Supervised Learning for Pathology Classifiers

Literature Review

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

ABSTRACT

In Pathology, orthopedic diagnoses require a team of trained professionals to perform, which is often not feasible. These trained professionals are also prone to classification inaccuracies due to human error. Machine Learning algorithms have proved to be substantially accurate in image recognition tasks which are the husk of those performed by the medical professionals. The problem lies in the fact that the same amount of data that is needed for these algorithms is not available in the field of medicine and thus the algorithms cannot perform with the same accuracy. This review evaluates and critically compares state of the art models in Supervised Learning, and the subsequent methods used to increase the efficiency of classification tasks, namely: Transfer Learning and DA. Furthermore, this paper assesses contributions made to the field of Pathology by Deep learning techniques. We conclude that ConvNeXts exhibit the highest image classification capability of the proposed models but DenseNets are the most suited for medical image analysis tasks. Furthermore, that Deep Learning models can best be optimized by using MrTrAdaBoost instance pre-training and fine-tuning methods while using RandAugment DA within a constrained set of augmentations that will preserve medical image label integrity.

1 INTRODUCTION

Deep Learning, a subset of Artificial Intelligence, uses highly dimensional datasets as well as high performance computer architectures to generalise patterns in computer vision tasks. Supervised learning is a deep learning technique which uses labelled datasets to train neural networks to recognise patterns in images and classify those images to the predefined labels, after which these neural networks are then tested in their ability to categorise unseen data [25]. Deep Learning is synonymous with Convolutional Neural Networks (ConvNets), a deep learning architecture specifically for computer vision [32]. Neural networks are a type of learning algorithm which uses units that hold values determined by some activation a and parameters $\theta = W, B$, where W is a set of weights and B a set of biases. The neurons are connected in a layered structure. The convolution refers to the layers in the layered architecture responsible for feature detection of the input image [22]. Recent studies in the field have regarded Vision Transformers (ViTs) as the state of the art deep learning model in terms of image classification. ViTs are deep learning models that rely solely on self-attention as opposed to convolution to compute its input and output representations. Self-attention relates different elements of an input sequence by computing them simultaneously as opposed to ConvNets which compute in input sequences one element at a time. Furthermore, seminal work on SwinTransformers, a ViT variant for image classification using shifted windows [40] addressed the pitfalls of earlier transformers, caused an unprecedented shift in the field from the use of ConvNets to ViTs [37, 41]. Medical image analysis in clinical diagnoses have been supplemented with Deep Learning tools and have shown highly successful results [35]. However, the continued success of Natural Image Processing has seen is largely attributed to the availability of labelled images which is not the case for medical imaging datasets. Open source medical imaging datasets are magnitudes smaller with sizes ranging from 267 to 65,000 subjects. Small dataset sizes lead to low statistical confidence, high error margins, and overfitting [15]. Overfitting is a phenomenon in Deep Learning in which a model memorizes the training dataset instead of learning the fundamental characteristics of the data [39]. Overfitting can be mitigated either by increasing the size of the dataset or through modifications to a model architecture. Enhancing the dataset to decrease the effect of overfitting is done through Data Augmentation (DA). DA is the technique of producing new data points to enhance a dataset while still preserving labels. This can be done with extremely small datasets in order to improve classification accuracy [9]. DA can include simple transformations such flipping, cropping and colour space augmentations. These are often categorised as Generic Augmentations. Generic Augmentation is a set of computationally inexpensive geometric and photometric transformations [25]. DAs can also present themselves in more complex methods. The most influential of these include adversarial training, generative methods, Smart Augmentation and Neural Augmentation [5]. Other techniques employed to prevent overfitting involve optimizing the model architecture. This has subsequently influenced advances in Network Design, its tools, guidelines, and principles. Examples of the byproducts include DenseNet, EffNet, and RegNet[13, 17, 21, 42]. These are model architectures are tailored to a broad range of visual recognition tasks. Manually tuning parameters and network architectures requires extensive knowledge of neural network model architecture, is complex and takes a long time. This has lead to research in Automatic Machine Learning(Auto-ML) which are techniques in which the model optimizes its own neural topology and parameters using stochastic optimization and neuroevolutionary methods. [28, 32]. Examples of these include ADAM [8] and Neuro-evolution of Augmenting Topologies (NEAT) [30]. Other optimization strategies include Dropout [27], Batch normalization [31], Transfer Learning [19], Pre-training [11] and finally One-shot and Zero-shot learning [29]. This paper will only consider DA, Transfer learning and Pre-training. Transfer learning and Pre-training are similar in the fact that both techniques operate by leveraging the transfer of knowledge from a larger domain to augment the knowledge of a smaller one. Transfer learning differs whereby both the network architecture and the weights must be transferred whereas, Pre-training only transfers weights and the network architecture remains the same [5, 11, 19]. This literature review evaluates recently proposed: ConvNet, SwinNetv2, EfficientNetv2, ResNet-RS and DenseNet supervised learning models and

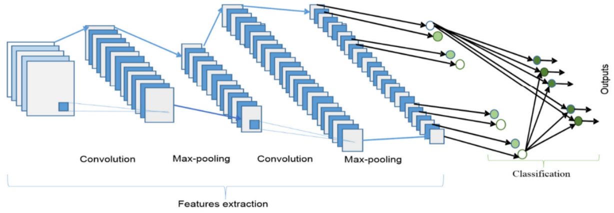

Figure 1: The overall general structure of a Convolutional Neural Network [2].

their contributions to image classification tasks. In addition, this review explores the contemporary use of deep learning techniques in the field of Pathology, its successes, pitfalls and opportunities. Furthermore, it discusses effective methods of pre-training, transfer learning and DA techniques. In culmination, this paper provides a hypothesis on the most appropriate model to use with regards to research and experiments on orthopedic datasets, favourable DA techniques to remedy overfitting, as well as suggestions of opportunities for research in the field of Deep Learning.

2 SUPERVISED LEARNING MODELS

As shown in Figure 1 above, ConvNets generally consist of two parts, a collection of feature extractors and a classifier. The feature extractors are split into convolution and max-pooling layers which are grouped into a plane. The output for each plane is used as input for the next. High level features are extracted from features propagated in the lower layers. Gradually as the layer levels increase, the dimensions of the features are reduced and the number of feature maps increased. The final and fully connected later is the classification later, where the extracted features received as inputs from the previous layer are used to calculate a decisive output which classify the given input image [2]. Different families of ConvNets have different network architectures which satisfy different task needs. The following section includes explanations of each architecture that will be reviewed, supplemented with graphical representations for architectures that are distinct.

2.1 Convolutional Neural Network

Despite the early 2020's renaissance in supervised learning, induced by the emergence of ViTs, shifting the research paradigm from the use of ConvNets to ViTs for computer vision tasks, just last year in 2022, Liu et. al [42] at Facebook published a controversial paper on ConvNets still being superior to ViTs. The research proposes a modified ResNet-50 model: ConvNeXt, which borrows design aspects from ViTs, while still maintaining a pure ConvNet architecture. The experiment results presented in the paper show that ConvNeXt outperforms the SwinTransformer, a hierarchical transformer with the highest ImageNet image classification accuracy. However, this is done at a higher Floating Point Operation (FLOP) average. FLOP is an field standard indicator of the energy usage and latency of a model [36]. This performance trend follows on all general computer vision tasks namely image classification, object detection/segmentation and semantic segmentation on both ImageNet-1k and ImageNet-22k datasets. ConvNeXt is also proven to exhibit likewise scaling behaviour boasted by ViTs, in relation to the magnitude of the training set. This is important since transformers were shown to outperform ResNets with the availability of large datasets. ConvNets generally outperform ViTs with higher resolution images. This is due to the fact that previously, ViTs had a global attention design that scaled on a quadratic complexity with regards to input size and thus were precarious with higher resolution images. This was addressed in SwinTransformer v2. In the experiment described in the paper, the original ResNet-50 architecture was used and its architecture improved by leveraging the hierarchical design from SwinTransformerv2. The experiment compared three objectives: image classification, object detection/segmentation and semantic segmentation. These objectives cover a wide range of computer vision tasks and are therefore standard in the field to test a model's robustness. The paper also boldly utilized the same inference throughput measurement method from the paper proposing SwinTransformer v2 [41]. The strengths in this research is in its contribution of the ConvNeXt model, its highly substantiated statement of ConvNeXt's superiority in image classification, and its generalizable methods which follows standards and thus simple to replicate. Despite these positives, the paper carries some latent naive assumptions and contradictions. The paper proposes that the SwinTransformer sliding window strategy allowed it to behave similarly to ConvNets, and uses this to infer that convolution is still desired in the field and not becoming redundant as marketed in recent field research. The authors use this perspective to deduce that ViTs are aiming to bring back convolutions but in less effective ways, those of which ConvNeXts already optimize. This is contradictory to the authors' design choice in borrowing the hierarchical structured strategy from ViTs. This paper contributes a highly competitive model, one that could potentially cause another paradigm shift back into the reign of ConvNets for research and applicable use. Although the paper contributes to generic tasks within computer vision, the paper focuses on scaling up with regards to dataset sizes and not on which model's inductive biases performs more optimally on small datasets.

2.2 Swin Network

Liu et al. presented SwinTransformer [40] and then shortly afterwards SwinTransformerv2 [41], a hierarchical ViT with the aim of reducing the functional divide between vision tasks and those in Natural Language Processing (NLP). This work was inspired by the revolutionary emergence of AlexNet [20] into the field of computer vision and its unprecedented success on the ImageNet-1K challenge at the time, which subsequently lead to ConvNets dominating the field as the optimal choice as the backbone for computer vision tasks. In the same likeness, SwinTransformer caused a radical shift in the field from ConvNets to ViTs being more prevalent. This was attributed to attention in transformer models and its ability to compute long range dependencies within the data. The strengths in this article is in the contribution of its paradigm shifting model. Subsequently, the paper is highly regarded. The key attribute of SwinTransformer v2 is their ability to scale with regards to image resolution. This is pivotal in the context of the topic of this literature review as medical images often come in high resolutions. In addition, SwinTransformers use self-supervised learning techniques in pre-training to be less reliant on large labelled datasets.

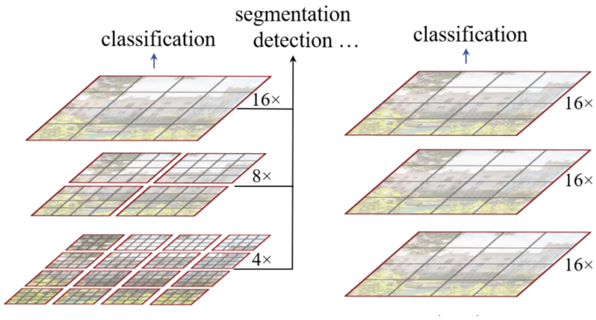

Figure 2: This figure shows the proposed SwinTransformer (on the left) and vanilla ViT feature maps (on the right). The SwinTransformer curates its hierarchical feature maps by merging image patches (the individual blocks separated by grey lines) in the deeper layers, and it achieves linear complexity in with regards to image size by computing self-attention only within its local domain (shown by the red lines).[40].

This corroborates further in its suitability in image classification for orthopaedic datasets. As depicted in Figure 2 above, SwinTransformers are ViT network architectures which are hierarchical in design and whose input sequence to output representation is computed with shifted windows. This shifted window strategy remedies the issue prevalent in other ViTs which have a quadratic computational complexity with regards to image size. SwinTransformer has a linear complexity. This significant decrease in complexity is achieved by the local computation of self-attention within non-overlapping windows that partition an image. There is limited research in the scaling of vanilla ViTs. Vanilla ViTs also require huge amounts of training data which SwinTransformers require less of. Furthermore, The vanilla ViT models also only apply to image classification problems, whereas SwinTransformer is compatible with a much larger range of computer vision tasks, most notably image classification, object detection, dense prediction and semantic segmentation [1]. Although it is a point to note that the SwinTransformer v2 model requires high Graphical Processing Unit (GPU) memory consumption and long training times on average. This is less of a problem with small datasets. A minor concern with the research objectives of SwinTransformers is that the research aims to close the gap between language models and vision models and thus fails to notice the inherent difference between the two distinct tasks, and further appreciate and optimize models to those tasks. This problem of dedicated models is visited and considered in the design of ConvNeXt [42].

2.3 Efficient Network

Tan and Le [26] proposed EfficientNet v2 (EffNet v2) which builds on from the ConvNet variant EffNet [17]. The EffNet architecture is one which aims to approach computer vision tasks with the intention of solving cases where larger training datasets scale to longer deep learning model training times. Long model training times is a concern as it decreases the ability of the model to be retrained or to be modified and tested. EffNetv2 presents faster training speeds and better parameter efficiency that previous models. This improvement is attributed to the proposed model using training aware neural architecture search and scaling to optimize speed. The scaling refers to progressive training in which the settings of the network are dynamically edited. Progressive training is synonymous with and prevalent in GANs, transfer learning, adversarial learning and language models. The difference between the progressive training used in these examples and the progressive training in EffNet v2 is that prior progressive training exclusively scaled up image sizes in order to optimize speed, while EffNet v2 introduces progressive regularization which adaptively adjusts the image using DA to obtain new data points while simultaneously also scaling the image size. This improves the training speed as well as the accuracy of the model. Early epochs in the model are trained with smaller images sizes and weak regularization. Progressively the size and strength of the regularization are increased. The three different forms of regularization approaches are the main topic of the study article. The first is Dropout, which is a regularization technique implemented at the network level and seeks to reduce co-adaptation by randomly removing channels [12]. Thus, the dropout rate can be modified. The second technique, known as RandAugment, involves adding an adjustable magnitude to each image's data [9]. The third method, called Mixup, is a methodology for cross-image DA. The mixup this paper uses, follows the ratio $x_i = x_j + (1-\lambda)x_i$ and $y_i = y_j + (1-\lambda)y_i$ to combine two images with the labels (x_i, y_i) and (x_j, y_j). The mixup ratio can be changed as necessary during training. While other works have also dedicated research to increasing training efficiency such as ConvNeXt [42] and ViTs [1, 37, 40], these models come with the draw back of having a large amount of parameters. EffNetv2 was evaluated on four transfer learning datasets namely: CIFAR-10, CIFAR-100, Flowers and Cars. The test dataset used to result in the model's summative prediction accuracy was the ImageNet ILSVRC2012 dataset after being pretrained on the ImageNet-21k dataset. Simple cutout DA was used. EffNetv2 boasts an 87.3 percent, top-1 accuracy on the test set. This result outperformed the most successful by accuracy ViT at the time: ViT-L/16, by 2 percent accuracy while also having trained the same data 5 to 11 times faster. Computing resources were kept controlled for all models in the experiment. The cumulative results of the study deduces that training with large images is a slow process, that depth wise convolutions are especially slow in the early layers of the neural network, and finally that equally scaling up each stage is not the most optimal solution. The strengths of this paper is in its contribution of the EffNet v2 model which performs at a highly competitive level with regards to classification accuracy, as well as the novelty in its progressive training strategy. The methods of the experiment are also highly general and comparatively simple to replicate. The model considers large image sizes and has inherent DA in its network design. Despite this, in the context of medical datasets, research in EffNet architectures in general concerns optimising the speed of training models which are tasked with image classification of large datasets, which medical imaging datasets do not have the pleasure of. Although this is the case, the model still produces a high classification accuracy score which is highly compelling.

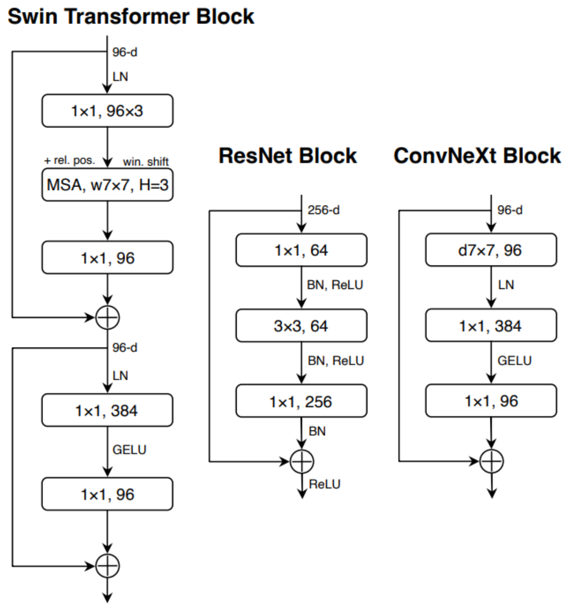

Figure 3: This figure depicts the block designs for aforementioned SwinNets, ConvNeXts and ResNets. The number on the left of the comma defines the convolutions of the layer and the number on the right, the depth of the block [42].

2.4 Residual Network

Bello et. al [18] revisited consistently well-performing deep learning architecture ResNets, with the aim of introducing novel ideas around training methods into the field of deep learning, as well as to contribute a ResNet architecture which can match state of the art models on popular standardized benchmark tests. ResNets are deep networks in which data flows from one layer into the next without any proxy from intermediate layers. Figure 3 above depicts the block design of a ResNet and how it compares to the aforementioned model designs. Gradual contributions to the ResNet family have consistently produced competitive results, but this paper argues that while these architectures which draw attention to themselves based on their statistics, the model is often the attribute of the study that is most advertised to have improved while the training methods, study strategies and hyper parameters are ignored. In publications, this is often exacerbated by the fact that new models with new training methods are compared to and benchmarked against old models with old training methods. This can be very misleading since training methods can contribute more to the effectiveness of computer vision tasks than the data architectures. The reason this bias is perpetuated is since training methods do not generalize as easily as data architectures do, and so are not easily repeated and tested in other experiments. Thus grouping new methods with new models can be misleading as to which attribute contributes more to the effectiveness of the results. Therefore, the paper suggests that to compare different data architectures, it should follow that training methods should be kept constant. In the field, this is often done through parameters and FLOPs, but the paper motivates that latencies and memory consumption are more relevant [17].

Bello et al. [18] found that scaling and training strategies generally had a more profound effect to effectiveness than architectural changes to a model, and that scaling strategies depended on the training regime. The paper offers and tests two scaling strategies. To arrive at these two strategies, models were exhaustively trained for the full duration, at 350 epochs instead of 10 epochs. Through this, the dependency between scaling strategy and the training regime which specifically included the number of epochs, model size and dataset size, was realized. After which, the two strategies were made explicit as: 1) To scale model depth in regimes where overfitting is prevalent, for example: small data sets, 2) To increase image resolution slowly. By using these scaling strategies and modern training strategies used in EffNets, a new family of ResNet architectures was proposed, ResNet-RS. ResNet-RS is arrived at by applying modern training methods to a a canonical ResNet, this is then measured, after which a few common and modern architectural improvements are made and the model is tested again. Results from the paper showed that the newly proposed family of architectures are $1.7x - 2.7x$ faster than EffNets while achieving very similar accuracies. Through the scaling experiments, the paper corroborated that DA generally works better for small datasets or models with high epochs, but the specific augmentations applied to the dataset to optimize efficiency can be class dependent. Furthermore, that optimal scaling strategies transfer well between tasks in which outfitting is an issue, but this largely depends on the training regime. Using the ImageNet dataset for training, it was evident that in the case of a high-end amount of epochs, data architectures should be deeper, and in the case of a low-end amount of epochs, more width should be applied to the model. This paper contributes a substantial caveat in data architectures overshadowing novel training methods and supporting strategies when it comes to evaluating the performance of a model. From this starting point, Bello et. al [18] experimented on canonical ResNets with cutting edge training methods and strategies and were not only able to enforce the importance of training methods and scaling strategies, but to make explicit the dependencies between them. By doing so this paper provides strong ground for researchers to reuse the paper's methods and to take hidden method biases into considerations when publishing publications. A minor drawback of this paper is in its experiment design, where it was only compared to EffNets, where readers could have benefited from more comparative references, especially with other models around the same computational speed.

2.5 Dense Convolutional Network

Densely Connected Convolutional Networks (DenseNet) made their appearance to the field of deep learning by Huang et. al's [13] contribution which proposes a convolutional network architecture whereby each later is connected to every other layer in a feed-forward fashion, hence the name. For each layer, the feature maps of all preceding layers are used as in its inputs, and it follows that its own feature map is used as input in all subsequent layers. This dense coupling promotes feature reuse, which in turn significantly reduces the number of parameters since there is no need to learn redundant feature maps. This is in contrast with ResNet parameters which are larger and whose architecture has additive identity

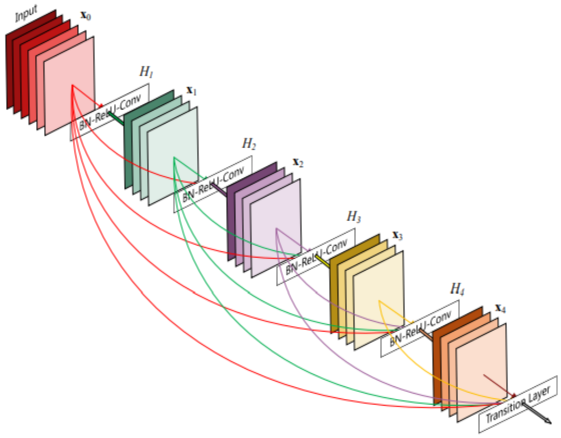

Figure 4: This figure shows the convolution layers of a DenseNet network architecture. This specific example shows a 5-layer dense block, with a growth rate of k = 4, since the transition layer takes all 4 preceding convolution layer feature maps as input. Each layer takes in all its preceding feature maps as input as well [13].

transformations per layer. Although, in various ResNets, the contribution by the layers are low but each layer is larger because it has its own weights [14]. DenseNets differ in that they make explicit the information that is retained and information that is added. As depicted in Figure 4 above, DenseNets are more coupled than conventional ConvNets where a network of L layers would have L connections, one between each layer and the subsequent layer, in this model each layer and has $L(L+1)/2$ direct connections. DenseNets exploit feature reuse by concatenation of the feature maps retained by preceding layers which results in an increase in the variation of inputs in the subsequent layers and ultimately improves efficiency. This relationship is mathematically modelled using $x_0, \ldots, x_{‘-1}$ as input by:

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

where $[\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{-1}}]$ refers to the concatenation feature maps produced in layers 0, ..., $l-1$.

DenseNets are therefore easier to train since layers can use the can reuse the gradients from the loss function and from the input string. In addition, DenseNets have an inherent regularizing effect which reduces overfitting, benefiting tasks with smaller training datasets. The proposed DenseNets were able to produce results superseding state of the art status and outperforming ResNet [14] by an error rate difference of 10.04% while requiring less computation. This paper's primary contribution was in the introduction of DenseNets, which in addition provided a computer vision data architecture solution which could scale to fewer parameters, includes regularization by design, and promotes the notion of reuse in network design and the reduction of redundancy. An unfortunate miss on the paper was a benchmark test on the widely used ImageNet benchmark, which would allow the model to directly be compared to more models more generally.

3 DEEP LEARNING IN PATHOLOGY

The inception of deep learning techniques being used in the field of pathology can be dated back to the 1970s where computers were used to automate image analysis using pixel processing and mathematical modelling to construct generic AI if-else rule base systems to solve niche tasks [22]. It was only later in the 1990s that supervised learning techniques specifically were introduced into the field. Around the same time research on ConvNets began but as aforementioned, it was the success of newly proposed AlextNet in 2012, winning the ImageNet competition by an impressive margin that ConvNets become largely prevalent in academics and a few years later practically in the field of medicine as well. Image classification was one of the first areas of interest for this new technology. In pathology, the task usually has one or multiple inputs or images, multiple images being collectively called an exam or a subject. Currently the areas in which deep learning techniques are most prevalent in medical image analysis include: segmentation, classification, abnormality detection and computer aided detection [35]. The output of the task is usually binary, returning either disease present or not. Bar et. al presented their work on deep learning in chest pathology [3] with novel methodology that is now prevalent in the field. The paper presents an experiment on a ConvNet model being tested on a small medical image dataset of 433 images, while having been pretrained on the non-medical ImageNet dataset. The novelty of this experiment was in its transfer learning of non-domain specific knowledge. This study concluded with favourable results and showed the feasibility and potential for using deep learning techniques in pathology and for using larger, non-medical training data.

3.1 Medical Datasets

Dataset sizes in the filed of medicine are often much smaller than those in computer vision but this is not always the case and sometimes not always the issue in classification inaccuracy. One example is the Picture Archiving and Communication System (PACS) in radiology which is a repository of over a million images in radiology and continues to grow. In the west there are also large publicly available medical datasets. Another issue distinct to medical datasets is that medical imaging classification is often presented as a binary task whereas in many cases any image can be a member of multiple classes of labels. This causes phenomena in which models perform well in general cases but fail on extremely rare ones. Another niche issue is class imbalance where data or images for classes of rare cases may be difficult to find or simply do not exist. These dataset issues and alike in the field provide opportunity for research in DA methods. It is currently research in GANs that are expected to solve many of the problems inherent to medical datasets.

3.2 Medical Image Analysis Evaluation Methods

Network architectures are often measured by classification accuracy compared to standard benchmarks, while medical image system analyses are measured using much more practical metrics which include F-1 score, precision, recall. These are calculated as:

$$F1score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{2}$$

where Precision and Recall are defined as:

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

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

And Accuracy is calculated as:

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

Where TP represents true positives, TN represents true negatives FP represents false positives, and FN represents false negatives [23, 35]. Precision and Recall is then used to calculated the Precision under Recall curve (PR AUC). Furthermore, to evaluate multi-class identification problems, macro-averaging is used which combines the PR AUC for each class to retrieve one value.

3.3 Caveats in Deep Learning for Pathology

Litjens et. al [22] present a survey on a corpus of over 300 papers on the use of deep learning techniques in pathology. The most substantial point that can be made by the survey is that modifying architectures to become deeper, effectively adding more layers and increasing parameters does little to increase a model's accuracy. It has been shown that it is possible for researchers to use the same architecture with the same methods but different datasets and produces wildly different results. The key contributor to the classification accuracy in medical datasets is very domain specific knowledge, and leveraging this knowledge to handle the particular dataset being dealt with. In addition to this, data processing and DA models that go beyond modifying the model are important as well. This domain specific approach also brings into consideration the quality of the training data. The quality of training data is often jeopardized by dataset bias, an alarming area of concern proposed by Varoquaux and Cheplygina [15]. Medical imaging publications often do not report the demographics of the data, and the results of the experiments do not often represent the distribution it should. This issue is due to data sampling, which is first often biased towards patients that are diseased, since they are more prevalent in hospitals, and then spectrum bias, in which a cohort may not be accurately represented. Further dataset availability for certain conditions distort research, as highly available datasets are more likely to be studied extensively. Varoquaux and Cheplygina [15] introduce many overlooked overheads in medical image analyses and thus their paper provides a strong argument suggesting redesign in methods used in academic pursuits to involve deep learning with pathology. Although, this paper overlooks the privacy concerns of the participant data studied by deep learning and medical researchers, and the ethical implications of disclosing such or using the information. Singh et. al [33] in a recent study on COVID-19 screening using deep learning emphasized that it is also important to note that in medical image classification, not all stats are equal, some are more severe than others. An example being one in which false negatives are far more expensive than false positive errors since the failure to diagnose a deadly and or contagious disease carries large risk. A huge positive for using deep learning techniques is in that they are largely open source, where the code is made available on source control platforms. This comes with the benefit to researchers to comparatively quickly use and compare different models and apply different techniques to analyse various medical image datasets and classifications of problems.

4 OPTIMIZATIONS

Small datasets are often the consequence of domain areas where large amounts of labelled data has not been accumulated, does not exist, or is too expensive to curate such as the cases prevalent in medical image analysis. The more data a machine learning model has access to, the better it performs. This is especially corroborated by Google's [16] contribution of their trillion word corpus and its improvement of text based models in the field of machine learning. Small datasets are prone to the inability to generalize. Generalizability refers to the ability of a model to categorize a test dataset after being training on a training dataset. Models with poor generalizability are said to have overfit the training data [5]. Overfitting degrades classification accuracy. There are a myriad of strategies to decrease overfitting. These either include enhancing the data set, such is are called DA methods, or by training and or tweaking the model. Any combination of these can be applied. Optimization strategies include Dropout [27], Batch normalization [31], Transfer Learning [19], Pre-training, DA[11] and finally One-shot and Zero-shot learning [29], although this section will only consider DA, Transfer learning and Pre-training.

4.1 Transfer Learning

. There are two classes of effective methods of transfer learning methods standard in practice explained by Weiss et. al [19] in their exhaustive survey done on transfer learning techniques. These are Homogeneous Transfer and Heterogeneous Transfer. Pre-training and fine-tuning transfer learning solutions prevalent in medical image analysis [3, 22] are categorised in Homogeneous Transfer methods. Homogeneous Transfer Learning for a given feature space X and models S and T is defined as $X_S = X_T$, this is the case where the source and the target have the same or similar features, thus are in the same domain. This is evident in Pre-training where a model is trained on a large amount of unsupervised data similar to the training dataset. This allows the model to learn general representations of the input data and to initialize starting weights which reduces the amount of training data needed. The process of fine-tuning follows on whereby the initial parameters of the pre-trained model are transferred and the model parameters updated when trained on task specific data. Heterogeneous Transfer learning for a given feature space X and models S and T is defined as $X_S \neq X_T$, this is the case where the source and the target have different metrics or involve different deep learning tasks. The paper categories a myriad of Homogeneous Transfer Learning strategies into: Parameter, Instance, Asymmetric Feature, Symmetric Feature and Relational. The strategy most applicable to medical image analysis is Instance Transfer Learning. The source and target tasks in this method are the same, but the domains are distinct. The weights are transferred, the model is pre-trained on a sizable labelled dataset from the source domain, and then the model is trained on a smaller labelled dataset from the target domain [4]. An example of this in medical image analysis as aforementioned, is exhibited in a paper by Bar et. al [3],

	Top-1 accuracy (%)	Top-5 accuracy (%)
Baseline	48.13±0.42	64.50±0.65
Flipping	49.73±1.13	67.36±1.38
Rotating	50.80±0.63	69.41±0.48
Cropping	61.95±1.01	79.10±0.80
Color Jittering	49.57±0.53	67.18±0.42
Edge Enhancement	49.29±1.16	66.49±0.84
Fancy PCA	49.41±0.84	67.54±1.01

Table 1: This table depicts Taylor and Nitschke's [25] DA experiment results on the Caltech101 dataset.

in which a ConvNet was trained on non-domain specific knowledge and is tasked with classifying medical images. The experiment discussed in the paper was successful and thus proved technique worthwhile for these cases. The paper by Weiss et. al [19] surveys two methods of Instance Transfer Learning, the first being 2SW-MDA [4] and the second being MsTrAdaBoost [38]. 2SW-MDA is optimized for unlabelled data in the target and MsTrAdaBoost is optimized for targets with limited labelled training datasets and is therefore preferable to apply in supervised learning methods.

4.2 Data Augmentation

DAs are a set of techniques used to supplement deep learning models by enhancing the magnitude and the quality of their training datasets. These can generally be grouped into generic and complex augmentations. Generic augmentations further specialized into geometric and photonumeric augmentations [5].

4.2.1 Generic Data Augmentations. Geometric transformations, also referred to as Traditional Transformations, are the set of DAs that are based on adjusting data points based on their x and y coordinates. Photonumeric transformation include modifying the colour or pixel content of the input image. This includes flipping, colour space, cropping and random cropping, rotations, translations and noise injections in which images are injected with a matrix of random values derived from a Gaussian distribution. For each image in the training set, a duplicate is made, the duplicate is transformed and added back into the dataset. For a dataset of size N, the resulting augmented dataset is of size $2N$. As shown in Table 1 above, of generic transformations, cropping generally outperforms all other methods.

4.2.2 Smart Data Augmentations. Other effective forms of DA are dependent on deep learning. These include Feature Space Augmentation, Adversarial training, Neural Style Transfer and Meta-learning DAs [5]. Automated Augmentation strategies search for an optimal augmentation policy within a set of a limited number of geometric transformations with various degrees of distortion [5]. Cubuk et. al [6] test a newly proposed automated augmentation technique coined RandAugment, amongst state of the art automated augmentation methods and found that RandAugment outperformed Automated Augmentation [10], Fast Automated Automated Augmentation [34] and Population Based Augmentation [7] on all tested datasets, while having a significantly reduced search space. Most notably RandAugment led to a $1,0-1.3\%$ improvement over base-line augmentation. RandAugment achieves this by its reduced search space, allowing it to be applied to the target, bypassing any proxies. A search space defines the amount of possible augmentations usable. It is also portable enough to be applied to all dataset sizes and across different tasks. These features make it a promising augmentation method to use for medical datasets. Wang and Perez [24] measure the effectiveness of traditional augmentations, GANs and neural augmentation permutations in a controlled experiment. The research concluded that Neural Augmentation without content loss performed significantly well compared to no augmentation and performed second only to traditional augmentation in certain tasks. Neural Style Transfer or Neural Augmentation is an approach that takes random images from the dataset and net maps them through a ConvNet to generate new images [5]. The research also found that traditional augmentations performed almost as well as neural augmentations in the cases where it was not the optimal choice, while at a smaller time expense. The paper suggests that the traditional and neural augmentations paired up could outperformed all observed results. This is an area further research could build on.

5 DISCUSSION

ConvNets are the leading architecture in the field with regards to the accuracy of image classification, often on large datasets. Over the last few years there has been debate whether transformers were better suited for classification tasks and ConvNets whether ConvNets are becoming redundant. Recently published and highly cited papers have presented firm evidence to suggest the supremacy of ConvNets. ConvNeXt, the modified ResNet proves not only the versatility of ConvNets but also is currently regarded to be the leading architecture for all general computer vision tasks. ConvNeXt exhibits likewise scaling behaviour to ViTs in terms of dataset magnitude and classification accuracy. This is important since this is the same attribute of ViTs that lead to its success. ConvNeXt also outperform transformers on datasets with higher resolution images. This is a preferred attribute for medical image classification tasks as medical images are often in high resolution. Possible cons of ConvNeXt are that it does not consider smaller dataset sizes in its design and thus does not have much build in regularization in mind or account for any inductive biases. ConvNeXt's performance is a result of large datasets which do not exist in medical image analysis. While ConvNeXt may outperform all other models while tested on large datasets, it may not be optimal for small and unbalanced medical datasets. This provides an opportunity for growth with the model, modifying it to scale up and down in terms of training set magnitude.

SwinTransformers for a sought to close the divide between language and computer vision models, were successful in doing so, revolutionized how researchers approached the network design of deep learning models, and for a brief moment, were the leading architecture of choice for all computer vision tasks. An early pitfall of the SwinTransformer v1 was its quadratic time complexity with regards to the size of an image. This was corrected in SwinTransformer, the leading ViT, with a linear time complexity in this regard. Due to the notion of attention in SwinTransformers, it has the ability to compute long range dependencies in images and thus can be applied generally to a wide range of deep tasks, not just those in computer vision. This contributed to its ability to scale impressively well with high resolution images which is key in medical image analysis. In addition, SwinTransformers use self-supervised learning techniques in pre-training to be less reliant on large labelled datasets. Despite this the model still requires large datasets to be effective and this method falls outside the scope of supervised learning. Discretion should be considered with the use of SwinTransformers as the model requires high GPU memory consumption and long training times, although this is less of a problem in medical image analysis as datasets are small. Another concern with SwinTransformers is that the model is intended to be a general purposed architecture for language and vision tasks and thus often fails to optimize the architecture to one dedicated task. This is considered and remedied in the design of ConvNeXts.

The EffNet architecture family addresses long model training times as a concern. By designing small models which correct this issue, EffNets increase the modifiability and testability of deep learning models. EffNet v2 presents faster training speeds and better parameter efficiency than previous models. EffNet v2 also introduces progressive regularization into its design which uses DA in its scaling technique. This improves the architectures training speed and classification accuracy simultaneously. This inherent regularization makes the model well suited for small dataset sizes prevalent in medical image analysis. It is a point to note that while medical imaging datasets can benefit from the DA, the reduction of training speeds is not a large benefit because of the lack of training data, although this may be beneficial in experiments where weights are tuned and the model is re-test often.

ResNets have consistently proven themselves to be a reliable choice for a wide range of computer vision tasks and are often the backbone of many newly proposed models. ResNet-Rs currently performs $1.7x - 2.7x$ faster than EffNets with similar accuracies. The model also has built in DA. It considers small datasets and scales the model depth accordingly which none of the other models discussed do as well. This also remedies the pitfall that deep residual networks face in relying on large training sets. The gradual image scaling also scales well with high resolution images. ResNet-RS shows promising potential for its ability to generalize medical imagery. It scales with training set size as well as image resolution.

DenseNet significantly reduces the number of parameters of a model. ResNet parameters are larger, the model is larger and has a low contribution per layer whereas DenseNets exploit feature reuse which improves the efficiency of the model. The models are therefore easier to training since and consume less energy. DenseNets have an inherent regularization effect which reduces overfitting thus benefiting tasks with smaller datasets. This would aid medical imaging datasets.

Medical imaging datasets are evidently small, are prone to overfitting and are unbalanced. Its typically the case that it is expensive to label or collect data. Sometimes enough data simply does not exist. Domain specific knowledge is needed to properly increase the accuracy of a model to classify medical images, as well as techniques above adjusting the model parameters, such as DA or transfer learning techniques. Papers which have a serious interest in contributing to the field of Pathology and not only improve computer vision tasks, should consider domain knowledge and dataset biases when conducting experiments, designing models and their accompanying methods. Models trained to categorize medical data benefit from two methods of transfer learning the most. Pre-training and fine-tuning. The homogeneous learning strategy most useful to medical imaging is instance transfer learning as it is an image classification task and there are large repositories of non-medical images publicly available. It is also comparatively simple. The method of instance transfer learning used should be MrTrAdaBoost as it is optimized for targets with limited labelled datasets and is therefore preferable to apply in supervised learning methods. The more data a machine learning model has access to, the better it performs. The most popular form of increasing classification accuracy is by augmenting the datasets. This is done independently of the model and is therefore typically used in conjunction with other optimization techniques. Any combination of DA methods can be applied Any combination of these can be applied. Generic or Traditional transformations are simple and very effective. Of the lot, cropping outperforms all other geometric and photonumeric transformations. Smart DAs are expensive in time and are complex but generally outperform generic transformations . There is a lot of potential in this field. Of the smart augmentations RandAugment performs considerably well against baseline augmentations and is also light weight in its class of augmentations because of its low search space and it is also portable enough to be applied to all dataset sizes and across different tasks and is therefore perfect fro medical imaging dataset. neural augment without content loss performs well and is a novel approach, there is potential for neural augmentations to be researched further especially in conjunction with traditional augmentation methods.

6 CONCLUSIONS

Conclusively, ConvNets are the leading deep learning model for image classification. ConvNeXt outperforms all other models in the ConvNet family, including transformers, but these models require large datasets to fully exploit their potential. ConvNets are often designed to distinct problems sets are there is therefore no silver bullet model architecture. ResNets are typically the most versatile models with respect to dataset size and image resolution which exclusively of dataset imbalance. DenseNets which mirror ResNets more but efficiently should therefore the best fit for medical image analysis. Medical imaging benefit greatly from DA and transfer learning optimization techniques. These optimization techniques should also aid in drawing out the some of the full potential of ConvNeXt and SwinTransformers when trained on small datasets. The most preferable transfer learning techniques for medical image analysis are instance pre-training and fine-tuning techniques, most specifically MrTrAdaBoost when working with labelled datasets. Traditional DA methods are simple, effective and considerably effective in increasing the magnitude of a dataset. Cropping proves to outperform other generic transformations. Smart augmentation methods have high potential and generally outperform traditional augmentations. Auto augmentation methods, specifically RandAugment is extremely portable and effective for a myriad of datasets. Neural augmentation methods are complex and effective by small amount but there is promising opportunity for research in using neural augmentation methods in conjunction with traditional augmentation methods.

REFERENCES

[1] Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Neil Houlsby Alexey Dosovitskiy, Lucas Beyer. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2020). https://doi.org/10.48550/arXiv.2010.11929 arXiv:arXiv:2010.11929v2

[2] Md. Zahangir Alom, Tarek Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Nasrin, Mahmudul Hasan, Brian Van Essen, Abdul Awwal, and Vijayan Asari. 2019. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics* 8 (03 2019), 292. https://doi.org/10.3390/electronics8030292

[3] Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan. 2015. Chest pathology detection using deep learning with non-medical training. *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* (2015), 294–297.

[4] Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. 2012. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data* 6, 4 (Dec. 2012). https://doi.org/10.1145/2382577.2382582

[5] Taghi M.Khoshgoftaa Connor Shorten. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 60 (April 2019), 1–48. https://doi.org/10.1186/s40537-019-0197-0

[6] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. 2020. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18613–18624. https://proceedings.neurips.cc/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf

[7] Ion Stoica Pieter Abbeel Xi Chen Daniel Ho, Eric Liang. 2019. Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules. (2019). https://doi.org/10.48550/arXiv.1905.05393 arXiv:arXiv:1905.05393

[8] Jimmy Ba Diederik P. Kingma. 2014. Adam: A Method for Stochastic Optimization. (2014). https://doi.org/10.48550/arXiv.1412.6980 arXiv:https://arxiv.org/abs/1412.6980v9

[9] Serge Dolgikh. 2021. Analysis and Augmentation of Small Datasets with Unsupervised Machine Learning. (2021). https://doi.org/10.1101/2021.04.21.21254796

[10] Dandelion Mane Vijay Vasudevan Quoc V. Le Ekin D. Cubuk, Barret Zoph. 2018. AutoAugment: Learning Augmentation Policies from Data. (2018). https://doi.org/10.48550/arXiv.1805.09501 arXiv:arXiv:1805.09501

[11] D Erhan, Y Bengio, A Courville, PA Manzagol, P Vincent, and S Bengio. 2010. Why does unsuervised pre-training help deep learning? Journal of Machine Learning Research. (2010).

[12] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.

[13] Laurens van der Maaten Gao Huang, Zhuang Liu. 2018. Densely Connected Convolutional Networks. (2018). https://doi.org/10.48550/arXiv.1608.06993 arXiv:arXiv:1608.06993v5

[14] Zhuang Liu Daniel Sedra Kilian Weinberger Gao Huang, Yu Sun. 2016. Deep Networks with Stochastic Depth. (2016). https://doi.org/10.48550/arXiv.1603.09382 arXiv:arXiv:1603.09382

[15] Veronika Cheplygina Gaël Varoquaux. 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ*, Article 48 (April 2022). https://doi.org/10.1038/s41746-022-00592-y

[16] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12. https://doi.org/10.1109/MIS.2009.36

[17] Ross Girshick Kaiming He Piotr Dollar Ilija Radosavovic, Raj Prateek Kosaraju. 2020. Designing Network Design Spaces. (2020). https://doi.org/10.48550/arXiv.2003.13678 arXiv:arXiv:2003.13678v1

[18] Xianzhi Du Ekin D. Cubuk Aravind Srinivas Tsung-Yi Lin Jonathon Shlens Barret Zoph Irwan Bello, William Fedus. 2021. Revisiting ResNets: Improved Training and Scaling Strategies. (2021). https://doi.org/10.48550/arXiv.2103.07579 arXiv:arXiv:2103.07579v1

[19] DingDing Wang Karl Weiss, Taghi M. Khoshgoftaar. [n. d.]. A survey of transfer learning. *Journal of Big Data* 3, 9 (April [n. d.]), 1–40. https://doi.org/10.1186/s40537-016-0043-6

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[21] Ross Girshick Kaiming He Piotr Dollar lija Radosavovic, Raj Prateek Kosaraju. 2020. Designing Network Design Spaces. (2020). https://doi.org/10.48550/arXiv.2003.13678 arXiv:arXiv:2003.13678v1

[22] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42 (2017), 60–88. https://doi.org/10.1016/j.media.2017.07.005

[23] Anil Yuce1 Samaneh Abbasi-Sureshjani Simon Schönenberger Paolo Ocampo Konstanty Korski Fabien Gaire Luca Deininger, Bernhard Stimpel. 2022. A comparative study between vision transformers and CNNs in digital pathology. (2022). arXiv:https://arxiv.org/pdf/2206.00389.pdf

[24] Jason Wang Luis Perez. 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. (2017). https://doi.org/10.48550/arXiv.1712.04621 arXiv:arXiv:1712.04621v1

[25] Geoff Nitschke Luke Taylor. 2018. Improving Deep Learning with Generic Data Augmentation. In *IEEE Symposium Symposium Series on Computational Intelligence* (Bengaluru, India) *(SSCI 2018)*. IEEE, Cape Town, CT, SA, 1542–1546. https://doi.org/10.48550/arXiv.1708.06020

[26] Quoc V. Le Mingxing Tan. 2021. EfficientNetV2: Smaller Models and Faster Training. (2021). https://doi.org/10.48550/arXiv.2104.00298 arXiv:arXiv:2104.00298v3

[27] Srivastava Nitish, Hiton Geoffrey, Krizhevsky Alex, Sutskever Ilya, Salakhutdinov Ruslan, et al. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15, 1 (2014), 1929–1958.

[28] Mitchell Stern Noam Shazeer. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. (2018). https://doi.org/10.48550/arXiv.1804.04235 arXiv:arXiv:1804.04235v1

[29] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot Learning with Semantic Output Codes. In *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), Vol. 22. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf

[30] Elliot Meyerson Aditya Rawal-Dan Fink Olivier Francon Bala Raju Hormoz Shahrzad Arshak Navruzyan Nigel Duffy Babak Hodjat Risto Miikkulainen, Jason Liang. 2017. Evolving Deep Neural Networks. (2017). https://doi.org/10.48550/arXiv.1703.00548 arXiv:arXiv:1703.00548v2

[31] Christian Szegedy Sergey Ioffe. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015). https://doi.org/10.48550/arXiv.1502.03167 arXiv:arXiv:1502.03167

[32] Liam Toledo Geoff Nitschke Shane Acton, Sasha Ambramowitz. 2020. Efficiently Coevolving Deep Neural Networks and Data Augmentations. In *IEEE Symposium Symposium Series on Computational Intelligence* (Canberra, ACT, Australia). IEEE, Cape Town, CT, SA, 8 pages. https://doi.org/10.1109/SSCI47803.2020.9308151

[33] Rajeev Kumar Singh, Rohan Pandey, and Rishie Nandhan Babu. 2021. COVID-Screen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays. *Neural Computing and Applications* 33, 14 (2021), 8871–8892. https://doi.org/10.1007/s00521-020-05636-6

[34] Taesup Kim Chiheon Kim-Sungwoong Kim Sungbin Lim, Ildoo Kim. 2019. Fast AutoAugment. (2019). https://doi.org/10.48550/arXiv.1905.00397 arXiv:arXiv:1905.00397

[35] Adnan Qayyum Muhammad Awais Majdi Alnowami Muhammad Khurram Khan Syed Muhammad Anwar, Muhammad Majid. 2014. Medical Image Analysis using Convolutional Neural Networks: A Review. *J Med Syst* 42 (2014), 226–239.

[36] Raphael Tang, Ashutosh Adhikari, and Jimmy Lin. 2018. Flops as a direct optimization objective for learning sparse neural networks. *arXiv preprint arXiv:1811.03060* (2018).

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] Yi Yao and Gianfranco Doretto. 2010. Boosting for transfer learning with multiple sources. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), 1855–1862.

[39] Xue Ying. [n. d.]. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series* 1168, 2 (Feb. [n. d.]), 7 pages. https://doi.org/10.1088/1742-6596/1168/2/022022

[40] Yue Cao Han Hu Yixuan Wei Zheng Zhang Stephen Lin Baining Guo Ze Liu, Yutong Lin. 202q. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. (202q). https://doi.org/10.48550/arXiv.2103.14030 arXiv:arXiv:2103.14030v2

[41] Yutong Lin Zhuliang Yao Zhenda Xie Yixuan Wei Jia Ning Yue Cao Zheng Zhang Li Dong Furu Wei Baining Guo Ze Liu, Han Hu. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. (2022). https://doi.org/10.48550/arXiv.2111.09883 arXiv:arXiv:2111.09883v2

[42] Christoph Feichtenhofer1 Trevor Darrell2 Saining Xie1 Zhuang Liu1, Hanzi Mao1 Chao-Yuan Wu1. 2022. A ConvNet for the 2020s. (2022). https://doi.org/10.48550/arXiv.2207.03620 arXiv:arXiv:2201.03545v2