Deep Learning, Small Data Pathology Classifiers

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

Tomas Slaven
SLVTOM001@myuct.ac.za
University of Cape Town
Cape Town, South Africa

Shaylin Chetty
CHTSHA042@myuct.ac.za
University of Cape Town
Cape Town, South Africa

1 INTRODUCTION

This project's major purpose is to analyze and compare various predictive machine learning techniques, including Supervised, Self-Supervised, and Auto-Machine Learning approaches on their effectiveness in Medical Image Analysis tasks. The initiative intends to accomplish this by combining orthopaedic X-ray and CT scan datasets. The datasets will be used to determine which models in each Deep Learning (DL) technique delivers the highest degree of classification success. The project will begin with the collection and pre-processing of a number of datasets. The data will be cleaned and processed during the pre-processing step to ensure that it is ready for analysis. Following this, the datasets will be divided into training and testing sets, with the training set subjected to a variety of machine-learning approaches. Unsupervised learning, supervised learning, and auto machine learning approaches will be evaluated. Data augmentation techniques will also be used to improve the dataset and the performance of the machine learning models. The machine learning models will be tested on the testing set after they have been trained. The outcomes will be analyzed, and the most effective method will be made explicit to further evaluate its advantages and disadvantages, and offer suggestions for enhancing its performance. The most robust method that can be used across multiple datasets will be determined in the final evaluation. The project's overall goal is to find the most effective method for predicting orthopaedic conditions using machine learning techniques. The overall aim of this project is therefore to aid medical practitioners in deciding which DL methods to consider for computer vision tasks in the field of Pathology.

The analysis of pathological data is tedious, expensive and error-prone, often relying on the experience of domain experts. However, even experts are subject to classification inaccuracies on account of human error. Like most domains, the application of Machine Learning technologies to solve these problems has led to a lot of excitement - machine learning has the ability to speed up data processing times, unlock deeper insights and make decisions in a fraction of the time it would take a specialist to do so, supplementing their work. While there are opportunities for machine learning technologies to be applied, pathology data is still plagued by one major issue: data availability.

The availability and diversity of medical data are leading factors preventing the wide-scale uptake of DL technologies in pathology [4, 21, 45]. Current datasets are small and unbalanced and, by extension, not representative of the true population [9]. Various techniques have been proposed to overcome small datasets. An obvious approach would be to incorporate additional subjects into datasets but this is not feasible: the time and cost constraints of curating and labelling data, combined with the patient confidentiality issues that arise, make this method impractical. This results in some data being left unlabeled, which could be used in self-supervised learning methods such as MoCo [5, 15], as later discussed in this paper. Data Augmentation methods have also been proposed. Data Augmentation is the process of enhancing the size of datasets [9, 48] by performing geometric [21, 38] or photometric transformations on data [24, 38]. This helps improve the accuracy of models by creating a larger set of training data that is diverse in the conditions of images (for example, rotated images can be accounted for). However, these methods don't account for stochastic variations in images nor do they explore the true population meaning the enhanced datasets are still imbalanced and non-representative. Lastly, methods have been proposed, most notably in Convolutional Neural Networks (CNN) and Vision Transformers (ViT), to transfer knowledge from other systems to the current system, but this technique does not translate well to medical datasets. Clearly, an alternative approach is needed to address the problem of small datasets in medical domains.

This research concerns working around the data availability issue through three different viewpoints, namely using DL models on existing datasets, using Generative Adversarial Networks (GANs) to artificially increase the size of datasets which will be applied to a self-supervised learning classifier and investigating the use of Neural Architecture Searching (NAS) algorithms.

1.1 Related Work

1.1.1 Enhancing Small Datasets in Computer Vision with GANs. The prominent challenge in computer vision tasks is a result of limited labeled data, particularly in niche application domains. Capturing custom data is an option, but it poses data cleaning and capturing challenges, and using small datasets may lead to overfitting [22]. Data augmentation techniques improve models, but they only make models invariant to certain conditions and do not enhance data diversity [18]. To address this, the development of GANs is introduced. Using GANs for synthetic image generation allows for the creation of entirely new images that confidently follow the same distribution as the original data [12]. This enhances the size and quality of the training dataset, leading to the development of more robust models.

GANs are a powerful tool for synthetic image generation that can address imbalances in skewed data and add to the size and diversity of small datasets. They consist of two neural networks, the Generator and the Discriminator, that compete in an adversarial game [12]. Vanilla GAN has limitations, and improved versions such as WGAN and StyleGAN2 address these limitations [40] [41]. However, when introduced to medical applications, GANs face obstacles such as unbalanced and less annotated datasets, high computational costs, and training complexity.

Overall, GANs provide powerful techniques that improve the quality of small datasets, particularly in data-constrained environments such as medical imagery. While Vanilla GANs provide a good

basis for generative models, improvements made by WGAN-GP and StyleGAN2 make them better suited for certain applications. GANs can also be used for discriminatory tasks by using the Discriminator. However, further research is needed to develop medicine-inclined evaluation metrics.

1.1.2 Supervised learning and Data Augmentation for medical datasets. There exist a multitude of various DL models for image classification tasks, however, their applications to medical imaging datasets are not extensively explored. We proceed to discuss some of the popular and strong performing supervised learning models, while also presenting the issues brought upon by utilizing small datasets and strategies for overcoming them.

Convolutional Neural Networks (CNNs) are currently regarded as the leading architecture for computer vision tasks, and the modified ResNet [17], ConvNeXt [49], is considered to be the best for general computer vision tasks. SwinTransformers [47] are successful in bridging language and computer vision models and have the ability to compute long-range dependencies in images. The EffNet [32] architecture family focuses on addressing long model training times, while DenseNets [11] reduce the number of parameters in a model. These models constitute a variety of strengths and weaknesses, and while there is no one-size-fits-all model architecture for CNNs, we do observe ResNets being the most versatile for dataset size and image resolution, and DenseNets being best for medical image analysis.

Data augmentation (DA) and transfer learning techniques are effective in improving the performance of CNNs in medical image analysis [37]. This is because medical datasets are typically small, unbalanced, and prone to overfitting, making domain-specific knowledge and transfer learning techniques crucial for improving classification accuracy. Pre-training and fine-tuning are the most useful transfer learning strategies for medical imaging datasets, and traditional DA methods such as cropping are effective[20, 30]. Smart augmentation methods, such as auto-augmentation, have high potential for a variety of datasets, while neural augmentation methods are complex and effective but require further research [6, 10].

1.1.3 Automated Machine Learning as a tool for Medical Image Classification. As mentioned, the challenges of medical image classification using current machine learning models centres around the unavailability of large labelled data sets. Data augmentation is proposed as a solution to increase data sets, but it is not sufficient by itself [39]. CNNs have been the predominant model in computer vision for decades until the recent emergence of ViTs, which has put this free-standing position to the test. However, issues with ViTs, when compared to CNNs, do include: the need for quadratic computational complexity according to token size [31]; ViTs typically generalize worse than CNNs when trained on small data sets [26]; and ViTs has not yet been well optimized and integrated into hardware platforms in industry [35]. For the goal of medical image classification, CNNs then appear to be our model of choice.

The concept of automated machine learning (AutoML) can be used to automate the entire process of creating an ML model. AutoML, currently, does not present a feasible option due to its complexity and number of stages, but Neural Architecture Search (NAS), a sub-topic of AutoML that automates the architecture design process of ML models, does provide a feasible solution. NAT-M4 [27] currently presents the NAS method with the highest potential for medical image classification on sparse data sets. However, this is yet to be empirically tested and other promising NAS methods such as DeepMAD [35], ShapleyNAS [44], and ZenNAS [25] should still be considered.

In conclusion, data augmentation and CNNs present viable solutions for medical image classification with limited labelled data sets. Furthermore, automated machine learning and NAS, specifically NAT-M4, present a feasible option for improving image classification models through automation. However, more research is needed in developing specialized NAS methods for medical image classification and creating a widely used benchmark for medical image classification.

1.2 Research Objectives

In order to evaluate various predictive models, the research will be split into three core research components with associated research objectives namely supervised, self-supervised and auto machine learning methods.

(1) To compare and evaluate the performance of various state-of-the-art (SOTA) DL models on image recognition tasks on medical datasets. Therefore to ascertain the most efficient model for medical image recognition tasks.

This research objective aims to provide valuable insights into the performance of several SOTA image prediction supervised learning models on medical datasets. This information will assist researchers and practitioners in selecting the most suitable model for their specific needs. We will test a combination of CNN and ViT supervised learning models with two experimental variables: data augmentation and transfer learning. Additionally, we will test models with varying architecture sizes to measure the effect of model size on the quantitative evaluation metrics made explicit in Section 3. A more detailed experiment design on the supervised learning component is described in Section 2.1.

(2) To determine whether GANs improve the performance of self-supervised classification models on pathology data.

To overcome the aforementioned small data issues that are present in pathology data, we want to test the effectiveness of generative models in improving these datasets. Since Generative Adversarial Networks (GANs) produce unlabeled data, each model's output will be tested on the self-supervised classification task. In total, 3 GANs will be tested, with each GAN implementing a different loss function and architecture as explained in section 2.2. Besides testing the impact of GANs on downstream classification tasks, we explicitly evaluate the images created by each GAN through qualitative (manual) and quantitative evaluation as detailed in Section 3.

(3) To investigate the effectiveness of SOTA NAS methods in creating CNNs for binary, medical image classification tasks. Therefore, to identify the best NAS method for medical image recognition tasks as well as evaluate the use of NAS methods in this context.

The nature of deep learning (multiple layers in a Neural Network) has opened the doors to a vast search space when designing a CNN and promoted the development of better-performing models over time. The vastness of this search space indicates an almost infinite scope for improvement, with the main restriction being the actual design process of a CNN. Traditional, manual CNN design methods are non-trivial and laborious, resulting in a failure to exploit the vast search space of Neural Network design. Neural Architecture Search supplies an automated method to generate architectures for a given task. This provides us with the ability to exploit a much larger degree of the search space of CNN design and prompt the development of better performing ML models as a consequence. This will be tested by comparing and evaluating the ability of four SOTA NAS methods such as NAT-M4 [27], DeepMAD [35], ShapleyNAS [44], and ZenNAS [25] to redesign a ResNet based CNN architecture for medical image classification. Further detail is provided in Section 2.3.

(4) To evaluate the advantages and disadvantages of the observed optimal solution and offer suggestions for its performance.

The objective stated above is secondary to all three research components: supervised, self-supervised, and auto-machine learning. Its aim is to offer medical practitioners additional information to consider when deciding which method to use in practice.

2 PROCEDURES AND METHODS

Generic procedures for all research components undergo dataset procurement from a medical professional at the Groote Schuur Hospital, and the prepossessing of the medical datasets for image prediction tasks. The justification for the choice of the supervised, self-supervised and auto machine learning methods are described in Section 1.1 Related Work.

2.1 Supervised Learning

Provisionally, the supervised learning experiment will test the following three models in three sizes of varying architecture depth in order to measure the effect of model sizes on performance.

2.1.1 ConvNeXt. ConvNeXt [49] is a modified ResNet-50 CNN which borrows the hierarchical design from the design aspect of Vision Transformers (ViTs). ConvNeXt boasts the most competitive performance results across computer vision tasks and exhibits impressive scaling behaviours with regard to both dataset size and image resolution. Despite this, the ConvNeXt Model requires large amounts of data to perform optimally and operates at a high Floating Point Operation (FLOP) average which is an indicator of high energy usage. This experiment will test ConvNeXt-T with a size of 29 million parameters, ConvNeXt-S with a size of 50 million parameters, and CovNeXt-B with a size of 89 million parameters, all model settings as proposed by Liu et al. [49].

2.1.2 SwinNet. SwinV2 [47] is a hierarchical ViT which caused an unprecedented paradigm shift in computer vision, virtue to the publication of seminal work by [46] Liu et. al in the early 2020's renaissance of supervised learning, shifting research interest from CNNs to ViTs. The efficacy of SwinNet derives from its ability to compute long-range dependencies within data, using a technique coined attention. SwinV2 computes shifted windows which remedies the most prominent issue in precious implementations of ViTs which would have a quadratic time complexity with regard to image size. However, SwinV2 requires high Graphical Processing Unit (GPU) memory consumption as well as long training times. This experiment proposes testing SwinV2-T with a magnitude of 49 million parameters, SwinV2-S with a magnitude of 28 million parameters, and SwinV2-B with a magnitude of 88 million parameters, all model configurations as by Liu et. al [47].

2.1.3 DenseNet. DenseNets [11] are a CNN solution in which each layer is connected to every other layer in a feed-forward fashion. This network design choice therefore exploits feature reuse by having additive identity transformations per layer. This significantly decreases the number of parameters in the model. DenseNets are thus simpler to train since layers may reuse gradients from the loss function and the input text. DenseNets also possess an intrinsic regularization effect that mitigates overfitting, which is particularly beneficial for tasks with limited training datasets such as those typical in Medical Image Analysis. This experiment will be subject to training DenseNet-121 with approximately 8 million parameters, DenseNet-169 with approximately 14 million parameters, and DenseNet-201 with approximately 20 million parameters, all model settings as put forward by Huang et al. [11].

2.1.4 Other models. Under favourable circumstances and if time permits, the experiment will expand to further include Le and Tan's EffNetv2 CNN in its small, medium and large architectures [32] and ResNet18, ResNet50 and ResNet152 CNNs proposed by Le et. al [19]. The models will be set up as per their respective papers.

2.1.5 Augmentations and Transfer Learning. To test the effectiveness on DL augmentations and transfer learning on the performance on models trained on medical datasets, the models will be tested under a combination of states dependent on two experimental variables. Each model will have its performance evaluated and compared without DA, with cropping as implemented by Taylor and Nitschke [30], Neural Augment used by Wang and Perez [29], and RandAugment as by Cubuk et. al [7]. Likewise, under the Transfer Learning states, without pre-trained weights used, with pre-trained weights used, and with pre-trained weights used in addition to AdaBoost an instance-based transfer learning technique described in [20]. We find the that models with precisely labelled data most benefit from boosting as a supplementary transfer learning technique.

2.1.6 Evaluation Process. The models will be sourced from Pytorch libraries. The DA and boosting will used the frameworks made available by the Torch Vision libraries. The datasets will be split into a 70-15-15 percentage split for training, validation and testing respectively. Learning algorithms have been applied to the dataset and the output obtained will be compared with the actual output in order to calculate F-1 score, precision, recall, accuracy and PR AUC, more information on these is in Section 3. Evaluation Metrics.

2.2 Generative Models and Self-Supervised Learning

A review of medical literature points to three variants of GANs that show the most promise on medical datasets, namely Vanilla GAN, WGAN-GP and StyleGAN2. Each of these will be investigated independently of the other with the generated results subject to the same evaluation criteria.

2.2.1 Vanilla GAN. Vanilla GAN, as originally proposed by Goodfellow et al [13], is the simplest class of GANs and, as such, will be used as the baseline model in our experiments. The GAN consists of two multi-layer perceptions that are trained simultaneously in an adversarial game in order to produce images that are similar to the original data. The first multi-layer perceptron is known as the Generator and is responsible for approximating the underlying distribution of the training images and generating images based on this distribution while the second multi-layer perceptron, the Discriminator, will classify the data as either real or synthetic. The implementation will follow the original implementation.

2.2.2 WGAN-GP. Vanilla GAN implements a Jensen-Shannon Divergence (a sigmoid-based curve) as a loss function which has been shown to lead to slow training times or result in the divergence of the approximated and actual distributions, both of which impact the stability of the model. WGAN-GP [3, 14] aims to reduce instability by using the Wasserstein distance as a loss function with gradient penalties to ensure the model converges. We will implement the model according to the specifications proposed by Gulrahani et al. [14]

2.2.3 StyleGAN2. In domains where small variations in images could affect its classification, generative models need to be able to create these variations confidently. StyleGAN2 [23] accounts for stochastic variation by replacing the multi-layer perceptron with a CNN and introducing noise at each convolution. This property makes it desirable in applications where images have variant aspects where the smaller variations need to be captured. The implementation will follow the model specifications proposed by Karras et al. [23]

2.2.4 MoCo v3. We employ a contrastive-based self-supervised classification framework in the form of MoCo v3 - a simpler, more accurate and stable version of MoCo [15] as proposed by Chen et al [5]. Version 3 gets rid of the memory queue used by MoCo and implements a symmetrized loss function. The model follows that of Chen et al. [5] with a ResNet-50 backbone, LARS optimizer, batch size of 4096, a learning rate of 0.3, a weight decay of 1.5e-6 and the temperature parameter set to 1. The momentum coefficient used to train the key autoencoder is set to 0.996.

2.2.5 Test Data. The dataset will be provided by Prof. Nitschke and will be used without any preprocessing.

2.2.6 Experiment Design. Each model will be created according to their original specification as published by the original authors, without any hyperparameter optimization. Compared to Vanilla GAN, two variables will be changed: WGAN-GP changes the loss function while Style-GAN2 changes how noise is applied to images (while maintaining the WGAN-GP loss function). Each model will be run independently from the others on the same computer architecture. The images generated by each GAN will be subject to 4 evaluation metrics, namely visual Turing tests, FID, GAN-train and GAN-test. This will determine which GAN produces the best quality and diversity of images.

To measure the effect that each GAN has on classification, 4 classifiers will be trained. Each GAN will create a dataset consisting of original and generated data which will then be used to train their respective classifier and there would be one classifier trained using the original (unaugmented) dataset. Each of the classifiers will then be tested on the original dataset. This will help us determine the improvement (or reduction) in classifier accuracy caused by each GAN. The GAN that results in the largest improvement in accuracy would be the most suitable for pathology tasks.

2.3 Automated Machine Learning

A review of literature points to four NAS methods that show the most promise on medical datasets, namely NAT-M4, DeepMAD, ShapleyNAS, and ZenNAS. Each of these will be investigated independently of the other with the generated results subject to the same evaluation criteria.

2.3.1 NAT-M4. The basis of NAT-M4 is Neural Architecture Transfer (NAT). NAT automates the discovery of optimal architectures and their associated weights for image classification tasks. NAT uses multiple subnets, creating a supernet, that train simultaneously via weight sharing. This is accomplished through an alternating procedure be- tween supernet adaptation and evolutionary search. This discovers custom neural networks optimized for conflicting objectives. This method is much more efficient than running Neural Architecture Search (NAS) from scratch for each new task, as NAT is able to obtain multiple custom neural networks that cover the entire range of objectives. The resulting supernet then spans the trade-off front of the objectives and can be used for all future deployment-specific NAS without any additional training [27].

2.3.2 DeepMAD. Mathematical Architecture Design for Deep CNNs (DeepMAD) describes a procedure for designing high-performance convolutional neural network models in a systematic manner. DeepMAD is based on recent advancements in DL theories and utilizes a constrained Mathematical Programming (MP) problem to optimize the architecture of CNN models. This approach yields optimized structural parameters, including network width and depth. The MP problem in DeepMAD has a low dimension of a few dozen, which means it is easily solvable on CPUs without the need for customized MP solvers. Therefore, the need for a GPU or creating a deep model in memory is voided. This makes DeepMAD extremely fast, even on servers with limited memory. Once the MP problem is solved, the optimized CNN architecture can be derived from the solution [35].

2.3.3 Shapley-NAS. Shapley-NAS introduces a novel approach for neural architecture search (NAS). It is built upon the findings of DARTS and aims to overcome the issues associated with DARTS. Shapley-NAS evaluates the contribution of individual components within a supernet to the validation accuracy of a neural network by utilizing the Shapley value. Unlike previous methods that consider

only the magnitude of architecture parameters, Shapley-NAS also takes into account their practical influences on task performance. The Shapley value is an effective tool for handling complex relationships between individual elements by quantifying the average marginal contribution of all possible combinations of operations. Monte Carlo sampling is used to approximate the Shapley value efficiently and employs a momentum update mechanism to reduce fluctuations caused by the sampling process. Shapley-NAS outperforms previous methods, such as DARTS, by showing a higher correlation with task performance. The proposed method achieves superior results on different data sets and search spaces, including a 2.43% error rate on CIFAR-10 and a top-1 accuracy of 23.9% on ImageNet under the mobile setting [44].

2.3.4 ZenNAS. The design of high-performance deep neural networks is challenging and Neural Architecture Search (NAS) methods can help. NAS algorithms consist of two key components: an architecture generator and an accuracy predictor. However, building a high-quality accuracy predictor is computationally expensive. To address this, a new NAS algorithm called Zen-NAS is proposed which uses a cost-effective proxy called Zen-Score to measure the expressivity of a deep neural network. Zen-NAS is a zero-shot method that does not optimize network parameters during search and achieves SOTA performance on various datasets. This approach is inspired by recent studies that show deep models are more expressive and can reduce bias error [25].

2.3.5 Experiment Design. For the research objective regarding the investigation of different NAS methods, we will be conducting experiments to explore their impact on optimizing the architecture of a ResNet-based CNN. Each method will be presented with a ResNet CNN foundation. We will then split our dataset into an 80:20 ratio of training and test data. The training data split will then be used by the different NAS methods to optimize each of their ResNet CNN foundations via each method's training scheme. Upon completion of training and similarly having discovered an optimal architecture, each method's discovered architecture will be tested on the completely unseen test split of our dataset. These architectures will be evaluated based on their performance on the test dataset according to various metrics, such as accuracy, F1-Score, search time, number of parameters, and computational cost (FLOPs).

3 EVALUATION METRICS

3.1 Quantitative Evaluation

Classification accuracy is often used to evaluate network architectures against standard benchmarks, whereas medical image system analyses rely on more practical metrics such as F-1 score, precision, and recall. These metrics are computed using the following formulas:

$$F1score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{1}$$

where Precision and Recall are defined as:

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

And Accuracy is calculated as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

Where TP represents true positives, TN represents true negatives FP represents false positives, and FN represents false negatives [28, 37]. 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.

As is the standard in Generative Adversarial Network (GAN) literature, FID scores [16] will be used. Although based on InceptionNet, and thus ImageNet, Woodland et al. [42] observed a negative correlation between human perception and FID scores on medical data meaning this metric holds valid on medical datasets. However, we need another metric whose validity is not based on observation. We propose using a metric that is not dependent on another image database but is rather dependent on the original set of images. This would make such a metric attractive in niche domains. We use an approach closely related to that proposed by Shmelkov et al. [36]: use classifiers to measure the recall and precision of generated images. This gives us two metrics: GAN-train and GAN-test. Although originally used for Conditional GANs [33, 36], and less frequently used than FID and Inception scores, it does show promise for other GAN variants and, recently, has been applied to medical applications [1, 2, 43]. GAN-train measures recall (that is, how diverse is the generated data). It does this by training a CNN classifier on generated data and testing it with a validation set of real images - a high classification rate shows that the generated data captured the diversity of data. GAN-test measures precision (that is, the quality of images). The CNN classifier is trained on real data and evaluated on generated data. A high accuracy rate means that generated samples follow a similar distribution to the original data. To benchmark the performance, a CNN classifier will be trained and tested on real data.

3.2 Qualitative Evaluation

Generated images will be subject to qualitative evaluation through visual Turing tests approach[8, 34, 42]. A set of images containing equal numbers of real and synthetic images will be presented to a pathology specialist who will rate the fidelity of each image on a scale of 1-10, where 10 is for the best quality images. This exercise will be done for all GANs. The average fidelity of real images will act as an upper bound on the fidelity of generated images to detect overfitting. We will report the average fidelity across all GANs. Simultaneously, we will whether an image is being perceived as real or fake based on the fidelity score (score $< 5 \rightarrow$ fake, score $\geq 5 \rightarrow$ real) assigned by the specialist. This will determine the accuracy of the GAN where a lower accuracy means that real and fake images are hard to discern indicating a better-performing GAN. Caution is needed, however, as these metrics are intrinsically subjective. Hence, quantitative evaluation measures are needed.

3.3 Implementation Tools

The following tools are crucial for the successful implementation of the experiment design and will be utilized extensively throughout the duration of the project.

(1) Pytorch [1]
(2) Pytorch-Lightning for Boosting Algorithm [2]
(3) Torch Vision for DA [3]
(4) ImageNet1K-V1 for TL [4]
(5) Centre for High Performance Computing (CHPC) cluster [5]

4 ETHICAL, PROFESSIONAL AND LEGAL ISSUES

Data provided by the Department has been anonymized and contains no sensitive or personally identifiable information. The data has received ethics clearance from Dr Kruger of Groote Schuur Hospital. All models and the respective code base will remain in property of the University of Cape Town.

5 ANTICIPATED OUTCOMES

5.1 Contributions

The research this proposal outlines aims to make the following contributions to the body of knowledge in the science of DL in Pathology. General contributions include the trained models with the updated weights.

5.1.1 Supervised Learning.

- The primary contribution this section of the project makes is to compare and present the effectiveness of different SOTA supervised learning models on medical datasets.
- To evaluate and report the association between prediction accuracy and architecture size for small pathology datasets.
- Furthermore the project aims to present comparisons of the most affluent DA techniques in the field of Pathology and their effectiveness in augmenting medical imaging datasets.
- In addition to the methods, the marginal effects of TL and subsequently boosting are set to be presented.

5.1.2 Generative Adversarial Networks and Self-supervised Learning.

- We aim to experimentally determine the most suitable GAN for medical imagery synthesis and provide an evaluation on the cost (complexity) vs model improvement of GANs to determine whether the added benefit in downstream classification tasks is justified.
- We aim to provide an analysis of the strengths and weaknesses of the various GANs studies and provide possible recommendations for the improvement of these models.
- The data provided by the Department has been anonymised and contains so sensitive patient data. All data used has been given ethical clearance.

5.1.3 Auto-Machine Learning.

- Will provide empirical testing and results on whether NAS methods can effectively optimize and improve ResNet CNN architectures for medical image classification on small datasets.
- Will empirically compare which SOTA NAS method is best for optimizing ResNet CNN.
- Provide insight into the suitable architecture configurations for well-performing ML models on small datasets.
- Identify the strengths and weaknesses of implementing NAS methods to improve ResNet CNN classification on small, medical datasets

5.2 Impact

5.2.1 Supervised Learning. Overall this research will therefore provide insights as to the most effective DL model for image prediction tasks, the best way to increase its performance, and aid researchers and practitioners in choosing the most optimal solution. for their needs. Conclusively this section aims to add to academic discussion and to form a basis for further research in the use of supervised learning methods in medical image analysis.

5.2.2 Generative Adversarial Networks and Self-supervised Learning. The proposed project aims to prove the effectiveness of generative adversarial networks in creating synthetic data that confidently fit the underlying distribution of original data on medical datasets. Additionally, we aim to encourage further research into two main areas: firstly, GAN evaluation metrics for medical data need to be established either through the creation of new metrics or by proving that existing metrics hold. Secondly, research into configuring GANs such that it is suitable for medical data.

5.2.3 Auto-Machine Learning. The proposed project aims to improve the effectiveness of CNNs when it comes to binary image classification on small, medical datasets. Additionally, we aim to provide empirical testing of the performance of NAS-optimized CNN models on small-sized datasets, which has not been widely investigated in the field of Neural Architecture Search.

5.3 Success Factors

5.3.1 Supervised Learning.

- Successfully trained models with updated weights on the medical dataset.
- Clear evidence of DA techniques increasing the performance of models.
- Clear evidence of Transfer Learning techniques increasing the of performance of models.
- Results of the evaluation metrics point to a decisive model conclusively optimal for medical datasets where DA and Transfer Learning techniques corroborate methods of increasing the models performance.
- Optimal model with a prediction accuracy above 90 percent and therefore usable in the problem domain.

5.3.2 Generative Adversarial Networks and Self-supervised Learning.

- Successfully incorporating the generated unlabeled images with a self-supervised classification model to increase the accuracy of the model.

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

- The stable generation of realistic images that accurately captures the underlying distribution of original data to increase the training size
- The balancing of medical datasets using GANs

5.3.3 Automated Machine Learning.

- Successful training and implementation of the various NAS methods.
- Production of sufficient experimental results that can be used to draw valid conclusions on NAS for improving CNNs in medical image classification.
- Optimized models provide an increase in accuracy when compared to standard ResNet models.

6 PROJECT PLAN

This section of the project proposal outlines the project management plan. The plan includes an assessment of Risks, Timelines, Resources, Deliverables, Milestones, and Work Allocation.

6.1 Risks

The projected risks identified in this project are made explicit in Appendix A which outlines the risk, the probability of observing the risk, the impact of the risk, and the risk's associated mitigation strategy.

6.2 Timeline

This project starts on April 25th with the submission of the project proposal and ends with the IT Showcase on October 24th. The project goes through the following concurrent stages:

(1) Project Proposal
(2) Data Collection
(3) Pre-processing
(4) Core Implementation
(5) Training
(6) Evaluation
(7) Research Paper
(8) Demonstrations

A detailed Gantt chart is displayed in Appendix B which depicts project phases, tasks, milestones and their associated dates.

6.3 Resources

- Personal computer with an Integrated Development Environment
- UCT High Performance Cluster
- GPUs as provided by the Department of Computer Science
- Chest and Elbow X-Ray and CT scans as provided by Dr. Kruger

6.4 Deliverables

The core research deliverables are summarized below. This excludes smaller deliverables, that constitute core deliverables, as requested by our supervisor. Refer to Appendix B for a timeline of events.

- Final research report
- Literature reviews
- Project proposal
- Software feasibility demonstration
- Final code submission
- Final project demo
- Project website
- Project poster

6.5 Milestones

The aforementioned deliverables are seen as project milestones and summarized in Appendix B.

6.6 Work Allocation

Each team member will be allocated one of the three initial objectives, with objective four shared across members. Each member will be responsible for the development, testing and evaluation of their respective models. Data collection, ethics-related work, presentations and the creation of the upcoming website/poster will be shared between members equally. Each member will be responsible for their own final report.

Objectives are allocated as follows:

- Winner Bryan Kazaka - Supervised learning models for medical image classification (Section 1.2, objective one)
- Shaylin Chetty - Generative Adversarial Networks and Self-Supervised Classifiers (Section 1.2, objective two)
- Tomas Slaven - Neural Architecture Search in Convolutional Neural Networks (Section 1.2, objective three)

REFERENCES

[1] Ibrahim Saad Aly Abdelhalim, Mamdouh Farouk Mohamed, and Yousef Bassyouni Mahdy. 2021. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. *Expert Systems with Applications* 165 (March 2021), 113922. https://doi.org/10.1016/j.eswa.2020.113922

[2] Paolo Andreini, Simone Bonechi, Monica Bianchini, Alessandro Mecocci, and Franco Scarselli. 2020. Image generation by GAN and style transfer for agar plate image segmentation. *Computer Methods and Programs in Biomedicine* 184 (Feb. 2020), 105268. https://doi.org/10.1016/j.cmpb.2019.105268

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. http://arxiv.org/abs/1701.07875 arXiv:1701.07875 [cs, stat].

[4] Lorenzo Brigato and Luca Iocchi. 2021. A Close Look at Deep Learning with Small Data. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 2490–2497. https://doi.org/10.1109/ICPR48806.2021.9412492 ISSN: 1051-4651.

[5] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. http://arxiv.org/abs/2104.02057 arXiv:2104.02057 [cs].

[6] 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

[7] 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

[8] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. http://arxiv.org/abs/1506.05751 arXiv:1506.05751 [cs].

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

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

[11] 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

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. http://arxiv.org/abs/1406.2661 arXiv:1406.2661 [cs, stat].

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved Training of Wasserstein GANs. http://arxiv.org/abs/1704.00028 arXiv:1704.00028 [cs, stat].

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. http://arxiv.org/abs/1911.05722 arXiv:1911.05722 [cs].

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. http://arxiv.org/abs/1706.08500 arXiv:1706.08500 [cs, stat].

[17] 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

[18] Jiwoong J Jeong, Amara Tariq, Tobiloba Adejumo, Hari Trivedi, Judy W Gichoya, and Imon Banerjee. 2022. Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. *Journal of Digital Imaging* 35, 2 (2022), 137–152.

[19] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. 2015. Deep Residual Learning for Image Recognition. (2015). https://doi.org/10.48550/arXiv.1512.03385 arXiv:arXiv:1512.03385

[20] 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

[21] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. http://arxiv.org/abs/2006.06676 arXiv:2006.06676 [cs, stat].

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems* 33 (2020), 12104–12114.

[23] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. http://arxiv.org/abs/1812.04948 arXiv:1812.04948 [cs, stat].

[24] Parvinder Kaur, Baljit Singh Khehra, and Er. Bhupinder Singh Mavi. 2021. Data Augmentation for Object Detection: A Review. In *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, Lansing, MI, USA, 537–543. https://doi.org/10.1109/MWSCAS47672.2021.9531849

[25] Ming Lin, Pichao Wang, Zhenhong Sun, Hesen Chen, Xiuyu Sun, Qi Qian, Hao Li, and Rong Jin. 2021. Zen-nas: A zero-shot nas for high-performance image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 347–356.

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11976–11986.

[27] Zhichao Lu, Gautam Sreekumar, Erik Goodman, Wolfgang Banzhaf, Kalyanmoy Deb, and Vishnu Naresh Boddeti. 2021. Neural architecture transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 9 (2021), 2971–2989.

[28] 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

[29] 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

[30] 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

[31] Sachin Mehta and Mohammad Rastegari. 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178* (2021).

[32] 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

[33] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. http://arxiv.org/abs/1411.1784 arXiv:1411.1784 [cs, stat].

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. http://arxiv.org/abs/1606.03498 arXiv:1606.03498 [cs].

[35] Xuan Shen, Yaohua Wang, Ming Lin, Yilun Huang, Hao Tang, Xiuyu Sun, and Yanzhi Wang. 2023. DeepMAD: Mathematical Architecture Design for Deep Convolutional Neural Network. *arXiv preprint arXiv:2303.02165* (2023).

[36] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. 2018. How good is my GAN? http://arxiv.org/abs/1807.09499 arXiv:1807.09499 [cs].

[37] 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.

[38] Luke Taylor and Geoff Nitschke. 2018. Improving Deep Learning with Generic Data Augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, Bangalore, India, 1542–1547. https://doi.org/10.1109/SSCI.2018.8628742

[39] Luke Taylor and Geoff Nitschke. 2018. Improving deep learning with generic data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 1542–1547.

[40] Lilian Weng. 2019. From gan to wgan. *arXiv preprint arXiv:1904.08994* (2019).

[41] McKell Woodland, John Wood, Brian M Anderson, Suprateek Kundu, Ethan Lin, Eugene Koay, Bruno Odisio, Caroline Chung, Hyunseon Christine Kang, Aradhana M Venkatesan, et al. 2022. Evaluating the Performance of StyleGAN2-ADA on Medical Images. In *Simulation and Synthesis in Medical Imaging: 7th International Workshop, SASHIMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*. Springer, 142–153.

[42] McKell Woodland, John Wood, Brian M. Anderson, Suprateek Kundu, Ethan Lin, Eugene Koay, Bruno Odisio, Caroline Chung, Hyunseon Christine Kang, Aradhana M. Venkatesan, Sireesha Yedururi, Brian De, Yuan-Mao Lin, Ankit B. Patel, and Kristy K. Brock. 2022. Evaluating the Performance of StyleGAN2-ADA on Medical Images. Vol. 13570. 142–153. https://doi.org/10.1007/978-3-031-16980-9_14 arXiv:2210.03786 [cs, eess].

[43] Yong Xia, Wenyi Wang, and Kuanquan Wang. 2023. ECG signal generation based on conditional generative models. *Biomedical Signal Processing and Control* 82 (April 2023), 104587. https://doi.org/10.1016/j.bspc.2023.104587

[44] Han Xiao, Ziwei Wang, Zheng Zhu, Jie Zhou, and Jiwen Lu. 2022. Shapley-NAS: Discovering Operation Contribution for Neural Architecture Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11892–11901.

[45] Xin Yi, Ekta Walia, and Paul Babyn. 2019. Generative Adversarial Network in Medical Imaging: A Review. *Medical Image Analysis* 58 (Dec. 2019), 101552. https://doi.org/10.1016/j.media.2019.101552 arXiv:1809.07294 [cs].

[46] 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

[47] 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

[48] Yu Zheng, Zhi Zhang, Shen Yan, and Mi Zhang. 2022. Deep AutoAugment. http://arxiv.org/abs/2203.06172 arXiv:2203.06172 [cs].

[49] 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

A RISKS

Risks	Probability	Impact	Mitigation
A team member falls ill	Medium	Medium	An efficient work allocation plan and the segregation of work amongst team members will ensure that the work of other members can continue in the event that one member falls ill.
A team member misses deadlines	Low	Medium	Regular meetings between team members and the project supervisor will ensure that this does not happen or that contingencies can be made in the event that it does happen.
Loadshedding	High	High	The development and training of models will have to be planned according to the loadshedding schedule however extra time has been accounted for in the planned activities to account for power outages.
Success factors not being met	Medium	High	Through extensive research, the team will be able to diagnose the cause of any problems and provide recommendations for improvement in the final project paper.
Limited access to the UCT High-Performance Cluster	Medium	High	Sufficient planning ensures that we can book cluster time in advance.
Delays in obtaining medical data	Low	Medium	Data recruitment has already begun and, as such, data should be ready before models need to be trained. Data has already been cleared for ethics by Dr. Kruger of Groote Schuur Hospital.
Delays in getting manual evaluation of generated images in GANs	Medium	High	Contact will be made with specialists as soon as possible to build a professional connection and to give a rough time-frame of when manual inspections will take place.

B GANTT CHART