Aus der Abteilung für Biomedizinische Informatik
(Zentrum für Präventivmedizin und Digitale Gesundheit)
der Medizinischen Fakultät Mannheim
(Kommissarischer Leiter: Dr. med. Fabian Siegel)

Efficient Deep Learning at Inference Time for Gram Stained Image Classification

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von
Hee E. Kim

aus
Daegu, Republic of Korea
2024

Dekan: Prof. Dr. med. Sergij Goerdt
Referent: Prof. Dr. med. Thomas Ganslandt

Preface

This dissertation is cumulatively written and all publications are centered around the theme of efficient deep learning for Gram stain classification. The scope of this dissertation, in particular, encompasses efficient model selection and model optimization for efficient processing. Chapter 3 is a survey paper that provides insights into efficient model selection while considering the characteristics of medical image data. In terms of model selection, it fosters transfer learning (TL) and discourages designing yet another novel model architecture. The majority of studies have utilized pre-trained models as feature extractors, implying that only the last fully connected layers need to be re-trained using custom medical datasets. TL proves to be more efficient than searching for state-of-the-art neural architectures often requiring substantial computational resources.

Based on the lessons learned from Chapter 3, Chapter 4 and Chapter 5 apply transfer learning to pre-trained models. Chapter 4 examined models based on convolutional neural networks (CNN), while Chapter 5 investigated visual transformer (VT) models. Both chapters focus on model optimization techniques including pruning and quantization to ensure efficient data processing during inference time. Chapter 4 delved into the investigation of an optimal TL configuration by unfreezing model layers gradually. This inquiry unveiled that a higher accuracy was achieved as more layers were fine-tuned. This result indicates that re-training layers deeper could capture the characteristics of Gram-stained images. Guided by the insights gained from Chapter 4, all models were re-trained from scratch and model optimization was focused in Chapter 5. In this chapter, a broad spectrum of VT models were evaluated by test accuracy, model size and time efficiency at inference time and compared to CNN models. A comparative analysis with CNN models was also conducted.

Peer-reviewed publications are listed in this cumulative dissertation. For each publication, a complete list of type of article, authors, title, journal, journal impact factor and published date is provided. The first author is indicated in bold and more detailed personal contributions to each publication are documented in Table Preface.

Chapter 3 (A literature review)

- Title: Transfer learning for medical image classification: a literature review
- Authors: **Hee E. Kim**, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E Maros, Thomas Ganslandt
- Journal: BMC medical imaging
- Impact factor: 2.795
- Published date: 13 April 2022

Chapter 4 (An original research paper)

- Title: Rapid Convolutional Neural Networks for Gram-Stained Image Classification at Inference Time on Mobile Devices: Empirical Study from Transfer Learning to Optimization
- Authors: **Hee E. Kim**, Mate E Maros, Fabian Siegel, Thomas Ganslandt
- Journal: Biomedicines
- Impact factor: 4.757
- Published date: 4 November 2022

Chapter 5 (An original research paper)

- Title: Interoperable and Lightweight Vision Transformer for Gram-Stained Image Classification
- Authors: **Hee E. Kim**, Mate E Maros, Thomas Miethke, Maximilian Kittel, Fabian Siegel, Thomas Ganslandt
- Journal: Biomedicines
- Impact factor: 4.757
- Published date: 30 April 2023

Table Preface: A summary table of personal contributions to the publications.

Work Steps	Chapter 3	Chapter 4	Chapter 5
Conception (%)	90	90	90
Literature search (%)	100	100	100
Ethics proposal (%)	N/A	90	90
Animal experimentation proposal (%)	N/A	N/A	N/A
Data collection (%)	70	90	90
Data analysis (%)	100	100	100
Interpretation of results (%)	90	100	100
Manuscript writing (%)	80	80	80
Revision (%)	90	85	75
Indicate which figures and tables resulted from your dissertation work.	100	100	100

CONTENTS

Page

List of Figures

List of Tables

List of Acronyms

AI	Artificial intelligence
AR	Augmented reality
ASIC	Application specific integrated circuit
AUC	Area under the receiver operating characteristic curve
BS	Batch size
CAD	Computer-aided diagnosis
CNN	Convolutional neural networks
CT	Computed tomography
DL	Deep learning
DNN	Deep neural networks
EEG	Electroencephalogram
FC	Fully connected
FDA	Food and drug administration
FPGA	Field-programmable gate arrays
FPS	Frames per second
GAN	Generative adversarial networks
GCD	Greatest common divisor
GPU	Graphics processing unit
HOG	Histograms of oriented gradients
ILSVRC	ImageNet large scale visual recognition challenge
IoT	Internet of things
KI	Künstliche Intelligenz
LBP	Local binary pattern
mHealth	Mobile health
MHU	Medical faculty Mannheim, Heidelberg University
MLP	Multilayer perceptron
MRI	Magnetic resonance imaging
MSA	Multi-head self-attention layers
NLP	Natural language processing
OCT	Optical coherence tomography
OHMD	Optical head-mounted display
PACS	Picture archiving and communication system
QC	Quantized per channel
QT	Quantized per tensor
RAM	Random access memory
ReLU	Rectified linear unit
SPECT	Single-photon emission computed tomography
Swin	Shifted windows
TL	Transfer learning
TPU	Tensor processing unit
VT	Vision transformers
WSI	Whole slide image

Chapter 1

Introduction

Gram stain analysis is a laboratory procedure that rapidly classifies microbial pathogens into two classes: Gram-positive or Gram-negative. The goal of Gram stain analysis is to reduce the time to treatment required for accurately identifying the specific bacteria causing e.g. sepsis. In other words, the objective is to minimize the time from the onset of symptoms to diagnosis and targeted treatment. While physicians typically promptly administer antibiotics to sepsis patients, the rapid identification of the microorganism for personalized treatment remains crucial for determining patient survival. Currently, the course of this procedure relies on medical specialists, however, this need not be the case if this issue is reframed as a computer vision problem. This is a field where numerous machine learning (ML) and deep learning (DL) researchers and practitioners have contributed significantly during the last decades, similar to the advancements in natural language processing. A partial automation of the procedure can be achieved by a computer, while the final decision can be made by physicians.

Studies related to ML and DL in the medical domain have gained prominence as emerging research topics over the last decade, leading to numerous studies being published daily. For instance, as of June 20, 2023, the PubMed database indexes 271,735 studies with the following search terms ”(machine learning) OR (deep learning) OR (artificial intelligence)”. However, despite this influx of studies, there are still relatively few that have transitioned into routine care implementation. While research-to-practice gaps are not uncommon across disciplines, we are living in a time where artificial intelligence (AI) is available for everyone. This initiative is called “democratization of AI” contributed by numerous researchers and global tech companies. Its aim is to enable individuals, including those new to machine learning, to train deep learning models using their own custom datasets, without relying on highly skilled practitioners or researchers. The democratization of AI is characterized by publicly accessible data, open-source frameworks, pre-trained models and free online education.

ImageNet [Den+09a], for example, is the largest publicly available dataset containing over 14 million images with manually labeled annotations. Fei-Fei Li, a pioneer in AI democratization and a Stanford University associate professor, played a pivotal role in its creation. Other widely used open datasets for pre-training models are as follows: COCO (common objects in context) [Lin+14], CIFAR (Cana-

dian institute for advanced research) [KNH10], MNIST (modified national institute of standards and technology) [LeC98], UCI machine learning repository [AN07], Reddit dataset [HYL17], and Wikimedia Commons [VK14]. Several ML frameworks are publicly available, and the choice of framework is often a matter of personal preference. TensorFlow [Aba+16b], Keras [Cho+18], PyTorch [Pas+19], and Theano [Al-+16] are popular choices, while Open Neural Network Exchange (ONNX) [BLZ+19] is an emerging framework that facilitates model interoperability across DL frameworks and enables deployment on various hardware and operating systems. ONNX was initially released in September 2017 by Facebook, now Microsoft, and numerous hardware vendors and research institutes have contributed to the interoperable DL ecosystem. For instance, a model created in TensorFlow can be seamlessly converted and integrated into Keras or PyTorch, and a model trained in Python can be deployed in a C++ application on embedded hardware. This interoperability is achievable because the semantics of tensor-oriented computations in current AI frameworks are consistent, allowing for conversion to a standardized set of operators and syntax [Ahm+21]. In addition to ONNX, Liu et al. [Liu+20b] proposed Model Management for deep neural networks (MMdnn), a graph-oriented conversion tool, for nine DL frameworks in 2020, although it did not gain significant attention from the community and the project became stale. OpenVINO [Gor+19], a compiler and runtime suite, addresses general problems and fosters interoperable DL ecosystems. In addition, the choice of a framework provides pre-trained deep learning models. While each framework archive and offer pre-trained models, Model Zoo [1] offers a comprehensive overview of pre-trained models across frameworks and tasks, complete with code examples. The ONNX Model Zoo [2] standardizes the pre-trained model format, organizing them alongside codes and research papers. Both resources are user-friendly as they categorize models based on specific problems and tasks. Lastly, a diverse range of competencies is necessary to develop a successful deep learning product. From those new to data science to experienced researchers, individuals can acquire deficient skills through open online courses, known as massive open online courses (MOOC). Numerous prestigious universities, educational institutions, and online learning platforms offer a wide array of learning materials for unrestricted participation, often free of charge or at a small subscription fee. Major contributors to the MOOC landscape are Coursera [3], edX [4], Udemy [5], Udacity [6], Khan Academy [7], and Google Cloud Skill Boost [8]

However, despite the bold statement of "AI is available for everyone", research institutions such as medical faculties or small technology-driven organizations still struggle with customizing and deploying deep learning models for their custom tasks. For instance, You et al. [You+19] reported that training ResNet50 for 90 epochs with ImageNet-1k on an NVIDIA M40 GPU took 14 days. Hence, the de-

[1] https://modelzoo.co, (Accessed on June 20, 2023).
[2] https://github.com/onnx/models, (Accessed on June 20, 2023).
[3] https://www.coursera.org, (Accessed on June 20, 2023).
[4] https://www.edx.org, (Accessed on June 20, 2023).
[5] https://www.udemy.org, (Accessed on June 20, 2023).
[6] https://www.udacity.org, (Accessed on June 20, 2023).
[7] https://www.khanacademy.org, (Accessed on June 20, 2023).
[8] https://www.cloudskillsboost.google/journeys, (Accessed on June 20, 2023).

mocratization of AI remains a hollow statement without adequate infrastructure. Cloud computing appears to be a convenient and appealing solution because investment in an expansive infrastructure is not required. However, this comes with its own set of concerns. Surrendering control over infrastructure could potentially compromise data privacy. For instance, Cambridge Analytica accessed and collected personally identifiable information from 87 million Facebook users without consent in 2013 and provided data analysis assistance to United States Senator Ted Cruz for Trump's presidential campaigns in 2016 [IH18]. Similarly, in 2021, data from 700 million LinkedIn users were leaked [Gib+21]. Hence, a strategy on how to train and deploy AI models with constrained hardware must be considered in advance. Model training time should be reasonable and accountable without requiring expensive infrastructure or reliance on cloud computing services as well as model performance during inference time should be reliable when models are deployed on limited hardware resources. Otherwise, constrained infrastructure would hinder the provision of meaningful patient services in routine care.

This thesis, therefore, advocates for the proactive consideration of the computational costs in advance and cumulatively presents the optimal utilization of efficient deep learning through an empirical case study on Gram stain classification. Chapter 2 provides the background and nomenclature commonly used in the thesis. Chapter 3 is a literature review on transfer learning (TL) of convolutional neural networks (CNN) for medical image classification. Despite data scarcity, the potential of TL has been recognized for its capability to reduce computational costs and time without degrading the predictive power. The content of this chapter is based on the article [Kim+22b]. Chapter 4 demonstrates the utility of pruning and quantization in reducing model size and inference time without compromising model quality. Three CNN models were empirically tuned for Gram stain classification and their performance was evaluated on two Android smartphones. The content of this chapter is based on the article [Kim+22a]. Chapter 5 presents a comparative analysis of six visual transformers (VT) models and two CNN models for Gram stain classification. The comparison was carried out using various configurations, including different model sizes, training epochs, quantization schemes and datasets with varying amounts of data. The content of this chapter is based on the article [Kim+23]. Chapter 6 concludes the thesis and provides overarching discussions. Finally, Chapter 7 foresees potential future work related to efficient deep learning in the medical domain.

Chapter 2

Background

To understand what deep learning (DL) is, a couple of terminologies such as Artificial intelligence (AI), and machine learning (ML) need to be clarified together. AI and ML are terminologies associated with DL and they are interchangeably used among people. Since they are closely related to one another, often researchers and experienced practitioners encounter confusion. AI is the overarching umbrella term covering a wider range of subsets such as ML, DL, robotics, natural language processing and more. AI integrates those subsets harmonically in order to simulate human-like general intelligence in a machine that is able to perceive, process and respond to event inputs dynamically. Both ML and DL are subsets of AI, and they are mathematical expressions constructed based on data without involving explicit programming. DL is a branch of ML and DL models specifically employ numerous neurons. This chapter introduces the basics of DL in computer vision and common notations used in the following chapters. Three pillars of computer vision applications in the medical domain are also introduced with corresponding tasks and representative DL architectures. Finally, the landscape of efficient deep learning techniques is explained based on the lifecycle of a deep learning model.

2.1 Deep Learning in Computer Vision

The goal of computer vision is to enable computers to understand images or videos by analyzing, interpreting and understanding visual data. In other words, it imitates human vision to make a certain decision out of a given visual data. Before the deep learning era, hand-crafted and manual feature extraction was the essential step to understanding visual information. Features are high-level information about images including edge, corner, texture, color, edge and more. With these features, computers are able to recognize important objects or areas in a given image. However, hand-crafted methodologies were primitive and not able to compete with the capability of the human visual system.

Owing to the recent flourishing of deep learning technologies, interest in computer vision has gained momentum across domains such as medical image analysis [Kim+22b], autonomous vehicles [Jan+20], pose estimation in robotics [Sün+18], 3D image modeling or augmented/virtual reality [Abu+18] and more. In fact, deep

learning and CNN are not new artifacts from the 21st century. The first definition of machine learning was introduced in 1959 by Arthur Samuel [Sam59], and the first publication of neural networks with multi-layer perceptrons was published by Ivakhnenko and Lapa in 1967 [ILL67]. However, training deep and sophisticated neural networks requires a large amount of data and computational power and data were insufficient and computer features were limited in the 50s and 60s. The early stages of computers were not capable to store data and process them. Remarkable use cases of deep learning were developed only recently with the emergence of a large amount of data and big data technology.

Neuron is the core element of deep neural networks and it is inspired by the neurons in the human brain. The mechanism of a neuron imitates the communication of brain cells. For instance, an activated presynaptic cell carries a signal to the synapse and fires postsynaptic cells with neurotransmitters. This sequence of how neurons are jointly influencing one another is imitated as follows: A single neuron receives real numbers from the neurons in the previous layer, generates a real number and then transmits it to the neurons in the next layer. A postsynaptic cell is denoted as a neuron, neurotransmitters are denoted as real numbers, and likewise, presynaptic cells are neurons in the previous layer. Figure 2.1 depicts the model of a single neuron. A linear combiner sums up all input signals multiplied by weight values, then adds a bias. The output of the linear combiner is fed into a non-linear activation function. Finally, the output is transmitted to the next neuron in the subsequent layer as an input.

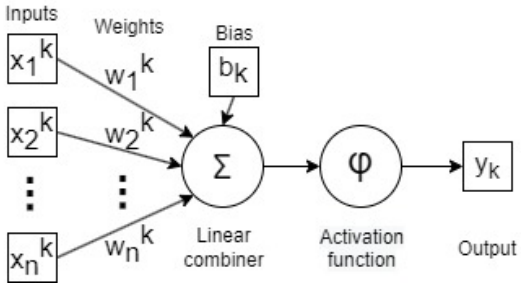

Figure 2.1: Representation of a single neuron. k is the index of a neuron in a network. A popular activation function among many others is rectified linear unit (ReLU) [Xu+15], which outputs 0 or the input value when it is positive meaning that it suppresses all negative vectors to black color.

Numerous deep-learning architectures were proposed by researchers and practitioners to address tailored tasks. The choice of techniques depends on the task and data type. Two mainstream architectures in computer vision are as follows: CNN and VT. Figure 2.2 shows an example of a shallow CNN model consisting of input layer, hidden layers and output layer. Usually, the hidden layer is very deep consisting of numerous convolutional layers and pooling layers. The layers near the input capture generic features (e.g. edge or color), whereas the layers near the output detect more specific features (e.g. eyes in a face) in images. While general neural networks consist of fully connected layers only, CNN contain at least one convolutional layer which extracts the local features of a given image

On the other hand, figure 2.3 is the figure from the original VT [Dos+20] taking 16x16 patches that are equivalent to the encoder block of the transformers model [Vas+17]. Both CNN and VT are able to process grid-like topology data (i.e. image data). While CNN process image data by nature, VT need a step to flatten grid-like data into a sequence of tokens. The core element of VT is the attention mechanism which is relatively new in the field of computer vision. It was initially introduced as an alignment model for natural language processing [BCB14], and then the original attention-based model was proposed by Vaswani et al. from Google in 2017 [Vas+17]. In 2020, Dosovitskiy from Google utilized the encoder block of the transformers model for image classification problem [Dos+20]. The self-attention mechanism for image analysis captures the global relationships of each image patch to attend to all other patches in a given image. Compared to CNN, it requires larger data for training and more computational resources because the relevance of every image patch needs to be computed during model training.

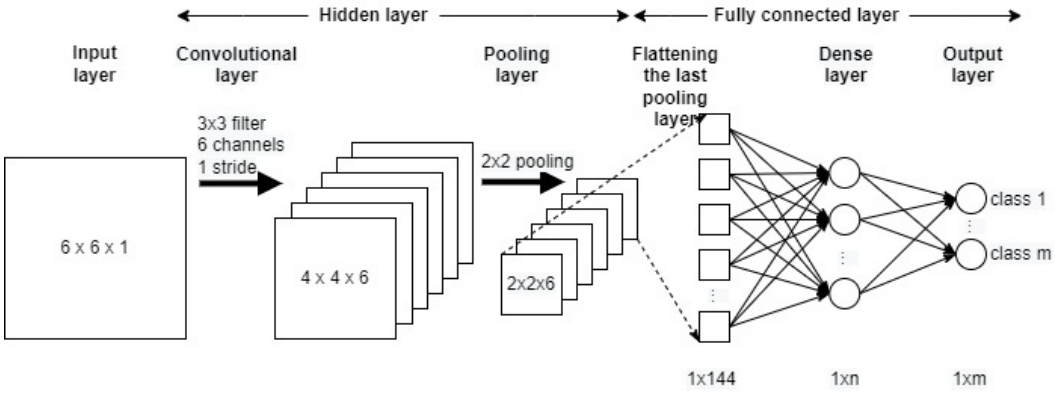

Figure 2.2: An example of convolutional neural networks illustrating two convolutional layers, pooling layers and fully connected layers.

2.2 Computer Vision Applications and Tasks in Healthcare

Computer vision applications refer to a set of tasks or challenges, each targeting specific objectives through the extraction and analysis of visual data from images or videos. These tasks are varying levels of granularity and types of visual information, ranging from image classification [Kim+22b] to intricate image segmentation [Min+21]. For instance, image classification provides a generalized understanding of an entire image, while segmentation operates at the granularity of individual pixels, offering more detailed insights. These tasks are not isolated from one another; for instance, within the domain of object detection [Jia+19], model architectures encompass components of both classification and regression [Syk93]. In the medical domain, DL has been widely facilitated to address a variety of computer vision applications from computer-aided diagnosis to medical report generation.

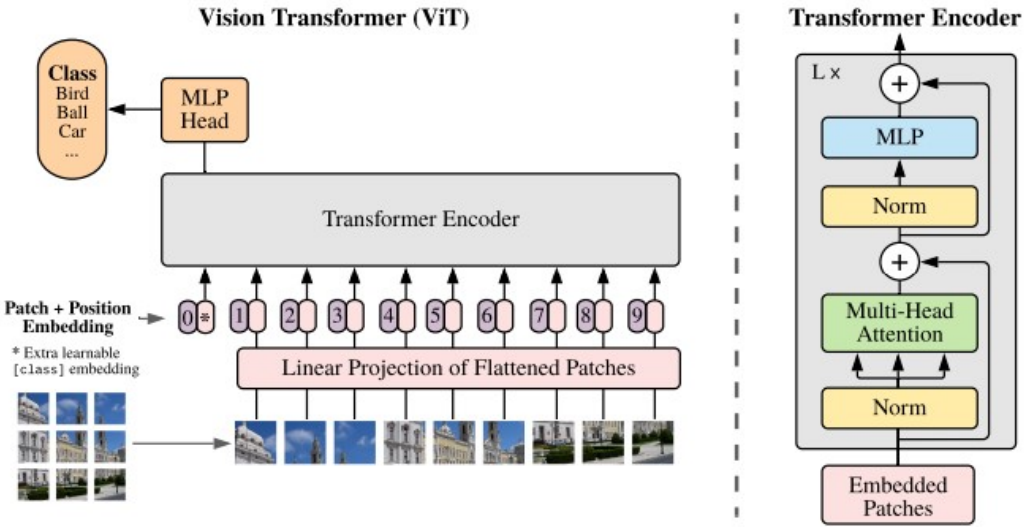

Figure 2.3: Vision transformer depicted by [Dos+20] inspired by the transformer encoder [Vas+17]

2.2.1 Computer-Aided Diagnosis

Computer-aided diagnosis (CAD) [CHS20] is designed to offer support to caregivers by automating medical image analysis. More precisely, CAD aims to augment medical doctors' decision-making processes through comprehensive medical image analysis. This augmentation manifests in multiple ways: enhancing the diagnostic capacities of healthcare professionals, expediting decision-making, and potentially relieving them of more routine or intricate tasks, thus enabling them to spend more time on patient care. A pivotal exemplar of CAD tasks is medical image classification, where the objective is to assign a single label to an entire medical image. This task, which constitutes a foundational element of computer vision across diverse domains, has seen extensive exploration in the medical field. Illustrative instances include classifying Gram-stain images as positive or negative, categorizing X-ray images as normal or abnormal, segregating MRI images based on specific clinical conditions and more. The number of neurons in the output layer should correspond to the count of distinct labels and the softmax function [GBC16a] is the preferred activation function in the output layer, as it can predict probabilities associated with each class label. Alternatively, the sigmoid function (logistic function) [Nwa+18] can be employed when the desired output needs to be normalized within the range of 0 to 1.

In contrast to medical image classification, highlighting regions of interest concerning lesions, organs, and anatomical substructures are much more complex tasks, nonetheless, they can provide more granular information to radiologists or pathologists. Object detection [Jia+19] facilitates bounding boxes to the region of interest and assigns a class to each box, while segmentation masks image pixels directly on the given images and provides pixel-level classification. To elucidate further, highlighting the location of microorganisms within an entire slide image yields more intricate insights compared to a mere classification task. The approach for lesion

segmentation (or detection) remains consistent, with variations only in the anatomical parts of interest. The case studies span from segmenting malignant lung nodules to the segmentation of multi-tissue nuclei, brain tumors, retinal vessels, and pancreatic structures.

2.2.2 Image Enhancement and Reconstruction

DL made remarkable progress in image enhancement and image reconstruction. These two applications share a common aim, namely, to enrich input image data and ultimately improve the model performance of subsequent tasks. Despite this overarching aim, they diverge in their objectives and functionalities. While image enhancement directs its focus towards refining specific attributes of a given image (e.g. such as contrast, sharpness, saturation, and brightness), image reconstruction aims to restore a missing or corrupted image, which is a task that resonates with image synthesis endeavors. Notably, the applications of image reconstruction tend to be more intricate and complex.

An illustrative example lies in medical imaging where radiation is employed. In such cases, a sophisticated balance between image quality and radiation hazards must be considered. Consequently, employing image reconstruction techniques can mitigate the exposure of patients to excessive radiation doses. Additionally, challenges like data scarcity and imbalanced datasets can be effectively tackled through image reconstruction, particularly in the form of image synthesis. While generative adversarial networks (GAN) [Goo+14a] are renowned for generating previously unseen images, they can also be tailored and trained for tasks encompassing image enhancement or reconstruction. Similarly, architectures such as U-Net [RFB15] and autoencoders [GBC16b] present viable strategies to address these challenges. However, the selection of an appropriate DL architecture hinges upon the data characteristics inherent to a given application context.

2.2.3 Image Registration

Medical image registration [MF93] is concerned with the alignment of multiple medical images by determining optimal spatial coordinates between them. Over the past decade, DL has achieved substantial strides in advancing the field of medical image registration; however, it remains less popular in comparison to other applications. This subject encompasses a multitude of applications, each offering distinct insights. Representative examples include the alignment of multi-modal images (e.g. such as magnetic resonance imaging (MRI) and single-photon emission computed tomography (SPECT)) belonging to a single patient. Other significant examples are the alignment of uni-modal images derived from multiple patients or sequences of images captured from a single patient at distinct time intervals. The latter, in particular, holds considerable utility for longitudinal studies, enabling the tracking of disease progression across time.

Regardless of the specific task, the fundamental objective of aligning multiple images necessitates a minimum of three data points along the x and y coordinates

within the given images. Given the numerical nature of these data points, the integration of a linear function in the output layer is requisite for facilitating alignment. Further contributions of DL to medical image registration are anticipated.

2.2.4 Medical Report Generation

Medical report generation aims to automate the creation of medical reports from corresponding medical images. In other words, it is a multimodal DL application that is capable of processing heterogeneous data and striving to establish semantic associations among them. For the training of such models, diagnostic reports and medical images extracted from the picture archiving and communication system (PACS) are leveraged. Medical report generation is a repetitive, time-consuming and error-prone task that still relies on the medical service providers. However, AI-powered solutions have the potential to alleviate this burden on clinicians, equipping them with the means to render swift and precise decisions. The examples encompassed by medical report generation span a spectrum from relatively straightforward tasks to intricate ones. While a simple task could extend to attributing categories (e.g. shape, margin, and density) to mammographic images, it is capable to generate professional medical reports that are comparable to those written by trained radiologists. The reports encompass diagnoses, nuanced descriptions of impressions, and comprehensive findings. This subject area remains relatively underexplored, yet it holds considerable promise due to the substantial amount of medical image data and its corresponding diagnostic reports that are archived in PACS systems.

Chapter 3

Transfer Learning for Medical Image Classification

DOI: 10.1186/s12880-022-00793-7

3.1 Introduction

Medical image analysis is a robust subject of research, with millions of studies having been published in the last decades. Some recent examples include computer aided tissue detection in whole slide image (WSI) and the diagnosis of COVID-19 pneumonia from chest images. Traditionally, sophisticated image feature extraction or discriminant handcrafted features (e.g. histograms of oriented gradients (HOG) features [DT05] or local binary pattern (LBP) features [HW90]) have dominated the field of image analysis, but the recent emergence of deep learning (DL) algorithms has inaugurated a shift towards non-handcrafted engineering, permitting automated image analysis. In particular, convolutional neural networks (CNN) have become the workhorse DL algorithm for image analysis. In recent data challenges for medical image analysis, all of the top-ranked teams utilized CNN. For instance, the top-ten ranked solutions, except one team, had utilized CNN in the CAMELYON17 challenge for automated detection and classification of breast cancer metastases in whole slide images [Ban+18a]. It has also been demonstrated that the features extracted from DL surpassed that of the handcrafted methods by Shi et al. [Shi+18].

However, DL algorithms including CNN require—under preferable circumstances—a large amount of data for training; hence follows the data scarcity problem. Particularly, the limited size of medical cohorts and the cost of expert-annotated data sets are some well-known challenges. Many research endeavors have tried to overcome this problem with transfer learning (TL) or domain adaptation [WDG19] techniques. These aim to achieve high performance on target tasks by leveraging knowledge learned from source tasks. A pioneering review paper of TL was contributed by Pan and Yang [PY10] in 2010, and they classified TL techniques from a labeling aspect, while Weiss et al. [WKW16] summarized TL studies based on homogeneous and heterogeneous approaches. Most recently in 2020, Zhuang et al. [Zhu+20] reviewed more than forty representative TL approaches

from the perspectives of data and models. Unsupervised TL is an emerging subject and has recently received increasing attention from researchers. Wilson and Cook [WC20] surveyed a large number of articles on unsupervised deep domain adaptation. Most recently, generative adversarial networks (GAN)-based frameworks [Goo+14b; Zhu+17; Zha+19b] gained momentum, a particularly promising approach is DANN [Gan+16]. Furthermore, multiple kernel active learning [Wan+19] and collaborative unsupervised methods [Zha+20] have also been utilized for unsupervised TL.

Some studies conducted a comprehensive review focused primarily on DL in the medical domain. Litjens et al. [Lit+17] reviewed DL for medical image analysis by summarizing over 300 articles, while Chowdhury et al. [Cho+21] reviewed the state-of-the-art research on self-supervised learning in medicine. On the other hand, others surveyed articles focusing on TL with a specific case study such as microorganism counting [Zha+21], cervical cytopathology [Rah+20a], neuroimaging biomarkers of Alzheimer's disease [Aga+21] and magnetic resonance brain imaging in general [Val+21].

In this paper, we aimed to conduct a survey on TL with pretrained CNN models for medical image analysis across use cases, data subjects and data modalities. Our major contributions are as follows:

- (i) An overview of contributions to the various case studies is presented;
- (ii) Actionable recommendations on how to leverage TL for medical image classification are provided;
- (iii) Publicly available medical datasets are compiled with URL as supplementary material.

The rest of this paper is organized as follows. Section 2 covers the background knowledge and the most common notations used in the following sections. In Sect. 3, we describe the protocol for the literature selection. In Sect. 4, the results obtained are analyzed and compared. Critical discussions are presented in Sect. 5. Finally, we end with a conclusion and the lessons learned in Sect. 6. Figure 3.1 is the main diagram that presents the whole manuscript.

3.2 Background

3.2.1 Transfer Learning

Transfer learning (TL) stems from cognitive research, which uses the idea, that knowledge is transferred across related tasks to improve performances on a new task. It is well-known that humans are able to solve similar tasks by leveraging previous knowledge. The formal definition of TL is defined by Pan and Yang with notions of domains and tasks. "A domain consists of a feature space $\mathcal{X}$ and marginal probability distribution $\mathrm{P(X)}$, where $\mathrm{X} = \{x_1, ..., x_n\} \in \mathcal{X}$. Given a specific domain denoted by $\mathrm{D} = \{\mathcal{X}, \mathrm{P(X)}\}$, a task is denoted by $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ where $\mathcal{Y}$ is a label space and $f(\cdot)$ is an objective predictive function. A task is learned from the pair $\{x_i, y_i\}$ where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Given a source domain $\mathbf{D_S}$ and source task

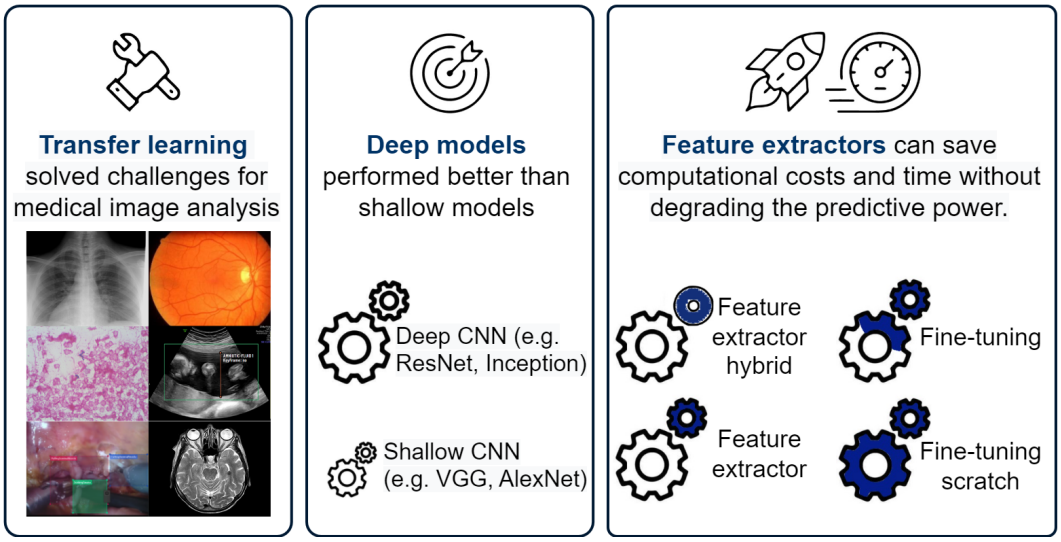

Figure 3.1: Visual abstract summarizing the scope of our study.

$\mathbf{T_S}$, a target domain $\mathbf{D_T}$ and learning task $\mathbf{T_T}$, transfer learning aims to improve the learning of the target predictive function $f_T(\cdot)$ in $\mathbf{D_T}$ by using the knowledge in $\mathbf{D_S}$ and $\mathbf{T_S}$" [PY10].

Analogously, one can learn how to drive a motorbike $\mathbf{T_T}$ (transferred task) based on one's cycling skill $\mathbf{T_S}$ (source task) where driving two-wheel vehicles is regarded as the same domain $\mathbf{D_S} = \mathbf{D_T}$. This does not mean that one will not learn how to drive a motorbike without riding a bike, but it takes less effort to practice driving the motorbike by adapting one's cycling skills. Similarly, learning the parameters of a network from scratch will require larger annotated datasets and a longer training time to achieve an acceptable performance.

3.2.2 Convolutional Neural Networks using ImageNet

CNNs are a special type of deep learning that processes grid-like topology data such as image data. Unlike the standard neural network consisting of fully connected layers only, CNN consists of at least one convolutional layer. Several pretrained CNN models are publicly accessible online with downloadable parameters. They were pretrained with millions of natural images on the ImageNet dataset (ImageNet large scale visual recognition challenge; ILSVRC).

In this paper, CNN models are denoted as backbone models. Table 3.1 summarizes the five most popular models in chronological order from top to bottom. LeNet [Lec+98] and AlexNet [KSH12] are the first generations of CNN models developed in 1998 and 2012 respectively. Both are relatively shallow compared to other models that are developed recently. After AlexNet won the ImageNet large scale visual recognition challenge (ILSVRC) in 2012, designing novel networks became an emerging topic among researchers. VGG [SZ15], also referred to as OxfordNet, is recognized as the first deep model, while GoogLeNet [Heg+19], also known as Inception1, set the new state of the art in the ILSVRC 2014. Inception introduced the novel block concept that employs a set of filters with different sizes, and its deep

networks were constructed by concatenating the multiple outputs. However, in the architecture of very deep networks, the parameters of the earlier layers are poorly updated during training because they are too far from the output layer. This problem is known as the vanishing gradient problem which was successfully addressed by ResNet [He+16] by introducing residual blocks with skip connections between layers.

The number of parameters of one filter is calculated by (a * b * c) + 1, where a * b is the filter dimension, c is the number of filters in the previous layer and added 1 is the bias. The total number of parameters is the summation of the parameters of each filter. In the classifier head, all models use the Softmax function except LeNet-5, which utilizes the hyperbolic tangent function. The Softmax function fits well with the classification problem because it can convert feature vectors to the probability distribution for each class candidate.

Table 3.1: Overview of five backbone models are listed in chronological order.

Model type	Model	Released year	Parameters (FE[a] only)	Parameters (all)	Layers (FE+FC[b])	Dataset
Shallow and linear	LeNet5	1998	1.7 K	60 K	4 (2+2)	MNIST[c]
	AlexNet	2012	3.7 M	62.3 M	8 (5+3)	ImageNet[d]
	VGG16	2014	14.7 M	134.2 M	16 (13+3)	
Deep	GoogLeNet	2014	5.3 M	5.3 M	22 (21+1)	
	ResNet50	2015	23.5 M	25.6 M	51 (50+1)	

[a]Feature extraction
[b]Fully connected layers
[c]Database of handwritten digits with 60 K training and 10 K test images.
[d]Database of over 14 M hand-annotated images for visual object recognition research.

3.2.3 Transfer Learning of Convolutional Neural Networks

TL with CNN is the idea that knowledge can be transferred at the parametric level. Well-trained CNN models utilize the parameters of the convolutional layers for a new task in the medical domain. Specifically, in TL with CNN for medical image classification, a medical image classification (target task) can be learned by leveraging the generic features learned from the natural image classification (source task) where labels are available in both domains. For simplicity, the terminology of TL in the remainder of the paper refers to homogeneous TL (i.e. both domains are image analysis) with pretrained CNN models using ImageNet data for medical image classification in a supervisory manner.

Roughly, there are two TL approaches to leveraging CNN models: either feature extractor or fine-tuning. The feature extractor approach freezes the convolutional layers, whereas the fine-tuning approach updates parameters during model fitting. Each can be further divided into two subcategories; hence, four TL approaches are defined and surveyed in this paper. They are intuitively visualized in Figure 3.2. Feature extractor hybrid (Figure 3.2a) discards the FC layers and attaches a machine learning algorithm such as SVM or Random Forest classifier into the feature

extractor, whereas the skeleton of the given networks remains the same in the other types (Figure 3.2bd). Fine-tuning from scratch is the most time-intensive approach because it updates the entire ensemble of parameters during the training process.

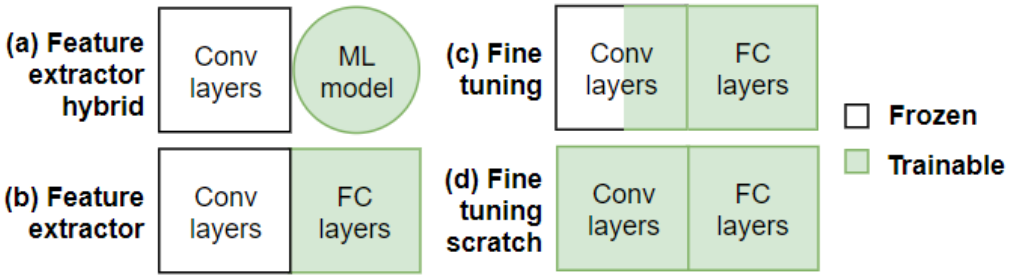

Figure 3.2: Four types of transfer learning approach. The last classifier block needs to be replaced by a thinner layer or trained from scratch (ML: Machine learning; FC: Fully connected layers).

3.3 Methods

Publications were retrieved from two peer-reviewed databases (PubMed database on January 2, 2021, and Web of Science database on January 22, 2021). Papers were selected based on the following four conditions: (1) convolutional or CNN should appear in the title or abstract; (2) image data analysis should be considered; (3) "transfer learning" or "pretrained" should appear in the title or abstract; finally, (4) only experimental studies were considered. The time constraint is specified only for the latest date, which is December 31, 2020. The exact search strings used for these two databases are denoted in Appendix A. Duplicates were merged before screening assessment. The first author screened the title, abstract and methods in order to exclude studies proposing a novel CNN model. Typically, this type of study stacked up multiple CNN models or concatenated CNN models and handcrafted features, and then compared its efficacy with other CNN models. Non-classification tasks, and those publications which fell outside the aforementioned date range, were also excluded. For the eligibility assessment, full texts were examined by two researchers. A third, independent researcher was involved in decision-making in the case of discrepancy between the two researchers.

3.3.1 Methodology Analysis

Eight properties of 121 research articles were surveyed, investigated, compared and summarized in this paper. Five are quantitative properties and three are qualitative properties. They are specified as follows: (1) Off-the-shelf CNN model type (AlexNet, CaffeNet, Inception1, Inception2, Inception3, Inception4, Inception-Resnet, LeNet, MobileNet, ResNet, VGG16, VGG19, DenseNet, Xception, many or else); (2) Model performances (accuracy, AUC, sensitivity and specificity); (3) Transfer learning type (feature extractor, feature extractor hybrid, fine-tuning, fine-

tuning or many); (4) Fine-tuning ratio; (5) Data modality (endoscopy, CT/CAT scan, mammographic, microscopy, MRI, OCT, PET, photography, sonography, SPECT, X-ray/radiography or many); (6) Data subject (abdominopelvic cavity, alimentary system, bones, cardiovascular system, endocrine glands, genital systems, joints, lymphoid system, muscles, nervous system, tissue specimen, respiratory system, sense organs, the integument, thoracic cavity, urinary system, many or else); (7) Data quantity; and (8) The number of classes. They fall into one of three categories, namely model, transfer learning or data.

3.4 Results

Figure 3.3 shows the PRISMA flow diagram of paper selection. We initially retrieved 467 papers from PubMed and Web of Science. 42 duplicates were merged from two databases, and then 425 studies were assessed for screening. 189 studies were excluded during the screening phase, and then full texts of 236 studies were assessed for the next stage. 114 studies were disqualified from inclusion, resulting in 121 studies. These selected studies were further investigated and organized with respect to their backbone model and TL type. The data characteristics and model performance were also analyzed to gain insights regarding how to employ TL.

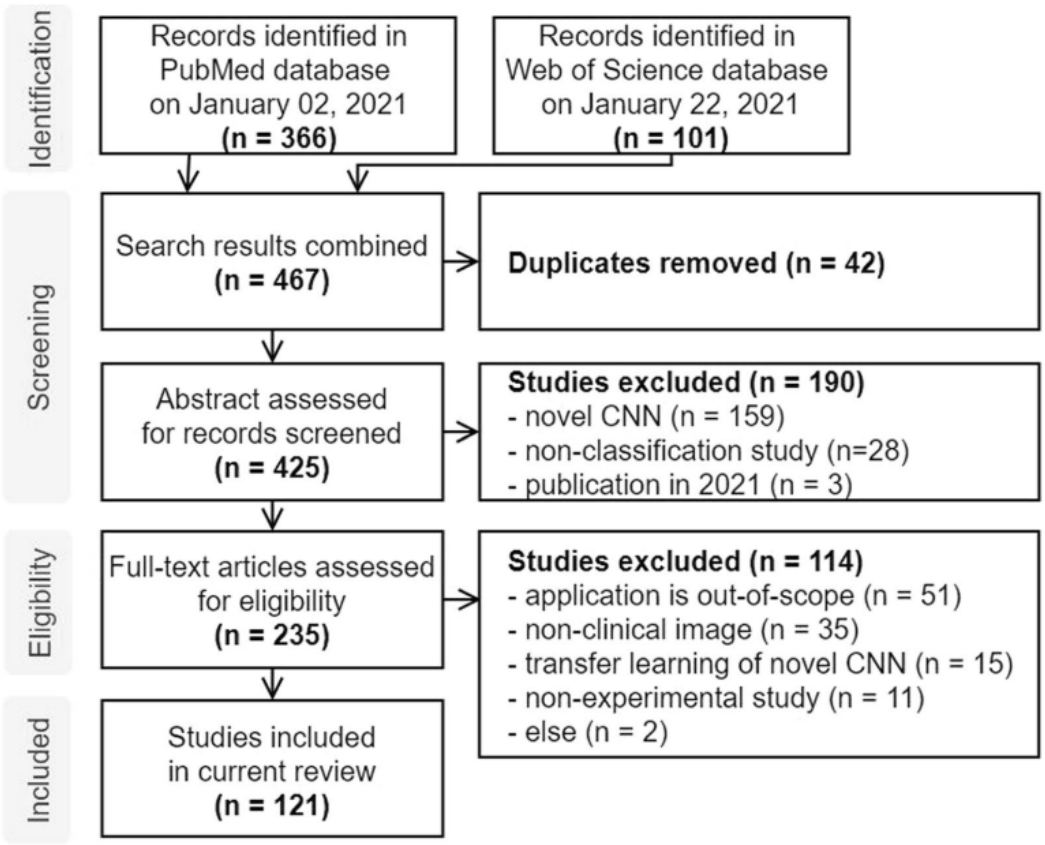

Figure 3.3: Flowchart of the literature search.

Figure 3.4a shows that studies of TL for medical image classification have emerged

since 2016 with a 4-year delay after AlexNet [KSH12] won the ImageNet Challenge in 2012. Since then the number of publications grew rapidly for consecutive years. Studies published in 2020 seem shrinking compared to the number of publications in 2019, because the process of indexing a publication may take anywhere from three to six months.

3.4.1 Backbone Model

The majority of the studies (n = 57) evaluated several backbone models empirically as depicted in Figure 3.4b. For example, Rahaman and his colleagues [Rah+20b] contributed an intensive benchmark study by evaluating fifteen models, namely: VGG16, VGG19, ResNet50, ResNet101, ResNet152, ResNet50V2, ResNet101V2, ResNet152V2, Inception3, InceptionResNet2, MobileNet1, DenseNet121, DenseNet169, DenseNet201 and XceptionNet. They concluded that VGG19 presented the highest accuracy of 89.3%. This result is exceptional because other studies reported that deeper models (e.g. Inception and ResNet) performed better than the shallow models (e.g. VGG and AlexNet). Five studies [Bur+18; Che+19b; Lak17; Yan+18a; Yu+19a] compared Inception and VGG and reported that Inception performed better, and Ovalle-Magallanes et al. [Ova+20] also concluded that Inception3 outperformed compared to ResNet50 and VGG16. Finally, Talo et al. [Tal+19] reported that ResNet50 achieved the best classification accuracy compared to AlexNet, VGG16, ResNet18 and ResNet34.

Besides the benchmark studies, the most prevalent model was the Inception (n = 26) that consists of the least parameters shown in Table 3.1. AlexNet (n = 14) and VGG (n = 10) were the next commonly used models although they are shallower than ResNet (n = 5) and InceptionResnet (n = 2). Finally, only a few studies (n = 7) used a specific model such as LeNet5, DenseNet, CheXNet, DarkNet, OverFeat or CaffeNet.

3.4.2 Transfer Learning

Similar to the backbone model, the majority of models (n = 46) evaluated numerous TL approaches, which are illustrated in Figure 3.4c. Many researchers aimed to search for the optimal choice of TL approach. Typically, grid search was applied. Shin and his colleagues [Shi+16] extensively evaluated three components by varying three CNN models (CifarNet, AlexNet and GoogLeNet) with three TL approaches (feature extractor, fine-tuning from scratch with and without random initialization), and the fine-tuned GoogLeNet from scratch without random initialization was identified as the best performing model.

The most popular TL approach was feature extractor (n = 38) followed by fine-tuning from scratch (n = 27), feature extractor hybrid (n = 7) and fine-tuning (n = 3). Feature extractor takes the advantage of saving computational costs by a large degree compared to the others. Likewise, the feature extractor hybrid can profit from the same advantage by removing the FC layers and adding less expansive machine learning algorithms. This is particularly beneficial for CNN models with heavy FC layers like AlexNet and VGG. Fine-tuning from scratch was the second

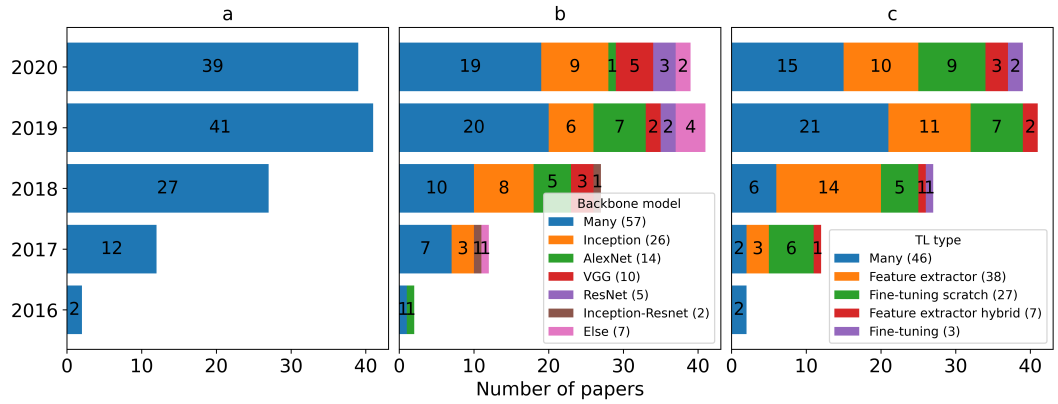

Figure 3.4: Studies of transfer learning in medical image classification over time (y-axis) with respect to (a) the number of publications, (b) applied backbone model and (c) transfer learning type.

most popular approach despite it being the most resource-expensive type because it updates the entire model. Fine-tuning is less expensive compared to the fine-tuning from scratch as it partially updates the parameters of the convolutional layers. Additional file 2: Table B in Appendix B presents an overview of four TL approaches which were organized based on three dimensions: data modality, data subject and TL type.

3.4.3 Data Characteristics

As the summary of data characteristics is depicted in Figure 3.5, a variety of human anatomical regions has been studied. Most of the studied regions were breast cancer exams and skin cancer lesions. Likewise, a wide variety of imaging modalities contained a unique attribute of medical image analysis. For instance, computed tomography (CT) scans and magnetic resonance imaging (MRI) are capable of generating 3D image data, while digital microscopy can generate terabytes of whole slide image (WSI) of tissue specimens.

Figure 3.5b shows that the majority of studies consist of binary classes, while Figure 3.5c shows that the majority of studies have fallen into the first bin which ranges from 0 to 600. Minor publications are not depicted in Figure 3.5 for the following reasons: the experiment was conducted with multiple subjects (human body parts); multiple tasks; multiple databases; or the subject is non-human body images (e.g. surgical tools).

3.4.4 Performance Visualization

Figure 3.6 shows scatter plots of model performance, TL type and two data characteristics: data size and image modality. The Y coordinates adhere to two metrics, namely area under the receiver operating characteristic curve (AUC) and accuracy. Eleven studies used both metrics, so they are displayed on both scatter plots. The X coordinate is the normalized data quantity, otherwise it is not fair to compare the

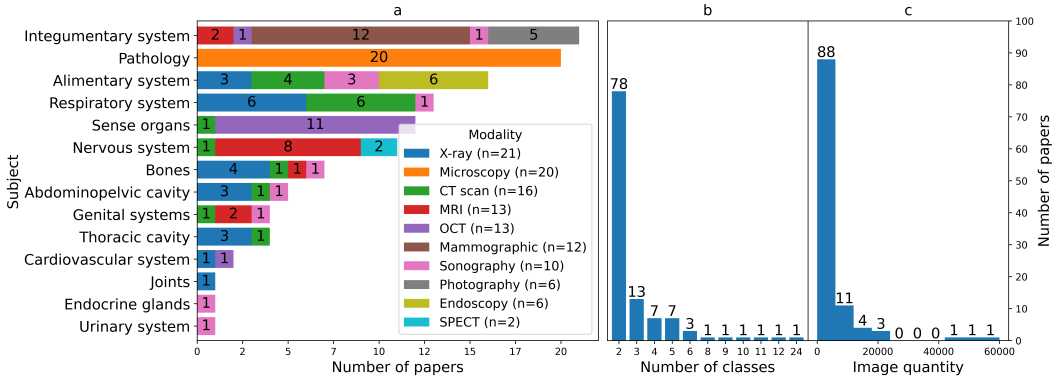

Figure 3.5: The overview of data characteristics of selected publications. (a) The correlation of anatomical body parts and imaging modalities. (b) The number of classes (c) The histogram of the quantity of medical image datasets.

classification performance with two classes versus ten classes. The data quantities of three modalities—CT, MRI and Microscopy—reflect the number of patients.

For the fair comparison, studies employed only a single model, TL type and image modality are depicted (n = 41). Benchmark studies were excluded; otherwise, one study would generate several overlapping data points and potentially lead to bias. The excluded studies are either with multiple models (n = 57), with multiple TL types (n = 14) or with minor models like LeNet (n = 9).

According to Spearman's rank correlation analyses, there were no relevant associations observed between the size of the data set and performance metrics. Data size and AUC (Figure 3.6a, c) showed no relevant correlation ($r_{sp} = 0.05, p = 0.03$). Similarly, only a weak positive trend ($r_{sp} = 0.13, p = 0.17$) could be detected between the size of the dataset and accuracy (Figure 3.6b, d). There was also no association between other variables such as modality, TL type and backbone model. For instance, the data points of models, such as feature extractors that were fitted into optical coherence tomography (OCT) images (purple crosses, Figure 3.6a, b) showed that larger data quantities did not necessarily guarantee better performance. Notably, data points in cross shapes (models as feature extractors) showed decent results even though only a few fully connected layers were being retrained.

3.5 Discussion

In this survey of selected literature, we have summarized 121 research articles applying TL to medical image analysis and found that the most frequently used model was Inception. Inception is a deep model, nevertheless, it consists of the least parameters (Table 3.1) owing to the 1 × 1 filter [LCY14]. This 1 × 1 filter acts as a fully connected layer in Inception and ResNet and it lowers the computational burden to a great degree [Sze+14]. To our surprise, AlexNet and VGG were the next popular models. At first glance, this result seemed counterintuitive because ResNet is a more powerful model with fewer parameters compared to AlexNet or VGG. For instance, ResNet50 achieved a top-5 error of 6.7% on ILSVRC, which

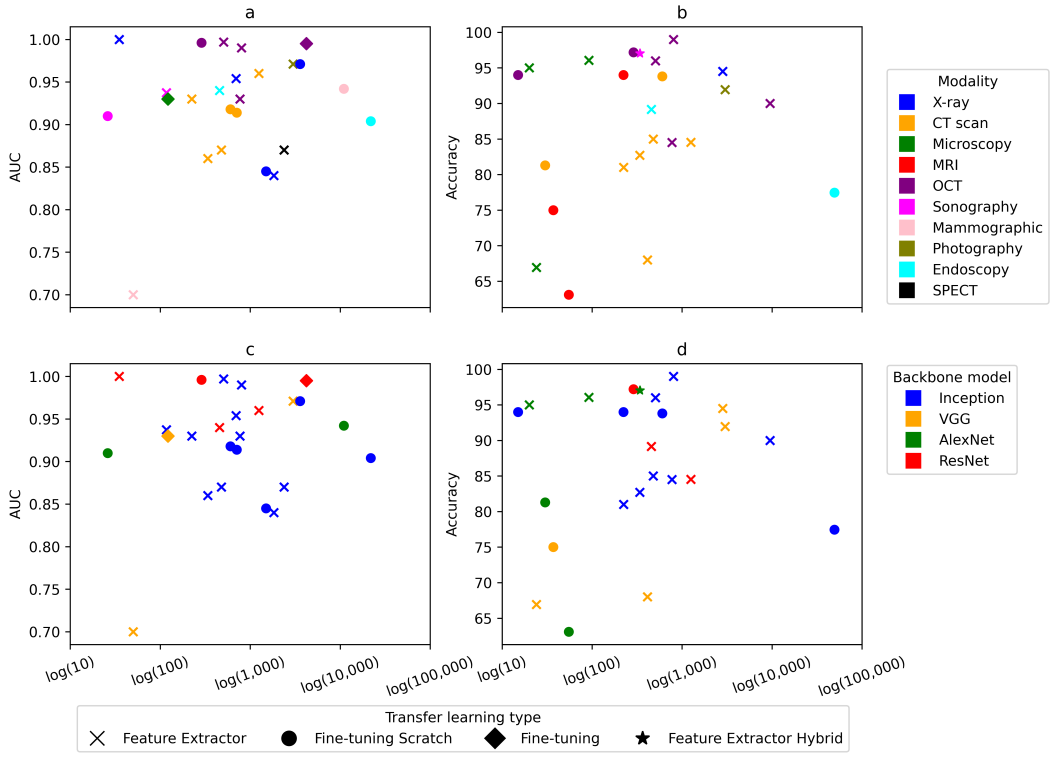

Figure 3.6: Scatter plots of model performance with data size, image modality, backbone model and transfer learning type. Color keys in (a) and (b) indicate the medical image modality, whereas color keys in (c) and (d) represent backbone models. Transfer learning types are in any of four marker shapes for all subfigures.

was 2.6% lower than VGG16 with 5.2 times fewer parameters and 9.7% lower than AlexNet with 2.4 times fewer parameters [He+16]. However, this assumption is valid only if the model was fine-tuned from scratch. The number of parameters significantly drops when the model is utilized as a feature extractor as shown in Table 3.1. He et al. [He+18] performed an in-depth evaluation of the impact of various settings for refining the training of multiple backbone models, focusing primarily on the ResNet architecture. Another assumption was that AlexNet and VGG are easy to understand because the network morphology is linear and made up of stacked layers. This stands against more complex concepts such as skip connections, bottlenecks, convolutional blocks introduced in Inception or ResNet.

With respect to TL approaches, the majority of studies empirically tested as many possible combinations of CNN models with as many as possible TL approaches. Compared to previously suggested best practices [Cho21], some studies determined fine-tuning arbitrarily and ambiguously. For instance, [Hem+20] froze all layers except the last 12 layers without justification, while [Val+19a; Han+18] did not clearly describe the fine-tuning configuration. Lee et al. [Lee+20] partitioned VGG16/19 into 5 blocks, unfroze blocks sequentially and identified the model fine-tuned with two blocks that achieved the highest performance. Similarly, fine-tuned CaffeNet by unfreezing each layer sequentially [Zha+16]. The best results were obtained by the model with one retrained layer for the detection task and with two retrained

layers for the classification task.

Fine-tuning from scratch (n = 27) was a prevalent TL approach in the literature, however, we recommend using this approach carefully for two reasons: firstly, it does not improve the model performance as shown in Figure 3.6 and secondly, it is the computationally most expensive choice because it updates large gradients for entire layers. Therefore, we encourage one to begin with the feature extractor approach, then incrementally fine-tune the convolutional layers. We recommend updating all layers (fine-tuning from scratch), if the feature extractor does not reflect the characteristics of the new medical images.

There was no consensus among studies concerning the global optimum configuration for fine-tuning. [Sin+19] concluded that fine-tuning the last fully connected layers of Inception3, ResNet50, and DenseNet121 outperformed fine-tuning from scratch in all cases. On the other hand, Yu et al. [Yu+19b] found that retraining from scratch of DenseNet201 achieved the highest diagnostic accuracy. We speculate that one of the causes is the variety of data subjects and imaging modalities addressed in Sect. 4.3. Hence, investigating the medical data characteristics (e.g. anatomical sites, imaging modalities, data size, label size and more) and TL with CNN models would be interesting to investigate, yet it is understudied in the current literature. Morid et al. [MBD20] stated that deep CNN models may be more effective for the following image modalities: X-ray, endoscopic and ultrasound images, while shallow CNN models may be optimal for processing these image modalities: OCT and photography for skin lesions and fundus. Nonetheless, more research is needed to further confirm these hypotheses.

TL with random initialization often appeared in the literature [KCC17; Kim+20b; Lee+17; Tan+20]. These studies used the architecture of CNN models only and initialized the training with random weights. One could argue that there is no transfer of knowledge if the entire weights and biases are initialized, but this is still considered as TL in the literature.

It is also worth noting that only a few studies [Zha+18; Xio+19] employed native 3D-CNN. Both studies reported that 3D-CNN outperformed 2D-CNN and 2.5-CNN models, however, Zhang et al. [Zha+18] set the number of the frames to 16 and Xiong et al. [Xio+19] reduced the resolution up to 21*21*21 voxels due to the limitation of computer resources. The majority of the studies constructed 2D-CNN or 2.5D-CNN from 3D inputs. In order to reduce the processing burden, only a sample of image slices from 3D inputs was taken. We expect that the number of studies employing 3D models will increase in the future as high-performance DL is an emerging research topic.

We confirmed (Figure 3.5c) that only a limited amount of data was available in most studies for medical image analysis. Many studies took advantage of using publicly accessible medical datasets from grand challenges (https:// grandchall enge. org/ chall enges). This is a particularly beneficial scientific practice because novel solutions are shared online allowing for better reproducibility. We summarized 78 publicly available medical datasets in Additional file 3: Suppl. Table C (Appendix C), which were organized based on the following five attributes: data modality, anatomical part/region, task type, data name, published year and the link.

Although most evaluated papers included only brief information about their hardware setup, no details were provided about training or test time performance. As most medical data sets are small, usually consumer-grade GPUs in custom workstations or seldom server-grade cards (P100 or V100) were sufficient for TL. Previous survey studies have investigated how DL can be optimized and sped up on GPUs [MV19] or by using specifically designed hardware accelerators like field-programmable gate arrays (FPGA) for neural network inference [Guo+18]. We could not investigate these aspects of efficient TL because execution time was rarely reported in the surveyed literature.

This study is limited to surveying only TL for medical image classification. However, many interesting task-oriented TL studies were published in the past few years, with a particular focus on object detection and image segmentation [Sun+20], as reflected by the amount of public data sets (see also Additional file 3: Appendix C., Table 3). We only investigated off-the-shelf CNN models pretrained on ImageNet and intentionally left out custom CNN architectures, although these can potentially outperform TL-based models on certain tasks [Rah+21; Alz+21]. Also, we did not evaluate aspects of potential model improvements leveraged by the differences of the sourceand the target domain of the training data used for TL [Alz+20]. Similarly, we did not evaluate vision transformers (ViT) [Dos+21], which are emerging for image data analysis. For instance, Liu et al. [Liu+21b] compared 22 backbone models and four ViT models and concluded that one of the ViT models exhibited the highest accuracy trained on cropped cytopathology cell images. Recently, Chen et al. [Che+21] proposed a novel architecture that is a parallel design of MobileNet and ViT, in view of achieving not only more efficient computation but also better model performance.

3.6 Conclusions

We aimed to provide actionable insights to the readers and ML practitioners, on how to select backbone CNN models and tune them properly with consideration of medical data characteristics. While we encourage readers to methodically search for the optimal choice of model and TL setup, it is a good starting point to employ deep CNN models (preferably ResNet or Inception) as feature extractors. We recommend updating only the last fully connected layers of the chosen model on the medical image dataset. In case the model performance needs to be refined, the model should be fine-tuned by incrementally unfreezing convolutional layers from top to bottom layers with a low learning rate. Following these basic steps can save computational costs and time without degrading the predictive power. Finally, publicly accessible medical image datasets were compiled in a structured table describing the modality, anatomical region, task type and publication year as well as the URL for accession.

Chapter 4

Rapid Convolutional Neural Networks for Gram-Stained Image Classification at Inference Time on Mobile Devices

DOI: 10.3390/biomedicines10112808

4.1 Introduction

The number of mobile health (mHealth) apps is growing substantially. The number of mHealth apps in the Google Play store reached over 54,603 in the second quarter of 2022 [Cec22b], while there were 52,406 in the Apple App Store [Cec22a]. According to Roth [Rot13], mHealth apps can be classified into four categories: information apps, which provide a recent trend in healthcare and allow users to find medical information; diagnostic apps, which process data to support physicians in diagnostic decisions; control apps, which control basic functionalities such as the power switch of another medical device; and adapter apps, which adapt smartphones to perform a medical function.

The application developed and evaluated in this study is a diagnostic app, which automates Gram-stained analysis. It is a laboratory procedure that classifies microbial pathogens as either Gram-positive or Gram-negative. It is a promising application in a microbiology laboratory because this task still relies on humans. Physicians and trained medical technical assistants need to navigate the whole slide images manually. This problem can be leveraged by recent advances in deep learning (DL) methodologies, in particular, convolutional neural networks (CNN) which have emerged as the de facto DL methodology in the field of image analysis. For instance, a whole slide image can be distinguished into major species of microorganisms or the position of microorganisms can be highlighted directly on the image. In this manner, the system could enhance the competencies of caregivers with less human intervention. This could lead to rapid initial medical care for patients who suffer from infectious diseases.

However, deploying a DL solution is a non-trivial problem and deploying to resource-limited and battery-powered devices such as smartphones is challenging. For instance, Smith et al. reported that it took 9 min to classify a single whole slide image with a workstation powered by Nvidia GTX 1070 GPU [SKK18]. Moreover, Netflix announced in 2012 that they failed to deploy the winner solution of the "1 million-dollar Netflix Challenge" due to engineering costs of the complex machine learning solution [Ama13]. One of the major obstacles is the computational burden because DL models consist of millions of parameters. For instance, ResNet152 and AlexNet models consist of 60 million parameters and 132 million parameters for the VGG16 model, meanwhile, the latest Google glass enterprise Edition 2 released in May 2019 features only 3 GB of memory, 32 GB of storage and 800 mAh battery capacity, allowing for only 8 h of running time. Accordingly, resource utilization becomes a non-trivial issue because millions of arithmetic operations require longer processing time and drain the battery more quickly. Especially, battery-powered devices (e.g. mobile devices, internet of things and wearable devices) must be carefully considered when DL solutions are developed.

This challenge has led to compact and rapid DL as an emerging topic in recent years. Han [Han17] distinguished four types of research endeavors on this subject based on what and how to speed up DL models. The target to be accelerated is either training time or inference time; on the other hand, it can be achieved by introducing novel hardware or tuning algorithms optimally. Graphics processing unit (GPU) initially developed for accelerating computer graphics is now a core element of server infrastructure for rapid deep learning processing. Google developed an application-specific integrated circuit (ASIC) known as a tensor processing unit (TPU) [Jou+17], which is optimally designed to process deep learning solutions implemented by its own framework, TensorFlow. Within the realm of efficient algorithms, numerous approaches have been proposed; for example, Chollet et al. [Cho17] speeded up the training time with little accuracy degradation by introducing an innovative model architecture with depth-wise separable convolutional neural networks. Smith et al. [Smi+17] and Goyal et al. [Goy+17] shortened training time by applying a large batch size (BS). Numerous normalization approaches [IS15; SK16; Che+18a; Kla+17] and regularization techniques such as early stopping [Pre98] and structure sparsity regularization by suppressing irregular memory access successfully accelerated training time. On the other hand, model compression methods such as pruning [ZG17] and quantization [Jac+18] are able to expedite the inference time. Pruning removes the low-impact parameters incrementally, while quantization scales down the bit representation from 32-bit floating-point numbers to lower-bit representation.

The contributions of this study are as follows: identify the optimal transfer learning configuration of CNN models to Gram-stained image classification; accelerate the inference time by model optimization methods; and deploy and evaluate the execution speed of the optimized models on two Android devices.

4.2 Materials and Methods

4.2.1 Efficient Convolutional Neural Networks

CNN [KSH12] is a class of deep neural networks that are designed to solve various computer vision problems. CNN constitutes fully connected layers and convolutional layers. The former is the classical layer where all neurons are interconnected to one another in adjacent layers, while the latter is a core element of CNN which generates generic feature maps from the previous layer. In terms of computational complexity, the convolutional layer is less expensive because neuron weight sharing reduces the number of connections between neighbor layers.

The pre-trained CNN models are tuned to the efficient models in three steps. The overview is illustrated in Figure 4.1. The first step is transfer learning (TL), which is a technique particularly widely adopted technique for medical image analysis owing to its capability of model adaptation towards new tasks [Kim+22b]. TL is inspired by the learning mechanism, in which the knowledge acquired before can leverage the learning procedure to learn similar tasks. Since TL can reuse weights of pre-trained CNN models, TL is able to reduce the computational burdens to a large degree. Pruning zeros out non-significant connections in neural networks. It gradually eliminates low-impact parameters based on magnitude without decreasing model accuracy. Unlike dropout [Sri+14], it ignores some nodes randomly during the training phase but pruning eliminates model parameters (connections). This attribute makes models require less storage overhead and reduces the memory footprint. Quantization converts 32-bit floating-point numbers to lower-bit representations such as 8-bit integer numbers. An intuitive example of quantization is converting floating-point numbers to integer numbers (e.g. 1.245 to 1). Unlike pruning being applied during the training phase, quantization is a post-production method because it is typically applied during the post-modeling phase.

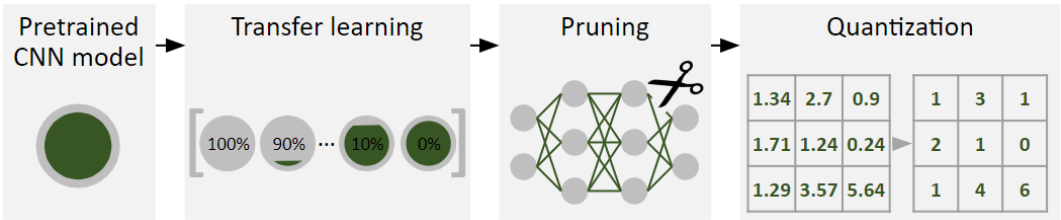

Figure 4.1: A flowchart diagram depicts the process to optimize naïve CNN models to the efficient CNN models. Transfer learning adapts CNN models pre-trained from natural images to the custom image dataset. Pruning trims out non-significant weights while quantization drops floating-point numbers by rounding a given value to the nearest integer number.

4.2.2 Data Set

Eight thousand five hundred Gram-stained images with two labels (positive vs. negative) were taken from sepsis patients who suffered from at least one microbial in-

fection such as Staphylococcus, Escherichia, or Streptococcus. Images with both labels (two types of germs appeared on a single image) were excluded from this study (n = 446) in order to make a binary image classification. Given images were cropped areas of interest containing stained microorganisms from a whole slide microscopy image. The size of the images varied from 800-pixel by 600-pixel to 1920-pixel by 1080-pixel. Exemplary sample images and labels are shown in Figure 4.2.

Gram-positive images were two-fold more frequent (n = 5962) than Gram-negative images (n = 2766). Therefore, class balancing needed to be applied. Otherwise, the models were conditioned to predict the majority labels and abandon the minority class. Hence, Gram-negative images were augmented to balance the class proportion by rotating the given images. After the augmentation, the dataset was enriched from 8728 to 10,994 images. For the sake of a fair evaluation, the test dataset and validation dataset was isolated from the training set. This study split the given data into 80% for training, 10% for validation and 10% for testing.

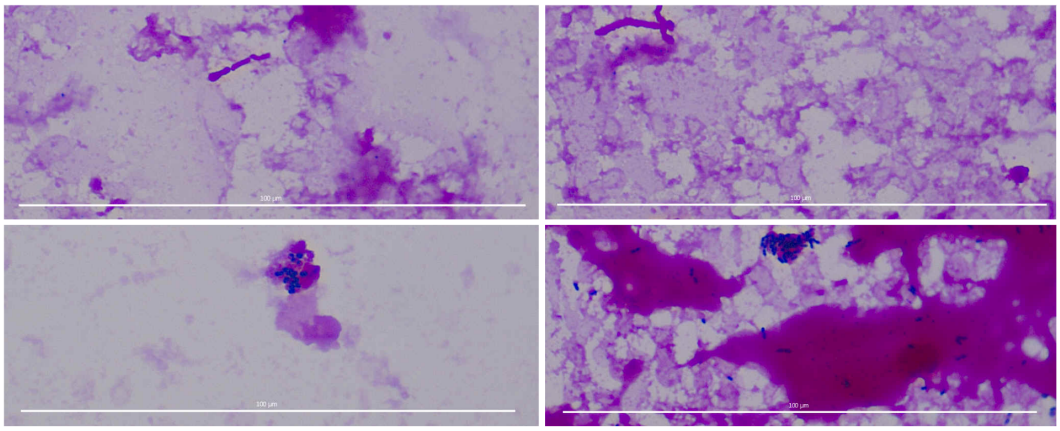

Figure 4.2: Sample images of Gram-stained data. Two Gram-negative images are shown on the top, and two Gram-positive images are shown on the bottom. Some pathogens are distinctive with a high contrast of a clean background whereas often other pathogens are blurred and/or have bloodstains in the background and/or low brightness level. Scale bar represents 100 µm.

4.2.3 Study Design

The machine learning task in this study is binary image classification. The implemented models will predict whether the image is Gram-positive or -negative. Three pre-trained models were utilized in order to avoid model selection bias. They are, namely, Inception [Sze+15], ResNet [He+16], and MobileNet [How+17]. Inception was chosen because it is the most prevalent model utilized in the medical domain according to Morid et al. [MBD20] and Kim et al. [Kim+22b]. Furthermore, ResNet is the most widely used backbone model for other tasks such as object detection and segmentation [Lee+19]. Finally, MobileNet was selected because it was explicitly designed to be deployed to resource-constrained-devices [How+17]. Each

model was calibrated to the Gram-stained analysis and then optimized and evaluated. The consecutive steps performed are as follows: TL; pruning; quantization; and evaluation.

The main objective of TL is to identify the best accuracy setup and others are to reduce the model size and minimize the inference time without accuracy loss. Primarily, pre-trained models were tuned to Gram-stained images because the given models were trained with the ImageNet dataset [Den+09b] containing natural images only. The optimal fine-tuning ratio was determined by exploring numerous configurations. The number of model layers was binned into 10 buckets and each bucket was incrementally fine-tuned from the shallow strategy (feature extraction) to the deep strategy (fine-tuning from scratch). The former strategy is also referred to as feature extraction and it updates no convolutional layers except the fully connected layers, while the latter updates all layers from scratch. This study iteratively walked through 11 different settings from the shallow strategy (re-training 0%) to the deep strategy (re-training 100%).

Once models were transferred to Gram-stained images, one of the model compression methods, pruning was applied. Pruning trims the low impact parameters incrementally. In other words, model parameters were iteratively pruned from 10% up to 90%. Similar to the fine-tuning method, nine target sparsity values were evaluated gradually from the dense model (pruned 10%) to the very sparse model (pruned 90%). Following this, another model compression method, quantization was applied. In this study, we scaled down the default 32-bit representation to three lower bit-schemes, namely 16-bit floating-point numbers, 16-bit mixed numbers (floating and integer) and 8-bit full integer numbers.

4.2.4 Metrics

Accuracy evaluates the quality of models; however, it fails to provide insight into model behaviors when it is deployed to production. Computational costs and model size should be considered especially when it is deployed to resource-constrained devices. Hence, this study evaluated models not only with the classical metrics (accuracy) but also with model size and inference time. For the sake of statistical stability, model accuracy was tested 10 times while inference time was tested 50 times and the average values were reported in this paper.

4.2.5 Apparatus

TensorFlow and TensorFlow Lite were the chosen frameworks for deep learning solutions in this study. Both frameworks are open-source tools developed by the Google Brain team [Aba+16a]. TensorBoard was used as a model-debugging tool and to graphically track all execution history. All models were processed at the data

center of the Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health Baden-Württember, Medical Faculty Mannheim. Regarding the reproducible research, hardware was virtualized by Docker for a controlled development environment. Each container was configured with one Intel

Xeon Silver 4110 CPU, one NVIDIA Tesla V100 32 GB GPU and 189 GB of shared memory. The inference time of the compressed models was evaluated on two android mobile devices: Samsung Galaxy A20E and S10. The quantized models need to be tested on devices with ARM-based CPU and not x86-based CPU workstations because integer arithmetic is optimized for the ARM CPU architecture. The averaged inference time was measured by a C++ binary tool developed by Google via a command-line interface called Android Debug Bridge allowing communication with mobile devices. This study utilized only one CPU thread on mobile devices. All other active processes were deactivated during the testing, and the network state was switched off.

4.3 Results

4.3.1 Transfer Learning

Twelve models (three models with four different batch sizes) were evaluated. Figure 4.3 illustrates the results of the three pre-trained models. Regardless of model and batch size, there was a noticeable trend shown in Figure 4.3 in which accuracy dropped when only a few layers were re-trained (approximately 10 to 20% of the total number of layers of the respective model/architecture), but it recovered when more layers (>50%) were re-trained. The highest accuracy for Inception3 and MobileNet was achieved when the model was re-trained from scratch (100%) with 64 minibatch, while ResNet50 attained the best accuracy with the fine-tuning ratio of 80% and 32 minibatch.

All execution histories were reported in our GibHub repository and they are publicly accessible at: (accessed on 1 November 2022). The average training time for TL was roughly 145 min (220 min for ResNet50, 160 min for Inception 3, and 60 min for MobileNet) when the number of epochs was 100. The exact training time was not reported in this section because the scope of this paper was to compare the inference time.

4.3.2 Pruning

Twenty-seven models (three models with nine different pruning ratios) were pruned and evaluated in this phase. Each setup was trained and tested 10 times and the averaged accuracy values are depicted in Figure 4.4. The result shows that pruning was able to compress the model up to 15 times (Figure 4.4. Bar chart) as compared to the baseline model (0% sparsity) without or with only a minor loss of model accuracy (Figure 4.4, line chart). Only MobileNet (colored in green) with a high sparsity ratio suffered from a substantial decrease in accuracy.

4.3.3 Quantization

The weights and activation of pruned models were converted from 32-bit float to 16-bit float, 16-bit integer and 8-bit integer numbers. Accuracy was not dropped

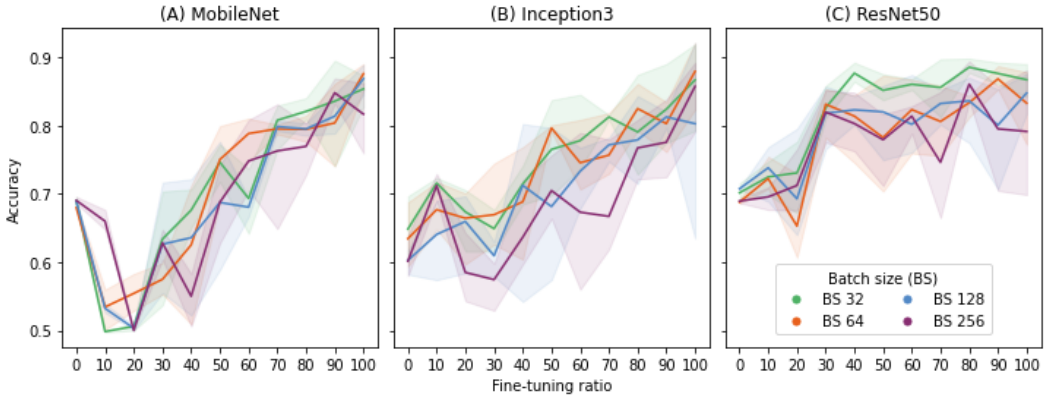

Figure 4.3: Results of test accuracy based on the combination of 11 tuning ratios and four batch sizes (BS). Each setup/ratio has been repeated and tested 10 times for the sake of statistical analysis. The average is shown as a bold line, while the minimum and maximum accuracy are shown as areas in a lighter color.

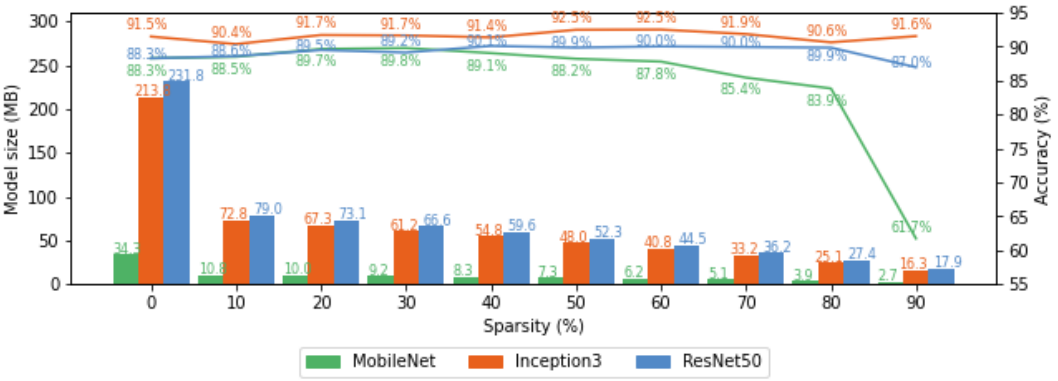

Figure 4.4: Results of models with 0% sparsity (leftmost) are the baseline where pruning was not applied. The bar chart depicts that all pruned models except MobileNet were successfully compressed without sacrificing the model accuracy shown in the line chart. The accuracy deteriorated when MobileNet pruned more than 70%.

for all models despite the model size having been significantly reduced. Figure 4.5 shows that the size of models converted to integer-type was reduced from at least 3 times and up to 4.3 times (Figure 4.5A–C) with accuracy loss at most 1.1% to accuracy gain up to 0.9% (Figure 4.5D–F).

4.3.4 Evaluate Inference Time on Mobile Devices

The three clusters represent different pruning ratios of models from 0% to 50% to 90%, as shown in Figure 4.6 (x-axis). The leftmost cluster is the baseline model to which pruning was not applied. On a cluster-to-cluster basis comparison, there was no remarkable difference among clusters in terms of the inference time. The latency of 50% and 90% sparse models on mobile devices was similar to that of the baseline model.

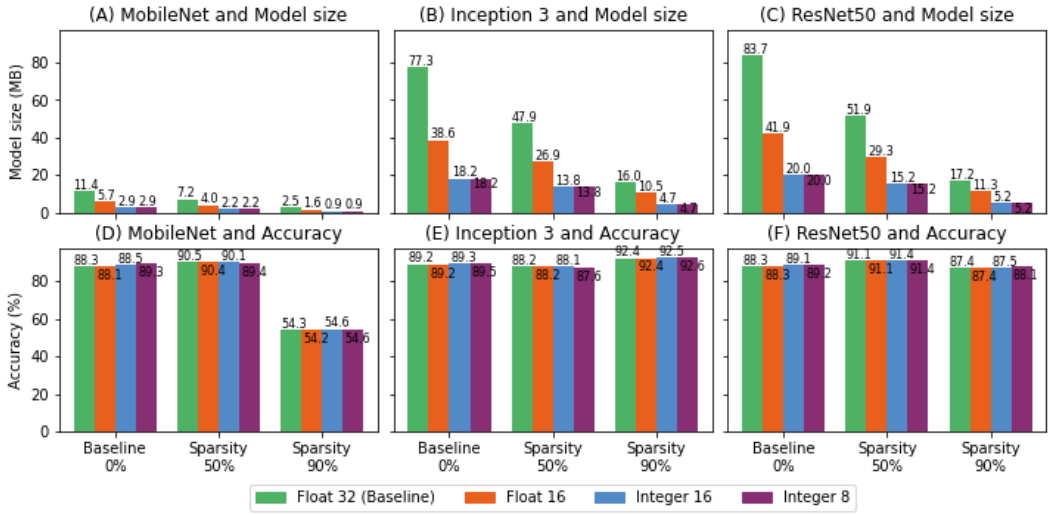

Figure 4.5: Quantization method reduced model size (A–C) with minor accuracy loss (D–F). Models in 16-bit float type were 2 times smaller than the baseline model and models in integer type were 4 times smaller than the baseline model.

Each cluster consists of four bars in four colors indicating different bit schema from float 32, float 16 and integer 16 to integer 8. On a bar-to-bar basis comparison, quantization sped up the inference time to at least 1.9 times to 2.8 times faster. The improvement of the execution time was more distinctive on Galaxy S10 than A20E.

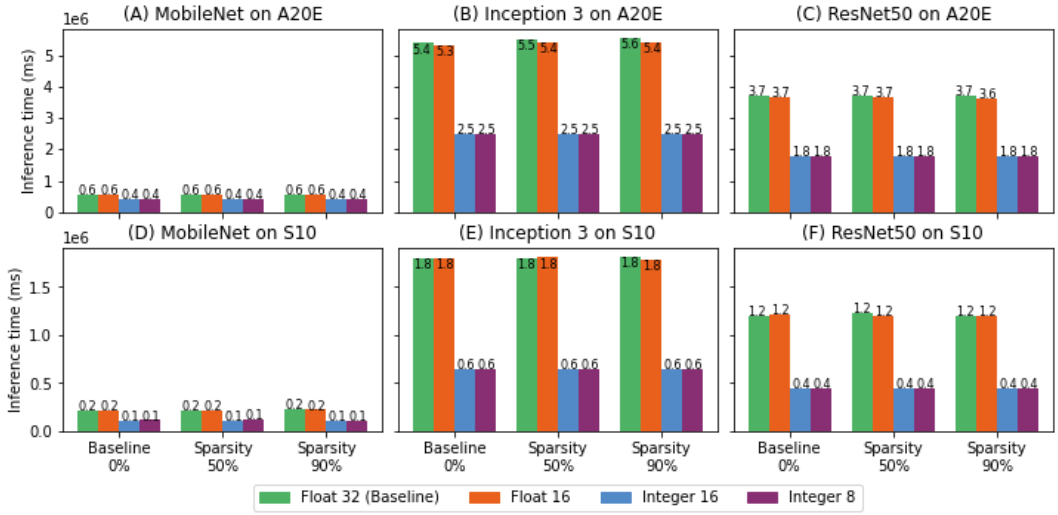

Figure 4.6: Inference time on Galaxy A20E and S10. The latency of integer-type models (blue and purple) is at least 1.9 times faster and at most 2.8 times faster than the float-type models (green and orange).

4.4 Discussion

The performance of the fine-tuning method was not much influenced by batch size. An empirical study by Wilson et al. [WM03] stated that a large batch size leads to a decrease in performance; however, we did not observe a significant accuracy drop in this study. In fact, a large batch size requires fewer iterations to converge the respective model at the expense of using more memory, but it was only a marginal gain (2 to 5 s faster) in training time. On the other hand, the performance was highly sensitive to the fine-tuning ratio due to the heterogeneity of features between Gram-stained images and natural images. In this study, the highest accuracy was attained by re-training convolutional layers by at least 80%. Hence, in order to capture the characteristics of Gram-stained images, we recommend re-training as many model layers as possible.

Model size and accuracy were affected by the pruning method to a large margin as shown in Figure 4.4. Although the model comprised many fewer parameters, pruning did not decrease the model accuracy, except for the MobileNet. The ResNet50 model with 90% fewer parameters was 13 times smaller than the baseline model; nonetheless, the accuracy increased by a small margin. However, the 90% sparse MobileNet suffered from low accuracy as it dropped to 61.7% (Figure 4.4, Line chart). We further investigated with the 90% pruned MobileNet whether the accuracy can be recovered by extending the training steps. For this, we trained the model for 100,000 epochs which took 36 days at the workstation described in the Method-Apparatus section. Figure 4.7 shows that the accuracy recovered from 76% to 83%; however, it would be hard to justify such extensive training for only marginally better accuracy.

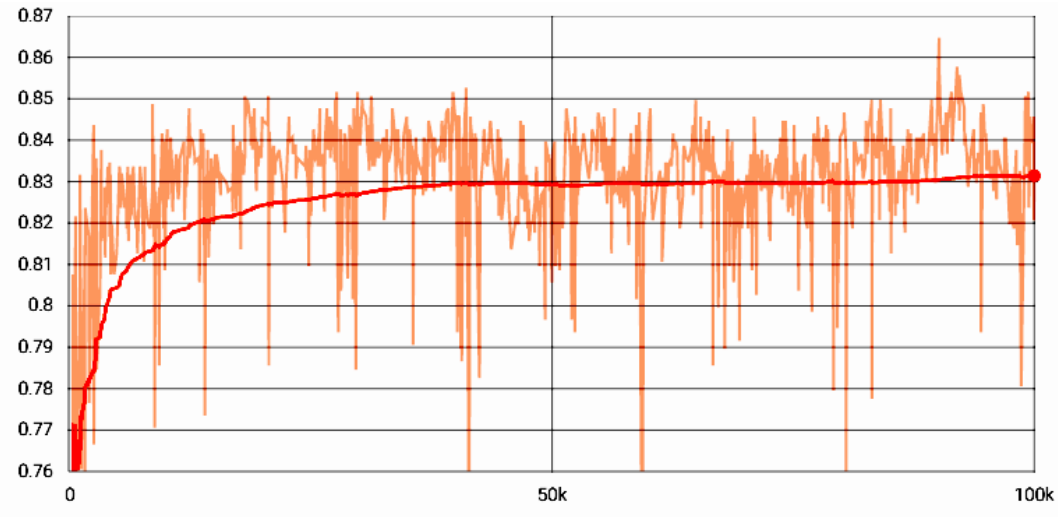

Figure 4.7: MobileNet with 90% sparsity was trained for 100,000 epochs. Accuracy recovered (83%), but it was not as good as the baseline model (88%). The noisy data points were smoothed by a moving average method, which calculates a series of averages of subsets of data points.

Similar to pruning, quantization reduced the model size up to 4.3 times without

losing accuracy. To our surprise, both quantized and pruned models occasionally gained accuracy by a small margin. We assume that removing unnecessary parameters and lowering bit representation might restrict the DL models not to overfit the training dataset. With regard to the inference time, no significant differences were reported among the same data type (e.g. float 32 and float 16; integer 16 and integer 8). It is because the major operations (matrix multiplication and backpropagation) are still carried out using 32-bit in spite of the input and output being quantized into lower bit representation. Matrix multiplications process multiple 8-bit or 16-bit operands that require more bits to process and store. On the other hand, backpropagation with a lower bit could not nudge the subtle updates for weights and biases. Both accumulator and backpropagation are the cornerstone tasks of convolution and therefore require more computational costs.

The inference time on the smartphone Galaxy S10 was more distinctive than the smartphone Galaxy A20E, as illustrated in Figure 4.6. The major reason is cache memory, where data are frequently accessed by the CPU. Unlike the smartphone Galaxy S10, which consists of three cache memories, the smartphone Galaxy A20E does not. Therefore, Galaxy A20E is less efficient, although the size of random access memory (RAM) of Galaxy A20E is large enough to host compressed models.

The integer quantized models had to be evaluated on ARM-based CPU devices (e.g. Android and iPhone devices) because the static execution plan was optimized to the integer arithmetic operator at conversion time. Therefore, when an x86-based CPU workstation attempts to process the quantized models, it conveys irregular computation patterns. For instance, the quantized Inception model at our server with Tesla V100 took more than an hour to process a single image.

We intentionally did not employ a mixture of augmentation because it did not make sense due to the characteristics of Gram-stained images. We refrained from applying scaling or any distortion techniques because magnification on a microscope is already fixed. Cropping is not allowed because it could easily trim out the microorganisms in the images. The color intensity of the images, however, might have been harmonized; nevertheless, we intentionally did not change color at default to increase variability and robustness because the color is the most critical feature for Gram-stained analysis. Finally, employing a mixture of data augments in real time slowed the training time by a large margin.

Deep learning applications on the Internet of things (IoT) for healthcare create many opportunities because they can collect, harmonize and process data from multiple sources in real time. This will support caregivers to provide better treatments with lower costs at the right time. For instance, several successful applications were developed during the COVID19 pandemic. Drew et al. [Dre+20] recruited about 2 million users and predicted geographical hotspots in advance of official public health reports. Alkhodari and Khandoker [AK22] developed COVID detection tools and demonstrated the potential of telehealthcare. However, there are still several challenges that need to be addressed. The disadvantages of IoT are security and privacy concerns due to the lack of holistic information security approaches for the IoT [MT19]. Cloud computing in healthcare has paved the way for rapid and low-cost healthcare services; however, the risk of healthcare data breaches has also been

aroused. According to reference [Seh+20], 3912 data breach cases were confirmed in the healthcare domain from 2005 to 2019 in the United States. Hence, utilizing deep learning compression techniques and processing data in a local device could reduce the risk of data breach because data are not transmitted to the cloud server.

Compressed DL solutions were tested in general purpose devices only (smartphones) in this study. Although the smartphone is one type of device that can host an augmented reality application by overlaying information on the display incorporating a built-in camera, deploying the solution to a body-worn device such as smart glasses would be more intuitive because such a device is able to project information directly through an optical head-mounted display (OHMD). Kim and Choi [KC21] surveyed 57 academic papers on the applications of smart glasses and stated that smart glasses are most often used in the healthcare domain (n = 21, 37%). Evaluating the performance of caregivers with and without augmented reality would be an interesting prospective study. Google glass would be the choice of the device because the models developed in this study could be seamlessly deployed and evaluated on other android devices like Google glass. Beyond Gram-stained image classification, more complex experiments can be conducted. For instance, Zielinski et al. [Zie+17] classified 33 different genera and species of bacteria with 660 images, and genus level image classification can be carried out with the same dataset used in this study. It would be interesting to see how the rapid DL methodology can improve the inference time compared to the published solutions.

4.5 Conclusions

Despite many publications proving the success of DL in medical applications, deploying a DL solution to resource constraint devices is a hard problem. This paper emphasized that DL models must be carefully designed with consideration of resource-limited devices. We investigated a rapid and compact DL model and evaluated the model performance on two mobile devices. The lessons learned and empirical guidelines drawn out of this study are as follows: we observed that the behavior and performance of models heavily rely on the tuning ratio but not on the batch size. For Gram-stained image classification, re-training more convolutional layers achieved higher accuracy. With respect to model compression, plain models were compressed successfully with minor or no accuracy loss. Pruning was the successful element for model size reduction, while inference time was mainly accelerated by quantization.

The philosophy of the collaboration of humans and computers shall be the right path for artificial intelligence (AI) computers that amplify human competencies, not replace them. We anticipate that the rapid AR application of smart glasses or mobile devices can support caregivers for better and faster clinical decisions and can also be used for education purposes or assisting operations.

Chapter 5

Lightweight Visual Transformers Outperform Convolutional Neural Networks for Gram-Stained Image Classification: An Empirical Study

DOI: 10.3390/biomedicines11051333

5.1 Introduction

The progress of deep learning (DL) and artificial intelligence is astonishing, and it attracts numerous researchers and practitioners from multidisciplinary domains. Although tremendous literature regarding DL applications has been published in the medical domain [Kim+22b], it is uncommon that DL applications are actually deployed in the clinical routine. In addition to common issues such as strict medical device regulations [Pit+20], interoperability and responsibility of DL models [AET18], researchers and practitioners also face multiple technical challenges in utilizing DL solutions, e.g. hardware capacities in hospitals are limited, and cloud computing or edge networks are also uncommon because of data privacy concerns [Ryo+13]. Therefore, medical DL applications are often deployed to resource-limited devices, resulting in performance degradation.

Lightweight DL is an especially crucial subject when it comes to infectious diseases. According to Seymour et al. [Sey+17], in-hospital mortality could be lowered if antibiotics were administered within an initial three-hour window of sepsis care, which is remarkably time-sensitive. Gram-stain analysis is a rapid laboratory test that classifies bacterial species into two groups: either Gram-positive or Gram-negative [Coi06]. It aims to shorten the time needed to correctly classify the underlying bacteria in sepsis patients and ultimately aims to decrease the time to a targeted treatment, which is the interval from symptom onset and diagnosis to the application of the therapy of the disease. Although a physician instantly prepares antibiotic therapy for a patient in the practice, precise and rapid identification of an exact microorganism still matters for tailored treatment. This task currently

relies on medical professionals [Cen+22] and it can be partially automated by DL solutions [Kom+18]. At this time, only a few studies utilize DL for Gram-stain analysis. Liu et al. [Liu+21a] utilized six machine-learning algorithms and identified two species of Gram-positive bacteria, *B. megaterium* and *B. cereus*, by harnessing spectral features of Gram-stained images. Smith et al. [SKK18] proposed a classification model by means of a convolutional neural networks (CNN) model, however, the solution took 9 min to classify a whole-slide image comprising 4.1 million pixels. Recently, our research group has demonstrated that by applying pruning and quantization model size (15×) and inference time (3-4×) of CNN can be substantially reduced and accelerated on limited edge devices such as smartphones without sacrificing accuracy [Kim+22a]. However, visual transformers (VT) models, the state-of-the-art methodology in computer vision, have not yet been investigated.

CNN had been the de-facto DL architecture in the computer vision community since AlexNet [KSH17] won the ImageNet Challenge in 2012. However, a marked paradigm shift occurred in 2020 when Google Brain Team introduced vision transformer (ViT) [Dos+20]. In fact, ViT is not a novel model architecture, but it has developed from the standard transformer encoder [14] from the natural language processing (NLP) domain. The performance of transformer models attained higher accuracy compared with the best-performing CNN model (e.g. ResNet [He+16]) on classification. The mechanism for understanding images differs considerably between CNN and VT. CNN captures a certain type of spatial structure present in the given dataset because they utilize spatial inductive biases that allow them to learn the local representations [Rag+21]. Inductive bias is a set of assumptions that can generalize a dataset and does not require large datasets compared with transformerers based models. On the other hand, VT learn global representations by using self-attention mechanisms [Rag+21]. Multiple studies demonstrated that global representations triumph over inductive bias when trained on sufficiently large scales of datasets, as ViT surpassed ResNet with 300 M images [Dos+20].

This paper aimed to provide a guideline to researchers and practitioners on VT model selection as well as optimal model configuration for Gram-stained image classification. For this, six VT models were investigated using target metrics such as accuracy, inference time and model size and were benchmarked against two well-established CNN models. All models were compressed to 8-bit and were interoperable using the ONNX framework.

5.2 Background

A VT comprises three major components, as shown in Figure 5.1. (1) Linear projection takes input images and outputs joint embeddings. It splits images into predetermined-size patches that are flattened to linear patch embeddings added by positional embeddings. The transformer encoder takes these joint embedding vectors as input, also referred to as tokens, and returns the same length of weighted vectors as output. A class embedding is also attached to the input embeddings for the classification. The key element of the encoder is (2) Multi-head self-attention layers (MSA) and it takes three vectors, namely, query, key and value, while “self

" indicates that query, key and value are identical. Attention is a weighted sum of value vectors and the weight is the inner product of a single query vector and a set of key vectors. Multi-head indicates that multiple attention modules process data in parallel. Finally, (3) Multilayer perceptron (MLP) is a fully connected neural network that classifies input images. The corresponding mathematical notation is found in the original paper [Vas+17].

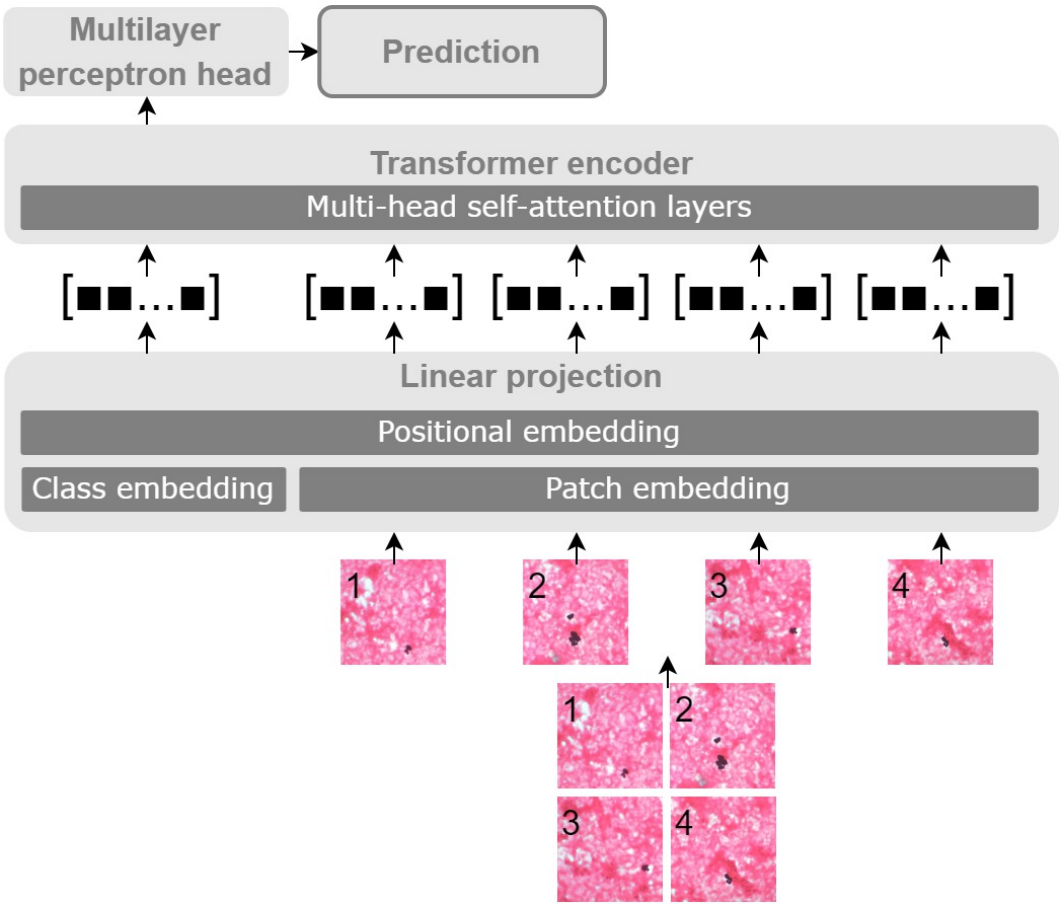

Figure 5.1: Process of a visual transformer where data flow from the bottom to the top. An input image is split into four patches in this figure for visibility. Each patch is encoded into a predefined size of vectors added by positional vectors and class vectors. The class vector is propagated to the multilayer perceptron head for a decision

Since the introduction and great success of the ViT model [Dos+20] by Dosovitskiy et al., numerous VT models and their applications have been proposed in the computer vision community. Despite VT models growing rapidly, they fall into one of five architecture categories and each architecture distinctively differs from one another. ViT is the (1) original VT model and its architecture is identical to the encoder block of the transformer model inherited from Natural Language Processing [Vas+17]. The authors demonstrated that ViT outperformed CNN, however, it required quadratic time complexity with respect to input image size and large data (300 million images) to pre-train. Therefore, many researchers have proposed

innovative architectures to tackle the problem of the ViT model. (2) Multistage models introduced limited size of attention such as localized attention or sparse attention and processed feature vectors gradually and progressively. This mechanism was able to lighten the computational burden and resulted in linear computational complexity. Such an archetype model was the hierarchical vision transformer using shifted windows (Swin) [Liu+21c] introduced by Liu et al. from Microsoft in 2021. Similarly, pyramid ViT (PVT) [Wan+21] and focal transformer models [Yan+21] are hierarchical VT models that introduce spatial reduction attention inspired by CNN's backboned pyramid structure for dense prediction tasks [Lin+17]. More recently, Hassani and Shi proposed a hierarchical VT based on neighborhood attention that can capture a more global context [Lin+18]. (3) Knowledge distillation is another solution that is capable of training VT efficiently. Tourvron et al. from Facebook AI designed data-efficient image transformers (DeiT) that utilized distillation tokens to learn from a teacher agent. On the other hand, Ren et al. introduced a cross inductive bias distillation (CiT) [Ren+22] with an ensemble of multiple lightweight teachers instead of a single heavy and highly accurate teacher agent. Unlike conventional knowledge distillation models that are matching teacher to student in a one-to-one spatial relationship, Lin et al. proposed a one-to-all spatial matching knowledge distillation VT [24], which surpassed other models by a large margin. The (4) self-supervised model was inspired by BERT [Dev+18] and rooted in the NLP domain. It slices a given image into multiple patches referred to as "visual tokens" and randomly drops some patches. The model learns the generic features of images in an unsupervised manner by recovering the eliminated visual tokens. The generative pre-training from pixels (imageGPT) [Che+20] is the same as GPT-2 [Rad+19] except for the activation and normalization layers. It outperformed a supervised model, ResNet. The drawback of imageGPT is the time complexity because its architecture learns images based on pixels instead of image patches. Bidirectional encoder representation from image transformers (BEiT) [Bao+21] is the most cited self-supervised model proposed by Bao et al. in Microsoft. It surpassed imageGPT by a large margin with much fewer parameters while concurrently outperforming two supervised VT models (ViT and DeiT). Finally, (5) hybrid type captures local and global representations by incorporating one or more components from CNN that could save on the computation burden by a large margin [Ram+19]. The idea of integrating inductive bias into global representations attracted numerous researchers. Multiple studies such as BoTNet [Sri+21], CMT [Guo+22], CvT [Wu+21], LeViT [Gra+21] and ViTc [Xia+21] improved accuracy and computational efficiency by combining convolutional layers to the VT model. MobileViT [MR22] was designed by Apple for efficient computation on mobile devices, however, it is more similar to CNN models than VT. Furthermore, models such as PiT [Heo+21] and PoolFormer [Yu+22] achieved competitive results by incorporating pooling layers without attention layers or convolutional layers.

Based on their properties and the Gram-staining classification task at hand, we included the following VT models (BEiT, DeiT, MobileViT, PoolFormer, Swin and ViT) for systematic analyses and evaluation (Figure 5.2).

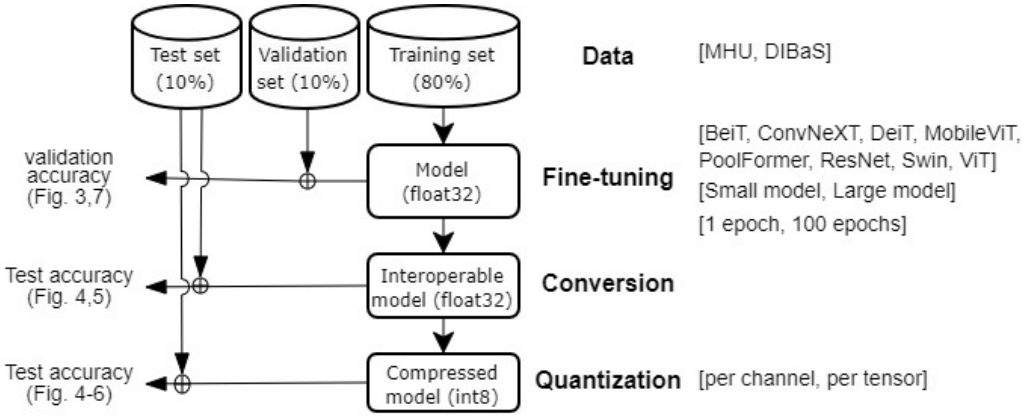

Figure 5.2: Overview of the study design. Eight models with minimum and maximum parameters were fine-tuned to two custom datasets with two epochs strategies during the fine-tuning phase, while each model was quantized either by channel or tensor during the quantization phase. In total, 128 models were evaluated, which is the Cartesian product of eight model architectures with two parameters on two datasets for two epochs and then two quantization schemes.

5.3 Materials and Methods

5.3.1 Data Set

Two Gram-stained image datasets were utilized in this study. One is the domestic dataset from Medical Faculty Mannheim, Heidelberg University (MHU) and the other is a publicly accessible dataset named DIBaS [Zie+17], the acronym for Digital Image of Bacterial Species. The MHU dataset consists of 8500 Gram-stained images collected from 2015 to 2019. The resolution of the images varied from 800 pixels by 600 pixels to 1920 pixels by 1080 pixels. In the given dataset, Gram-positive images (n = 5962) were two times more prevalent than Gram-negative images (n = 2766). On the other hand, the image size of DIBaS is identical to 1532 pixels by 2048 pixels. DIBaS contains only 660 images (20 images for 33 microorganisms) and it is also an unbalanced dataset where Gram-positive images (n = 280) and Gram-negative images (n = 194) are available. Therefore, an oversampling method [ZCL15] was applied to both of the datasets. For the MHU dataset, the number of Gram-negative images increased from 2766 to 5032 by applying rotation, while for the DIBaS dataset, we applied split and/or rotation to both classes and augmented Grampositive images from 280 to 448 and negative images from 194 to 410. The augmented and balanced datasets were split into a training set (80%), a validation set (10%) and a test set (10%). Statistical evaluation methods such as cross-validation are uncommon among AI researchers because they are resource-intensive and time-consuming. Both datasets contain images cropped from whole slide images and contain one microorganism such as Staphylococcus, Escherichia or Streptococcus. The size of the images was rescaled to the same resolution as the pre-trained images (224 × 224 or 256 × 256) during the fine-tuning phase.

5.3.2 Study Design

We examined 128 models by accuracy, inference time and model size. The overview of the study design is shown in Figure 5.2. During the fine-tuning phase, 64 models were re-trained based on the combination of different models, epochs and datasets. Then, each model was compressed by two quantization strategies during the quantization phase. Briefly, eight models with minimum parameters and maximum parameters were fine-tuned to two custom Gram-stained image datasets with two epochs strategies, and then models were quantized either by channel or tensor.

The eight models included six VT models and two CNN models. Each model represents a distinctive architecture, which is summarized in Table 5.1. We chose the most cited model implementation among the same architectures. The two CNNs, ConvNeXT [Liu+22] and ResNet, served as baselines to be compared with VT models. ResNet was chosen because it is known to be a versatile and well-performing CNN architecture on various tasks [Elh+22], while ConvNeXT is a ResNet variation with hyperparameters that are similar to the ViT model. Furthermore, ConvNeXT outperformed the ViT model in a similar study classifying Gram-positive bacteria in a previous study [Liu+22].

Table 5.1: Overview of the eight investigated neural network architectures in alphabetical order.

Model	Architecture traits	Image size	Patch$_a$ size	# Attention heads	# Params (min)	# Params (max)
BEiT	Self-supervised VT	224	16	12; 16	86 M	307 M
ConvNeXT	CNN	224	N/A	N/A	29 M	198 M
DeiT	Knowledge distillation VT	224	16	3; 12	5 M	86 M
MobileViT	Hybrid	256	2	4	1.3 M	5.6 M
PoolFormer	Hybrid	224	7,3,3,3	N/A	11.9 M	73.4 M
ResNet	CNN	224	N/A	N/A	11 M	60 M
Swin	Multi-stage VT	224	4	3,6,12,24; 4,8,16,32	29 M	197 M
ViT	Original VT	224	16	12; 16	86 M	307 M

[a]Patch size and attention heads are shown as a single value unless they differ from the parameters.

All models were pre-trained on the ImageNet-1k dataset, which is a collection of 1.3 million images of subjects such as dogs and cats with 1000 classes. Note that each model can be used in various sizes (e.g. MobileViT-xxs, -xs and -s). They share the same architecture but differ in the number of model components (e.g. attention

heads, encoder blocks, etc.). We examined each model with minimum (small) parameters and maximum (large) parameters. Furthermore, models were re-trained either for a single epoch or 100 epochs to examine the impact of the number of epochs on model accuracy during the fine-tuning phase. Two quantization strategies were applied to the models: (i) either the entire tensor (QT) as a whole or (ii) each channel separately (QC) was quantized from 32-bit float to 8-bit integer representation.

5.3.3 Metrics

The generalization capability of the models was evaluated by accuracy, which cares about the quantity of right or wrong decisions in unseen data. Accuracy is the most employed metric to measure the quality of a classifier, usually defined as true positives + true negatives divided by all samples. The F1-score [SF07] is often employed in conjunction with accuracy as a complementary metric for evaluating classifiers. Accuracy evaluates the quantity of right or wrong outcomes, whereas F1-score is a harmonic mean of precision and recall, which provides insight into whether a model is skewed to a certain class or not. Results of the F1-score are reported in Appendix A.1.

5.3.4 Apparatus

To ensure reproducibility, all our analyses were performed in a containerized environment using a docker. The model tuning and evaluation were conducted in the following virtual environment: One NVIDIA Tesla V100 32 GB GPU was assigned to the docker container and one Intel Xeon Silver 4110 CPU and 189 GB of memory were shared from the host server. HuggingFace Optimum [Wol+20] v1.3.0 was utilized for re-training, model conversion and quantization.

5.4 Results

5.4.1 Fine-Tuning Progress

The history of the fine-tuning progress of all pre-trained models is visualized in Figure 5.3. The purple lines are the history of the models with minimum parameters referred to as “small model”, while the gray lines indicate models with maximum parameters referred to as “large model” respectively. Subplots in Figure 5.3a show that accuracy gradually increased over the learning cycle, especially the accuracy slope of MobileViT, ResNet and ViT, which rapidly gained accuracy compared with other models as the evaluation accuracies at the beginning of the epoch and the last stage of the epoch differ by a large degree on those three models. Moreover, the evaluation accuracy of BEiT and DeiT was depicted as relatively lower than other models during the fine-tuning phase, while ConvNeXT was the highest during the fine-tuning phase. With regard to the model size, the large models demonstrated higher accuracy compared with the small models, except for BEiT and ViT. All models encountered rapid overfitting when they were fine-tuned on the DIBaS data

set (Figure 5.3b). In particular, ConvNext, DeiT and PoolFormer models immediately jumped to 100% validation accuracy regardless of the model size, while other models also attained 100% accuracy at the last epoch.

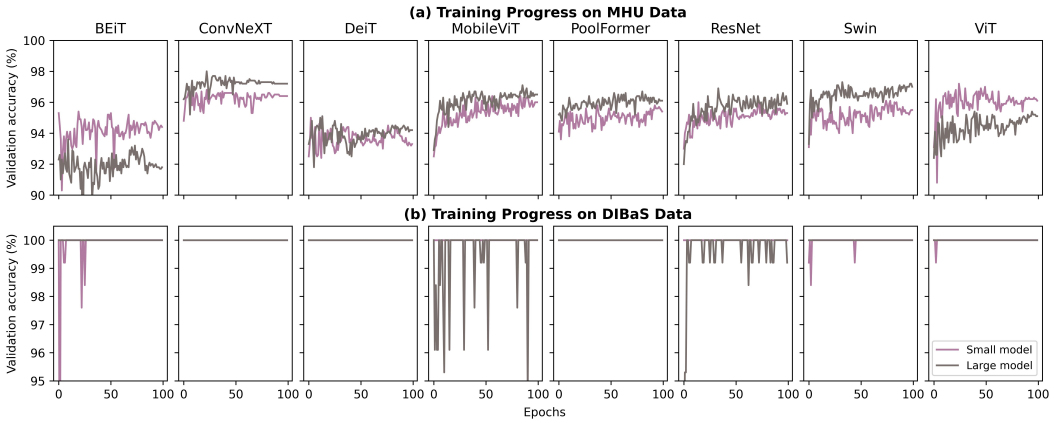

Figure 5.3: Fine-tuning history for 100 epochs on MHU (a) and DIBaS (b) datasets. Subplots are organized in correspondence with the alphabetic order of the model name. Parameters for each model architecture are colored either purple (model with minimum parameters) or gray (model with maximum parameters).

5.4.2 Accuracy and Quantization

The results of models re-trained for one epoch are illustrated in Figure 5.4. On the MHU dataset, the best accuracy was achieved by PoolFormer as follows: 93.2% for 1 epoch and 95.1% for 100 epochs. The highest accuracy on the DIBaS dataset was achieved by BEiT for 1 epoch (95.0%), respectively by ViT for 100 epochs (98.3%). Model accuracies were in the following range: BEiT (84.2–97.8%), ConvNeXT (49.4–92.8%), DeiT (80.6–92.3%), MobileViT (49.4–89.2%), PoolFormer (50.0–95.1%), ResNet (45.8–91.7%), Swin (49.4–93.2%) and ViT (85.7–98.3%). Overall, ViT showed the most well-rounded performance (always $>$85%) in these four settings (Figure 5.4a–d). Large BEiT and DeiT models suffered from performance degradation when undergoing channel-wise quantization (Figure 5.4d). Other models were sensitive to the dataset as they achieved competitive accuracy on the MHU dataset, but not on the DIBaS dataset. In particular, MobileViT large, PoolFormer small, ResNet and Swin small were sensitive to both dataset and epoch as they attained accuracy higher than 87.6% when they were re-trained for 100 epochs on the MHU dataset only

5.4.3 Time, Size and Trade-Offs

We found no difference between model performances on the two datasets (MHU and DIBaS) in terms of inference time regardless of the model architecture and quantization approach (Figure 5.5). However, there were large differences mainly influenced by the model size. Frames per second (FPS) of small models consistently outperformed those of large models by a considerable margin, which was

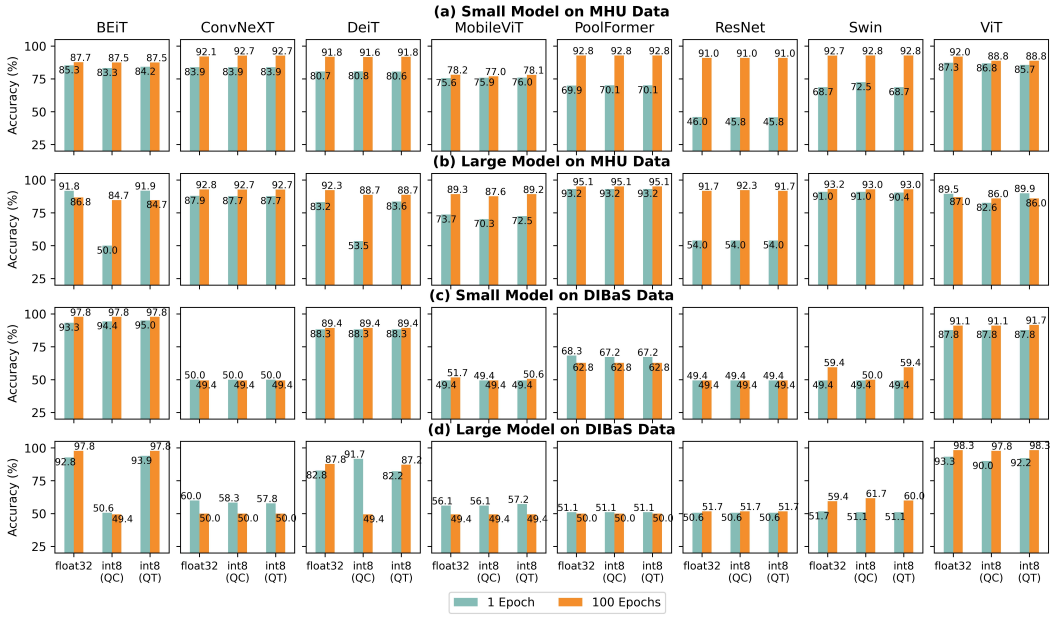

Figure 5.4: Accuracy of eight models with two parameter setups tuned on MHU and DIBaS datasets. Blue bars indicate the models re-trained for one epoch, whereas orange bars are models re-trained for 100 epochs. Subplots in (a,b) are the results on the MHU dataset, while (c,d) are the results on the DIBaS dataset. Models are organized from BEiT to ViT in alphabetic order in the columns with minimum parameters depicted in (a,c), while those with maximum parameters are shown in (b,d). Abbreviations: QC, per-channel quantization; QT, per-tensor quantization. Underlining results indicate the overall best models.

expected by design. The DeiT small model was able to process two times more images than the large model (5.9 images/s vs. 2.9 images/s). Models gained a minor improvement in FPS if they were quantized to integer8. BEiT, ConvNeXT, DeiT, PoolFormer and Swin accelerated 0.2–0.5 FPS, 0–0.5 FPS, 0.3–1 FPS, 0–1.2 FPS and 0–0.3 FPS, respectively. DeiT and ResNet small models were able to process at least five images per second (i.e., the results underlined in Figure 5.5), on the other hand, small BEiT and ViTs could process less than three images per second.

Next, we compared the overall evaluation of model size, accuracy and inference time visually using bubble charts (Figure 5.6). We notice that model accuracies on the MHU dataset outperformed compared to those on the DIBaS as the nodes consistently surpass 80% (yaxis, Figure 5.6a,b), whereas the position of the nodes varied from 50% to 98% accuracy (y-axis, Figure 5.6c,d). On the other hand, FPS was almost identical among similarly sized models, regardless of the datasets (x-axis, Figure 5.6). While the dispersion of FPS of small models (Figure 5.6a,c) was wider than that of the large models (Figure 5.6b,d). With regards to inference time, DeiT and ResNet classified more images than other models as they were consistently plotted on the upper-right quadrant of the plots

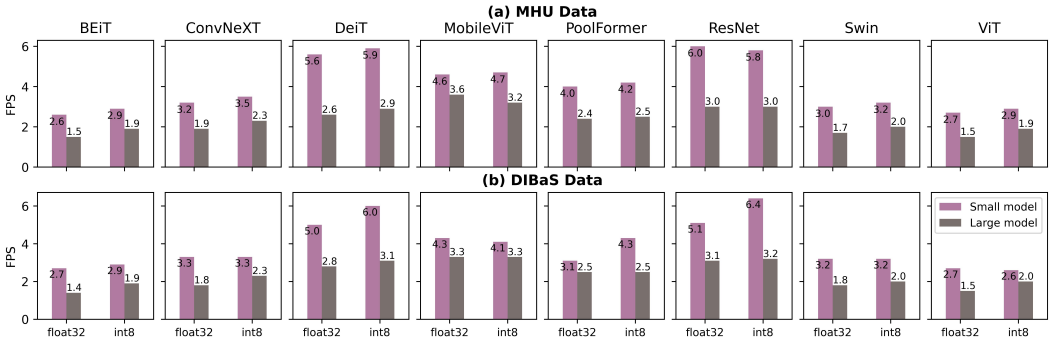

Figure 5.5: Bar charts of model throughputs on MHU (a) and DIBaS (b) datasets. Models are color-coded based on the number of parameters (small (purple) vs. large (gray)) and grouped by their bit representation (float32 vs. int8). The y-axis represents the throughput (inference time) measured as the number of processed frames per second (FPS), while int8 indicates per-tensor quantized models. Underlining results indicate high-throughput models, which can process at least five images per second.

5.5 Discussion

In this study, we performed a comprehensive comparison of six VT models and compared them to two CNN models. We examined their applicability to automated Gram-stained classification. Overall, VT models outperformed CNN models with fewer epochs and on a smaller dataset. Especially, VT models with ViT backbone (i.e., BEiT, Deit and ViT) were outstanding among other models. However, our findings have shown that model performances were determined not only by the model architecture but also by model configuration (e.g. epochs and quantization schemas) and the custom dataset. Hence, we advocate that the model architecture should be empirically determined by considering all of these parameters above.

With regard to the fine-tuning progress shown in Section 4.1, all models highly overfitted the DIBaS dataset. The high validation accuracy (Figure 5.3) did not guarantee high test accuracy (Figure 5.4) as five out of eight models (i.e., ConvNeXT, MobileViT, PoolFormer, ResNet and Swin) made a random guess on the DIBaS dataset for the testing phase (Figure 5.4c,d). We found that deep learning models suffer from the overfitting problem if the available data quantity is $<$1000 images. Regularization techniques (e.g. weight decay, weight normalization and batch normalization) have been previously proven to generalize models and address the overfitting problem. Weight decay [KH91] penalizes a large magnitude of coefficients, while batch normalization [IS15] rescales the layer's input, and similarly, weight normalization [SK16] regulates the magnitude of learnable parameters. In addition to regularization techniques, early stopping [Pre12] of the training process is also a widely applied strategy to avoid overfitting. It ends training if there is no improvement during the training-validation phase.

Both CNNs and VT classifiers achieved better results on a larger dataset (MHU) than on a smaller dataset (DIBaS). We found that BEiT, DeiT and ViT achieved

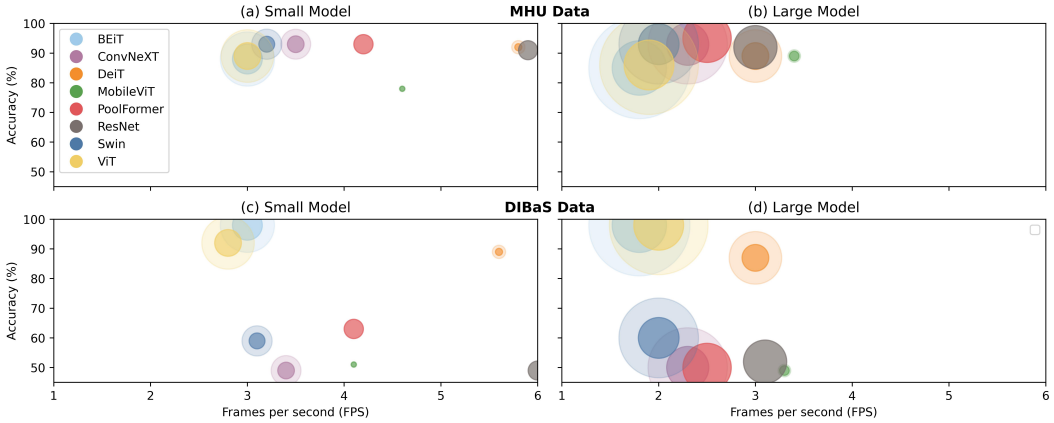

Figure 5.6: Results of accuracy, quantization, inference time and model size of eight models with minimum and maximum parameters as bubble charts on MHU dataset (a,b) and DIBaS dataset (c,d), respectively. The transparency of colors indicates the model quantization where the semi-transparent color represents the float32 models, while the opaque color represents integer8 models.

high accuracies, regardless of the number of epochs and model size. This might be explained by their common backbone model: ViT [Dos+21]. We assume that global information learned by selfattention layers surpasses the value of learning local information by CNN. Architectures combining CNNs and VT showed competitive accuracy results under certain conditions in this study. PoolFormer, which completely lacks an attention layer, showed the best accuracy on the MHU dataset when it was fine-tuned for 100 epochs, however, MobileViT, which consists of three VT blocks while having six CNN blocks and two additional convolutional layers, showed the lowest accuracy performances in average among VT models. BEiT and DeiT suffered from a considerable accuracy drop when they were quantized per channel (QC). These results are counterintuitive to the general belief [Wu+20] as QC is expected to obtain higher accuracies. QC provides a better and more sophisticated prediction because it consists of more parameters to train compared with QT models. It is possible that more intensive quantization made models overfit our custom dataset and failed to generalize. It might also explain the accuracy refinement from float32 models to int8 models, although the improvement was marginal. We assume that removing or reducing the number of model parameters conveyed a similar impact as regularization techniques on a relatively small dataset.

Accuracy is a metric that captures the first impression of models, however, more insights could be gained when used with other metrics such as inference time and model size [49]. They are, in fact, non-trivial aspects of DL models in the context of deployment to resource-limited devices such as mobile devices (smartphones) without dedicated GPU resources. This is especially the case for patients suffering from an infectious disease because minimizing the time to diagnosis and the time to treatment is crucial for them [5]. This study demonstrated that the inference time in FPS units and the throughputs were enhanced on the models with smaller parameters and with the lower-bit presentation, as shown in Figure 5.5. These gains do not seem trivial, however, the optimization solutions can be scaled out when a model

classifies a whole slide image (10k × 6.4k) which is equivalent to 1.3k cropped images (224 × 224). An overview of all results is depicted in Figure 5.6 which might provide a guideline for a model selection on a selected dataset. As public Gram-stained data sets are extremely scarce (besides our local dataset (MHU), we found only one more additional public Gram-stained image dataset (DIBaS)), we could not perform systematic statistical comparisons. For this, most approaches require at least five (ideally) separate data sets to be able to infer non-parametric rank-based statistics [Dem06; Her20].

Our study has other limitations. The scope of this study was limited to the image classification problem, although VT models have also made great progress on different problems such as object detection and segmentation [Kha+22]. Carion et al. from Facebook AI proposed DETR [Car+20] for object detection which consists of both CNN and VT models. YOLOS [Fan+21] is another successful model for object detection inspired by DETR. SegFormer [Xie+21] is a hierarchical transformer encoder and a lightweight perceptron decoder for image segmentation. ViTMAE [He+21] proposed by He et al. is a scalable self-supervised learner for computer vision. It learned the general presentation of images by masking 75% of the image patches and reconstructing the missing pixels. This study covered only Gram-stain image classification. Although we examined several VT models on two Gram-stain datasets, it might not be enough to draw generalizable conclusions about the effectiveness of visual transformers. In fact, numerous research endeavors have uncovered successful VT applications in the medical domain. Shamshad et al. [Sha+22] conducted a comprehensive survey paper recently that summarized studies utilizing VT in the medical domain and over 400 studies were classified based on the problems (e.g. classification, segmentation, registration, etc.) and further categorized by specific tasks (e.g. COVID-19 diagnosis, multi-organ segmentation, etc.) Some researchers have devoted efforts to constructing a novel network architecture or concatenating multiple machine learning models, however, the majority of studies utilize pre-trained transformers models and replace the decision layer for their custom task without modifying the network morphology. The explosive number of publications indicates that VT has permeated every sector of the medical domain and this suggests great potential to develop innovative medical applications. For instance, image-to-text converters have great potential in the medical domain. Tanwani et al. from Google [TBF22] proposed RepsNet which generates automated medical reports in natural language from medical images. Regardless of the advanced architectures of novel deep learning models such as VT, simple statistical methods or shallow ML algorithms often outperform these models or offer at least a sufficient enough performance, especially on limited, medical domain-specific tasks as demonstrated for anomaly detection in neuroimaging [59] and cross-lingual radiological report classification [Mar+21]. Deploying lightweight DL models to an augmented reality (AR) device [Mon+22] also has promising applications. For instance, doctors could wear an AR device during surgery to obtain augmented information on a patient, or they can be utilized for training purposes by taking some guidance from the AR device. Lee et al. [Lee+23] proposed a transformers-based model that classified one of the three body movements by harnessing electroencephalogram (EEG) signal data and graphics simulated by a head-mounted device.

Their model works in a virtual driving environment, and it is feasible to convert it for the medical field in future studies.

5.6 Conclusions

We encourage using VT models for Gram-stained image classification because they could learn the custom images with fewer epochs compared with CNN. With consideration of the model accuracy, models with ViT backbone are recommended as BEiT, DeiT and ViT were outstanding in this study. With regard to the inference time, DeiT small is recommended as the int8 model was able to process six images per second. Finally, the most compact model was MobileViT small, however, we do not recommend using it because of the low accuracy. We recommend the second most compact model, DeiT small in int8, as the accuracy was not degraded regardless of the number of parameters and quantization schemes. Overall, we recommend the DeiT model when we consider test accuracy, inference time and model size for Gram-stained classification. We also advocate using a dataset with 1k or more images, otherwise deep learning models encounter serious overfitting problems. Regarding quantization, per-tensor quantization showed more stable accuracy performances compared with per-channel quantization. We hope this study provides insight to researchers so that they may save time and computational resources in selecting a VT model and determining an optimal configuration, especially for a time-critical application such as Gram-stained image classification.

Chapter 6

Discussion and Outlook

This dissertation argues that accelerating innovation with AI in the medical domain is feasible without using cloud services or cutting-edge infrastructure. It is achieved by a plan involving transfer learning strategies, model selection that considers inference time, model size, and model compression techniques such as pruning and quantization. The central focus lies in the optimal utilization of efficient deep learning, as exemplified by the case study of Gram stain classification.

A comprehensive overview of transfer learning techniques for classification within the medical domain was demonstrated by numerous studies in Chapter 3. Rather than proposing novel network architectures, applying transfer learning emerges as a cost-effective approach that can save computational costs and time without degrading the predictive power. A practical starting point involves reusing pre-trained deep models (e.g. ResNet or Inception) as feature extractors, subsequently updating only the final fully connected layers. In order to achieve higher accuracy, one could gradually unfreeze convolutional layers using a low learning rate to facilitate further model adaptation. The configuration of model adaptation would vary based on the characteristics and quantity of the custom medical dataset.

The insights of the optimal configuration for re-training convolutional neural network (CNN) models and the performance of applying model compression techniques were demonstrated in Chapter 4. The accuracy of the investigated models heavily relies on the tuning ratio. Re-training a greater number of layers resulted in enhanced accuracy for Gram stain image classification. This insight holds promise for other medical datasets, as medical images often bear little resemblance to the datasets used for initial model pre-training. The combination of pruning and quantization demonstrated its effectiveness in reducing model size and inference time while maintaining model quality. Pruning predominantly contributes to reducing model size, while quantization accelerates inference time. These findings underscore the relevance of model compression techniques for the successful deployment of deep learning (DL) solutions on resource-limited devices.

The potential of visual transformers was presented in Chapter 5. Four visual transformer models were fine-tuned for the Gram stain classification task and compared and evaluated alongside two convolutional neural network models. Numerous models were empirically evaluated across diverse conditions using two distinct datasets.

VT models demonstrated an ability to capture features in unseen custom images with fewer epochs compared to CNN models. Furthermore, they consistently outperformed CNN for Gram-stain classification in most settings when dealing with smaller datasets. A comprehensive analysis of trade-offs between model performance metrics, including accuracy, inference time and model size was visualized and the shallow DeiT model, quantized to int8 emerged as the optimal choice due to its capacity to process six images per second without deteriorating accuracy. Concerning general conditions for adapting pre-trained models, reliable model performance was achieved through per-tensor quantization configuration and a dataset with more than 1,000 images.

Numerous research endeavors are addressing the challenges posed by constrained computing environments in the medical domain. Federated learning [Rie+20], for instance, is also an emerging idea and especially captures the attention of researchers in the medical domain. Its mechanism enables it to train models across multiple hospitals behind their firewalls. This mechanism facilitates model training across various hospitals, each protected by its own firewall. By constructing a large concatenated model, often referred to as a global model, utilizing distributed hospital networks. Concerns over data privacy are resolved because federated learning leverages internal data, transmitting only model parameters to the master node while patient data are isolated within their respective isolated data centers. Online federated learning is understudied, where the life cycle of a global model could be automatically adjusted based on its performance across different hospitals. This could preserve the model quality over time by adapting to the evolving data characteristics of each site.

Accelerators aware deep learning emerges as an important research subject. Jain et al. [Jai+20] proposed an augmented compiler approach in order to address the challenges of executing quantized models across diverse hardware with varying types of accelerators. The models tested, namely ResNet, Inception, and MobileNet, align with those discussed in the preceding chapter 4. Moreover, The introduction of the first open-source compilation framework for optimizing deep learning accelerators, Open Neural Network Compiler [Lin+19], initiated by Microsoft, provides an opportunity for researchers to delve into deep learning solutions at the system level. Performance of the inference time varies based on the type of accelerator architecture because the number of arithmetic logic units differs from accelerators (e.g. central processing unit, neural processing unit or digital signal processing) significantly influences the execution plan. During the doctoral study, unexpected computation patterns were also observed when quantized models were deployed on a workstation equipped with 32 cores of x86 CPU and Tesla V100. The processing time for a single image exceeded an hour, while mobile devices with a single ARM CPU achieved execution times of less than 6 ms. A profound understanding of heterogeneous hardware at the system level will help researchers and developers to accurately assess model behaviors during deployment.

Additionally, integrating and harmonizing lightweight AI models and compression techniques is a promising avenue for future research. Mishra et al. [MG23] conducted a comprehensive overview of literature on compressing deep neural networks for IoT applications, while this dissertation focused on the investigation of

two lightweight models, MobileNet and MobileViT. Both models were designed to accelerate inference times, while these models exhibit a tendency to overfit during adaptation. Investigating the dynamic combination of models, compression methods, and medical datasets can be insightful. Utilizing lightweight models could unlock the full potential of efficient deep learning, paving the way for a seamless transition from research outcomes to clinical practices.

There is also a high potential to investigate efficient models tailored for three-dimensional (3D) medical image data [Sin+20] such as computerized tomography, magnetic resonance imaging, and diffusion tensor imaging. These medical images are critical to modern medical practices, however, processing such data with conventional computers is challenging due to the large data volume. Only a few researchers studied 3D image datasets, however, an increase in contributions is anticipated, and augmented reality (AR) devices become more prevalent. Healthcare procedures through AR devices hold promise, yet these devices confront inherent hardware limitations. This subject remains future work to investigate further.

Artificial Intelligence (AI) has already become deeply ingrained in our daily lives, and its influence continues to expand. If adoption is inevitable, we must contemplate the careful integration of AI in the medical domain. However, the benefits of innovative technologies often elude small medical research institutes and healthcare units in developing nations. Hence, efficient DL can empower smaller institutions to develop AI healthcare solutions without acquiring a substantial infrastructure upgrade. This will allow computers to do simple and repetitive medical image analyses, while healthcare professionals spend more time with patients.

Chapter 7

Summary / Zusammenfassung

Summary

Deep learning (DL) and artificial intelligence (AI) are woven into the fabric of our daily lives, and they also hold/have shown promise in the medical domain. Despite numerous studies published in the last decade regarding AI application in medicine, DL models have yet to be widely implemented in daily clinical practice on a large scale. In the face of numerous obstacles on the path to a thriving healthcare AI landscape, this dissertation focuses specifically on technical issues related to constrained hardware resources. To address this problem, in this doctoral thesis, I investigated and demonstrated optimal DL techniques based on the use case of Gram-stain analysis for microorganism identification.

Efficient DL techniques such as transfer learning, pruning and quantization can be employed during model training and deployment strategies should be considered in advance. Particularly, I advocate for applying transfer learning to pre-trained models as feature extractors, as opposed to introducing novel model architectures. For Gram-stain classification, DL models could be compressed and test-time performance could be accelerated without compromising test accuracy or loss. While pruning contributed to the reduction in model size by 15×, quantizing the bit representation from 32-bit to 8-bit led to accelerated inference times by 3×. Taking into the quantization configuration, the findings demonstrated that quantization per channel outperformed tensor-wise quantization for the majority of DL models. This outcome contradicts conventional assumptions, however, intensive quantization may potentially hinder the generalization of DL models. Therefore, the most optimal configuration of DL models should be empirically determined depending on the custom task and data. In the majority of setups, vision transformers (VT) exhibited superior model performance compared to convolutional neural networks (CNN). Notably, among these configurations, DeiT tiny emerged as the fastest VT model in int8 configuration, processing six images per second.

By harnessing the investigated efficient DL techniques including transfer learning, pruning and quantization, this doctoral research might provide valuable insights for AI researchers to accelerate the pace of innovation in the medical domain and pave the way for the seamless integration of AI into everyday healthcare practices.

Zusammenfassung

Deep Learning (DL) und Künstliche Intelligenz (KI) sind fester Bestandteil unseres täglichen Lebens und sind im medizinischen Bereich vielversprechend. Dennoch wurden die im medizinischen Bereich veröffentlichten Studien bisher noch nicht im großen Umfang in die tägliche klinische Praxis umgesetzt. Angesichts zahlreicher Hindernisse auf dem Weg zu einer blühenden KI-Landschaft im Gesundheitswesen konzentriert sich diese Dissertation speziell auf technische Probleme im Zusammenhang mit begrenzten Hardware-Ressourcen. Um dieses Problem anzugehen, habe ich in dieser Doktorarbeit optimale Deep Learning-Techniken untersucht und dargestellt, basierend auf dem Anwendungsfall der Gram-Färbung-Analyse zur Identifizierung von Mikroorganismen.

Effiziente DL-Techniken wie Transferlernen, Pruning und Quantisierung können während der Modell-Trainingsphase nutzen und frühzeitig Einsatzstrategien in Betracht ziehen werden. Insbesondere befürworte Ich die Anwendung des Transferlernens auf vorab trainierten Modellen in Form von Merkmalsextraktoren, anstatt neue Modellarchitekturen einzuführen. Für die Klassifizierung von Gram-Färbungen könnten DL-Modelle komprimiert und die Testzeit-Performance beschleunigt werden, ohne die Testgenauigkeit oder den Verlust zu beeinträchtigen. Während das Pruning zur Verringerung der Modellgröße um das 15-fache beitrug, führte die Quantisierung der Bit-Repräsentation von 32 Bit auf 8 Bit zu beschleunigten Inferenzzeiten um das 3-fache. Unter Berücksichtigung der Quantisierungskonfiguration ergaben die Ergebnisse, dass die Quantisierung pro Kanal die Quantisierung pro Tensor für die Mehrheit der DL-Modelle übertraf, unabhängig davon, ob es sich um vision transformer (VT) oder convolutional neural networks (CNN) handelte. Dieses Ergebnis steht im Gegensatz zur gängigen Annahme, nichtsdestotrotz könnte die intensive Quantisierung die Verallgemeinerung von DL-Modellen für die Gram-Färbung-Klassifizierung potenziell behindern. Daher sollte die optimale Konfiguration von DL-Modellen abhängig von der individuellen Aufgabe und den Daten empirisch bestimmt werden. In den meisten Einstellungen wiesen VT eine überlegenere Modellleistung im Vergleich zu CNN auf. Besonders hervorzuheben ist, dass unter diesen Konfigurationen DeiT tiny als das schnellste VT-Modell in der int8-Konfiguration hervorging und sechs Bilder pro Sekunde verarbeitete.

Durch die Nutzung effizienter DL-Techniken und die Ausarbeitung einer umfassenden Strategie für die Modellbereitstellung werden KI-Forscherinnen und -forscher das Tempo der Innovationen im medizinischen Bereich beschleunigen. Diese Beschleunigung wird voraussichtlich den Weg für die nahtlose Integration von KI in den alltäglichen Gesundheitspraktiken ebnen, wertvolle Unterstützung für medizinische Dienstleister bieten und eine entscheidende Rolle bei der Weiterentwicklung der Patientenversorgung spielen.

Bibliography

[AA20] Saleh Albahli and Waleed Albattah. "Deep Transfer Learning for COVID-19 Prediction: Case Study for Limited Data Problems". In: *Current medical imaging* (2020).

[Aba+16a] Martín Abadi et al. "Tensorflow: A system for large-scale machine learning". In: *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 2016, pp. 265–283.

[Aba+16b] Martín Abadi et al. "Tensorflow: a system for large-scale machine learning." In: *Osdi*. Vol. 16. 2016. Savannah, GA, USA. 2016, pp. 265–283.

[Abi+18] Anas Z. Abidin et al. "Deep transfer learning for characterizing chondrocyte patterns in phase contrast X-Ray computed tomography images of the human patellar cartilage". en. In: *Computers in Biology and Medicine* 95 (Apr. 2018), pp. 24–33. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2018.01.008. URL: https://www.sciencedirect.com/science/article/pii/S0010482518300167 (visited on 05/20/2021).

[Abu+18] Hassan Abu Alhaija et al. "Augmented reality meets computer vision: Efficient data generation for urban driving scenes". In: *International Journal of Computer Vision* 126 (2018), pp. 961–972.

[AET18] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. "Interpretable machine learning in healthcare". In: *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 2018, pp. 559–560.

[Aga+21] Deevyankar Agarwal et al. "Transfer Learning for Alzheimer's Disease through Neuroimaging Biomarkers: A Systematic Review". In: *Sensors* 21.21 (2021). Publisher: Multidisciplinary Digital Publishing Institute, p. 7259.

[Ahm+21] Shakkeel Ahmed et al. "A Deep Learning framework for Interoperable Machine Learning". In: *The First International Conference on AI-ML-Systems*. 2021, pp. 1–7.

[Ahn+18] Jin Mo Ahn et al. "A deep learning model for the detection of both advanced and early glaucoma using fundus photography". In: *PloS one* 13.11 (2018). Publisher: Public Library of Science San Francisco, CA USA, e0207982.

[AK22] Mohanad Alkhodari and Ahsan H Khandoker. "Detection of COVID-19 in smartphone-based breathing recordings: A pre-screening deep learning tool". In: *PLoS One* 17.1 (2022), e0262448.

[Al-+16] Rami Al-Rfou et al. "Theano: A Python framework for fast computation of mathematical expressions". In: *arXiv e-prints* (2016), arXiv–1605.

[Alz+20] Laith Alzubaidi et al. "Towards a better understanding of transfer learning for medical imaging: a case study". In: *Applied Sciences* 10.13 (2020). Publisher: Multidisciplinary Digital Publishing Institute, p. 4523.

[Alz+21] Laith Alzubaidi et al. "Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data". In: *Cancers* 13.7 (2021). Publisher: Multidisciplinary Digital Publishing Institute, p. 1590.

[AM20] Ioannis D. Apostolopoulos and Tzani A. Mpesiana. "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks". In: *Physical and Engineering Sciences in Medicine* 43.2 (2020). Publisher: Springer, pp. 635–640.

[Ama13] Xavier Amatriain. "Big & personal: data and models behind netflix recommendations". In: *Proceedings of the 2nd international workshop on big data, streams and heterogeneous source Mining: Algorithms, systems, programming models and applications*. 2013, pp. 1–6.

[AN07] Arthur Asuncion and David Newman. *UCI machine learning repository*. 2007.

[Ban+18a] Peter Bandi et al. "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge". In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 550–560.

[Ban+18b] T. Banzato et al. "Use of transfer learning to detect diffuse degenerative hepatic diseases from ultrasound images in dogs: a methodological study". In: *The Veterinary Journal* 233 (2018). Publisher: Elsevier, pp. 35–40.

[Ban+19] Tommaso Banzato et al. "Accuracy of deep learning to differentiate the histopathological grading of meningiomas on MR images: A preliminary study". In: *Journal of Magnetic Resonance Imaging* 50.4 (2019). Publisher: Wiley Online Library, pp. 1152–1159.

[Bao+21] Hangbo Bao et al. "Beit: Bert pre-training of image transformers". In: *arXiv preprint arXiv:2106.08254* (2021).

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[BLZ+19] Junjie Bai, Fang Lu, Ke Zhang, et al. *ONNX: Open Neural Network Exchange*. https://github.com/onnx/onnx. 2019.

[Bor+20] Karol Borkowski et al. “Fully automatic classification of breast MRI background parenchymal enhancement using a transfer learning approach”. In: *Medicine* 99.29 (July 2020). ISSN: 0025-7974. DOI: 10.1097/MD.0000000000021243. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7373599/ (visited on 05/20/2021).

[Bur+17] Philippe Burlina et al. “Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis”. In: *Computers in biology and medicine* 82 (2017). Publisher: Elsevier, pp. 80–86.

[Bur+18] Jack Burdick et al. “Rethinking skin lesion segmentation in a convolutional classifier”. In: *Journal of digital imaging* 31 (2018), pp. 435–440.

[Byr+18] Micha\l Byra et al. “Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images”. In: *International journal of computer assisted radiology and surgery* 13.12 (2018). Publisher: Springer, pp. 1895–1903.

[Byr+19] Michal Byra et al. “Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion”. In: *Medical physics* 46.2 (2019). Publisher: Wiley Online Library, pp. 746–755.

[Car+20] Nicolas Carion et al. *End-to-End Object Detection with Transformers*. arXiv:2005.12872 [cs]. May 2020. DOI: 10.48550/arXiv.2005.12872. URL: http://arxiv.org/abs/2005.12872 (visited on 04/02/2023).

[Cec22a] L. Ceci. *Healthcare apps available Apple App Store 2022*. en. 2022. URL: https://www.statista.com/statistics/779910/health-apps-available-ios-worldwide/ (visited on 08/26/2022).

[Cec22b] L. Ceci. *Healthcare apps available Google Play 2022*. en. 2022. URL: https://www.statista.com/statistics/779919/health-apps-available-google-play-worldwide/ (visited on 08/26/2022).

[Cen+22] Franz-Simon Centner et al. “Comparative Analyses of the Impact of Different Criteria for Sepsis Diagnosis on Outcome in Patients with Spontaneous Subarachnoid Hemorrhage”. In: *Journal of Clinical Medicine* 11.13 (2022). Publisher: MDPI, p. 3873.

[Che+18a] Zhao Chen et al. “Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks”. In: *International Conference on Machine Learning*. PMLR, 2018, pp. 794–803.

[Che+18b] Phillip M. Cheng et al. “Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks”. en. In: *Abdominal Radiology* 43.5 (May 2018), pp. 1120–1127. ISSN: 2366-0058. DOI: 10.1007/s00261-017-1294-1. URL: https://doi.org/10.1007/s00261-017-1294-1 (visited on 05/20/2021).

[Che+19a] Chia-Hung Chen et al. “Computer-aided diagnosis of endobronchial ultrasound images using convolutional neural network”. In: *Computer methods and programs in biomedicine* 177 (2019). Publisher: Elsevier, pp. 175–182.

[Che+19b] Quan Chen et al. “A transfer learning approach for malignant prostate lesion detection on multiparametric MRI”. In: *Technology in cancer research & treatment* 18 (2019), p. 1533033819858363.

[Che+19c] Chi-Tung Cheng et al. “Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs”. In: *European radiology* 29.10 (2019). Publisher: Springer, pp. 5469–5477.

[Che+20] Mark Chen et al. “Generative pretraining from pixels”. In: *International conference on machine learning*. PMLR, 2020, pp. 1691–1703.

[Che+21] Yinpeng Chen et al. “Mobile-Former: Bridging MobileNet and Transformer”. In: *arXiv:2108.05895 [cs]* (Dec. 2021). arXiv: 2108.05895. URL: http://arxiv.org/abs/2108.05895 (visited on 02/03/2022).

[Chi+17] Jianning Chi et al. “Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network”. In: *Journal of digital imaging* 30.4 (2017). Publisher: Springer, pp. 477–486.

[Cho+17] Joon Yul Choi et al. “Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database”. In: *PloS one* 12.11 (2017). Publisher: Public Library of Science San Francisco, CA USA, e0187336.

[Cho+18] François Chollet et al. “Keras: The python deep learning library”. In: *Astrophysics source code library* (2018), ascl–1806.

[Cho+19a] Bum-Joo Cho et al. “Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network”. In: *Endoscopy* 51.12 (2019). Publisher: \copyright Georg Thieme Verlag KG, pp. 1121–1129.

[Cho+19b] Naweed I. Chowdhury et al. “Automated classification of osteomeatal complex inflammation on computed tomography using convolutional neural networks”. In: *International forum of allergy & rhinology*. Vol. 9. Issue: 1. Wiley Online Library, 2019, pp. 46–52.

[Cho+21] Alexander Chowdhury et al. “Applying Self-Supervised Learning to Medicine: Review of the State of the Art and Medical Implementations”. In: *Informatics*. Vol. 8. Issue: 3. Multidisciplinary Digital Publishing Institute, 2021, p. 59.

[Cho17] François Chollet. “Xception: Deep learning with depthwise separable convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[Cho21] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[CHS20] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. “Computer-aided diagnosis in the era of deep learning”. In: *Medical physics* 47.5 (2020), e218–e227.

[Cir+19] Marco Domenico Cirillo et al. “Time-independent prediction of burn depth using deep convolutional neural networks”. In: *Journal of Burn Care & Research* 40.6 (2019). Publisher: Oxford University Press US, pp. 857–863.

[Cla+20] Kadie Clancy et al. “Deep Learning Pre-training Strategy for Mammogram Image Classification: an Evaluation Study”. en. In: *Journal of Digital Imaging* 33.5 (Oct. 2020), pp. 1257–1265. ISSN: 0897-1889, 1618-727X. DOI: 10.1007/s10278-020-00369-3. URL: http://link.springer.com/10.1007/s10278-020-00369-3 (visited on 05/20/2021).

[CM17] Phillip M. Cheng and Harshawn S. Malhi. “Transfer Learning with Convolutional Neural Networks for Classification of Abdominal Ultrasound Images”. en. In: *Journal of Digital Imaging* 30.2 (Apr. 2017), pp. 234–243. ISSN: 0897-1889, 1618-727X. DOI: 10.1007/s10278-016-9929-2. URL: http://link.springer.com/10.1007/s10278-016-9929-2 (visited on 05/17/2021).

[Coi06] Richard Coico. “Gram staining”. In: *Current protocols in microbiology* 1 (2006). Publisher: Wiley Online Library, A–3C.

[CZA18] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. “Deep convolutional neural networks for breast cancer screening”. In: *Computer methods and programs in biomedicine* 157 (2018). Publisher: Elsevier, pp. 19–30.

[DA19] S. Deepak and P. M. Ameer. “Brain tumor classification using deep CNN features via transfer learning”. In: *Computers in biology and medicine* 111 (2019). Publisher: Elsevier, p. 103345.

[Dem06] Janez Demšar. “Statistical comparisons of classifiers over multiple data sets”. In: *The Journal of Machine learning research* 7 (2006). Publisher: JMLR. org, pp. 1–30.

[Den+09a] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[Den+09b] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[Dev+18] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[Dev+21] Liton Devnath et al. “Automated detection of pneumoconiosis with multilevel deep features learned from chest X-Ray radiographs”. eng. In: *Computers in Biology and Medicine* 129 (Feb. 2021), p. 104125. ISSN: 1879-0534. DOI: 10.1016/j.compbiomed.2020.104125.

[Dos+20] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. en. In: (Oct. 2020). DOI: 10.48550/arXiv.2010.11929. URL: https://arxiv.org/abs/2010.11929v2 (visited on 02/01/2023).

[Dos+21] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *arXiv:2010.11929 [cs]* (June 2021). arXiv: 2010.11929. URL: http://arxiv.org/abs/2010.11929 (visited on 02/01/2022).

[Dre+20] David A Drew et al. “Rapid implementation of mobile technology for real-time epidemiology of COVID-19”. In: *Science* 368.6497 (2020), pp. 1362–1367.

[DT05] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. Ieee, 2005, pp. 886–893.

[DYO19] Awwal Muhammad Dawud, Kamil Yurtkan, and Huseyin Oztoprak. “Application of deep learning in neuroradiology: Brain haemorrhage classification using transfer learning”. In: *Computational intelligence and neuroscience* 2019 (2019). Publisher: Hindawi.

[Elh+22] Omar Elharrouss et al. “Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches”. In: *arXiv preprint arXiv:2206.08016* (2022).

[Fan+21] Yuxin Fang et al. *You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection*. arXiv:2106.00666 [cs]. Oct. 2021. URL: http://arxiv.org/abs/2106.00666 (visited on 04/02/2023).

[Fuk+19] Ryohei Fukuma et al. “Prediction of IDH and TERT promoter mutations in low-grade glioma from magnetic resonance images using a convolutional neural network”. In: *Scientific reports* 9.1 (2019). Publisher: Nature Publishing Group, pp. 1–8.

[Gan+16] Yaroslav Ganin et al. “Domain-adversarial training of neural networks”. In: *The journal of machine learning research* 17.1 (2016). Publisher: JMLR. org, pp. 2096–2030.

[Gao+20] Jun Gao et al. “Lung Nodule Detection using Convolutional Neural Networks with Transfer Learning on CT Images.” In: *Combinatorial Chemistry & High Throughput Screening* (2020).

[GBC16a] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. Chap. 6, pp. 180–184.

[GBC16b] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. Chap. 14.

[Ges+18] Nils Gessert et al. “Automatic plaque detection in IVOCT pullbacks using convolutional neural networks”. In: *IEEE transactions on medical imaging* 38.2 (2018). Publisher: IEEE, pp. 426–434.

[Gib+21] Brandon Gibson et al. “Vulnerability in massive api scraping: 2021 linkedin data breach”. In: *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE. 2021, pp. 777–782.

[Góm+19] Juan J. Gómez-Valverde et al. “Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning”. In: *Biomedical optics express* 10.2 (2019). Publisher: Optical Society of America, pp. 892–913.

[Goo+14a] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[Goo+14b] Ian J. Goodfellow et al. “Generative Adversarial Networks”. In: *arXiv:1406.2661 [cs, stat]* (June 2014). arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661 (visited on 01/07/2022).

[Gor+19] Yury Gorbachev et al. “Openvino deep learning workbench: Comprehensive analysis and tuning of neural networks inference”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019.

[Goy+17] Priya Goyal et al. “Accurate, large minibatch sgd: Training imagenet in 1 hour”. In: *arXiv preprint arXiv:1706.02677* (2017).

[Gra+21] Ben Graham et al. *LeViT: a Vision Transformer in ConvNet’s Clothing for Faster Inference*. arXiv:2104.01136 [cs]. May 2021. URL: http://arxiv.org/abs/2104.01136 (visited on 03/28/2023).

[Guo+18] Kaiyuan Guo et al. “A Survey of FPGA-Based Neural Network Accelerator”. In: *arXiv:1712.08934 [cs]* (Dec. 2018). arXiv: 1712.08934. URL: http://arxiv.org/abs/1712.08934 (visited on 02/01/2022).

[Guo+22] Jianyuan Guo et al. *CMT: Convolutional Neural Networks Meet Vision Transformers*. arXiv:2107.06263 [cs]. June 2022. URL: http://arxiv.org/abs/2107.06263 (visited on 03/28/2023).

[Had+20] Mehdi Hadj Saïd et al. “Development of an Artificial Intelligence Model to Identify a Dental Implant from a Radiograph.” In: *International Journal of Oral & Maxillofacial Implants* 35.6 (2020).

[Han+18] Seung Seog Han et al. “Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network”. In: *PloS one* 13.1 (2018). Publisher: Public Library of Science San Francisco, CA USA, e0191493.

[Han17] Song Han. “Efficient methods and hardware for deep learning”. PhD Thesis. Stanford University, 2017.

[He+16] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[He+18] Tong He et al. “Bag of Tricks for Image Classification with Convolutional Neural Networks”. In: *arXiv:1812.01187 [cs]* (Dec. 2018). arXiv: 1812.01187. URL: http://arxiv.org/abs/1812.01187 (visited on 02/03/2022).

[He+19] Yifeng He et al. “Convolutional neural network to predict the local recurrence of giant cell tumor of bone after curettage based on presurgery magnetic resonance images”. In: *European radiology* 29.10 (2019). Publisher: Springer, pp. 5441–5451.

[He+21] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. arXiv:2111.06377 [cs] version: 2. Dec. 2021. URL: http://arxiv.org/abs/2111.06377 (visited on 04/02/2023).

[Heg+19] Roopa B. Hegde et al. “Feature extraction using traditional image processing and convolutional neural network methods to classify white blood cells: a study”. eng. In: *Australasian Physical & Engineering Sciences in Medicine* 42.2 (June 2019), pp. 627–638. ISSN: 1879-5447. DOI: 10.1007/s13246-019-00742-9.

[Hei+20] Morteza Heidari et al. “Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms”. In: *International journal of medical informatics* 144 (2020). Publisher: Elsevier, p. 104284.

[Hem+20] Ruben Hemelings et al. “Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning”. In: *Acta ophthalmologica* 98.1 (2020). Publisher: Wiley Online Library, e94–e100.

[Heo+21] Byeongho Heo et al. *Rethinking Spatial Dimensions of Vision Transformers*. arXiv:2103.16302 [cs]. Aug. 2021. URL: http://arxiv.org/abs/2103.16302 (visited on 03/28/2023).

[Her20] Steffen Herbold. "Autorank: A python package for automated ranking of classifiers". In: *Journal of Open Source Software* 5.48 (2020), p. 2173.

[Het+17] Jorden Hetherington et al. "SLIDE: automatic spine level identification system using a deep convolutional neural network". en. In: *International Journal of Computer Assisted Radiology and Surgery* 12.7 (July 2017), pp. 1189–1198. ISSN: 1861-6429. DOI: 10.1007/s11548-017-1575-8. URL: https://doi.org/10.1007/s11548-017-1575-8 (visited on 05/20/2021).

[HLG16] Benjamin Q. Huynh, Hui Li, and Maryellen L. Giger. "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks". In: *Journal of Medical Imaging* 3.3 (2016). Publisher: International Society for Optics and Photonics, p. 034501.

[How+17] Andrew G. Howard et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (2017).

[Hua+20a] J Huang et al. "An artificial intelligence algorithm that differentiates anterior ethmoidal artery location on sinus computed tomography scans". In: *The Journal of Laryngology & Otology* 134.1 (2020), pp. 52–55.

[Hua+20b] Kai Huang et al. "Assistant Diagnosis of Basal Cell Carcinoma and Seborrheic Keratosis in Chinese Population Using Convolutional Neural Network". In: *Journal of healthcare engineering* 2020 (2020). Publisher: Hindawi.

[Hut+18] Mikko J. Huttunen et al. "Automated classification of multiphoton microscopy images of ovarian tissue using deep learning". In: *Journal of biomedical optics* 23.6 (2018). Publisher: International Society for Optics and Photonics, p. 066002.

[HW90] Dong-Chen He and Li Wang. "Texture unit, texture spectrum, and texture analysis". In: *IEEE transactions on Geoscience and Remote Sensing* 28.4 (1990). Publisher: IEEE, pp. 509–512.

[HYL17] Will Hamilton, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs". In: *Advances in neural information processing systems* 30 (2017).

[IH18] Jim Isaak and Mina J Hanna. "User data privacy: Facebook, Cambridge Analytica, and privacy protection". In: *Computer* 51.8 (2018), pp. 56–59.

[ILL67] Alekse Grigorevich Ivakhnenko, Valentin Grigorevich Lapa, and Valentin Grigorevich Lapa. *Cybernetics and forecasting techniques*. Vol. 8. American Elsevier Publishing Company, 1967.

[IS15] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[Jac+18] Benoit Jacob et al. “Quantization and training of neural networks for efficient integer-arithmetic-only inference”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2704–2713.

[Jai+20] Animesh Jain et al. “Efficient execution of quantized deep learning models: A compiler approach”. In: *arXiv preprint arXiv:2006.10226* (2020).

[Jan+20] Joel Janai et al. “Computer vision for autonomous vehicles: Problems, datasets and state of the art”. In: *Foundations and Trends® in Computer Graphics and Vision* 12.1–3 (2020), pp. 1–308.

[Jia+19] Licheng Jiao et al. “A survey of deep learning-based object detection”. In: *IEEE access* 7 (2019), pp. 128837–128868.

[Jou+17] Norman P. Jouppi et al. “In-datacenter performance analysis of a tensor processing unit”. In: *Proceedings of the 44th annual international symposium on computer architecture*. 2017, pp. 1–12.

[Kaj+18] Tomohiro Kajikawa et al. “Automated prediction of dosimetric eligibility of patients with prostate cancer undergoing intensity-modulated radiation therapy using a convolutional neural network”. In: *Radiological physics and technology* 11.3 (2018). Publisher: Springer, pp. 320–327.

[Kan+20] Fahdi Kanavati et al. “Weakly-supervised learning for lung carcinoma classification using deep learning”. en. In: *Scientific Reports* 10.1 (Dec. 2020), p. 9297. ISSN: 2045-2322. DOI: 10.1038/s41598-020-66333-x. URL: http://www.nature.com/articles/s41598-020-66333-x (visited on 05/19/2021).

[Kat+19] Jakob Nikolas Kather et al. “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study”. en. In: *PLOS Medicine* 16.1 (Jan. 2019). Publisher: Public Library of Science, e1002730. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002730. URL: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002730 (visited on 05/19/2021).

[KC21] Dawon Kim and Yosoon Choi. “Applications of smart glasses in applied sciences: A systematic review”. In: *Applied Sciences* 11.11 (2021), p. 4956.

[KCC17] Sri Phani Krishna Karri, Debjani Chakraborty, and Jyotirmoy Chatterjee. "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration". In: *Biomedical optics express* 8.2 (2017), pp. 579–592.

[KH91] Anders Krogh and John Hertz. "A simple weight decay can improve generalization". In: *Advances in neural information processing systems* 4 (1991).

[Kha+22] Salman Khan et al. "Transformers in vision: A survey". In: *ACM computing surveys (CSUR)* 54.10s (2022). Publisher: ACM New York, NY, pp. 1–41.

[Kim+20a] Jong-Eun Kim et al. "Transfer learning via deep neural networks for implant fixture system classification using periapical radiographs". In: *Journal of clinical medicine* 9.4 (2020). Publisher: Multidisciplinary Digital Publishing Institute, p. 1117.

[Kim+20b] Young-Gon Kim et al. "Effectiveness of transfer learning for enhancing tumor classification with a convolutional neural network on frozen sections". In: *Scientific Reports* 10.1 (2020). Publisher: Nature Publishing Group, pp. 1–9.

[Kim+22a] Hee E. Kim et al. "Rapid Convolutional Neural Networks for Gram-Stained Image Classification at Inference Time on Mobile Devices: Empirical Study from Transfer Learning to Optimization". In: *Biomedicines* 10.11 (2022). Publisher: MDPI, p. 2808.

[Kim+22b] Hee E. Kim et al. "Transfer learning for medical image classification: a literature review". In: *BMC medical imaging* 22.1 (2022). Publisher: Springer, pp. 1–13.

[Kla+17] Günter Klambauer et al. "Self-normalizing neural networks". In: *Proceedings of the 31st international conference on neural information processing systems*. 2017, pp. 972–981.

[KM18] D. H. Kim and T. MacKinnon. "Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks". In: *Clinical radiology* 73.5 (2018). Publisher: Elsevier, pp. 439–445.

[KNH10] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. *Cifar-10 (canadian institute for advanced research)*. 2010.

[Kom+18] Matthieu Komorowski et al. "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care". In: *Nature medicine* 24.11 (2018). Publisher: Nature Publishing Group US New York, pp. 1716–1720.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (visited on 08/05/2020).

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[KWT18] Daniel H. Kim, Huub Wit, and Mark Thurston. "Artificial intelligence in the diagnosis of Parkinson's disease from ioflupane-123 single-photon emission computed tomography dopamine transporter scans using transfer learning". In: *Nuclear medicine communications* 39.10 (2018). Publisher: Wolters Kluwer, pp. 887–893.

[Lak17] Paras Lakhani. "Deep convolutional neural networks for endotracheal tube position and X-ray image classification: challenges and opportunities". In: *Journal of digital imaging* 30 (2017), pp. 460–468.

[LCY14] Min Lin, Qiang Chen, and Shuicheng Yan. "Network In Network". In: *arXiv:1312.4400 [cs]* (Mar. 2014). arXiv: 1312.4400. URL: http://arxiv.org/abs/1312.4400 (visited on 02/07/2022).

[Lec+98] Y. Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (Nov. 1998). Conference Name: Proceedings of the IEEE, pp. 2278–2324. ISSN: 1558-2256. DOI: 10.1109/5.726791.

[LeC98] Yann LeCun. "The MNIST database of handwritten digits". In: *http://yann. lecun. com/exdb/mnist/* (1998).

[Lee+17] Hyunkwang Lee et al. "Fully automated deep learning system for bone age assessment". In: *Journal of digital imaging* 30.4 (2017). Publisher: Springer, pp. 427–441.

[Lee+18] Jae-Hong Lee et al. "Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm". In: *Journal of dentistry* 77 (2018). Publisher: Elsevier, pp. 106–111.

[Lee+19] Youngwan Lee et al. "An energy and GPU-computation efficient backbone network for real-time object detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2019.

[Lee+20] Ki-Sun Lee et al. "Evaluation of Scalability and Degree of Fine-Tuning of Deep Convolutional Neural Networks for COVID-19 Screening on Chest X-ray Images Using Explainable Deep-Learning Algorithm". In: *Journal of Personalized Medicine* 10.4 (2020). Publisher: Multidisciplinary Digital Publishing Institute, p. 213.

[Lee+23] Po-Lei Lee et al. “Continual Learning of a Transformer-Based Deep Learning Classifier Using an Initial Model from Action Observation EEG Data to Online Motor Imagery Classification”. In: *Bioengineering* 10.2 (2023). Publisher: MDPI, p. 186.

[Lin+14] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[Lin+17] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[Lin+18] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *arXiv:1708.02002 [cs]* (Feb. 2018). arXiv: 1708.02002. URL: http://arxiv.org/abs/1708.02002 (visited on 10/03/2020).

[Lin+19] Wei-Fen Lin et al. “Onnc: A compilation framework connecting onnx to proprietary deep learning accelerators”. In: *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE. 2019, pp. 214–218.

[Lit+17] Geert Litjens et al. “A Survey on Deep Learning in Medical Image Analysis”. en. In: *Medical Image Analysis* 42 (Dec. 2017). arXiv: 1702.05747, pp. 60–88. ISSN: 13618415. DOI: 10.1016/j.media.2017.07.005. URL: http://arxiv.org/abs/1702.05747 (visited on 07/20/2020).

[Liu+20a] TY Alvin Liu et al. “Deep learning and transfer learning for optic disc laterality detection: Implications for machine learning in neuro-ophthalmology”. In: *Journal of Neuro-Ophthalmology* 40.2 (2020). Publisher: LWW, pp. 178–184.

[Liu+20b] Yu Liu et al. “Enhancing the interoperability between deep learning frameworks by model conversion”. In: *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2020, pp. 1320–1330.

[Liu+21a] Kunxing Liu et al. “Classification of two species of Gram-positive bacteria through hyperspectral microscopy coupled with machine learning”. In: *Biomedical Optics Express* 12.12 (2021), pp. 7906–7916.

[Liu+21b] Wanli Liu et al. “Is the aspect ratio of cells important in deep learning? A robust comparison of deep learning methods for multi-scale cytopathology cell image classification: from convolutional neural networks to visual transformers”. In: *arXiv:2105.07402 [cs]* (Nov. 2021). arXiv: 2105.07402. URL: http://arxiv.org/abs/2105.07402 (visited on 02/01/2022).

[Liu+21c] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[Liu+22] Zhuang Liu et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11976–11986.

[LJ20] Jae-Hong Lee and Seong-Nyum Jeong. “Efficacy of deep convolutional neural network algorithm for the identification and classification of dental implant systems, using panoramic and periapical radiographs: A pilot study”. In: *Medicine* 99.26 (2020). Publisher: Wolters Kluwer Health.

[LKJ20] Jae-Hong Lee, Do-Hyung Kim, and Seong-Nyum Jeong. “Diagnosis of cystic lesions using panoramic and cone beam computed tomographic images based on deep learning neural network”. en. In: *Oral Diseases* 26.1 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/odi.13223, pp. 152–158. ISSN: 1601-0825. DOI: https://doi.org/10.1111/odi.13223. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/odi.13223 (visited on 05/19/2021).

[Mar+18] Jon N. Marsh et al. “Deep learning global glomerulosclerosis in transplant kidney frozen sections”. In: *IEEE transactions on medical imaging* 37.12 (2018). Publisher: IEEE, pp. 2718–2728.

[Mar+21] Máté E. Maros et al. “Comparative analysis of machine learning algorithms for computer-assisted reporting based on fully automated cross-lingual RadLex mappings”. In: *Scientific Reports* 11.1 (2021). Publisher: Springer, pp. 1–18.

[Maz+18] Claudia Mazo et al. “Transfer learning for classification of cardiovascular tissues in histological images”. en. In: *Computer Methods and Programs in Biomedicine* 165 (Oct. 2018), pp. 69–76. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2018.08.006. URL: https://www.sciencedirect.com/science/article/pii/S0169260718305297 (visited on 05/19/2021).

[MBD20] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. “A scoping review of transfer learning research on medical image analysis using ImageNet”. In: *arXiv:2004.13175 [cs, eess]* (Nov. 2020). arXiv: 2004.13175. DOI: 10.1016/j.compbiomed.2020.104115. URL: http://arxiv.org/abs/2004.13175 (visited on 08/23/2021).

[MF93] Calvin R Maurer and J Michael Fitzpatrick. “A review of medical image registration”. In: *Interactive image-guided neurosurgery* 1 (1993), pp. 17–44.

[MG23] Rahul Mishra and Hari Gupta. “Transforming large-size to lightweight deep neural networks for iot applications”. In: *ACM Computing Surveys* 55.11 (2023), pp. 1–35.

[Min+20] Shervin Minaee et al. “Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning”. In: *Medical Image Analysis* 65 (Oct. 2020), p. 101794. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101794. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7372265/ (visited on 05/19/2021).

[Min+21] Shervin Minaee et al. “Image segmentation using deep learning: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3523–3542.

[Moh+18] Aly A. Mohamed et al. “A deep learning method for classifying mammographic breast density categories”. en. In: *Medical Physics* 45.1 (2018). _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12683, pp. 314–321. ISSN: 2473-4209. DOI: https://doi.org/10.1002/mp.12683. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12683 (visited on 05/19/2021).

[Mon+22] Nicola Montemurro et al. *Brain Tumor and Augmented Reality: New Technologies for the Future*. Issue: 10 Pages: 6347 Publication Title: International Journal of Environmental Research and Public Health Volume: 19. 2022.

[MR22] Sachin Mehta and Mohammad Rastegari. *MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer*. arXiv:2110.02178 [cs]. Mar. 2022. URL: http://arxiv.org/abs/2110.02178 (visited on 03/28/2023).

[MT19] Natalia Miloslavskaya and Alexander Tolstoy. “Internet of Things: information security challenges and solutions”. In: *Cluster Computing* 22 (2019), pp. 103–119.

[MV19] Sparsh Mittal and Shraiysh Vaishay. “A survey of techniques for optimizing deep learning on GPUs”. en. In: *Journal of Systems Architecture* 99 (Oct. 2019), p. 101635. ISSN: 1383-7621. DOI: 10.1016/j.sysarc.2019.101635. URL: https://www.sciencedirect.com/science/article/pii/S1383762119302656 (visited on 02/01/2022).

[NHW17] Aiden Nibali, Zhen He, and Dennis Wollersheim. “Pulmonary nodule classification with deep residual networks”. In: *International journal of computer assisted radiology and surgery* 12 (2017), pp. 1799–1808.

[Nis+18] Mizuho Nishio et al. “Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning”. In: *PloS one* 13.7 (2018). Publisher: Public Library of Science San Francisco, CA USA, e0200721.

[Nob+18] Raul Victor M. da Nobrega et al. “Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks”. In: *Neural Computing and Applications* (2018). Publisher: Springer, pp. 1–18.

[Nwa+18] Chigozie Nwankpa et al. “Activation functions: Comparison of trends in practice and research for deep learning”. In: *arXiv preprint arXiv:1811.03378* (2018).

[Ova+20] Emmanuel Ovalle-Magallanes et al. “Transfer Learning for Stenosis Detection in X-ray Coronary Angiography”. In: *Mathematics* 8.9 (2020). Publisher: Multidisciplinary Digital Publishing Institute, p. 1510.

[Par+20] P. Parmar et al. “An artificial intelligence algorithm that identifies middle turbinate pneumatisation (concha bullosa) on sinus computed tomography scans”. In: *The Journal of Laryngology & Otology* 134.4 (2020). Publisher: Cambridge University Press, pp. 328–331.

[Pas+19] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

[Pat+20] Ilaria Patrini et al. “Transfer learning for informative-frame selection in laryngoscopic videos through learned features”. In: *Medical & biological engineering & computing* (2020). Publisher: Springer, pp. 1–14.

[Pau+19] H. Yi Paul et al. “Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning”. In: *Pediatric radiology* 49.8 (2019). Publisher: Springer, pp. 1066–1070.

[Pen+20] Jie Peng et al. “Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging”. In: *European radiology* 30 (2020), pp. 413–424.

[Per+19] Shaked Perek et al. “Classification of contrast-enhanced spectral mammography (CESM) images”. eng. In: *International Journal of Computer Assisted Radiology and Surgery* 14.2 (Feb. 2019), pp. 249–257. ISSN: 1861-6429. DOI: 10.1007/s11548-018-1876-6.

[Pha20] Tuan D Pham. “A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks”. In: *Scientific reports* 10.1 (2020), pp. 1–8.

[Pit+20] Heikki Pitkänen et al. “European Medical Device Regulations MDR & IVDR”. In: *Business Finland* (2020).

[Pre12] Lutz Prechelt. “Early stopping—but when?” In: *Neural networks: tricks of the trade: second edition* (2012). Publisher: Springer, pp. 53–67.

[Pre98] Lutz Prechelt. “Early stopping-but when?” In: *Neural Networks: Tricks of the trade*. Springer, 1998, pp. 55–69.

[PSA20] Nikolaos D. Papathanasiou, Trifon Spyridonidis, and Dimitris J. Apostolopoulos. "Automatic characterization of myocardial perfusion imaging polar maps employing deep learning and data augmentation". In: *Hellenic Journal of Nuclear Medicine* 23.2 (2020), pp. 125–132.

[PY10] S. J. Pan and Q. Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359. ISSN: 1558-2191. DOI: 10.1109/TKDE.2009.191.

[Rad+19] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[Rag+21] Maithra Raghu et al. "Do Vision Transformers See Like Convolutional Neural Networks?" In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 12116–12128. URL: https://proceedings.neurips.cc/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html (visited on 04/14/2023).

[Rah+20a] Md Mamunur Rahaman et al. "A survey for cervical cytopathology image analysis using deep learning". In: *IEEE Access* 8 (2020). Publisher: IEEE, pp. 61687–61710.

[Rah+20b] Md Mamunur Rahaman et al. "Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches". In: *Journal of X-ray Science and Technology* 28.5 (2020), pp. 821–839.

[Rah+21] Md Mamunur Rahaman et al. "DeepCervix: A Deep Learning-based Framework for the Classification of Cervical Cells Using Hybrid Deep Feature Fusion Techniques". In: *arXiv preprint arXiv:2102.12191* (2021).

[Ram+19] Prajit Ramachandran et al. "Stand-alone self-attention in vision models". In: *Advances in neural information processing systems* 32 (2019).

[Ren+22] Sucheng Ren et al. "Co-advise: Cross inductive bias distillation". In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 2022, pp. 16773–16782.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[Rie+20] Nicola Rieke et al. "The future of digital health with federated learning". In: *NPJ digital medicine* 3.1 (2020), p. 119.

[RMS19] Jason Riordon, Christopher McCallum, and David Sinton. "Deep learning for the classification of human sperm". In: *Computers in biology and medicine* 111 (2019). Publisher: Elsevier, p. 103342.

[Rom+20] Miguel Romero et al. "Targeted transfer learning to improve performance in small medical physics datasets". en. In: *Medical Physics* 47.12 (2020). _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14507, pp. 6246–6256. ISSN: 2473-4209. DOI: https://doi.org/10.1002/mp.14507. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14507 (visited on 05/20/2021).

[Rot13] Vincent J. Roth. "The mHealth Conundrum: Smartphones & Mobile medical apps-How much FDA medical device regulation is required". In: *NCJL & Tech.* 15 (2013). Publisher: HeinOnline, p. 359.

[Ryo+13] Jungwoo Ryoo et al. "Cloud security auditing: challenges and emerging approaches". In: *IEEE Security & Privacy* 12.6 (2013). Publisher: IEEE, pp. 68–74.

[Sam+18] Ravi K Samala et al. "Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets". In: *IEEE transactions on medical imaging* 38.3 (2018), pp. 686–696.

[Sam+20a] Ravi K. Samala et al. "Generalization error analysis for deep convolutional neural network with transfer learning in breast cancer diagnosis". In: *Physics in Medicine & Biology* 65.10 (2020). Publisher: IOP Publishing, p. 105002.

[Sam+20b] Ravi K. Samala et al. "Risks of Feature Leakage and Sample Size Dependencies in Deep Feature Extraction for Breast Mass Classification". In: *Medical Physics* (2020). Publisher: Wiley Online Library.

[Sam59] Arthur L Samuel. "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

[SDS19] Neeru Singla, Kavita Dubey, and Vishal Srivastava. "Automated assessment of breast cancer margin in optical coherence tomography images via pretrained convolutional neural network". In: *Journal of biophotonics* 12.3 (2019). Publisher: Wiley Online Library, e201800255.

[Seh+20] Adil Hussain Seh et al. "Healthcare data breaches: insights and implications". In: *Healthcare*. Vol. 8. 2. MDPI. 2020, p. 133.

[Sey+17] Christopher W Seymour et al. "Time to treatment and mortality during mandated emergency care for sepsis". In: *New England Journal of Medicine* 376.23 (2017), pp. 2235–2244.

[SF07] Yutaka Sasaki and R. Fellow. "The truth of the F-measure, Manchester: MIB-School of Computer Science". In: *University of Manchester* (2007), p. 25.

[Sha+22] Fahad Shamshad et al. *Transformers in Medical Imaging: A Survey*. arXiv:2201.09873 [cs, eess]. Jan. 2022. URL: http://arxiv.org/abs/2201.09873 (visited on 04/03/2023).

[She+18] Xiaolei Shen et al. "An automatic diagnosis method of facial acne vulgaris based on convolutional neural network". In: *Scientific reports* 8.1 (2018). Publisher: Nature Publishing Group, pp. 1–10.

[She+19] Li Shen et al. "Deep learning to improve breast cancer detection on screening mammography". In: *Scientific reports* 9.1 (2019). Publisher: Nature Publishing Group, pp. 1–12.

[Shi+16] Hoo-Chang Shin et al. "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning". In: *IEEE Transactions on Medical Imaging* 35.5 (May 2016). Conference Name: IEEE Transactions on Medical Imaging, pp. 1285–1298. ISSN: 1558-254X. DOI: 10.1109/TMI.2016.2528162.

[Shi+17] Satoki Shichijo et al. "Application of Convolutional Neural Networks in the Diagnosis of Helicobacter pylori Infection Based on Endoscopic Images". en. In: *EBioMedicine* 25 (Nov. 2017), pp. 106–111. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2017.10.014. URL: https://www.sciencedirect.com/science/article/pii/S2352396417304127 (visited on 05/19/2021).

[Shi+18] Bibo Shi et al. "Prediction of occult invasive disease in ductal carcinoma in situ using deep learning features". In: *Journal of the American College of Radiology* 15.3 (2018). Publisher: Elsevier, pp. 527–534.

[Shi+19] Satoki Shichijo et al. "Application of convolutional neural networks for evaluating Helicobacter pylori infection status on the basis of endoscopic images". In: *Scandinavian Journal of Gastroenterology* 54.2 (Feb. 2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00365521.2019.1577486, pp. 158–163. ISSN: 0036-5521. DOI: 10.1080/00365521.2019.1577486. URL: https://doi.org/10.1080/00365521.2019.1577486 (visited on 05/19/2021).

[Sin+19] Varun Singh et al. "Assessment of critical feeding tube malpositions on radiographs using deep learning". In: *Journal of digital imaging* 32.4 (2019). Publisher: Springer, pp. 651–655.

[Sin+20] Satya P Singh et al. "3D deep learning on medical images: a review". In: *Sensors* 20.18 (2020), p. 5097.

[SK16] Tim Salimans and Diederik P. Kingma. *Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks*. en. Feb. 2016. URL: https://arxiv.org/abs/1602.07868v3 (visited on 04/02/2023).

[SKK18] Kenneth P. Smith, Anthony D. Kang, and James E. Kirby. "Automated interpretation of blood culture gram stains by use of a deep convolutional neural network". In: *Journal of Clinical Microbiology* 56.3 (2018). Publisher: Am Soc Microbiol, e01521–17.

[Smi+17] Samuel L. Smith et al. "Don't decay the learning rate, increase the batch size". In: *arXiv preprint arXiv:1711.00489* (2017).

[Sri+14] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014). Publisher: JMLR. org, pp. 1929–1958.

[Sri+19] Pradeeba Sridar et al. "Decision Fusion-Based Fetal Ultrasound Image Plane Classification Using Convolutional Neural Networks". en. In: *Ultrasound in Medicine & Biology* 45.5 (May 2019), pp. 1259–1273. ISSN: 0301-5629. DOI: 10.1016/j.ultrasmedbio.2018.11.016. URL: https://www.sciencedirect.com/science/article/pii/S0301562918305283 (visited on 08/23/2021).

[Sri+21] Aravind Srinivas et al. *Bottleneck Transformers for Visual Recognition*. arXiv:2101.11605 [cs]. Aug. 2021. URL: http://arxiv.org/abs/2101.11605 (visited on 03/28/2023).

[ST18] Sarmad Shafique and Samabia Tehsin. "Acute Lymphoblastic Leukemia Detection and Classification of Its Subtypes Using Pretrained Deep Convolutional Neural Networks". en. In: *Technology in Cancer Research & Treatment* 17 (Jan. 2018). Publisher: SAGE Publications Inc, p. 1533033818802789. ISSN: 1533-0346. DOI: 10.1177/1533033818802789. URL: https://doi.org/10.1177/1533033818802789 (visited on 05/19/2021).

[Sün+18] Niko Sünderhauf et al. "The limits and potentials of deep learning for robotics". In: *The International journal of robotics research* 37.4-5 (2018), pp. 405–420.

[Sun+19] Yue Sun et al. "Detecting discomfort in infants through facial expressions". In: *Physiological measurement* 40.11 (2019). Publisher: IOP Publishing, p. 115006.

[Sun+20] Changhao Sun et al. "Gastric histopathology image segmentation using a hierarchical conditional random field". In: *Biocybernetics and Biomedical Engineering* 40.4 (2020). Publisher: Elsevier, pp. 1535–1555.

[Swa+19] Zar Nawab Khan Swati et al. "Brain tumor classification for MR images using transfer learning and fine-tuning". In: *Computerized Medical Imaging and Graphics* 75 (2019). Publisher: Elsevier, pp. 34–46.

[Syk93] Alan O Sykes. "An introduction to regression analysis". In: (1993).

[SZ15] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". en. In: *arXiv:1409.1556 [cs]* (Apr. 2015). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556 (visited on 08/06/2020).

[Sze+14] Christian Szegedy et al. "Going Deeper with Convolutions". In: *arXiv:1409.4842 [cs]* (Sept. 2014). arXiv: 1409.4842. URL: http://arxiv.org/abs/1409.4842 (visited on 06/30/2021).

[Sze+15] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[Tal+19] Muhammed Talo et al. "Convolutional neural networks for multi-class brain disease detection using MRI images". In: *Computerized Medical Imaging and Graphics* 78 (2019). Publisher: Elsevier, p. 101673.

[Tal19] Muhammed Talo. "Automated classification of histopathology images using transfer learning". en. In: *Artificial Intelligence in Medicine* 101 (Nov. 2019), p. 101743. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2019.101743. URL: https://www.sciencedirect.com/science/article/pii/S0933365719307110 (visited on 05/19/2021).

[Tan+20] Yu-Xing Tang et al. "Automated abnormality classification of chest radiographs using deep convolutional neural networks". en. In: *npj Digital Medicine* 3.1 (May 2020). Number: 1 Publisher: Nature Publishing Group, pp. 1–8. ISSN: 2398-6352. DOI: 10.1038/s41746-020-0273-z. URL: https://www.nature.com/articles/s41746-020-0273-z (visited on 05/20/2021).

[TBF22] Ajay Kumar Tanwani, Joelle Barral, and Daniel Freedman. "RepsNet: Combining Vision with Language for Automated Medical Reports". In: vol. 13435. arXiv:2209.13171 [cs]. 2022, pp. 714–724. DOI: 10.1007/978-3-031-16443-9_68. URL: http://arxiv.org/abs/2209.13171 (visited on 04/06/2023).

[TLE18] Maximilian Treder, Jost Lennart Lauermann, and Nicole Eter. "Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning". In: *Graefe's Archive for Clinical and Experimental Ophthalmology* 256.2 (2018). Publisher: Springer, pp. 259–265.

[Val+19a] Mira Valkonen et al. "Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for ER, PR, and Ki-67". In: *IEEE transactions on medical imaging* 39.2 (2019). Publisher: IEEE, pp. 534–542.

[Val+19b] Mira Valkonen et al. "Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for ER, PR, and Ki-67". In: *IEEE transactions on medical imaging* 39.2 (2019), pp. 534–542.

[Val+21] Juan Miguel Valverde et al. “Transfer Learning in Magnetic Resonance Brain Imaging: A Systematic Review”. In: *Journal of Imaging* 7.4 (2021). Publisher: Multidisciplinary Digital Publishing Institute, p. 66.

[Vas+17] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 02/01/2023).

[VK14] Denny Vrandečić and Markus Krötzsch. “Wikidata: a free collaborative knowledgebase”. In: *Communications of the ACM* 57.10 (2014), pp. 78–85.

[Wan+19] Zengmao Wang et al. “Incorporating distribution matching into uncertainty for multiple kernel active learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.1 (2019). Publisher: IEEE, pp. 128–142.

[Wan+21] Wenhai Wang et al. “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 568–578.

[WC20] Garrett Wilson and Diane J. Cook. “A survey of unsupervised deep domain adaptation”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 11.5 (2020). Publisher: ACM New York, NY, USA, pp. 1–46.

[WDG19] Zengmao Wang, Bo Du, and Yuhong Guo. “Domain adaptation with neural embedding matching”. In: *IEEE transactions on neural networks and learning systems* 31.7 (2019). Publisher: IEEE, pp. 2387–2397.

[WKW16] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big data* 3.1 (2016). Publisher: SpringerOpen, pp. 1–40.

[WM03] D. Randall Wilson and Tony R. Martinez. “The general inefficiency of batch training for gradient descent learning”. In: *Neural networks* 16.10 (2003). Publisher: Elsevier, pp. 1429–1451.

[Wol+20] Thomas Wolf et al. “Transformers: State-of-the-art natural language processing”. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 2020, pp. 38–45.

[Wu+20] Hao Wu et al. *Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation*. arXiv:2004.09602 [cs, stat]. Apr. 2020. DOI: 10.48550/arXiv.2004.09602. URL: http://arxiv.org/abs/2004.09602 (visited on 02/09/2023).

[Wu+21] Haiping Wu et al. *CvT: Introducing Convolutions to Vision Transformers*. arXiv:2103.15808 [cs]. Mar. 2021. URL: http://arxiv.org/abs/2103.15808 (visited on 03/28/2023).

[Xia+21] Tete Xiao et al. *Early Convolutions Help Transformers See Better*. arXiv:2106.14881 [cs]. Oct. 2021. URL: http://arxiv.org/abs/2106.14881 (visited on 03/28/2023).

[Xie+21] Enze Xie et al. *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*. arXiv:2105.15203 [cs]. Oct. 2021. URL: http://arxiv.org/abs/2105.15203 (visited on 04/02/2023).

[Xio+19] Junfeng Xiong et al. "Implementation strategy of a CNN model affects the performance of CT assessment of EGFR mutation status in lung cancer patients". In: *IEEE Access* 7 (2019), pp. 64583–64591.

[Xu+15] Bing Xu et al. "Empirical evaluation of rectified activations in convolutional network". In: *arXiv preprint arXiv:1505.00853* (2015).

[Xu+19] Benjamin Y. Xu et al. "Deep learning classifiers for automated detection of gonioscopic angle closure based on anterior segment OCT images". In: *American journal of ophthalmology* 208 (2019). Publisher: Elsevier, pp. 273–280.

[Xue+20] Li-Yun Xue et al. "Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis". In: *European radiology* (2020). Publisher: Springer, pp. 1–11.

[Yam+19] Akira Yamada et al. "Dynamic contrast-enhanced computed tomography diagnosis of primary liver cancers using transfer learning of pretrained convolutional neural networks: Is registration of multiphasic images necessary?" en. In: *International Journal of Computer Assisted Radiology and Surgery* 14.8 (Aug. 2019), pp. 1295–1301. ISSN: 1861-6410, 1861-6429. DOI: 10.1007/s11548-019-01987-1. URL: http://link.springer.com/10.1007/s11548-019-01987-1 (visited on 05/19/2021).

[Yan+18a] Hao Yang et al. "Multimodal MRI-based classification of migraine: using deep learning convolutional neural network". In: *Biomedical engineering online* 17.1 (2018), pp. 1–14.

[Yan+18b] Yang Yang et al. "Glioma grading on conventional MR images: a deep learning study with transfer learning". In: *Frontiers in neuroscience* 12 (2018). Publisher: Frontiers, p. 804.

[Yan+21] Jianwei Yang et al. "Focal self-attention for local-global interactions in vision transformers". In: *arXiv preprint arXiv:2107.00641* (2021).

[Yin+19] Chris Ying et al. "Nas-bench-101: Towards reproducible neural architecture search". In: *International conference on machine learning*. PMLR. 2019, pp. 7105–7114.

[You+19] Yang You et al. “Fast deep neural network training on distributed systems and cloud TPUs”. In: *IEEE Transactions on Parallel and Distributed Systems* 30.11 (2019), pp. 2449–2462.

[Yu+18] Yang Yu et al. “Deep learning enables automated scoring of liver fibrosis stages”. en. In: *Scientific Reports* 8.1 (Oct. 2018). Number: 1 Publisher: Nature Publishing Group, p. 16016. ISSN: 2045-2322. DOI: 10.1038/s41598-018-34300-2. URL: https://www.nature.com/articles/s41598-018-34300-2 (visited on 05/17/2021).

[Yu+19a] ShaoDe Yu et al. “Transferring deep neural networks for the differentiation of mammographic breast lesions”. en. In: *Science China Technological Sciences* 62.3 (Mar. 2019), pp. 441–447. ISSN: 1674-7321, 1869-1900. DOI: 10.1007/s11431-017-9317-3. URL: http://link.springer.com/10.1007/s11431-017-9317-3 (visited on 05/19/2021).

[Yu+19b] Xiang Yu et al. “Utilization of DenseNet201 for diagnosis of breast abnormality”. In: *Machine Vision and Applications* 30.7 (2019). Publisher: Springer, pp. 1135–1144.

[Yu+22] Weihao Yu et al. *MetaFormer Is Actually What You Need for Vision*. arXiv:2111.11418 [cs]. July 2022. URL: http://arxiv.org/abs/2111.11418 (visited on 03/28/2023).

[Yua+19] Yixuan Yuan et al. “Prostate cancer classification with multiparametric MRI transfer learning model”. In: *Medical physics* 46.2 (2019). Publisher: Wiley Online Library, pp. 756–765.

[Zac+20] Robin Zachariah et al. “Prediction of Polyp Pathology Using Convolutional Neural Networks Achieves “Resect and Discard” Thresholds”. en-US. In: *Official journal of the American College of Gastroenterology — ACG* 115.1 (Jan. 2020), pp. 138–144. ISSN: 0002-9270. DOI: 10.14309/ajg.0000000000000429. URL: https://journals.lww.com/ajg/fulltext/2020/01000/prediction_of_polyp_pathology_using_convolutional.21.aspx?casa_token=cEky_PFhO8cAAAAA:wtEtbxoc6Bn7F3mc_9gPbsh1WkyK5djjEkNidfMq4FzaPbsO0Faqw2Xldgw-1GRDjVnBp_FAexKdXRf1YbaCsMlw1Js1zpM (visited on 05/19/2021).

[Zag+18] Gabriel Tozatto Zago et al. “Retinal image quality assessment using deep learning”. In: *Computers in biology and medicine* 103 (2018). Publisher: Elsevier, pp. 64–70.

[ZCL15] Zhuoyuan Zheng, Yunpeng Cai, and Ye Li. “Oversampling method for imbalanced classification”. In: *Computing and Informatics* 34.5 (2015), pp. 1017–1037.

[ZG17] Michael Zhu and Suyog Gupta. “To prune, or not to prune: exploring the efficacy of pruning for model compression”. In: *arXiv preprint arXiv:1710.01878* (2017).

[Zha+16] Ruikai Zhang et al. “Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain”. In: *IEEE journal of biomedical and health informatics* 21.1 (2016). Publisher: IEEE, pp. 41–47.

[Zha+18] Xiaofei Zhang et al. “Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks”. In: *IEEE transactions on nanobioscience* 17.3 (2018). Publisher: IEEE, pp. 237–242.

[Zha+19a] Shikun Zhang et al. “Computer-aided diagnosis (CAD) of pulmonary nodule of thoracic CT image using transfer learning”. In: *Journal of digital imaging* 32.6 (2019). Publisher: Springer, pp. 995–1007.

[Zha+19b] Tianyang Zhang et al. “Noise adaptation generative adversarial network for medical image analysis”. In: *IEEE transactions on medical imaging* 39.4 (2019). Publisher: IEEE, pp. 1149–1159.

[Zha+19c] Xinzhuo Zhao et al. “Deep CNN models for pulmonary nodule classification: model modification, model integration, and transfer learning”. In: *Journal of X-ray Science and Technology* 27.4 (2019). Publisher: IOS Press, pp. 615–629.

[Zha+20] Yifan Zhang et al. “Collaborative unsupervised domain adaptation for medical image diagnosis”. In: *IEEE Transactions on Image Processing* 29 (2020). Publisher: IEEE, pp. 7834–7844.

[Zha+21] Jiawei Zhang et al. “A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches”. In: *Artificial Intelligence Review* (2021). Publisher: Springer, pp. 1–70.

[Zhe+19] Qiang Zheng et al. “Computer-aided diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data by integrating texture image features and deep transfer learning image features”. In: *Journal of pediatric urology* 15.1 (2019). Publisher: Elsevier, 75–e1.

[Zhe+20] Ce Zheng et al. “Detecting glaucoma based on spectral domain optical coherence tomography imaging of peripapillary retinal nerve fiber layer: a comparison study between hand-crafted features and deep learning model”. In: *Graefe’s Archive for Clinical and Experimental Ophthalmology* 258.3 (2020). Publisher: Springer, pp. 577–585.

[Zhu+17] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[Zhu+19a] Yan Zhu et al. “Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy”. In: *Gastrointestinal endoscopy* 89.4 (2019), pp. 806–815.

[Zhu+19b] Zhe Zhu et al. “Deep learning analysis of breast MRIs for prediction of occult invasive disease in ductal carcinoma in situ”. In: *Computers in biology and medicine* 115 (2019). Publisher: Elsevier, p. 103498.

[Zhu+20] Fuzhen Zhuang et al. “A comprehensive survey on transfer learning”. In: *Proceedings of the IEEE* 109.1 (2020). Publisher: IEEE, pp. 43–76.

[Zie+17] Bartosz Zieliński et al. “Deep learning approach to bacterial colony classification”. In: *PloS one* 12.9 (2017). Publisher: Public Library of Science San Francisco, CA USA, e0184554.

Appendices

A Search Terms

The search terms used for PubMed were as follows: (”Convolutional neural network*”[Title/Abstract] OR ”CNN*”[Title/Abstract]) AND (”image processing, computer-assisted”[MeSH Terms] OR ”Diagnostic Imaging”[MeSH Terms] OR ”medical imag*”[Title/Abstract] OR ”clinical imag*”[Title/Abstract] OR ”biomedical imag*”) AND (”transfer learning”[Title/Abstract] OR ”pre-trained”[Title/Abstract] OR ”pretrained”[Title/Abstract]) NOT (”Review”[Publication Type] OR ”Letter”[Publication Type] OR ”meta-analysis”[Publication Type] OR ”Systematic Review”[Publication Type] OR ”Systematic Review”[Publication Type])
The search string applied in Web of Science database was as follows: TS=(”CNN” OR ”convolutional”) AND TS=(”medical imag*” OR ”clinical imag*” OR ”biomedical imag*”) AND TS=(”transfer learning” OR ”pre-trained” OR ”pretrained”) NOT TS=(”novel” OR ”propose”)

B Summary Table of Referenced Studies

Table B: A summary table of studies that utilized transfer learning in the medical domain.

Modality	Subject	Transfer Learning	Reference
CT scan	Abdominopelvic cavity	Feature extractor	[Hua+20a]
	Alimentary system	Feature extractor	[Yam+19; Pen+20]
		Fine-tuning scratch	[Had+20; LKJ20]
	Bones	Feature extractor	[Par+20]
	Genital systems	Fine-tuning scratch	[Kaj+18]
	Nervous system	Many	[DYO19]
	Respiratory system	Feature extractor	[Zha+19c]
		Feature extractor hybrid	[Nob+18]
		Fine-tuning scratch	[Zha+19a; NHW17; Pha20]
		Many	[Xio+19; Gao+20]
	Sense organs	Feature extractor	[Cho+19b]
	Thoracic cavity	Feature extractor	[Nis+18]
Endoscopy	Alimentary system	Feature extractor	[Zac+20; Zhu+19a]
		Fine-tuning scratch	[Cho+19a; Shi+17; Shi+19]
		Many	[Pat+20]
Mammographic	Integumentary system	Feature extractor	[Shi+18]
		Feature extractor hybrid	[Sam+20b]
		Fine-tuning scratch	[Yu+19a; Moh+18]
		Many	[Yu+19b; Zha+18; Per+19; Sam+18; HLG16; CZA18; Sam+20a; She+19]

Microscopy	Tissues	Feature extractor	[ST18; Yu+18; Hut+18; Tal19; Maz+18; RMS19; Mar+18]
		Fine-tuning	[Val+19b]
		Fine-tuning scratch	[Kan+20; Kat+19]
MRI	Bones	Many	[He+19]
	Genital systems	Feature extractor	[Che+19b; Yua+19]
	Integumentary system	Fine-tuning scratch	[Bor+20]
		Many	[Zhu+19b]
	Nervous system	Fine-tuning scratch	[Yan+18a; Fuk+19; Ban+19]
		Many	[Tal+19; Swa+19; Yan+18b; DA19]
OCT	Integumentary system	Feature extractor	[SDS19]
	Cardiovascular system	Many	[Ges+18]
	Sense organs	Feature extractor	[Ahn+18; TLE18; Zhe+20; Zag+18]
		Feature extractor hybrid	[Bur+17]
		Fine-tuning	[Hem+20]
		Fine-tuning scratch	[KCC17; Liu+20a]
		Many	[Cho+17; Góm+19; Xu+19]
Photography	Integumentary system	Feature extractor	[Bur+18; She+18; Cir+19]
		Fine-tuning	[Han+18]
		Fine-tuning scratch	[Hua+20b]
	Else	Fine-tuning scratch	[Sun+19]
Sonography	Abdominopelvic cavity	Feature extractor	[CM17]
	Alimentary system	Feature extractor	[Xue+20]
		Feature extractor hybrid	[Byr+18]
		Fine-tuning scratch	[Ban+18b]
	Bones	Feature extractor	[Het+17]
	Endocrine glands	Fine-tuning scratch	[Chi+17]
	Genital systems	Feature extractor hybrid	[Sri+19]
	Integumentary system	Many	[Byr+19]
	Respiratory system	Many	[Che+19a]
	Urinary system	Feature extractor hybrid	[Zhe+19]
SPECT	Nervous system	Feature extractor	[KWT18]
		Many	[PSA20]
X-ray	Abdominopelvic cavity	Feature extractor	[Che+18b]
		Feature extractor hybrid	[Dev+21]
		Many	[Sin+19]
	Alimentary system	Fine-tuning scratch	[Kim+20a; Lee+18; LJ20]
	Bones	Feature extractor	[Pau+19; KM18]
		Many	[Lee+17; Che+19c]
	Cardiovascular system	Many	[Ova+20]
	Joints	Many	[Abi+18]
	Respiratory system	Feature extractor	[Rah+20b; Hei+20; AA20; Min+20]
		Many	[Lee+20; AM20]
	Thoracic cavity	Fine-tuning scratch	[Lak17]
		Many	[Tan+20; Rom+20]
Many	Many	Many	[Shi+16; Cla+20]

C Summary Table of Public Medical Data

Table C: A summary table of public medical datasets. Abbreviations: C, Classification; D, Detection; R, Regression; Rg, Registration; S, Segmentation.

Modality	Anatomical Part/Region	Task Type	Data	Published Year	URL
CT scan	Abdomen	S	FLARE	2021	flare.grand-challenge.org
		S	KiTS21	2021	kits21.grand-challenge.org
		S	SLIVER07	2019	sliver07.grand-challenge.org

	Cardiac	C	orcaScore	2020	orcascore.grand-challenge.org
		S	CCTA	2020	asoca.grand-challenge.org
	Head and neck	S	INSTANCE	2022	instance.grand-challenge.org
		S	NucMM	2020	nucmm.grand-challenge.org
		S	StructSeg	2019	structseg2019.grand-challenge.org
		Many	CADA	2020	cada.grand-challenge.org
	Spine	S	VerSe	2020	verse2020.grand-challenge.org
	Thorax	C	STOIC	2021	stoic2021.grand-challenge.org/stoic-db
		D	LUNA16	2016	luna16.grand-challenge.org
		C	COVID19-CT	2020	covid-ct.grand-challenge.org
		R	LoDoPaB-CT	2021	lodopab.grand-challenge.org
		Rg	EMPIRE10	2010	empire10.grand-challenge.org
		S	COVID-19-20	2020	covid-segmentation.grand-challenge.org
		S	LOLA11	2011	lola11.grand-challenge.org
		Many	RibFrac	2020	ribfrac.grand-challenge.org
		Many	LNDb	2020	lndb.grand-challenge.org
	Many	S	Parse	2022	parse2022.grand-challenge.org
		Rg	CRC	2018	continuousregistration.grand-challenge.org
Endoscopy	Abdomen	D	EndoCV 2.0	2022	endocv2022.grand-challenge.org
	Pelvis	D	SARAS	2021	saras-mesad.grand-challenge.org
Microscopy	Tissues	C	BCNB	2021	bcnb.grand-challenge.org
		C	HEROHE	2020	ecdp2020.grand-challenge.org
		C	PatchCamelyon	2019	patchcamelyon.grand-challenge.org
		D	MIDOG	2021	midog2021.grand-challenge.org
		D	LYON	2019	lyon19.grand-challenge.org
		S	WSSS4LAUD	2021	wsss4luad.grand-challenge.org
		S	BCSS	2021	bcsegmentation.grand-challenge.org
		S	SegPC	2020	segpc-2021.grand-challenge.org
		S	PANDA	2020	panda.grand-challenge.org
		R	BreastPathQ	2019	breastpathq.grand-challenge.org
		R	LYSTO	2019	lysto.grand-challenge.org
		Many	CoNIC	2022	conic-challenge.grand-challenge.org
		Many	TIGER	2021	tiger.grand-challenge.org
		Many	DigestPath	2019	digestpath2019.grand-challenge.org
		Many	NuCLS	2021	nucls.grand-challenge.org
		Many	PAIP	2021	paip2021.grand-challenge.org
		Many	MoNuSAC	2020	monusac-2020.grand-challenge.org
		Many	ACDC	2019	acdc-lunghp.grand-challenge.org
		Rg	ANHIR	2019	anhir.grand-challenge.org
		Many	ICIAR	2018	iciar2018-challenge.grand-challenge.org
		Many	CAMELYON	2017	camelyon17.grand-challenge.org
MRI	Prostate	C	ProstateX	2018	prostatex.grand-challenge.org
		S	PROMISE12	2012	promise12.grand-challenge.org
	Brain	S	FeTA	2021	feta.grand-challenge.org
		S	BrainPTM	2021	brainptm-2021.grand-challenge.org
		S	crossMoDa	2021	crossmoda.grand-challenge.org/CrossMoDA
		S	Decathlon	2018	decathlon-10.grand-challenge.org
		S	SKI10	2010	ski10.grand-challenge.org
		Many	VALDO	2021	valdo.grand-challenge.org
OCT	Eyes	C	ROCC	2017	rocc.grand-challenge.org
		Many	AGE	2019	age.grand-challenge.org
		Many	iChallenges	2018	ichallenges.grand-challenge.org
		Many	RETOUCH	2017	retouch.grand-challenge.org
Photography	Eyes	C	AIROGS	2022	airogs.grand-challenge.org
		C	RIADD	2021	riadd.grand-challenge.org
		C	REFUGE	2020	refuge.grand-challenge.org
		C	PALM	2019	palm.grand-challenge.org
		C	ODIR	2019	odir2019.grand-challenge.org
		C	ADAM	2018	amd.grand-challenge.org
		C	IDRid	2018	idrid.grand-challenge.org
		S	DRIVE	2019	drive.grand-challenge.org

Sonography	Breast	Many	ABUS	2021	tdsc-abus2023.grand-challenge.org
	Brain	Many	CuRIOUS	2019	curious2019.grand-challenge.org
	Fetal	R	HC18	2018	hc18.grand-challenge.org
		Many	A-AFMA	2020	a-afma.grand-challenge.org
	Thyroid	Many	TN-SCUI	2020	tn-scui2020.grand-challenge.org
SPECT	Many	D	fastPET-LD	2021	fastpet-ld.grand-challenge.org
X-ray	Spine	R	AASCE	2019	aasce19.grand-challenge.org
	Thorax	C	CXR-COVID19	2021	cxr-covid19.grand-challenge.org
		D	NOCE21	2021	node21.grand-challenge.org
MRI; CT	Abdomen	S	CHAOS	2021	chaos.grand-challenge.org
	Many	S	QUBIQ	2021	qubiq21.grand-challenge.org
		Many	Learn2Reg	2021	learn2reg.grand-challenge.org
MRI; SPECT	Brain	Many	TADPOLE	2017	tadpole.grand-challenge.org
MRI; X-ray	Knee	C	KNOAP	2021	knoap2020.grand-challenge.org

D F1-Score

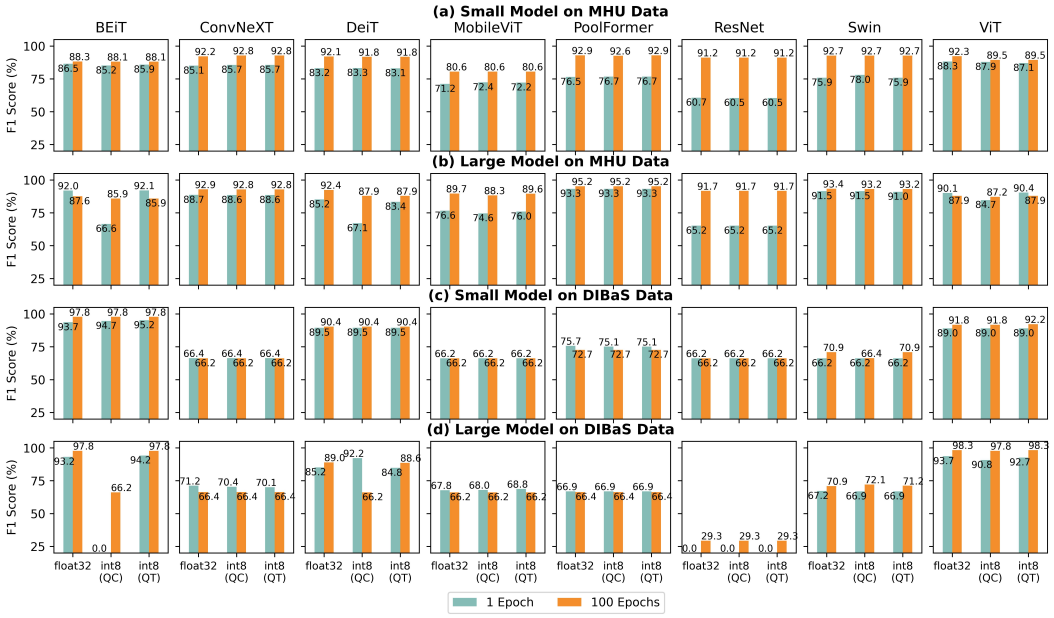

Figure D: F1-score of eight models with two parameters (small vs. large) on the MHU dataset (a,b) and DIBaS dataset (c,d). Abbreviations: QC, per-channel quantization; QT, per-tensor quantization.

Curriculum Vitae

PERSONAL DETAILS

Last- and Surname:	Kim, Hee Eun
Birthdate:	March 24, 1986
Birthplace:	Daegu

EDUCATION

Feb. 2019 - Present	**Doctor scientiarum humanarum (Dr. sc. hum.)** at Heidelberg University, Germany
Oct. 2013 - Aug. 2015	**Master of Science: Communication and Media Engineering** at Offenburg University of Applied Sciences Master thesis: evaluating Ayasdi's Topological Data Analysis for Big Data Grade: 1.2 (graduated with top honors)
Mar. 2005 - Aug.2010	**Bachelor of Science: Digital Media** at Ajou University Graduation project: a digital application for interactive television at Convention & Exhibition Center in Seoul Grade: 1.9 (Converted from Korean grade 3.35)
Feb. 2010 - Jun. 2010	**Exchange Semester** at Vilnius University, Lithuania.
Mar. 2002 - Feb. 2005	**High-school diploma** at Hangaram High School, Seoul.

PROFESSIONAL WORK EXPERIENCE

Nov. 2018 - Present	Research assistant at the Biomedical Informatics (DBMI), Medical Faculty Mannheim, Heidelberg University
Nov. 2015 - Oct. 2018	Research assistant at the Big Data Lab, Goethe University
Feb. 2011 - Sep. 2013	Software engineer at Samsung OpenTide Korea

Acknowledgement

First and foremost, I would like to thank Prof. Dr. Thomas Ganslandt for granting me the opportunity to be a research assistant. I was gracefully inspired and motivated by his interdisciplinary experiences and competencies as a medical doctor, engineer and researcher. He always paid attention to my ideas, allowed me to explore research questions and supported my work. I also personally appreciate his character in that he never put someone down; he only encouraged others. My Ph.D. journey was as challenging as others, but I was still able to enjoy my Ph.D. time because he had faith in me.

I express my sincere thanks to Dr. Mate Maros for his commitment to supporting my research works. He read all my manuscripts carefully, engaged in critical and in-depth discussions with me and provided extremely constructive advice. His suggestions were insightful and at the same time challenging to implement, but they helped me to shape my Ph.D. I feel fortunate to have received valuable advice from someone with both in-depth technical knowledge and domain expertise.

I owe special gratitude to Dr. Fabian Siegel. He provided me with valuable suggestions and guided me through complicated situations and challenges. I was also greatly inspired by his passion and enthusiasm for medical informatics along with his explosive energy in reconciling the domains of medicine and computer science. He is a unique figure who is highly intelligent and full of joy. I feel fortunate to be part of his team.

I would also like to extend my gratitude to Prof. Dr. Thomas Miethke from the Institute of Medical Microbiology and Hygiene, who provided me with factual domain knowledge and opportunities to collaborate with other microbiologists. I would also like to thank Prof. Dr. med. Michael Neumaier and Dr. Maximilian Kittel from the Institute for Clinical Chemistry for providing the Gram-stained image data and allowing me to publish the data to the public.

My appreciation goes to my colleagues in the Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health: Preetha Moorthy, Lukas Goetz, Nandhini Santhanam, Christian Palla, Geatan Wabo, Kerstin Gierend, Jacqueline Franßen, Mahboubeh Jannesari, Alejandro Cosa-Linan, Maximilian Fünfgeld and Kevser Fünfgeld. It was a pleasure working with all of you and sharing emotions. I am also thankful to Prof. Roberto Zicari from the Big Data Lab at Frankfurt Goethe University. He allowed me to join his team and paved the path for my research journey. I would like to thank Dr. Todor Ivanov, Dr. Karsten

Tolle and Naveed Mushtaq for their generous support when I was part of the Big Data Lab. I would also like to express gratitude to Stephan Trahasch and Yuri Demchenko, who provided encouragement and support during times of uncertainty.

I am grateful for my family; my mom and dad. Both of them provided me with incredible support. While my mom gave me unconditional love, my dad encouraged me to view challenges as invitations to grow. Without them, I would not have been able to manifest myself as fully as I have. I would also like to thank God for His graces and all the sisters and brothers in my church community. I feel blessed that many people pray for me to become a virtuous person.