DA applied.

Figure 10: The sample elbow image with Random Cropping applied.

Figure 11: The sample elbow image with Neural Augment applied.

Figure 12: One example of the sample image with RandAugment applied.

Figure 13: Another example of the same sample image with RandAugment applied.

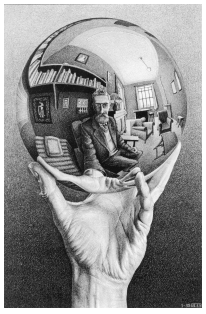


Figure 14: The style image chosen for the Neural Style augmentation.

Appendix D: Supplementary Results

This appendix includes extra results, such as macro-averaged outcomes, peak model performance, and supporting evidence for Default Validation's rationale. Using macro-averaged results helps with comparing our study to others in the field. Displaying the top models' performance data gives us a better understanding of each algorithm's strengths and weaknesses. These results are supplementary and do not directly address the primary research questions. They are included in this appendix as supporting information.

	Models					DA Methods				TL Methods	
	ConvNeXt	Swin-T	EffNet	**DenseNet**	ResNet	No DA	Random Crop	Rand Aug	**Neural Aug**	No TL	**TL**
Soft tissue swelling	81.66	81.59	81.39	81.34	81.61	81.38	81.64	81.58	81.46	81.67	81.52
Joint effusion	63.61	63.76	63.42	63.88	64.2	63.51	64.13	64.03	63.43	63.86	63.77
Distal humerus	94.67	94.67	94.67	94.67	94.67	94.67	94.67	94.67	94.67	94.67	94.67
Supracondylar	52.74	51.8	51.6	55.87	52.71	54.3	52.01	50.63	54.83	48.05	52.94
Medial epicondyle displaced	95.27	95.27	95.27	95.27	95.27	95.27	95.27	95.27	95.27	95.27	95.27
Lateral epicondyle displaced	89.35	89.35	89.35	89.35	89.35	89.35	89.35	89.35	89.35	89.35	89.35
Olecranon	96.45	96.45	96.45	96.45	96.45	96.45	96.45	96.45	96.45	96.45	96.45
Elbow dislocation anterior	98.82	98.82	98.82	98.82	98.82	98.82	98.82	98.82	98.82	98.82	98.82
Elbow dislocation posterior	98.82	98.82	98.82	98.82	98.82	98.82	98.82	98.82	98.82	98.82	98.82
Proximal radial	96.45	96.45	96.45	96.45	96.45	96.45	96.45	96.45	96.45	96.45	96.45
Radial head	95.27	95.27	95.27	95.27	95.27	95.27	95.27	95.27	95.27	95.27	95.27
Radial head subluxation	98.22	98.22	98.22	98.22	98.22	98.22	98.22	98.22	98.22	98.22	98.22
Proximal ulnar metaphysis	97.04	97.04	97.04	97.04	97.04	97.04	97.04	97.04	97.04	97.04	97.04
Normal	44.65	42.33	42.26	56.73	51.23	52.43	43.55	38.48	55.31	25.13	47.44
Macro Average Accuracy	85.93	85.7	85.65	**87.01**	86.44	86.57	85.84	85.36	**86.81**	84.22	**86.15**

Table 8: Elbow Dataset macro-averaged results. See Section 2.1 for model architectures and Section 2.2 for DA and TL methods.

	Models					DA Methods				TL Methods	
	ConvNeXt	Swin-T	**EffNet**	DenseNet	ResNet	No DA	Random Crop	Rand Aug	**Neural Aug**	No TL	**TL**
Alignment	58.47	56.52	56.29	55.5	54.52	56.42	55.74	55.5	57.38	54.1	56.26
Soft tissue swelling	80.89	88.87	89.32	89.36	89.36	86.63	88.54	87.73	87.34	85.8	87.56
Listhesis	80.85	86.21	87.19	87.23	87.23	85.17	85.6	85.42	86.77	84.27	85.74
Fracture	76.64	81.64	81.64	76.86	78.41	77.87	81.8	79.11	77.37	76.61	79.04
Dislocation	86.62	92.47	93.53	93.62	93.62	91.42	91.56	92.52	92.38	90.36	91.97
Spinous	87.19	93.79	94.41	94.68	94.68	92.98	92.84	93.16	92.84	91.33	92.95
Other pathogens	75.8	82.31	82.18	76.6	82.94	77.41	79.61	81.7	81.13	77.23	79.96
Normal	64.85	67.56	68.09	68.09	64.72	66.81	66.71	65.36	67.77	65.22	66.66
Macro Average Accuracy	76.41	81.17	**81.58**	80.24	80.69	79.34	80.3	80.06	**80.37**	78.12	**80.02**

Table 9: Neck Dataset macro-averaged results. See Section 2.1 for model architectures and Section 2.2 for DA and TL methods.

Elbow Top Models	Test Accuracy	Macro Avg. Acc	Precision	Recall	F1	AUC	Hamming
DenseNet121 with TL and No DA	**28.40**	88.72	50.32	29.10	36.88	62.72	11.28
DenseNet121 with TL and Neural Aug	27.81	**89.35**	**54.82**	**33.96**	**41.94**	**65.19**	**10.65**
EffNetv2-L with TL and Neural Aug	27.22	89.14	52.94	36.94	43.52	66.37	10.86
(Baseline) DenseNet121 with No TL and No DA	18.34	86.64	30.00	13.43	18.56	54.71	13.36
Neck Top Models	Test Accuracy	Macro Avg. Acc	Precision	Recall	F1	AUC	Hamming
ConvNeXt-B with TL and No DA	**41.49**	83.64	65.43	35.81	46.29	65.59	16.36
ConvNeXt-S with TL and Neural Aug	40.43	**84.18**	**68.83**	**35.81**	**47.11**	**65.92**	**15.82**
ConvNeXt-B with TL and Neural Aug	39.36	83.91	67.09	35.81	46.70	65.75	16.09
(Baseline) ConvNeXt-B with No TL and No DA	20.21	68.35	27.94	38.51	32.39	57.09	31.65

Table 10: Top model performances. Baseline shows the top model without enhancements. See Section 3.1 for evaluation metrics.

Elbow Top Models	Test Accuracy	Macro Avg. Acc	Precision	Recall	F1	AUC	Hamming
ConvNeXt-B with TL and Neural Aug	**30.00**	**88.66**	**49.02**	**40.00**	**44.05**	**67.39**	**11.34**
ResNet50 with TL and Neural Aug	27.50	88.57	48.37	35.60	41.01	65.41	11.43
DenseNet121 with TL and Random Cropping	27.50	86.88	36.25	23.20	28.29	59.04	13.12
(Baseline) ConvNeXt-B with No TL and No DA	9.38	83.04	9.38	6.00	7.32	49.36	16.96
Neck Top Models	Test Accuracy	Macro Avg. Acc	Precision	Recall	F1	AUC	Hamming
DenseNet121 with TL and Neural Aug	**48.44**	**83.20**	**62.50**	**39.22**	**48.19**	**66.68**	**16.80**
ResNet152 with TL and Neural Aug	48.44	80.08	56.25	35.29	43.37	64.23	18.36
ConvNeXt-T with TL and Neural Aug	46.88	83.20	62.50	39.22	48.19	66.68	16.80
(Baseline) DensetNet121 with No TL and No DA	39.06	82.03	57.81	36.27	44.58	64.84	17.97

Table 11: Top model performances with Default Validation. 'Baseline' shows the top model without enhancements. See Section 3.1 for evaluation metrics.

	Elbow		Neck	
	No Defaulting	Defaulting	No Defaulting	Defaulting
No Defaulting	x	✓$p = 3.38e-08$	x	✓$p = 8.44-42$
Defaulting	✓$p = 3.38e-08$	x	$p = 8.44-42$	x

Table 12: Statistical test results of the Mann-Whitney U Test on the models produced by the experiments, with and without Default Validation. Intersections marked by checks indicate where there was a statistically significant difference in test accuracies from the two groups ($N = 240$ models, $p <= 0.05$). See Section 6.5 for Default Validation.

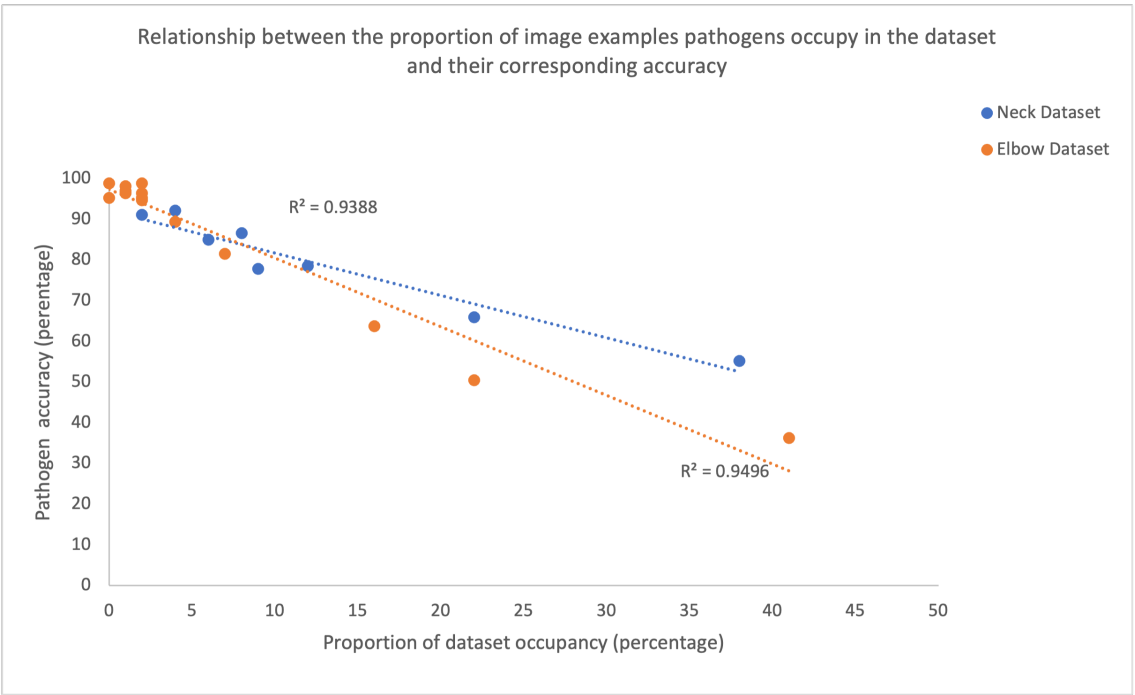


Figure 15: Relationship between Pathogen propensity and their associated accuracies for the Neck and Elbow datasets. See Appendix A for the dataset distributions and Table 8 and Table 9 in Appendix D for the per label accuracies.