Applying Transfer Learning in Classification of Ischemia from Myocardial Polar Maps in PET Cardiac Perfusion Imaging

Seyedmohammadreza Hosseini

Faculty of Medicine, Institute of Biomedicine

University of Turku, Turku PET Centre

Supervisors:

Dr. Jarmo Teuho

Adjunct Professor, Turku PET Centre, University of Turku

Prof. Riku Klén

Assistant Professor (Imaging instrumentation and detection technologies), Turku PET Centre, University of Turku

Research field: In vivo & clinical imaging

Abstract

Introduction: Ischemia is defined as the restriction of blood flow to a body organ, such as the heart, resulting in a cutback in oxygen supply. Myocardial ischemia is characterized by an imbalance between myocardial oxygen supply and demand, causing cardiac dysfunction, arrhythmia, myocardial infarction, and sudden death. Positron emission tomography myocardial perfusion imaging (PET-MPI) is an examination for accurately evaluating blood circulation to the heart muscle at stress and rest. Images obtained from this technique can be interpreted by experts or potentially classified by deep learning for the diagnosis of cardiac ischemia. Although deep learning has proved to be effective for medical image classification tasks, the challenge of small medical image datasets for model training remains to exist. Transfer learning is a state-of-the-art technique for resolving this challenge by utilizing pre-trained models for a new task. Pre-trained models are deep convolutional neural networks (CNNs) trained on a vast dataset, such as ImageNet, capable of transferring learned weights to a new classification problem. **Objective:** To study the effectiveness of image classification using transfer learning and benchmarking pre-trained CNN models for the classification of myocardial ischemia from myocardial polar maps in PET 15O-H2O cardiac perfusion imaging. **Subject and methods:** 138 JPEG polar maps from a 15O-H2O stress perfusion test from patients classified as ischemic or non-ischemic were used. Experiments for comparing a total of 20 pre-trained CNN models were performed. The results were compared against a custom CNN developed on the same dataset. Python programming language and its relevant libraries for deep learning were used. **Results and discussion:** Pre-trained models showed reliable performance compared to a custom-built CNN. VGG19, VGG16, DenseNet169, and Xception were superior among all pre-trained models. Ensemble learning improved overall performance, closest to the clinical interpretation level.

Keywords: Transfer learning, Deep learning, Myocardial Ischemia, PET Imaging, Convolutional neural networks

Table of Contents

1 Introduction 1
1.1 Preface 1
1.2 Thesis structure 2
1.3 Problem definition 3
1.4 Motivation 4
1.5 Research questions 5
1.6 Research objectives 6
2 Background 7
2.1 Myocardial ischemia 7
2.1.1 Introduction 7
2.1.2 Pathophysiology 8
2.1.3 PET imaging in the diagnosis of myocardial ischemia 9
2.1.4 Nuclear imaging and artificial intelligence 10
2.2 Transfer learning for image classification 11
2.2.1 Overview 11
2.2.2 Artificial intelligence and machine learning 12
2.2.3 Artificial intelligence in medicine 13
2.2.4 Machine learning 14
2.2.5 Machine learning and computer-aided diagnosis 14
2.2.6 Deep learning for medical image classification 15
2.2.7 Transfer learning 19
2.3 Similar investigations 22
3 Materials and methods 24
3.1 introduction 24
3.2 Patient population 24
3.3 PET/CT imaging and acquisition methods 25
3.4 Invasive coronary artery angiography 25
3.5 Analysis of the cardiac perfusion data 26
3.6 Polar map and label dataset description 26
3.7 Dataset configuration 27
4 Experiment design 29
4.1 Preface 29
4.2 Challenges 30
4.3 Workflow 30
4.3.1 Definitions 32

4.3.2 Training parameters and layer Adjustment 36
4.3.3 Evaluation metrics 40
4.3.4 Ensemble learning 42
5 Results and discussion 43
5.1 Results 43
5.1.1 Introduction 43
5.1.2 Pre-trained models 43
5.1.3 Comparison of transfer learning and a custom CNN 49
5.1.4 Ensemble learning 51
5.2 Discussion 52
5.3 Conclusion 53
6 Acknowledgments 55
7 References 56

List of abbreviations

Abbreviation	Meaning
AI	Artificial intelligence
AUC	The area under the "Receiver operating characteristic" curve
CAD	Coronary artery disease
CADx	Computer-aided diagnosis
CNN	Convolutional neural network
COVID-19	Coronavirus disease of 2019
CSC	Center for science Ltd.
CT	Computed tomography
CTA	Computed tomography angiography
CVD	Cardiovascular disease
DL	Deep learning
FDA	Food and drug administration
FFR	Fractional flow reserve
FN	False negative
FP	False positive
GPU	Graphics processing unit
ICA	Invasive coronary angiography
IDE	Integrated development environment
ILSVRC	ImageNet large scale visual recognition challenge
IV	Intravenous
KNN	K-Nearest neighbor
MBF	Myocardial blood flow
ML	Machine learning
MPI	Myocardial perfusion imaging
MRI	Magnetic resonance imaging
NLP	Natural language processing
PET	Positron emission tomography
RGB	Red, blue, and green
RO	Research objective
ROI	Regions-of-interest
RQ	Research question
SGD	Stochastic gradient descent
SVM	Support vector machine
SPECT	Single photon emission computed tomography
TN	True negative
TP	True positive
2D-OSEM	Two-dimensional ordered expectation-maximization algorithm

1 Introduction

1.1 Preface

Artificial intelligence (AI) and its subclasses, such as machine learning (ML) and deep learning (DL), have become the trend of research in the past couple of years in most fields of science. ML applications include every aspect of human lives, ranging from self-driving cars to intelligent health records and disease diagnosis techniques. It has been presumed that human intelligence cannot be conquered by machine intelligence for a long time because humans benefit from the power of intuition and emotions. However, during the last decade, groundbreaking advances in ML proved that AI, in general, can outperform human intelligence if provided with enough data and computation power.(1,2) Based on this premise, the research in medicine has shifted significantly from traditional methods towards AI techniques, specifically DL.(3) Unlike many other fields in the industry or science, the objective of AI is not to replace clinicians but rather to aid them with a confirmation of the diagnosis or treatment decisions.(1)

Medical imaging has been an essential concept in medical research and clinical practice due to the high accuracy and quantitative analysis power in visualizing physiologic and pathophysiologic conditions.(4) Additionally, since medical imaging is about processing and interpreting images obtained from a myriad of modalities such as microscopy or more complex methods, namely, nuclear imaging, it can use various machine learning algorithms in any task dealing with image data.(5)

This thesis is a study conducted at the intersection of computer science and medicine, where transfer learning is applied to a real-world problem in medicine: the classification of images obtained from nuclear imaging to diagnose myocardial ischemia. The results of this study will potentially provide insight into how transfer learning can be used in real-world practices by clinicians for a more accurate and effective diagnosis of diseases.

1.2 Thesis structure

The body of this thesis is organized into five chapters, including chapter 1 (introduction), chapter 2 (background), chapter 3 (materials and methods), chapter 4 (experiment design), and chapter 5 (results and discussion). Overall, the thesis begins with an introduction to the problem and the proposed solution, continues with the implementation section, and ultimately discusses the work results. The detailed structure of each chapter of the thesis is as follows:

Chapter 1 (Introduction) summarizes the main concepts discussed in this thesis. This chapter aims to familiarize the reader with the basic concepts of the study and introduce the key topics and aims of this research. Thesis structuring, problem definition, and approach positioning are the main contents of this chapter.

Chapter 2 (Background) focuses on defining the main topics of the thesis, such as myocardial ischemia and transfer learning. Plus, this chapter aims to provide scientific evidence supporting the problem statement. Finally, it attempts to articulate how previous studies have contributed to this research field and how this research can add value to the existing knowledge.

Chapter 4 (Materials and methods) discloses data gathering and processing details for this research while explaining the essential definitions in dataset preparation.

Chapter 5 (Experiment design) elucidates a comprehensive and detailed report on the experiment design process of the thesis. Additionally, the challenges in the experiment design and the terminology of the methods used in the study are clarified.

Chapter 6 (Results and discussion) unfolds the findings of this study and highlights the possible future directions in this research area.

1.3 Problem definition

Every year, more than 4.5 million people die because of cardiac diseases.(6) Myocardial ischemia has been the most lethal disease globally during the last decade.(6,7) Unlike many diseases, statistics of heat-related illnesses are not in favor of developed countries, where more than 38% of deaths every year result from cardiac diseases.(8) Therefore, it is essential to manage heart problems, including myocardial ischemia. In this regard, the early diagnosis of the disease, primarily through advanced imaging techniques such as nuclear imaging, can be beneficial and life-saving. (9)

Recently, nuclear imaging techniques such as PET imaging have gained significant attention in the diagnosis of myocardial ischemia.(9) The interpretation of PET-MPI images is typically performed by clinicians. With the help of AI techniques, clinicians can benefit from powerful computer-aided diagnosis (CADx) approaches in their interpretations.(10,11) In addition, AI approaches can be helpful in specific tasks, such as medical image classification, which is a critical problem in the diagnosis of diseases. For example, in the case of myocardial ischemia, "interpretation" of polar map images leading to a diagnosis decision is essentially a classification task. In such tasks, clinicians decide whether a specific polar map indicates healthy (high perfusion) or ischemic (low perfusion) condition.(9,12)

The field of AI-based diagnosis in medical imaging is a broad domain. AI approaches can vary significantly in algorithms and implementation based on the problem they seek to solve. (13)

In parallel, from the perspective of computer science, image classification is one of the fundamental problems that scientists have targeted to solve using AI methods. Because of the importance of image classification in various fields, from autonomous driving cars to medicine, considerable efforts have been made to develop more efficient and accurate AI techniques. As a result, advances in AI have led to significant paradigm shifts in image classification techniques, from classic ML methods to DL methods or, more precisely, CNNs.(10,11)

Starting from the ImageNet competition of the year 2012, CNNs have become one of the most popular techniques for image classification.(14) As a result, scientists in medical imaging have also tried to benefit from CNNs to solve medical image classification tasks for improved diagnosis accuracy.(15) A CNN is simply a predictive model that has been trained to understand and classify images from vast amounts of training data.(16,17) However, despite promising results of CNNs in most applications, they have been hardly adopted in practical medical image classification tasks where real-world data is often limited, and computing power is expensive. Additionally, accessing medical data from bioinformatic-related centers is ethically challenging.(15,18)

Transfer learning can be alternatively employed to tackle the challenges of developing a custom CNN. In transfer learning, the predictive power of a pre-trained CNN model gained from huge amounts of data from one task is transferred to a new task.(19–21) Various studies in the literature have proposed that transfer learning can leverage the performance and efficacy of DL in medical image classification tasks.(22) However, since pre-trained models have been initially trained on natural images, the relevance and practicality of transfer learning in medical image classification remain an attractive and challenging research topic.(23)

1.4 Motivation

From the perspective of image classification and ML in medical imaging, current studies focusing on transfer learning-based techniques have been primarily conducted to evaluate the performance of one or a few available pre-trained models. The objective of the conducted studies in this field has been to assess the validity and reliability of transfer learning in medical imaging by targeting a limited number of pre-trained models.(23–25)

Although previous studies have reviewed the feasibility of transfer learning in the classification of medical images, due to advances in the development of pre-trained models, an exciting research topic is to find the best pre-trained models from a medical imaging perspective.(23–29) In other words, existing studies have successfully addressed the viability of transfer learning in medical images, leaving many topics unaddressed, namely pre-trained model selection and benchmarking.

The novelty of this research is not limited to the implementation method and the methodology but also the imaging modality involved in the study. The field of nuclear medicine has been tied to DL, from image reconstruction and segmentation to image classification. However, as of mid-2022, unlike SPECT there is no research on PET-MPI polar map classification using transfer learning.(26)

Apart from the scientific value of the topic mentioned above, another motivation for this thesis and hopefully a future publication based on the results of this work is to practically bring value to the clinical process of diagnosis of myocardial ischemia. As the diagnosis of this condition relies on the opinion of doctors interpreting medical images, transfer learning can potentially act as applicable leverage to help elevate the accuracy and efficacy of clinicians' decision-making in diagnostic measures. Notably, it is not suggested that transfer learning or, in general, AI will replace clinicians. Instead, it means that doctors will benefit from the power of AI in their decision-making as an assisting tool for confirmation.(1,3)

1.5 Research questions

Although the knowledge in the field of transfer learning in medical image classification and the use of CNN models in the classification of images obtained from the PET-MPI technique is growing fast, the existing knowledge is insufficient in some important research topics. According to this, this thesis aims to fulfill the following research questions (RQ):

RQ1: Which pre-trained CNN models perform better in the binary classification of PET-MPI polar maps?

RQ2: How does transfer learning perform for the same dataset and task compared to a custom CNN developed and trained from the ground up?

RQ3: What is the optimum input image size of the PET-MPI polar maps for a pre-trained CNN model? Or, how does the size of input images for training relate to the performance of CNN models?

RQ4: How does an ensemble of best transfer learning models (found in RQ1 and RQ2) perform compared to every single model and the custom CNN trained on the same data?

1.6 Research objectives

According to the aforementioned research questions to be addressed, this thesis aims to achieve the following research objectives (RO):

RO1: Comparing the current state-of-the-art CNN models with each other and with a custom CNN trained from scratch on the same dataset.

RO2: Adjusting hyperparameters of available pre-trained CNN models and reviewing best settings in the architecture of models for improved performance in our task.

RO3: Building a new CNN model using ensemble learning from best-performing transfer learning models based on RO1 and RO2.

2 Background

2.1 Myocardial ischemia

2.1.1 Introduction

Even though the data suggest that myocardial ischemia is a lethal disease, it can be prevented and treated through practical approaches.(30–32) Given the significance of the global burden of coronary artery disease (CAD), developing accurate diagnostic techniques is of utmost importance. Non-invasive imaging techniques play an essential role in the CAD diagnosis and management and have demonstrated encouraging performance and results.(30,32–34) In the realm of cardiac imaging techniques, nuclear imaging modalities, including SPECT and PET, are modern approaches allowing for high specificity and high sensitivity myocardial perfusion imaging.(33,35) MPI technique provides anatomical and functional information about cardiac perfusion in patients with known or suspected CAD.(36)

Advances in AI and ML techniques have expanded to medical imaging territory wherein the recent past, only human cognition could perceive and translate abstract medical images into clinical interpretation. Over the past decade, CADx has been aiding doctors in making decisions based on ML-derived medical image analysis. In recent years, ML methods have been extensively used in predictive modeling to satisfy particular tasks such as image classification and object detection in medical images. Particularly, DL algorithms have gained significant attention due to their higher accuracy outperforming human cognition.(1,2,37,38) The use of DL in nuclear imaging has seen significant growth, with diverse investigations conducted mainly on the brain and cardiac PET/SPECT images demonstrating the feasibility of employing DL methods in image classification-based predictive modeling.(2,39) Despite the popularity of DL and its effectiveness in image analysis tasks, the fact that DL algorithms require massive training datasets for optimum performance restrains their usability in medical imaging tasks that suffer from a shortage of data. In addition to this, from the ground up training of DL

algorithms is computationally expensive and hardly accessible. The challenges of training from scratch have given rise to the implementation of transfer learning in medical image analysis with smaller datasets. (19–21,29)

Transfer learning is an ML technique that relies on transferring knowledge gained from one task to a relatively similar task. This approach is beneficial to DL problems requiring immense computational power, large datasets, and lengthy episodes to train neural networks.(19–22) A class of neural networks is CNN that is most predominantly used in DL algorithms and has been widely applied to medical imaging tasks such as image classification showing promising performance.(24) Nevertheless, training CNNs for medical image analysis tasks faces the challenge of insufficient data. Transfer learning aims to tackle this challenge by employing pre-trained CNN models trained on a huge dataset of natural images, supplying learned features similarly found in medical images.(29,40)

2.1.2 Pathophysiology

Myocardial ischemia denotes heart conditions arising from the restriction of blood flow to the heart muscle through coronary arteries. Various clinical ischemic manifestations are caused by obstruction of coronary blood flow by coronary stenosis, thrombosis, and/or hyper-constriction (vasospasm) of epicardial and microvascular coronary arteries.(41)

Myocardial ischemia is the consequence of unbalanced myocardial oxygen supply and demand. The reduced oxygen supply imputable to restricted myocardial blood flow causes reversible myocardial suffering. In chronic conditions, myocardial ischemia may cause irreversible injuries such as myocardial infarction and sudden death. The reason is that the myocardium suffers the loss of its full ability and capacity to pump blood.(12,30,41)

As the leading cause of mortality worldwide, cardiovascular diseases (CVDs) are a class of heart and vessel disorders, including several heart conditions such as CAD. Amongst all CVDs, CAD is the primary cause of death.(6,8,42)

2.1.3 PET imaging in the diagnosis of myocardial ischemia

A standard diagnostic test is cardiac imaging in patients with known or suspected CAD. Non-invasive cardiac imaging tests include magnetic resonance imaging (MRI), ultrasound imaging, computed tomography scan (CT), SPECT, and PET. All these tests can be carried out to demonstrate visual and interpretable information to evaluate myocardial perfusion.(9,12,33,36)

During the last two decades, nuclear medicine imaging modalities, including SPECT and PET-MPI, have received significant attention as accessible medical exams in administering ischemic heart problems. PET-MPI is a cardiac imaging technique that provides for the calculation of nuclear stress test and analysis of myocardial blood flow (MBF) simultaneously.(5,12,32)

PET imaging for diagnosing and managing CAD includes two main clinical applications. The first implementation concerns measuring MBF, while its second use is to evaluate myocardial metabolism in cases with the dysfunctional ischemic left ventricle. Interpretations obtained from both use cases of PET imaging in cardiology have contributed to a clearer insight into the characteristics of ischemia-associated heart diseases.(33,43,44)

The idea behind SPECT and PET imaging lies in tracking a certain radiotracer in the target organ inside the body and its uptake. For ischemic heart diseases, a specific radiotracer associated with certain physiologic events of the heart (i.e., perfusion) is injected into the patient's body. The quantification of the heart's specific functions is then calculated from the emission of the injected radiotracer and its reactions. These are commonly visualized in a polar plot (e.g., a polar map). Among nuclear cardiology imaging modalities, PET imaging is of higher demonstrative capabilities due to its higher radioactive count rates, increased spatial resolution, and lower radiation burden.(43,45)

To depict the physiological activity of a particular organ, short-lived PET radiotracers are injected into and tracked in the bloodstream of a patient. PET radiotracers consist of a positron-emitting radioisotope attached to an organic ligand which participates in a chemical reaction in the body, resulting in a characteristic distribution of the tracer throughout the tissue. PET imaging can produce 3D images of radionuclide distribution in the body.(45,46)

The name PET derives from the fact that the radioisotopes in the structure of radiotracers are unstable and decay by positron emission. An event of decay is when the radioisotope becomes stable. When the nucleus of the radioisotope decays, a positron is ejected, traveling only a short distance before encountering a nearby electron. The collision of a free electron with the emitted positron results in the annihilation of both particles, producing two high-energy gamma rays emitted in opposite directions. The detection of these high-energy emissions in opposite planes in a PET detector ultimately leads to the generation of images.(45–47)

Commonly used PET tracers in the clinic are 13N ammonia, 15O water, and 82 Rubidium. The selection of each tracer is based on the application and circumstances under which the imaging is operated. The reason is the different characteristics of PET tracers such as half-life time, manufacturing profile, uptake, and resolution of imaging induced by positron range.(48)

For quantitative flow assessments using PET imaging, an optimal candidate is 15O water because its uptake is not compromised due to adjacent metabolic activities in the organ. Based on the study of Manabe et al., the high extraction fraction of this tracer makes it possible to use it with low doses in short periods of stress and rest. Additionally, with 15O water, the exposure to radiation can be minimal. However, in clinical applications, 15O water is not the most commonly adopted tracer due to its relatively noisy images.(49)

2.1.4 Nuclear imaging and artificial intelligence

With the advent of AI, cardiovascular imaging, and nuclear cardiology such as PET, myocardial perfusion has been proved to be favored with the utilization of ML in the realm of image processing and image analysis.(12,31,32,35)

As a part of ML, DL uses multifold neural network layers to extract features from input data, including image data. DL algorithms have found special attention in nuclear cardiology image analysis due to direct cardiac image processing and myocardial ischemia identification and classification.(32,40,50)

Neural networks in DL have been inspired by networks of brain neurons and their functionality in learning processes. The foundation of neural networks revolves around layers of neurons in the order of input, hidden, and output layers, respectively. As their

name suggests, input layers comprise an extensive set of raw input data, namely "values" representing image pixels. Hidden layers are placed between the input and output layers and apply weights to the inputs. The last set of layers in a neural network is the output layer, where the final results for the targeted problem are obtained. (Figure 1) This problem can be, for example, image classification or object detection, which are everyday tasks in medical images for diagnostic and predictive purposes.(14,51–53)

A central factor in determining the preferred ML strategy for any task is the size of available data. Large datasets can benefit from a vast collection of algorithms for predictive modeling, while small datasets qualify to be engaged in fewer methods. Transfer learning is an ML strategy that can build predictive models from small datasets.(15,19–22)

2.2 Transfer learning for image classification

2.2.1 Overview

DL has proved to be a powerful tool for image classification tasks. Yet, to practically apply DL to medical image classification tasks, the challenge of small datasets and hardly accessible powerful graphics processing units (GPUs) should be addressed. The problem with small datasets in DL is that the model, initiating with random weights for training, is not fed with enough data to learn the features of an image represented by the modified weights during the training. In other words, the event of "learning" of a model heavily depends on the size of the training dataset. In the bargain, "training" a model is essentially an immeasurable number of calculations, making it highly computation intensive. To address the challenge of insufficient data and computation, transfer learning was introduced. Shortly, in transfer learning, instead of training a CNN with random weights, a pre-trained model taking in previously learned weights from another task is utilized.(20,22,23,54) Figure 1 represents the scope of transfer learning in the realm of AI and its subclasses.

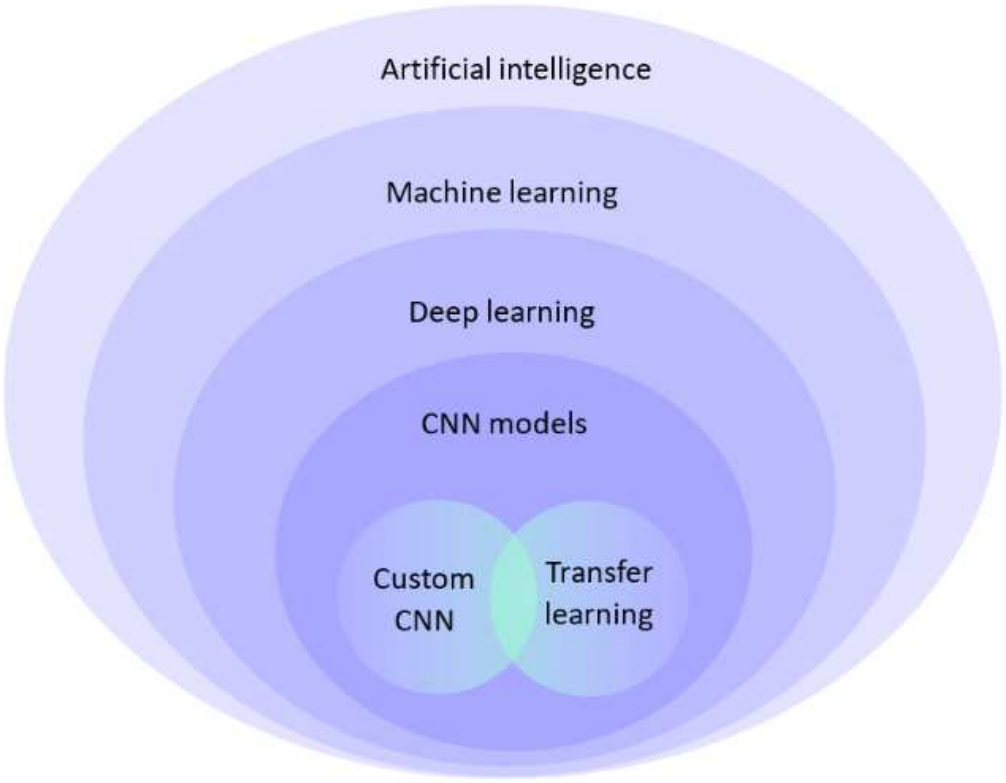

Figure 1. An overview on the scope of transfer learning in AI.

2.2.2 Artificial intelligence and machine learning

AI is not a new theory as it dates back to the 1950s when Alan Turing conceptualized the idea of human-level intelligence for machines in his literature, "Computing Machinery and Intelligence".(55) Turing proposed an innovative testing algorithm called the "imitation game" to evaluate a machine's ability in mimicking human-level perception. For the machine to pass the Turing test, a human evaluator should have failed to distinguish the response of a human from a machine in a text-based communication exam. Although Turing was the pioneer of human-inspired intelligence for computers, the term "artificial intelligence" was first coined by American computer scientist and cognitive scientist John Mccarthy, also known as the father of AI.(56) Mccarthy clarified his definition of AI as the science and engineering of building intelligent machines by describing “intelligence” as the computational ability to achieve goals. Contemporarily, AI is a general term that implies using a computer to model intelligent behavior with minimal human intervention.

Over the past several decades, with the advances in computer science and breakthroughs in the computing hardware industry in the 2000s, AI gained significant attraction and has evolved from simple "if-else" programming statements to intricate algorithms performing at a human level.(57)

2.2.3 Artificial intelligence in medicine

During the last 50 years, scientists have studied the prospects of applying AI in every possible healthcare domain. One of the first efforts in employing AI in a real-world medical problem was made in the diagnosis of abdominal pain through computer interpretation. In the last few decades, interest in AI applications in medicine has boosted aggressively, making traditional approaches in every field of medicine to reconstitute. Modern medicine has benefited from AI in disease diagnosis, therapeutics, and treatment. However, solving complicated problems in medical practice heavily relies on interpreting available medical data. AI aids in integral medical domains, including outcome prediction, treatment planning, and potentially preventive medicine and precision medicine. Ultimately, AI may improve diagnostic accuracy and facilitate disease treatment.(3,32,58)

The exponential growth of AI applications in modern medicine is primarily a result of ML and DL methods taking over classic AI approaches.(57) ML is a subclass of AI which has become eagerly popular in medicine as well as other fields of science. The effectiveness and effortlessness of ML approaches have made this topic an exciting space for researchers. As its name suggests, ML is based on the notion of machines being capable of automatically learning from input data, identifying patterns, and predicting outcomes with minimum human intervention. The concept of ML resembles human learning that revolves around learning from experience as well as trial-and-error in making decisions. DL is a specialized subset of ML that has dramatically been exerted by medical research, especially on image analysis tasks that require accurate information extraction from medical images. In DL systems, unlike classic programming techniques that rely on defining the relationship between an input and its output, both input and output are given to the computer with numerous examples as training datasets so the computer would learn and formulate the input-output relationship.(59,60)

2.2.4 Machine learning

ML is a class of AI that enables a computer to learn from data and make predictions. ML revolves around building predictive models recognizing patterns in data by running various mathematical functions. ML can be described as automated programming with minimum human involvement. However, human supervision can be of high importance in defining the main methodology of applying ML to a specific problem. Provided that, ML algorithms can be divided into supervised and unsupervised learning. In supervised learning, not only the input data but also the desired output data are given during the training phase, and the model learns to generate those outputs from the given inputs. On the other hand, unsupervised learning deals with unlabeled data and aims to analyze and cluster data for fundamentally different purposes.(59,61)

2.2.5 Machine learning and computer-aided diagnosis

Immediately after the dominance of computers with the infinite opportunities and possibilities they offered, the idea of building systems for automated medical image analysis was envisioned. During the 1970s to 1990s, researchers developed simple algorithms for certain medical image analysis tasks. At that instant, most of the automated systems were rule-based such as low-level pixel processors in image processing that served in edge and line detection. However, the performance of these so-called "good old-fashioned AI" systems has been reported as unreliable. Still, the rapid growth in the use of medical imaging modalities, bringing more and more medical images on screens, as well as the proven supremacy of ML methods, makes the vision of automated medical image analysis possible.(62)

One of the main areas in medical image analysis significantly affected by ML is CADx which aims to provide assistance and confirmation with disease diagnosis. During the last decades, classification tasks have been an exciting field for ML researchers to implement CADx algorithms in practice. It took 20 years for CADx systems to practically emerge in clinical applications and pipelines. One of the main issues associated with the CADx system has been the error of false-positive predictions compared to clinical interpretation. Because of this error, expenses and resources would be compromised. In other words, CADx systems would cause more expenses instead of decreasing costs of disease

management. However, with new advances in AI and the exponential growth in the power of computation, CADx systems are not far from reaching human-level interpretation.(62,63)

Medical image interpretation has been performed chiefly by human experts such as radiologists and physicians in the clinic. However, given wide variations in pathology and the potential fatigue of human experts, researchers and doctors have begun to benefit from computer-assisted interventions. However, it is emphasized that AI has not achieved human-level interpretation yet.

At present, the amount of medical data is enormous, but it is crucial to make good use of this substantial medical data to contribute to the medical industry. Although the amount of medical data is vast, there are still many problems: medical information is diverse, including maps, texts, videos, magnets, etc. Due to different equipment used, the quality of data varies considerably. Inconsistency in data types can be challenging for clinical interpretation.

2.2.6 Deep learning for medical image classification

2.2.6.1 Deep learning

From the perspective of advances in image classification techniques, ML can be divided into conventional ML and DL. The former is a classic subclass of ML methods with approaches such as SVM (support vector machine), random forests, and naive Bayes classifiers. Conventional ML methods have been used for many years to solve complex problems such as image classification. Although the performance of traditional ML has been reasonably confident for small-scale problems, more extensive input image data would be challenging. Alternatively, Cutting edge DL techniques dramatically advanced ML practice. The central advantage of DL over conventional ML methods is that it benefits from a multi-layered architecture, capable of extracting features from images without converting all pixel values into 1D arrays.(40,59,62)

The concept of DL dates back to the early 1940s when Walter Pitts and Warren McCulloch gave rise to a computer modeling system based on the human-inspired "neural networks".(64) However, the term "deep learning" was first introduced during research conducted in the Cognitive Systems Laboratory of the University of California by Rina

Decther in 1986.(65) Later in the early 2000s, the groundbreaking "artificial neural networks" were presented by Igor Aizenberg and colleagues to institute the very foundation of contemporary AI and DL.(66) The "deep" remark in “deep learning” denotes multiple layers of neural networks through which the data is transformed.

In traditional ML techniques, the primary approach was to develop pattern recognition algorithms that simply perform a statistical calculation on the raw data. For an image recognition problem, with up to millions of pixels as data, which is not necessarily meaningful, it is a challenging task to design a feature extractor to solely rely on pixel values. However, some conventional ML methods are still actively used for specific domains in image classification or in companion with DL models.

Since 2012, DL has become a popular branch of ML in computer vision problems. In DL, a multitude of layers stack up, building a predictive model empowered by connections of layers and input data.(14) DL is employed in a myriad of industries and scientific domains, from natural language processing (NLP) to computer vision. In computer vision problems, DL has proved to be the ultimate solution in specific tasks such as segmentation and classification. Image classification tasks take advantage of a layered structure of DL architectures to transform raw data into features in images. At the heart of this transformation, procedure are ConvNet layers. A ConvNet is a layer made up of many nodes or neurons, performing convolution operation on the input image data. A convolution operation scans an image using a filter (kernel), capable of extracting specific features from the image. For example, a filter can be an edge detector. Filters are simply matrices of numbers that manipulate pixel values of the input image size, resulting in different pixel values, hence different images. Each feature in an image can be filtered out with a specific matrix of numbers, and it is DL’s task to find those values. The process in which the DL model learns the numbers of the filters is known as “training”. During training, a model learns the values of the filters and stores feature extracting filters in each layer. The same process replicates in the subsequent layers, where the model finally learns to combine filters to extract high-level and complex features of an image. As an illustration, for an input image of a cat, the model first learns to extract basic features like colors and edges. Having learned basic features, the models learn to put together basic features to create patterns and then objects. (14,40,67)

Exponential growth in hardware development has also contributed to the unprecedented success of DL. The process of learning in a DL pipeline is essentially a series of

calculations. Without powerful central processing units (CPUs) and GPUs, developing DL models was hardly possible. Along with the advances in hardware, data generation began to grow, resulting in the creation of huge datasets that DL could potentially learn from.(14)

2.2.6.2 Image Classification

Despite humans' natural ability in extracting features from images and classifying them into certain categories by nature, it is a challenging task for a computer to distinguish patterns in images and annotate them based on predefined and given labels. In computer vision, image classification is a foundational task as it can be an elementary measure in other computer vision problems like segmentation.(67,68)

The emergence of Image processing provided diverse opportunities in the digital world where images can be considered as numbers and be subject to mathematical operations. The most valuable effect of mathematical operations in image processing is that information and data can be extracted, analyzed, and used for different purposes. One of the ways in which data is extracted and used for various applications is image classification.

Considering that the nature of a classification problem is predicting the image class, ML techniques can be employed to perform the classification operation. Given that, image classification is a problem in both image processing and ML that relies on the extraction of features from an image to predict its class conclusively. Mathematically, a prediction is simply the probability of an event which in this case is if a given image falls under a specific class. A class is practically a label such as 'animal', 'car', 'red', etc.

As an example, for an image classification task, a popular project is "handwritten image recognition," which is a multi-label classification problem. In this task, a dataset of handwritten numbers in various shapes is used to train a model. In this phase, handwritten digits are labeled with the desired output, which is the actual number each digit represents. The training images as well as their labels, are fed into the model for training. Having learned the features of handwritten images, the model then predicts what number is represented by an unknown given image. In other words, it calculates the probability of an unknown input image being a specific number from 0 to 9. In this example, each number represents a 'class'.(69)

Classification techniques developed during the years have shifted from classic ML methods, namely SVM, decision trees, and K-nearest-neighbor, towards artificial neural networks and, more specifically, DL methods. The turning point in the classification methods is the development of a deep convolutional neural network model called AlexNet that won the first prize in the ILSVRC (ImageNet Large Scale Visual Recognition Competition in the year 2012. AlexNet achieved an outstanding reduced top-5 error rate of 15/3% in the image classification task of the competition, outperforming the runner-up model by more than 10% improved performance. AlexNet was named after Alex Kirzhevsky, who published the paper "ImageNet Classification with Deep Convolutional Neural Networks" with his colleagues Ilya Sutskever and Geoffrey Hinton. This paper has been counted as the turning point of image classification techniques and a major milestone in CNN-based methods. As of mid-2022, the AlexNet paper has been cited more than 107 thousand times according to the google scholar database, making it one of the most influential publications in the history of computer vision. To explain the outstanding performance of their model, Kirzhevsky et al. primarily concluded that the depth of the CNN model was the key factor that resulted in the most accurate classification. Along with that, the possibility of training huge data on a deep CNN model was attributable to the fast GPU utilization for model training.(14,67,70)

In traditional ML techniques, the model could not learn neighbor information of a distinctive feature in an image, making the model inefficient for complex object recognition tasks. On the other hand, DL employs multiple deep and hidden layers of neural networks (i.e., CNNs) to learn high-level features of an image by comprehending neighbor information of pixels in an image.(14,40,57)

2.2.6.3 Convolutional neural networks

The power of DL in image classification tasks has peaked at a human-level interpretation that has roots in the architecture of CNNs. A CNN is a class of artificial neural networks, most commonly aimed at computer visual comprehension tasks. As their name suggests, CNNs apply a mathematical function called convolution operation to the images for “learning” not only from single pixels but also from surrounding pixels. Through consecutive convolutional operations, the model learns and extracts specific features in

an image and ultimately, based on the given input labels, classifies any given image into a certain class.(17,70,71)

A CNN model has a multi-layered architecture, including an input layer, convolution layers, max-pooling layers, fully connected layers, and the output layer. All layers before the fully connected layer are in charge of extracting features from images. Fully connected layers and the output layer are the classifier head of a CNN model. The performance of a CNN model relies on the arrangement of all layers involved in its architecture.(14,40)

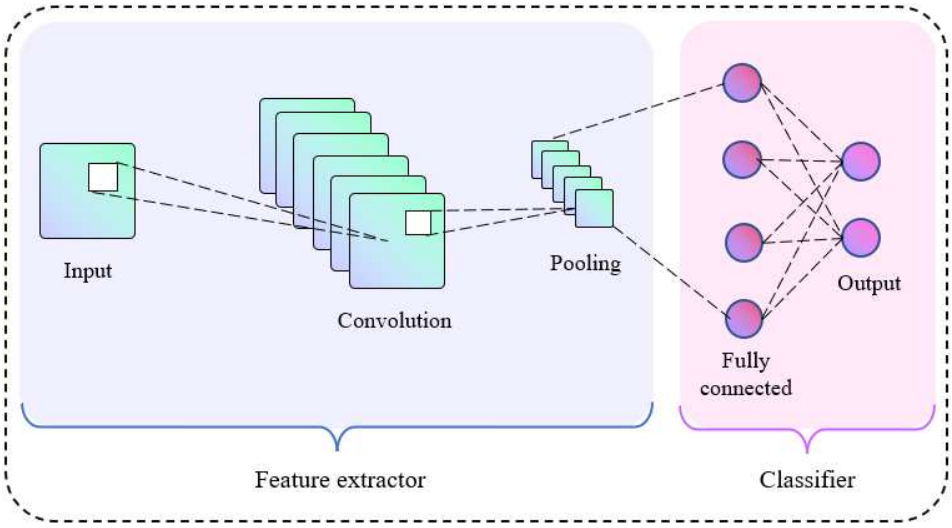

Figure 2. **A schematic view on the architecture of a CNN model.** On the left, feature extractor section of the model and on the right, the classifier top of the model is shown.

2.2.7 Transfer learning

The idea of transfer learning is originated from how humans learn. Prior knowledge stored in our memory accelerates the process of learning a new skill or subject. Typically, the more a new skill is related to our prior experiences, the faster learning transpires. For example, learning a musical instrument such as a guitar shrinks the learning curve of another related instrument like a violin. In other words, knowledge stored from one task is transferable to another task.

In ML, the learning process is defined by storing the correct values (weights) in feature extracting layers corresponding to a specific feature in the data. Therefore, the knowledge or weights learned from one task can be stored, and transferred to another task, hence transfer learning.(19–22)

Especially in DL, transfer learning has been adopted quickly because of the effortlessness it offers. DL algorithms are difficult to train. Firstly, training a model requires extensive resources from hardware and software to time and expert engineers with a deep understanding of mathematics and statistics behind DL algorithms. Secondly, in specific domains, accessing data is limited or available data is inadequate. To overcome the limitation of time and cost of training an entire model, transfer learning can be used. Figure 4 illustrates a general workflow of an image classification task using transfer learning compared to a conventional DL pipeline.(54,72,73)

2.2.7.1 Pre-trained models

Pre-trained models are simply CNN models developed on one task and available to deploy for a new task. The power of pre-trained models stems from the huge dataset provided to feed their hidden layers and the computation power on which they were trained.

Pre-trained models (sometimes referred to as networks) are open access and can be downloaded from online repositories such as Keras applications and TensorFlow hub. These networks can be applied to a range of ML problems, from image recognition to audio classification. Pre-trained models are evaluated based on their performance in the ImageNet competition. Every year, scientific teams from across the world participate in this competition to propose a CNN architecture for image recognition results on the ImageNet dataset.

2.2.7.2 ImageNet dataset

ImageNet dataset is a huge dataset of more than 14 million images in 20,000 classes. Each image in the dataset is “hand-annotated”. The annotation process aims to tag images with specific labels, so the class of the image is identified.(74) In the ImageNet competition, the target task is to classify images correctly and accurately in this database. As of mid-2022, pre-trained networks trained on this dataset have reached an 85.7% top-

1 accuracy and 97.5% top-5 accuracy (EfficientNetV2L).(75) Example of images in ImageNet dataset is presented in Figure 3.

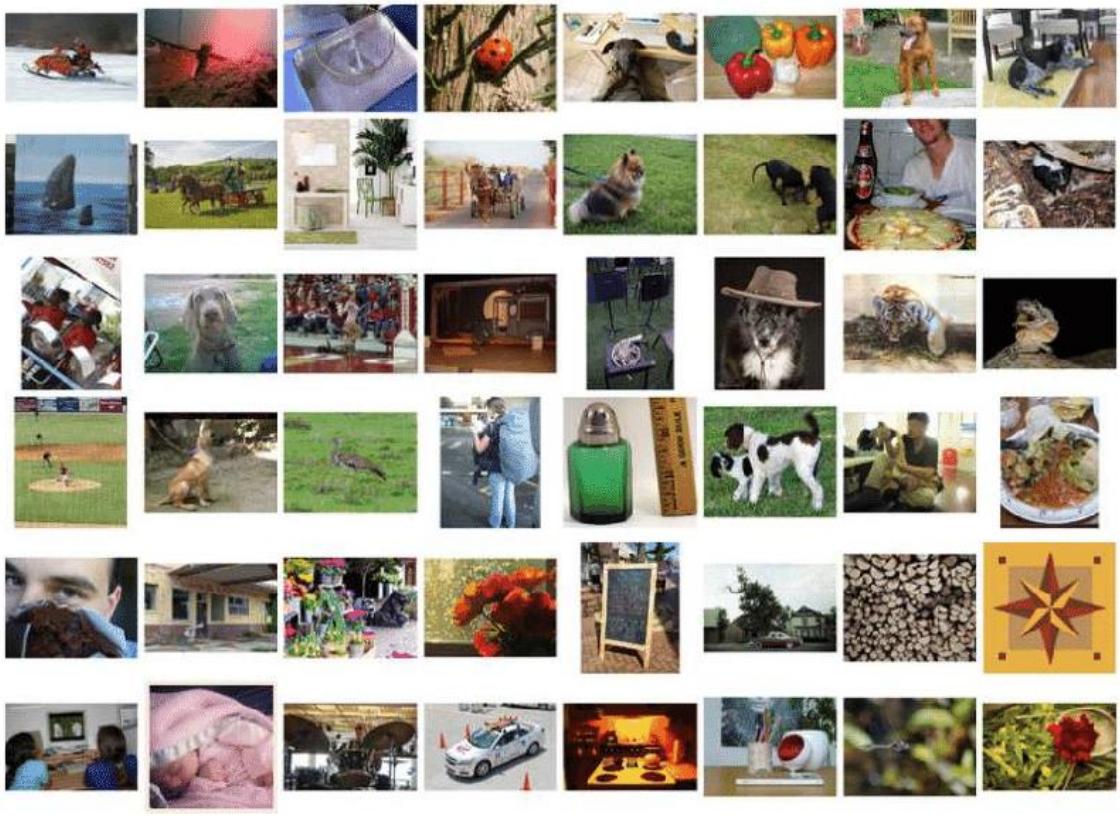

Figure 3. Sample images from ImageNet dataset.

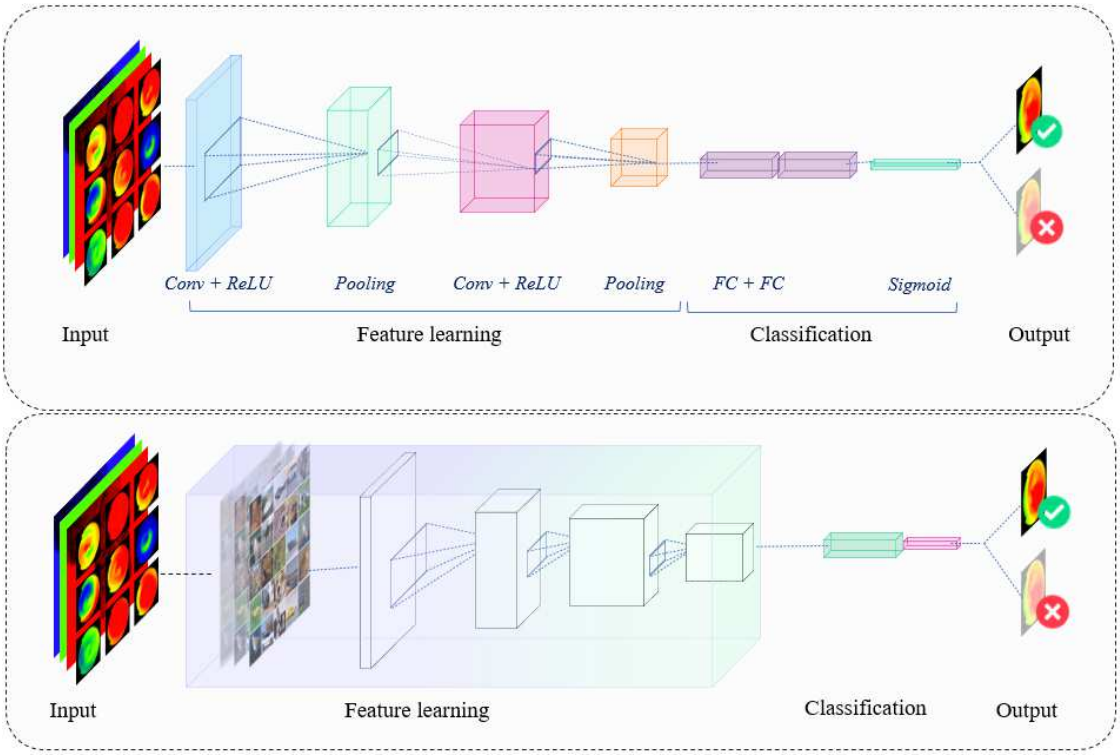

Figure 4. Training a custom CNN versus using a pre-trained model. The pipeline at the top represents the architecture of a CNN model, and the image on the bottom side represents transfer learning using ImageNet weights in the feature extractor base model.

2.3 Similar investigations

Owing to the development of DL, the shortcut approach to the same technique, transfer learning has also found an essential role in many applications including bioimage analysis. Several studies have been conducted on the use of transfer learning in medical image classification. For instance, two pioneering investigations were made on chest X-Rays using ResNet, DenseNet networks, and another work was done on ophthalmology using InceptionV3ResNet on images from the retina.(76) The latter was successfully able to receive the approval of FDA. Other investigations using transfer learning in the medical field include skin cancer identification and early diagnosis of Alzheimer's.(28)

In a recent study on SPECT myocardial perfusion imaging, transfer learning was used for the classification of abnormalities such as infarction and ischemia. Several pre-trained networks were employed in the research including AlexNet, DenseNet-201, GoogleNet, NASNetLarge, ResNet, VGG16, and VGG19. Researchers concluded that the results from transfer learning implementation were close to expert interpretation and could be used for second confirmation.(77)

Recently, a comprehensive study compared 15 available pre-trained networks in the identification of COVID-19 samples from chest X-Ray images. VGG pre-trained models performed best in that work with more than 89% accuracy in classification results.(78)

3 Materials and methods

3.1 introduction

ML techniques are intrinsically data-driven. For a model to learn and extract features from images, it is crucial to be supplied with enough data. In transfer learning, though, the size of a dataset can be relatively small because the pre-trained models are supposed to carry the learned weights from a previous task on a huge dataset. In this thesis, a total of 138 polar maps obtained from PET-MPI imaging using a PET/CT hybrid scanner were used. As this thesis requires data set organization for the training of the pre-trained models, the dataset configuration should be performed. Shortly, the arrangement of images in the dataset was set up entirely based on the original paper "Classification of ischemia from myocardial polar maps in 15O–H2O cardiac perfusion imaging using a convolutional neural network" by Teuho et al. where a custom CNN model was developed for the same polar map images. The purpose of mirroring the dataset arrangement from the work of Teuho et al. was to effectively compare the performance of pre-trained models in transfer learning against a custom CNN model with the same images and dataset preparation scenario. The rest of this chapter presents the details of image acquisition and data preparation.

3.2 Patient population

This study includes a total number of 138 subjects who had manifestations of obstructive CAD and were admitted to Turku University Hospital during the years 2007 to 2011. Symptoms of the patients were counted as stable chest pain or equivalent clinical manifestations. Plus, pre-test probability prediction for obstructive CAD was also made. Informed consent from all patients was collected. Additionally, the study was conducted based on the ethical instructions of the Declaration of Helsinki as well as securing the approval of the local ethics committee of the Hospital District of Southwest Finland.

Additional details on the study population can be found in the work of Stenström et al.(79) From the original 189 patients in the study, a selection of 138 cases was made who had available ICA and PET perfusion data in the form of polar maps from the stress imaging test. The dataset will be divided into subsequent training/validation and test datasets randomly.

3.3 PET/CT imaging and acquisition methods

In the first instance, the imaging device Discovery VCT PET/CT scanner (GE Healthcare Co., US) was used. The imaging protocol was computed tomography coronary angiography (CTA) plus MPI through PET/CT hybrid technique. Immediately after a CT-based attenuation correction, an adenosine stress perfusion PET was executed. Adenosine was launched 2 minutes before the start of the scan and was infused at 140 μg/kg body weight per minute. Oxygen-15 labeled water (900 to 1100 MBq) was administered (Radiowater Generator, Hidex Oy, Finland) as an IV bolus over 15 seconds. Subsequently, A dynamic mode acquisition for perfusion assessment of the heart was performed (14×5 seconds, 3×10 seconds, 3×20 seconds, and 4×30 seconds). Ultimately, image reconstruction was done on obtained images using a two-dimensional ordered expectation-maximization algorithm (2D-OSEM) applying a 35 cm field of view, 128x128 matrix size, 2 iterations, 20 subsets, and a 6.0 mm Gaussian post-filter.

3.4 Invasive coronary artery angiography

In a part of data gathering by clinicians, ICA or invasive coronary angiography was performed. Also, a measurement of FFR was executed for stenoses with intermediate severity, which is 30-80%. An experienced reader performed a quantitative analysis of ICA angiograms using automated edge-detection software. Thereby, obstructive CAD was defined as either >50% stenosis on ICA or FFR < 0.8. With FFR available, stenosis with FFR > 0.8 was classified as non-significant, regardless of the degree to which the coronary artery was narrowed.

3.5 Analysis of the cardiac perfusion data

For quantitative analysis of dynamic PET perfusion images, Carimas 2.9 software (Turku PET Centre, Finland) was used by a single reader blinded to the ICA results. Having defined the orientation of the heart manually, the myocardium was recognized automatically by the software. Plus, when required, the resulting regions of interest (ROIs) were manually adjusted. To obtain quantitative polar maps of stress MBF values in units of ml/g/min15, mathematical modeling was carried out, based on a single tissue compartment model. Additionally, having defined the threshold for ischemic stress MBF (< 2.3 ml/g/min), stress MBF values in polar maps were equally scaled from 0 to 3.5 ml/g/min using the Rainbow color scale.

3.6 Polar map and label dataset description

In the evaluation, a total of 138 polar maps for stress MBF along with corresponding ICA labels will be used. ICA data will be used as a gold standard in assigning the reference labels. Each polar map will be classified as ischemic (1) or non-ischemic (0) based on the ICA-defined obstructive CAD.

Polar map images were obtained from Carimas software in a high-resolution 2D JPEG format. Images were originally exported from Carimas with a size of 1024×1024 pixels, followed by automatic cropping and shrinking to 256×256 pixels in the processing pipeline. The pixel values were scaled and normalized between [0,1]. Polar maps were three-color RGB channeled, and all channels were used as input. Additionally, on the subject of image size, the input image sizes during model training, to meet the objectives of the research, the input image size for each experiment was set to be modified accordingly. The details of input image size modification, as well as other training hyperparameters, are discussed in chapter 6. Examples of polar map images are shown in Figure 5.

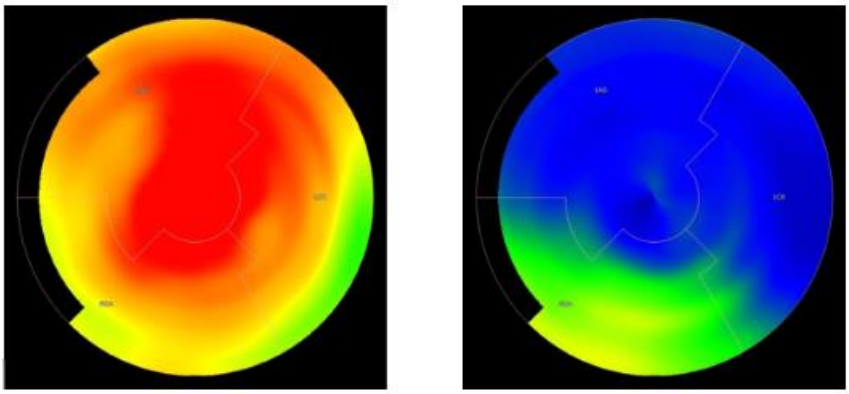

Figure 5. Example images of polar maps after processing. Red and yellow colors represent non-ischemic. Blue and green colors represent ischemic conditions.

Having polar maps processed, they were labeled as ischemic or non-ischemic. Labeling images is an essential step in supervised learning classification tasks. Each polar map was annotated with a label of 0 and 1 based on the ICA reference data. The labels were written in order as a .txt file.

3.7 Dataset configuration

In any standard DL process, one of the most important steps is to prepare the dataset. The dataset configuration is an essential setting required to be applied to the available data. The general purpose of the dataset is to avoid common errors in training, such as overfitting and data snooping.

In overfitting, the model returns significantly accurate results on the training data but the actual performance of the model on new and unknown data is meaningfully low. Data snooping or data fishing is another bias in which the data is misused to artificially create outstanding performance on the task.(80) To avoid such common statistical data exploitation practices, the dataset should be configured with appropriate portions of data in specific sets, namely the training set, validation set, and test set.

The training set is a set of data that is used during model training. This segment of the data is normally the biggest set of the whole available data. Training data is fed to the model consecutively. Thus, the model learns the patterns in the data. The provided training set should be diverse enough so that the model learns both image classes and their corresponding features.

The validation set is another segment of data involved in the model training. This means that using the validation set, the model performance is observed and monitored during the training process after each epoch. The validation set aims to provide information on the learning and loss function of the model during training. Training hyperparameters can be tweaked for better performance based on the learning information from loss function during the training, owing to the use of a validation set. Besides, the monitoring of validation set results during training can help avoid overfitting. In cases where the accuracy of a model on training data is significantly higher than the results on the validation set, there is a chance that the model is overfitted.

The test set or hold-out set is the third portion of the data, which is used completely after training, solely for testing the performance of the model. All the results demonstrating the performance of a model are obtained from the test set, which is unknown and unseen to the model.

In this thesis, 138 polar map images were split into a training set and validation set with 92 images in total and a test set with 46 images. The test set was only used during the evaluation of the models and was shown to the models during training. The training and validation split ratio of 1/3 was used, leaving 61 images in the training dataset and 31 images in the validation set. Figure 6 illustrates the workflow and organization of the polar map datasets and their role in each part of the experiments.

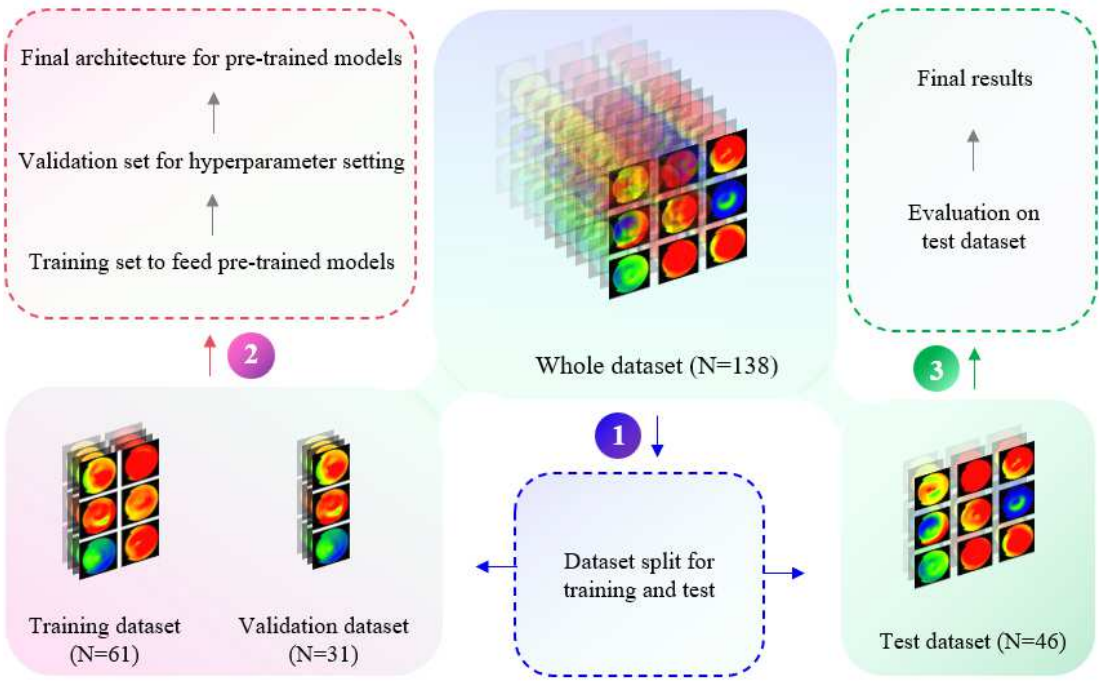

Figure 6. Dataset configuration workflow. 1) Whole dataset is split to training and test (hold-out) dataset. 2) The training and validation data splits applies. Training and validation sets are used during training. 3) test set is only used for performance evaluation.

4 Experiment design

4.1 Preface

In this chapter, the experiment design of this thesis is discussed. The experiment design of this thesis aims to fulfill the research objectives and research questions of the study that were described in chapter 1. A summary of the questions the experiments seek to answer is demonstrated in table 1.

Table 1. A summary of research questions.

Research question
RQ1. Which pre-trained models perform best for binary classification of PET-MPI polar maps?
RQ2. How does transfer learning perform compared to a custom CNN?
RQ3. What is the most optimum image size for pre-trained CNN models?
RQ4. How does an ensemble of best pre-trained CNN models perform for PET-MPI polar map classification?

The experiment design of the study should be formulated on certain principles to fulfill the mentioned research questions, establishing a reliable relationship between the variables in question and their result. The principles considered in this process are as follows:

1. In the experiment design process, the input image size as a dependent variable which is to be measured in terms of the effect on model performance, is separated from other hyperparameters making up the constant variables.

2. Other hyperparameters, including batch size, epoch number, fine-tuning, and optimizer selection, require tuning and adjustment to ensure optimum performance.

4.2 Challenges

The challenging section in the organization of experiments in this thesis was the number of variables in question. As the objective of this thesis was primarily to find the best pre-trained CNN model and the best input image size (roughly 4 image sizes for each of 20 models), the number of code executions excluding the remaining hyperparameters to be tuned would be supernumerary.

The other challenge of this study is about tuning hyperparameters of the models as well as finetuning the models for our task with their specific traits. Hyperparameters and network architecture values for each model sum up to a myriad of possible settings making it hardly possible to study all of the scenarios in one investigation. Each of the hyperparameters and network architecture values can be a separate subject of research, but since the objective of this research is to benchmark all models on a specific task and explore the effect of image size on the model performance, the remaining values and hyperparameters should be constant. Yet, the models must be tuned for the binary classification task and the dataset of this study.

4.3 Workflow

To tackle the challenges of experiment design, a pre-execution step was added to the study where the constant hyperparameters and the architecture of the models were tuned. This stage was performed on a trial-and-error basis, consequently, the best parameters and network settings were chosen for the rest of the study. As a result, the process and the results of each trial-and-error for this section of the experiments were eliminated from the research. After the pre-execution step, which is the hyperparameter adjustment stage, all the pre-trained models with different input image sizes are evaluated by consecutive executions of a unified code. Finally, the performance of each model with a specific input

image size is evaluated and reported. A summary of the experiment design workflow is demonstrated in Figure 7.

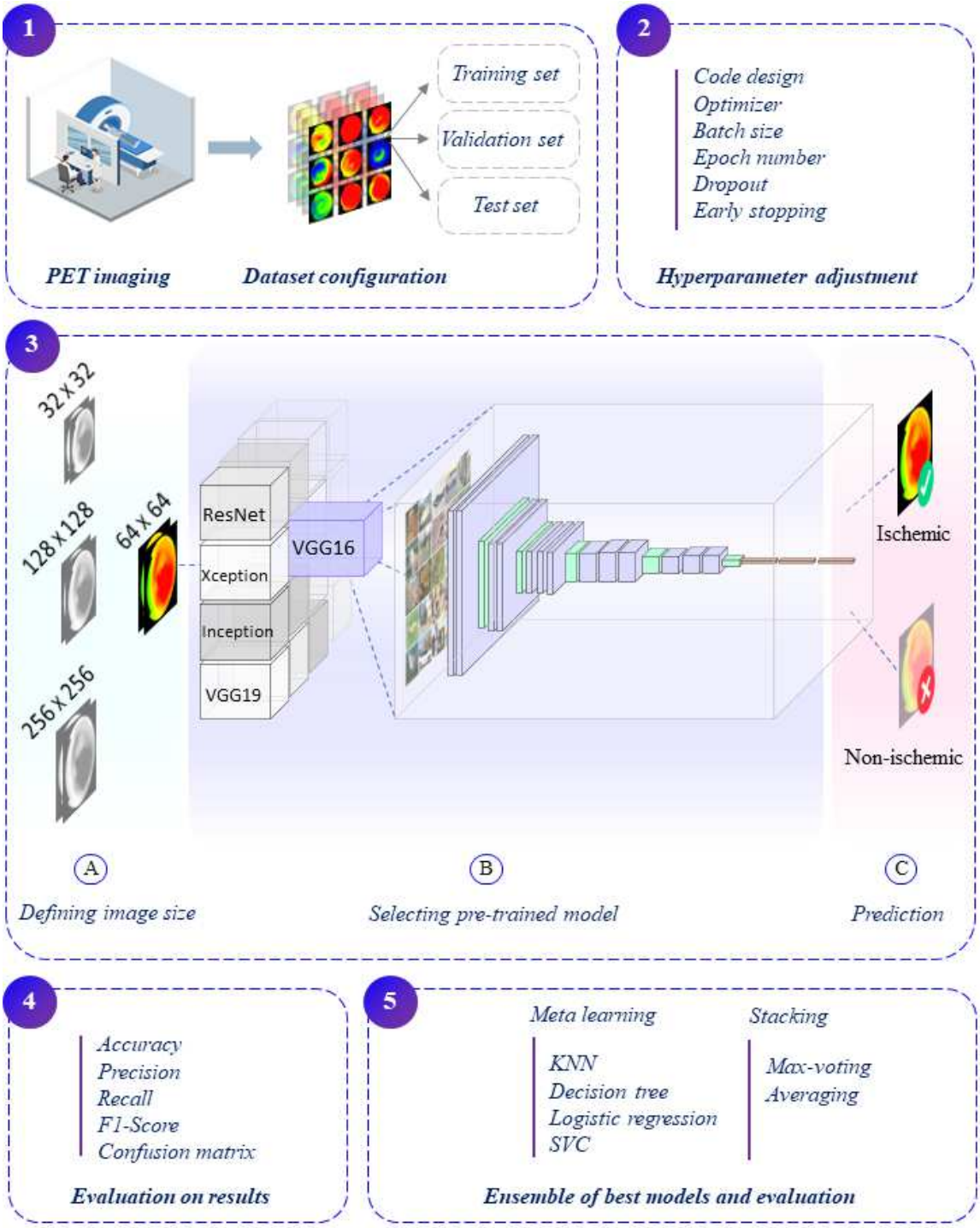

Figure 7. The entire workflow of the experiment design 1) Image acquisition and dataset configuration. 2) Hyperparameter adjustment and classifier architecture design. 3) Model selection and input image size definition for the main experiment to obtain prediction results. 4) Evaluating the results of experiment. 5) Ensemble learning.

4.3.1 Definitions

To better understand the process of experiment design in this thesis, it is helpful to inspect the representative terminology of transfer learning in a DL image classification task, used extensively in the literature as well as in this study.

I) **Trainable weights:** in the workflow of developing a CNN or using a pre-trained one, the knowledge of the model gained during the training is called weights. Trainability is an attribute of the weights in a CNN model. Usually, all weights are trainable (except for specific layers like the batch normalization layer) unless they are set to be non-trainable.

II) **Freezing and unfreezing:** the term freezing in DL denotes the act of locking knowledge stored in one or a set of ConvNet layers. By freezing a layer, the weights and biases learned through an initial learning process are locked and will not be compromised during a following backpropagation in training. In freezing, the weights of a layer are not updated, hence non-trainable.

III) **Base model:** the model trained with source data set on the source task, in this case being the ImageNet dataset for the image classification of natural images, is called the base model. Examples of base models are VGG19 and Xception models. To transfer the knowledge (weights) learned from the source task to the target task, the base model is used and modified by removing the top layers acting as classifiers. The top layers are task-specific layers, including neurons for as many classes as involved in the classification task.

IV) **Fine-tuning:** an optional step in a transfer learning workflow is to fine-tune the base model. Fine-tuning aims to improve the performance of a pre-trained model for the target task by unfreezing all or some ConvNet layers of the base model. However, for tasks with relatively small datasets, having layers frozen is the recommended approach.

V) **Feature extraction:** this term can be referred to as a method in transfer learning as well as a term in DL, meaning literally the act of using algorithms for dimensionality reduction to extract information from images. On the other hand, in transfer learning, feature extraction is considered an approach where the base model weights are respected, frozen, and not fine-tuned for the target task. Feature extraction is a fairly lightweight approach compared to fine-tuning a base model.

6.3 Implementation

In this section, the implementation of transfer learning for the classification of ischemia from polar map images is described. Overall, after choosing a selection of pre-trained models for this study, a unified python code taking in all 20 pre-trained models and the relevant libraries for training and evaluation of the models was developed. Afterward, the architecture of the final models, as well as training hyperparameters, were adjusted and finalized for the code execution. Eventually, after running the code, the results were evaluated and visualized. In the following, the process of each phase of the implementation is elucidated.

6.3.1. Model selection

The first step in applying transfer learning to any ML problem is pre-trained model selection. In this case, the pre-trained CNN models must be trained on an image classification source task. Pre-trained models for image classification are available in a myriad of libraries and open-access websites. Tensorflow Hub and Keras Applications are the two popular repositories building in a collection of state-of-the-art pre-trained models for image classification.

In this thesis, pre-trained models were downloaded from the Keras Applications repository. Based on published literature, specific models such as VGG16, VGG19, and Inception have been reported to demonstrate satisfying performance in medical image classification tasks. On the other hand, some models have received less attention in this domain, such as DenseNet and EfficientNetB family. Table 2 shows the selected pre-trained models and an overview of their top-5 and top-1 accuracy on the source task.

Table 2. A summary of selected pre-trained models from Keras website.

Model	Size (Mb)	Top-1 Accuracy	Top-5 Accuracy	Parameters
Xception	88	79.0%	94.5%	22.9M
VGG16	528	71.3%	90.1%	138.4M
VGG19	549	71.3%	90.0%	143.7M
ResNet50	98	74.9%	92.1%	25.6M
ResNet101	171	76.4%	92.8%	44.7M
ResNet152	232	76.6%	93.1%	60.4M
InceptionV3	92	77.9%	93.7%	23.9M
InceptionResNetV2	215	80.3%	95.3%	449
MobileNetV2	14	70.4%	90.1%	3.5M
DenseNet201	80	75.0%	93.6%	20.2M
DenseNet121	33	76.2%	92.3%	8.1M
DenseNet169	57	76.2%	93.2%	14.3M
EfficientNetB0	29	77.1%	93.3%	5.3M
EfficientNetB1	31	79.1%	94.4%	7.9M
EfficientNetB2	36	80.1%	94.9%	9.2M
EfficientNetB3	48	81.6%	95.7%	12.3M
EfficientNetB4	75	82.9%	96.4%	19.5M
EfficientNetB5	118	83.6%	96.7%	30.6M
EfficientNetB6	166	84.0%	96.8%	43.3M
EfficientNetB7	256	84.3%	97.0%	66.7M

6.3.2 Code development

Due to a large number of selected models, and the need for adjusting training hyperparameters, equivalently for all pre-trained models, a unified code encompassing all the models and the settings was developed.

The main framework for implementing the neural network layers, hyperparameters, and architecture modification of models was the Tensorflow library. Plus, the Keras library was also utilized for essential settings applied to pre-trained models. Python (version 3.8) and Anaconda were used as an interpreter and environment managers. Accordingly, all pre-trained models were called from the Keras Applications repository and imported into one .Py file. Necessary python packages, namely Tensorflow, Pandas, Numpy, and Sklearn, were used for training and evaluation of models using Anaconda as Python environment manager. The code was executed and developed on PyCharm community edition IDE and Windows 11 operating system.

Along with the necessary libraries, Argparse, a python command-line parsing module from the standard python library, was used. On top of Argparse, a user-friendly command-line interface was created in a batch (.bat) file. The batch file provided ease of use in changing hyperparameters and CNN model settings for the trial-and-error phase of the experiment as well as the final code execution. Figure 8 shows the interface of the batch file used for the experiment execution.

```
@ECHO OFF
TITLE PET Classification Experiments
ECHO Activate Environment
ECHO ###########################################################################################################
call conda activate polarmaps1
ECHO Run Experiments
ECHO ###########################################################################################################

REM Experiment design and running the code

python experiments.py --model_name Xception --input_size 128 --freeze_fe --batch_size 5 --epochs 35
ECHO ------------------------------------------------------------------------------
python experiments.py --model_name VGG16 --input_size 128 --freeze_fe --batch_size 5 --epochs 35
ECHO ------------------------------------------------------------------------------
python experiments.py --model_name VGG19 --input_size 128 --freeze_fe --batch_size 5 --epochs 35
ECHO ------------------------------------------------------------------------------
python experiments.py --model_name MobileNetV2 --input_size 128 --freeze_fe --batch_size 5 --epochs 35

pause
```

Figure 8. **The interface for experiment execution.** A batch file designed as a dashboard for running one code to execute all the scenarios.

The overall code structure for the implementation of pre-trained models is as follows:

a) Importing the libraries and downloading the desired models
b) Reading data
c) Feature extraction
d) Training the classifier layers
e) Evaluating models on the test data
f) Saving the experiment results

4.3.2 Training parameters and layer Adjustment

I. **Weights:** to import weights learned from the source task where pre-trained models were trained on, the base model has to convey the learned ImageNet weights.

II. **Activation function:** In the heart of CNNs are neurons. A neuron calculates a weighted sum of inputs from training and directs the results through an activation function, creating an output that either is the input for the subsequent layer or the probability distribution for each class as a final prediction.

Pre-trained models have been originally trained to classify natural images in the ImageNet competition. Consequently, the activation function to classify the final output of the whole CNN architecture was set to classify images into 1000 classes. Instead, for our task, a binary classification problem, it is critical to overwrite the most suitable activation function on the original setting. Among a variety of activation functions, including ReLU, Softmax, and Sigmoid. The standard activation function for binary classification to output the prediction from the last fully connected layer is the Sigmoid activation function, also called the logistic function. Sigmoid is a non-linear activation function that guarantees the output to be between 0 to 1. Also, a SoftMax activation function could be used by defining two classes. However, the results from the trial-and-error phase of the experiments showed higher classification accuracy while using the Sigmoid activation function.

III. **Fine-tuning:** In general, transfer learning approaches vary based on the number of CNN layers that are decided to be frozen in feature extraction and classification. In all strategies taken in transfer learning scenarios, the top layers acting as classifier layers are trained. Conversely, the extent to which the base model layers are trained, i.e., unfroze, defines the fine-tuning approach. Given that, there are three possible approaches to practice transfer learning:

a) Freezing all layers in the feature extractor convolutional base model.
b) Freezing a portion of the base model and training some layers of the convolutional base for feature extraction.
c) Unfreezing all layers of the base model to extract features through backpropagating down to the bottom layers of the CNN, relying solely on the model architecture and depth.

Picking the suitable strategy for the application of transfer learning depends on the similarity of the task dataset and source dataset as well as the size of the dataset. For a small dataset of 138 polar maps in this study, many trial-and-error experiments were done to achieve the best settings for fine-tuning of models.

During the trial-and-error phase, models were tested for performance by changing only the "trainable" attribute. The purpose was to determine if fine-tuning the base model could improve the performance of the test dataset. Based on the results of the trial-and-error phase, the best strategy was freezing all layers in the base model and only training the classifier.

IV. **Optimizer:** Pre-trained models are originally optimized for the source by the agency of various optimizers. For this thesis, though, one single optimizer is to be selected so the effect of different optimizers would not affect the evaluation of in question variables. To achieve a single best-performing optimizer, in a series of trial-and-error experiments, three optimizers suggested by recent publications were tested for performance. SGD, Adam, and RMSProb competed in our task, introducing Adam as the best generalizing and least overfitting optimizer. Therefore, for the rest of the study, Adam optimizer was employed.

V. **Batch size:** The number of training images fed to the model in every iteration (epoch). For DL tasks, defining the right batch size has remained to be challenging for researchers. Some studies in the literature have shown that bigger batch sizes help overcome overfitting and lead to higher generalization. Conversely, particular studies have proven the opposite for specific cases, suggesting that smaller batch sizes with corresponding learning rates for the training can improve generalization while avoiding overfitting.

In this thesis, different batch sizes of 30, 20, and 5 were tested in the trial-and-error phase. Bigger batch sizes were bound to misclassifying one class or compromising specific metrics such as F1-score while learning another class with high accuracy. On the contrary, the mini-batch size of 5, because of the increased chance of containing images from both classes during all iterations, therefore, increasing the chance of the model learning both classes, displayed more consistent results. As a result, the batch size for the rest of the work was set to 5.

VI. **Epochs and early stopping:** The number of epochs in ML is described by the number of times the entire training dataset is fed to the model. In a learning process of 30 epochs, each image sample has 30 chances to update the learned weights of the model during training. Excessive epochs can cause overfitting, meaning that the model memorizes the features of an image instead of learning them. This results in meaningful decreased accuracy in the test dataset. For this thesis, we set the epochs to 30, mirroring the settings of a custom CNN built for the same dataset in a study by Teuho et al. However, due to the fact that pre-trained models have different architectures. For instance, VGG16 is a significantly deeper model than other models. For a deep network with significantly richer parameters, learning can continue to advance with excessive epochs, unlike smaller models. In order to tackle the challenge of overfitting and underfitting while assigning one epoch number for all models, early stopping was used.

Having an early stopping technique in charge, the model does not iterate the training based on the pre-defined number of epochs but instead stops to run more epochs at a specific point. Early stopping benefits efficacy in training by

constantly monitoring accuracy on validation dataset and training dataset with reference to extra epochs. At a specific epoch, the performance on the validation set begins to degrade. This event is a signal for the model to stop training the remaining epochs to avoid overfitting.

VII. **Dropout:** Another technique to combat overfitting in a DL problem is to use dropout regularization, especially when training data is insufficient. In this approach, a number of neurons in the network are ignored and disconnected from other neurons, hence dropout. Using this method enables the network to improve its generalization capability by randomly adding noise to the layers. As a result, during each update of the layer on every training iteration, a different "view" of the layer is introduced to incoming connections. Srivastava et al. in their publication "Dropout: a simple way to prevent neural networks from overfitting" concluded that dropping out some neurons can break the "co-adaptations" of training data and weight because of the backpropagation process. Accordingly, in this study, given the small dataset for training and the arrangement of layers in the classifier head of the proposed architecture, the dropout approach was adopted. Dropout was applied to all fully connected layers before the last dense layer with the Sigmoid activation function. The dropout rate was set to 0.5. Figure 9 illustrates the arrangement of classifier layers, including the dropout layer.

VIII. **Shuffle:** To avoid overfitting due to small mini batch size used in the study, a shuffling regularization was also added to the previously discussed tactics. Through shuffling, data fed to the model in every epoch shuffle, so it would better represent the whole data population.

IX. **Classifier:** Having pre-trained models and the training parameters set up for the primary code execution, the final step is to configure the classification layers. The proposed arrangement of layers acting as classifiers on top of base models is as follows:

```
base_model,
tf.keras.layers.Flatten(),
tf.keras.layers.Dense(1024, activation='relu'),
tf.keras.layers.Dropout(0.5),
tf.keras.layers.Dense(512, activation='relu'),
tf.keras.layers.Dropout(0.5),
tf.keras.layers.Dense(256, activation='relu'),
tf.keras.layers.Dropout(0.5),
tf.keras.layers.Dense(1, activation='sigmoid')
```

Figure 9. The arrangement of classifier layers on top of base pre-trained models.

4.3.3 Evaluation metrics

To measure the performance of different pre-trained models, it is essential to define the evaluation methodology for a binary classification problem. Relying solely on the "accuracy" of a model in classifying input data does not yield a fair and thoroughly conclusive judgment of its performance. For example, with a data set containing 10% of training images labeled "*unhealthy*" (negative) and 90% labeled "*healthy*" (positive), a model predicting all unseen data as "*healthy*" is 90% accurate. Yet, this model totally misclassifies the "unhealthy" class, producing significant false-negative predictions. Considering a clinical interpretation being made based on this model, patients suffering from a disease remain undiagnosed and sent home without further examination and treatment.

In a predictive modeling problem, an important consideration is to calculate performance measurement parameters such as false negative (FN), true negative (TN), false positive (FP), and true positive (TP). Each parameter includes the terms “true” or “false” to represent the prediction result as well as “positive” or “negative”, indicating the predicted class. Furthermore, to demonstrate the prediction results from an ML model, a confusion matrix can be used. In a binary classification problem, a confusion matrix is a matrix of 2 × 2, including TP, FP, TN, and FN predictions made by a model. In a confusion matrix, rows are representative of classes, and columns display predictions. Figure 10 shows an example of a visualized confusion matrix.

To avoid misinterpretations caused by substandard evaluation of a model's performance, several standard evaluation metrics are introduced in the literature. In this thesis, evaluation metrics were selected based on similar investigations, as well as the task domain, being a binary classification of ischemic and non-ischemic images. In such problems, where a disease diagnosis is based on the predictions made by an ML model, avoiding false-negative cases is of utmost importance. In other words, sometimes, it is more effective to select a model with lower accuracy because of the predictive power for a specific problem. Accordingly, the metrics presented in table 3 were selected for the study to test the pre-trained models' performance.

Table 3. A description on evaluation metrics and their formula.

Metric	Question to be answered	Formula
Accuracy	How many total predictions were correct from all predictions?	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	How many predictions were correctly predicted as positive from all positive predictions? *(e.g., from all ischemic predictions, how many are truly ischemic?)*	$\frac{TP}{TP + FP}$
Recall	How many predictions were correctly predicted positive from all actual positive inputs? *(e.g., from all ischemic images in the test set, how many are correctly predicted as ischemic?)*	$\frac{TP}{TP + FN}$
F1-score	How good the quality of predictions are and how completely the model predicts the labels from inputs?	$2 \times \frac{Precision \times Recall}{Precision + Recall}$

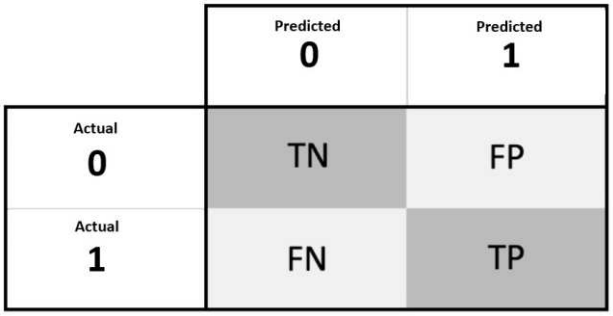

Figure 10. A schematic representation of a confusion matrix.

4.3.4 Ensemble learning

After obtaining the results from the experiment, based on the research objectives, a final step is to evaluate the effect of ensemble learning using the best pre-trained models.

Ensemble learning is an ML method that combines the predictive power of several models. Studies in the literature suggest that ensemble learning has proved to be beneficial in improving model performance. Based on ensemble learning, a combination of several models can perform better than any of the models used alone. Over the years, with the development of models and data pipelines for each ML technique, ensemble learning strategies have also become diverse. Stacking, boosting and meta ensemble learning are examples of common techniques in this domain.(81)

To assess the premise of “together” better than “one” in ensemble learning, this thesis aims to use 4 best pre-trained models to combine as one classifier and evaluate its performance.

Ensemble learning techniques used in this section include stacking techniques and meta learning techniques, such as logistic regression, decision tree, KNN and naïve Bayes.

5 Results and discussion

5.1 Results

5.1.1 Introduction

In this chapter, the results of all experiments (excluding the intensive trial-and-error code executions) are presented. To satisfy the research objectives mentioned in the first chapter of this thesis results from the experiments are reported in three sections. Firstly, performance results from all pre-trained models are presented and compared. Secondly, having introduced the best pre-trained models, the practicality of transfer learning is questioned by comparing pre-trained models to a custom CNN. Lastly, based on the results from the first section, the validity of ensemble learning using the 4 best pre-trained models is reviewed.

5.1.2 Pre-trained models

A total of 20 pre-trained models were selected for this study. Each of the models was trained on different input image sizes every time, from small sizes (32 × 32 and 64 × 64) to bigger images (128 × 128 and 256 × 256). However, in some cases, some image sizes would not fit the architecture of pre-trained base models due to pooling layers and the depth of the models. Also, since each model was originally trained on a specific image size, an additional input image size for each model was added to the experiments. Plus, From the entire EfficientNetB family, the EfficientNetB0 model was selected during the trial-and-error part of the experiment design. As a result, a total of 64 setups for 20 pre-trained models were tested for performance.

Visualizations of the results was done using Seaborn library which is a powerful tool among Python libraries for plotting graphs. In the following, the results of all pre-trained

models are presented and compared using scatter plot diagrams and tables. Additionally, Because of the large number of experiments (64 different model setups), benchmarking all models in one unambiguous illustration is challenging. Therefore, among all the different setups of every pre-trained model, the one with the best performing input image size is selected for visualization.

To compare the performance of models on test dataset and validation dataset (during training), the results from both training metrics and test metrics are presented.

5.1.2.1 Accuracy

As demonstrated in Figure 11, **VGG19** with input image size of 128 × 128 pixels achieved 85% accuracy on test dataset, followed by, **Xception** (input image size of 71 × 71), **VGG16** (input image size of 128 × 128), **DenseNet121** (input image size of 128 × 128), **DenseNet169** (input image size of 32 × 32) and **DenseNet201** (input image size of 224 × 224) with 83% accuracy on test dataset.

Interestingly, regarding the input image size variable, no meaningful trend in terms of accuracy was observed.

Except for ResNet50, ResNet101, and EfficientNetB0 models that performed poorly on the test dataset, achieving accuracies below 60%, other networks showed acceptable results of up to 80% accuracy.

Comparing the test and validation datasets, a slight difference can be observed in the accuracy achieved by several models. Overall, since most models were trained using regularization techniques such as dropout and early stopping, they are not prone to overfitting. The accuracy results of all models with all input image sizes involved, on both test dataset and validation dataset, are presented in the following graphs.

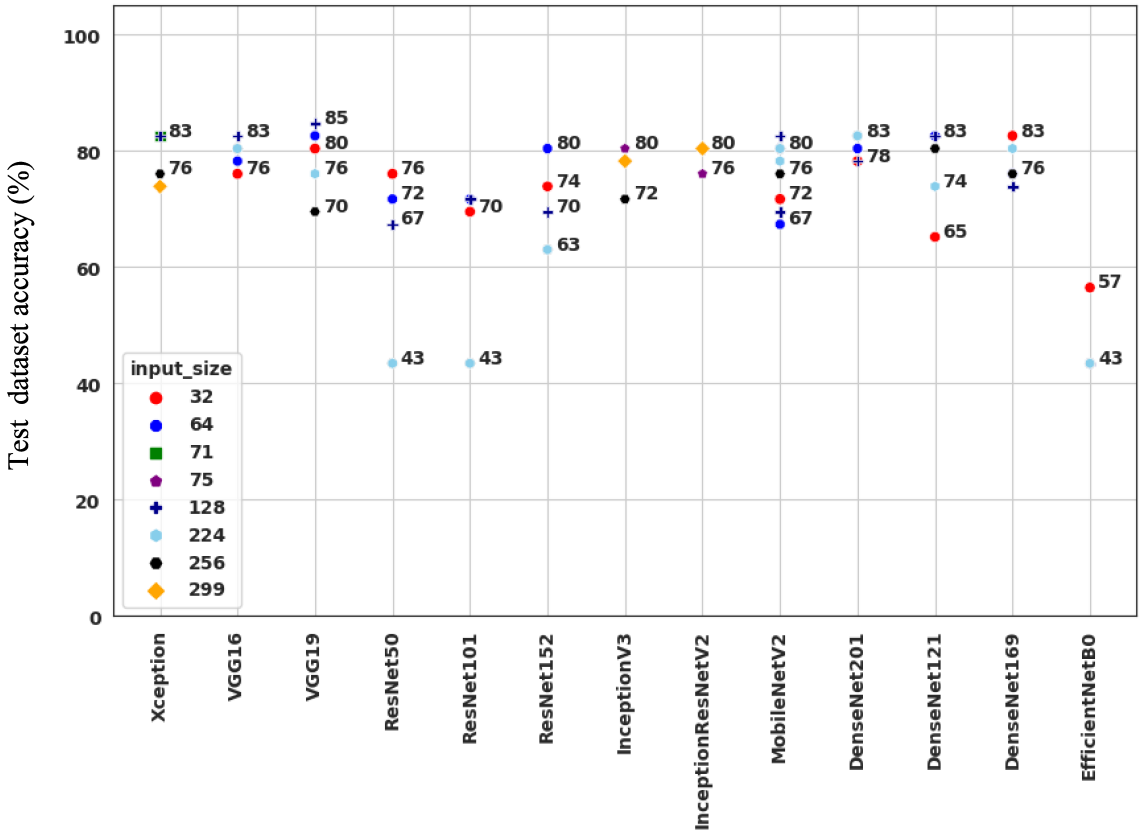

Figure 11. Accuracy of pre-trained models on test dataset.

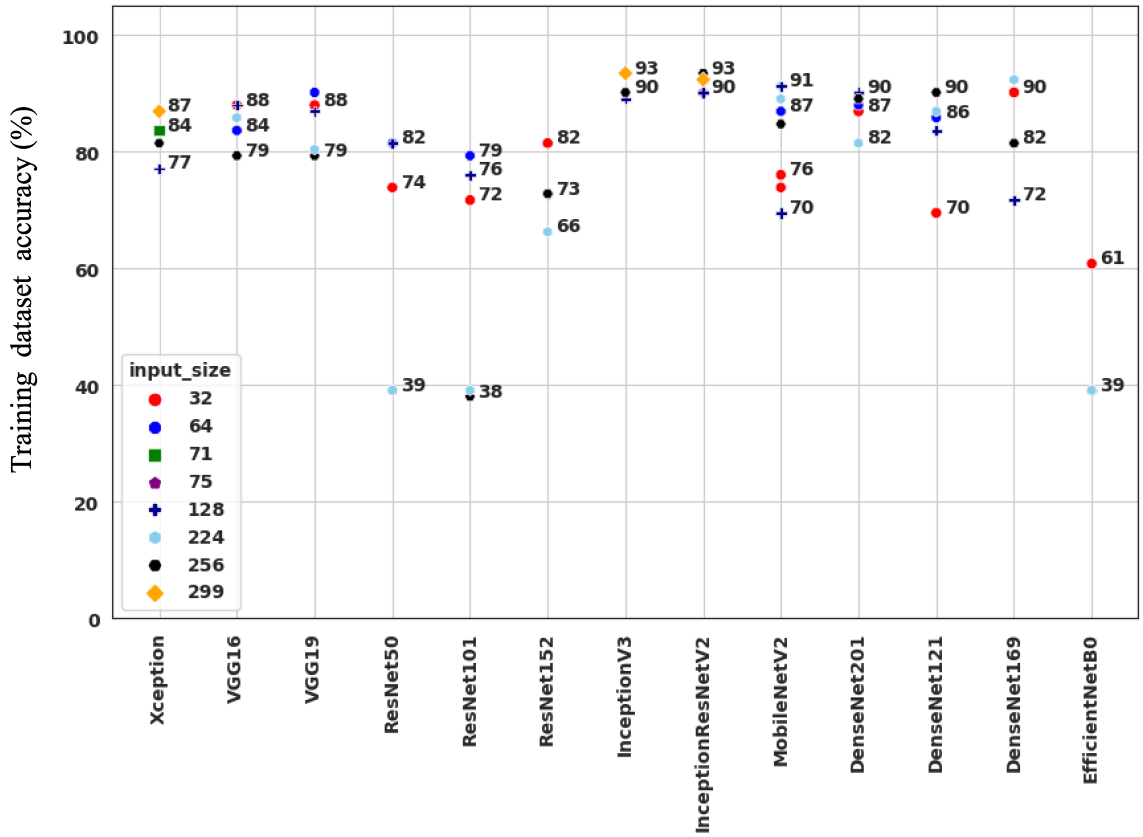

Figure 12. Accuracy of pre-trained models on training dataset.

5.1.2.2 Precision, Recall, F1-score

Pre-trained models achieved decent results on the test dataset with 100% precision and recall. However, the precision-recall tradeoff necessitates the use of the F1-score as well. Precise models tend to lose some information, producing more false negatives, while high-recall models introduce more false-positive predictions. F1-score balances the precision-recall tradeoff by calculating both metrics. In conclusion, a model with a high F1-score is more satisfying than one with excellent precision/recall results and a poor F1-score rating. Provided that, the best models in terms of F1-score are VGG19, DenseNet169, DenseNet201, and Xception, all with 80% F1-score. Illustrations below demonstrate how good pre-trained models performed in classifying ischemia from the test dataset in each metric.

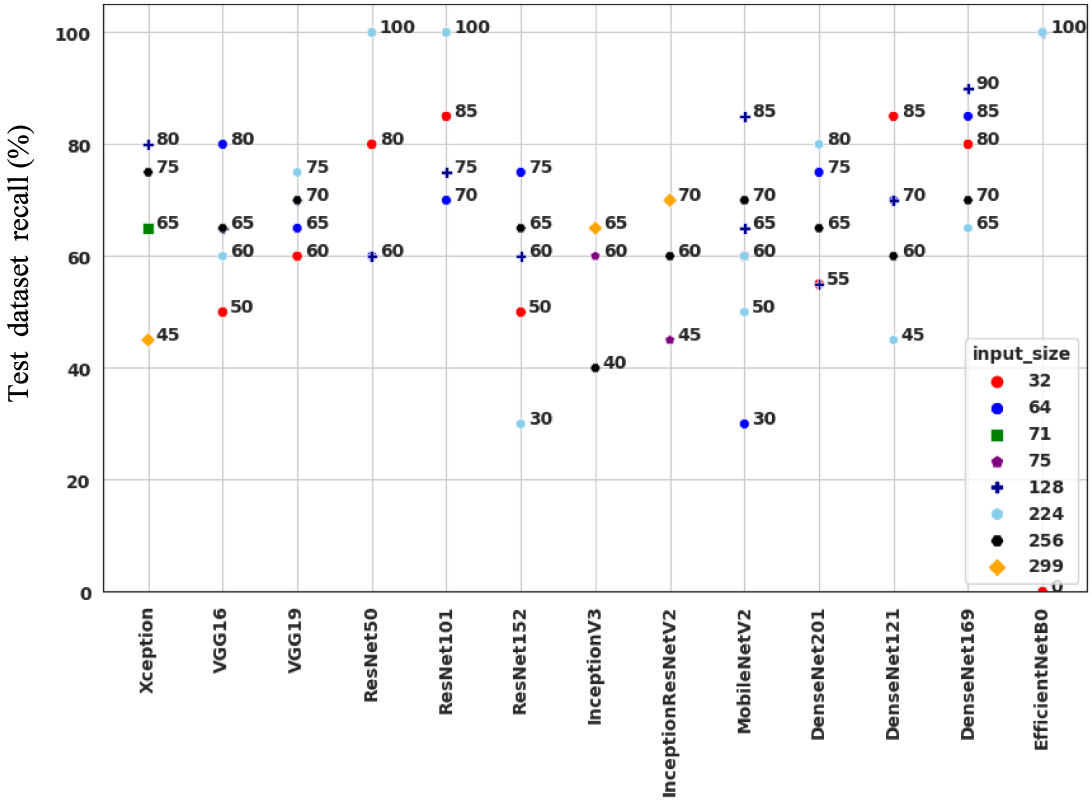

Figure 13. Recall of pre-trained models on test dataset.

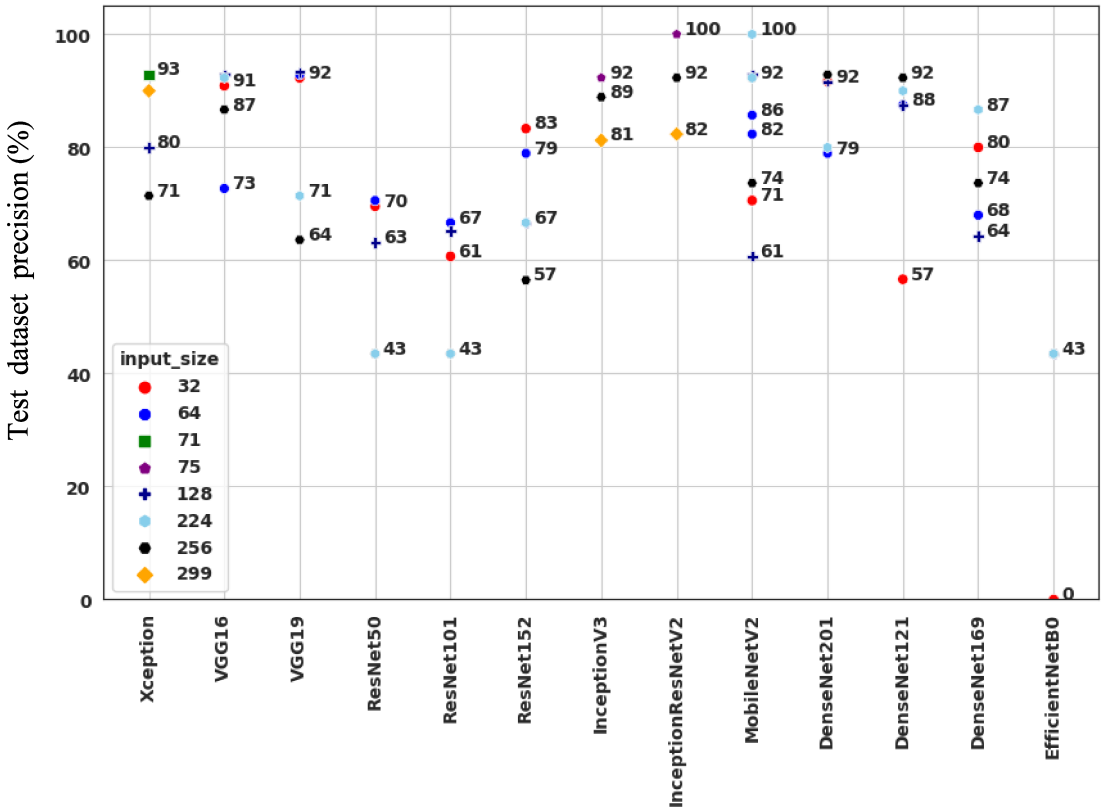

Figure 14. Precision of pre-trained models on test dataset.

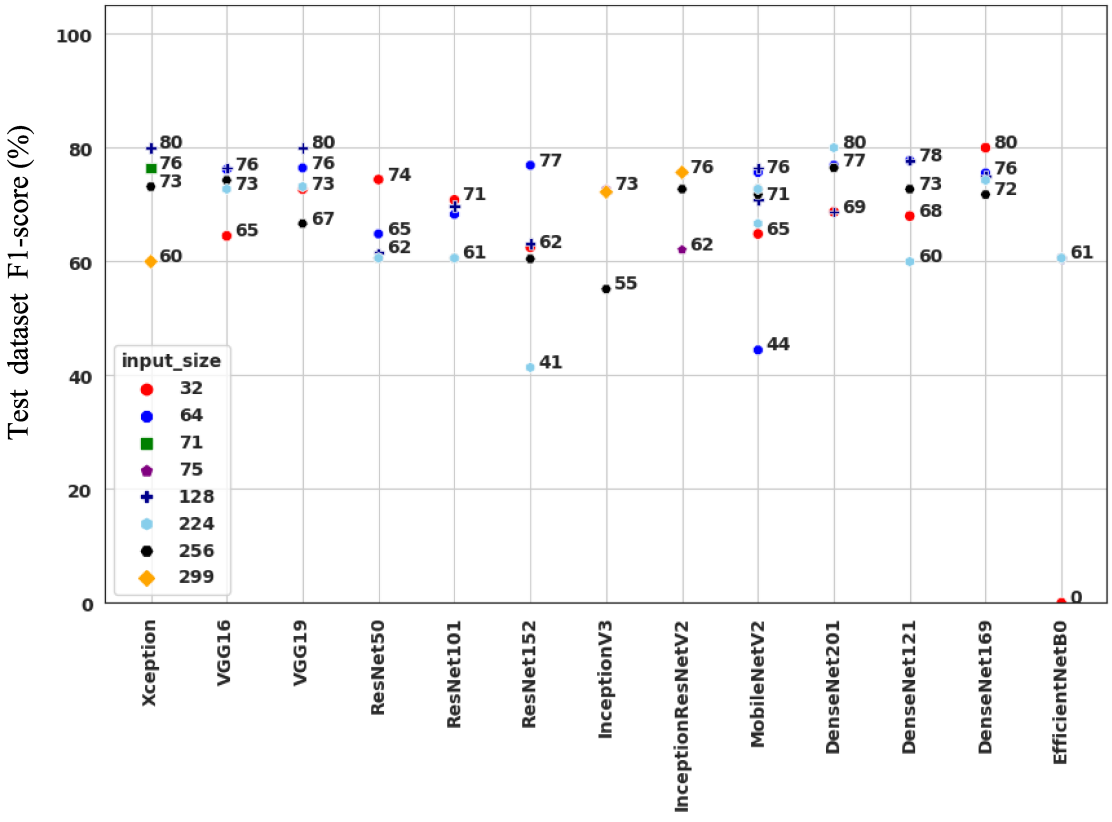

Figure 15. F1-score of pre-trained models on test dataset.

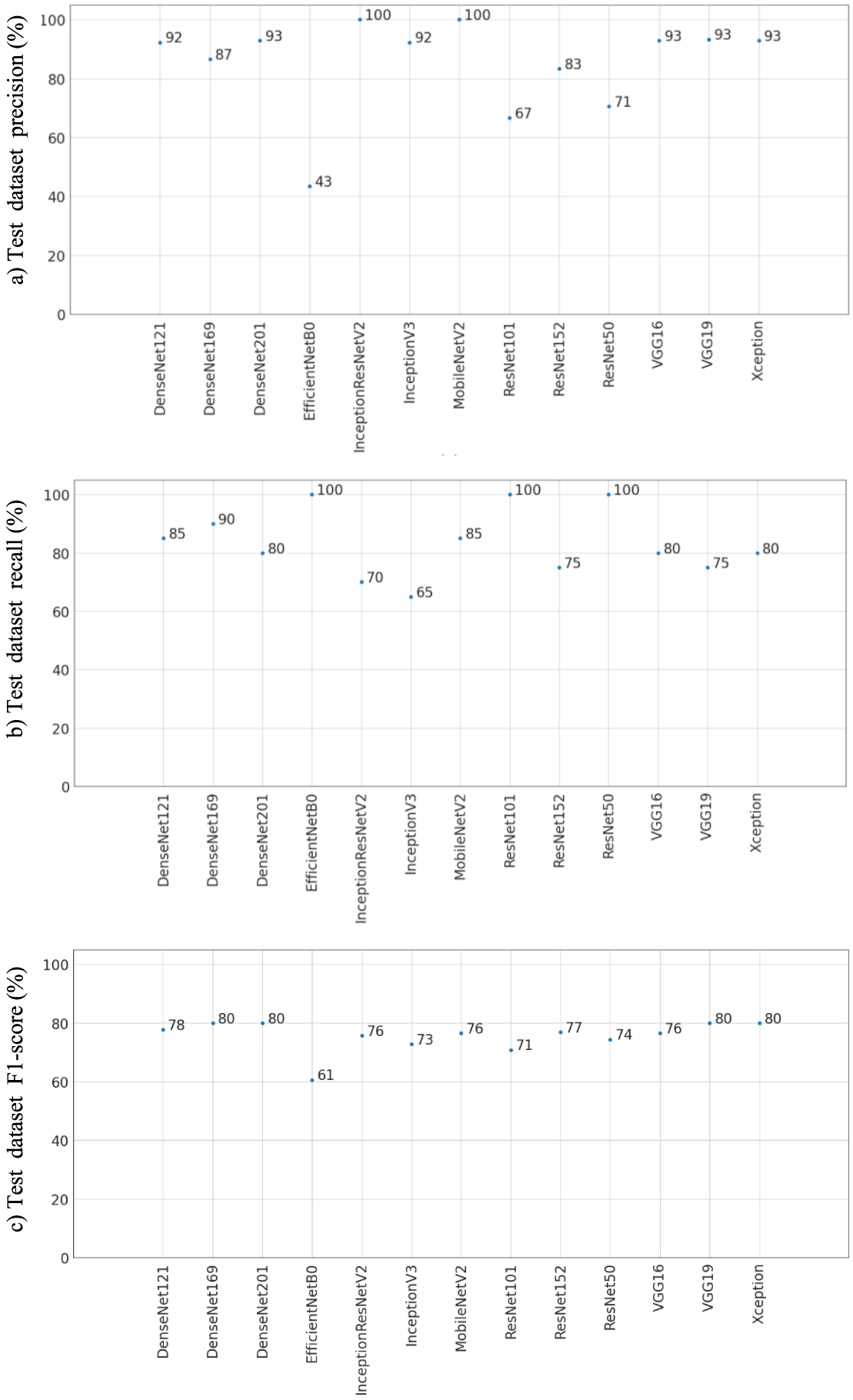

Figure 16. Precision, recall and F1-score of test dataset from highest value reported from any input image size. a) Precision of test dataset. b) recall of test dataset. c) F1-score of test dataset.

5.1.3 Comparison of transfer learning and a custom CNN

In this section, a selection of transfer learning models performing better than the rest are selected to be compared with a custom CNN developed on the same dataset and dataset configuration. The network architecture developed by Teuho et al. at Turku PET Centre, used a 256 × 256 input image size to train images in the batch sizes of 20 through 35 epochs.

Teuho et al. evaluated their proposed model based on accuracy, the area under curve (AUC), F1-score, sensitivity (recall), specificity, and precision. Having run their model over the data for 100 times, the reporting values for each metric are based on the calculated median. Table 4 briefly reports on the custom CNN performance.

Table 4. Evaluation metrics and CNN evaluation results from the study of Teuho et al.

Accuracy	83%
Precision	92%
Recall	65%
F1-score	76%
AUC	80%
Specificity	96%

To simplify the comparison between transfer learning models and the custom CNN model, similar metrics are compared. Additionally, to avoid complexity and repeating non-informative results, some evaluation metrics were not included in the transfer learning approach. In the following, an overall comparison between transfer learning models and the custom CNN is made in terms of accuracy, precision, sensitivity (recall), and F1-score.

The work of Teuho et al. included comparing the developed CNN model to the clinical diagnosis by doctors on the same test dataset, including 46 images. Based on this comparison, as well as detailed results achieved during all runs of test data, a CNN classifier proved to be in close agreement with clinical interpretation. However, in terms

of F1-score and recall, which is clinically valuable, the CNN model performs inferior to the clinical interpretation.

Table 5 reveals how transfer learning performs compared to the conventional custom CNN and clinical interpretation.

Table 5. Comparison between transfer learning, a custom CNN and clinical interpretation in myocardial ischemia classification results.

Approach		Accuracy	Precision	Recall	F1-score
Transfer learning	VGG19 (128 × 128)	85%	92%	75%	80%
	VGG16 (128 × 128)	83%	71%	75%	76%
	DenseNet201 (224 × 224)	83%	79%	80%	80%
	Xception (71 × 71)	83%	93%	65%	76%
Custom CNN		83%	93%	65%	76%
Clinical interpretation		87%	94%	75%	83%

As it can be understandable from the table, all approaches demonstrated a compactly similar result on the task. Xception model represented the exact same results from the custom CNN. The DenseNet201 and VGG16 also performed relatively close to the custom CNN. However, VGG19 network showed slightly improved performance compared to other models, including the custom CNN. Although each network has shown a higher peak in performance under specific settings, in a series of experiments, evaluation metrics are in favor of clinical interpretation.

5.1.4 Ensemble learning

Results from the comparison phase of the experiment revealed that the four models to be combined for ensemble learning are VGG19, VGG16, Xception, and DenseNet169. The input image size for all models was set as 128 × 128.

The results from most ensemble learning techniques were not informative due to similarity in performance. To report the results of this section of the experiment, from each category of ensemble learning techniques, one method was selected. From the stacking methods, max-voting showed excellent performance, and from meta-learning approaches, all methods yielded similar results. Tables 6 and 7 represent the best results of ensemble learning methods employed in this thesis.

Table 6. Performance results from ensemble learning on training dataset.

Method	Training accuracy	Training precision	Training recall	Training F1-score
Max-voting	89%	79%	97%	87%
Meta learning	94%	89%	97%	93%

Table 7. Performance results from ensemble learning on test dataset.

Method	Test accuracy	Test precision	Test recall	Test F1-score
Max-voting	**86%**	**93%**	**75%**	**83%**
Meta learning	84%	93%	70%	80%

Comparing the results from ensemble learning and clinical interpretation results from the study of Teuho et al., it can be concluded that ensemble learning can leverage the power of pre-trained models. Max-voting achieved the closest results to clinical interpretation in terms of accuracy and precision and achieved the highest F1-score of 83%, similar to clinical interpretation.

5.2 Discussion

One of the main objectives of this thesis was to compare pre-trained models for the classification of ischemia from polar maps. Having accomplished this objective, a number of superior pre-trained models were introduced. The best performing pre-trained model was the VGG16 model with input images of 128 × 128 pixels. Conversely, the EfficientNetB family exhibited unacceptable results for our task, despite their outstanding results on the source task.

To answer the second research question of this study, a selection of best pre-trained models was compared against a custom CNN developed on the same dataset. Interestingly, Xception network with a small input image size of 71 × 71, showed similar performance to the custom CNN. VGG19, marginally outperformed the custom CNN in most metrics. Yet, the majority of other pre-trained models performed inferior to the custom CNN model.

Changing input image size for the models did not meaningfully affect the overall performance of pre-trained models. According to several studies investigating the effect of image resolution on CNN performance, bigger image sizes were assumed to bring improved performance.(82–84) However, in our study, no meaningful trend was observed. In some networks (i.e., Xception), training on smaller images yielded better performance on most metrics. On the contrary, other networks such as InceptionV3 and InceptionResNetV2 performed more satisfactorily using larger input image sizes.

Ensemble modeling was used to aggregate the prediction power of each pre-trained model. The ensemble model outperformed all pre-trained models as well as the custom CNN, submitting the closest to human performance.

With reference to the comparison made between a custom CNN and pre-trained models, although some pre-trained models outperformed the custom CNN in most metrics, there are still some questions left to be addressed. Firstly, it is important to explain why pre-trained models with a multitude of layers and millions of parameters failed to perform close to a custom CNN. One explanation is that domain similarity between the source task to our task is not at an optimum level. However, this assumption can be challenged by VGG19's decent performance. Another assumption is that, due to the differences in the architecture of pre-trained models, some models performing better in extracting low-level features such as colors and edges were superior in our task. In polar map images,

low-level features are the dominant features of the images, therefore, models with more weights in the bottom layers of their architecture would perform better.

5.3 Conclusion

In this thesis, specific research questions were raised to fill the research gap in the domain of transfer learning in medical imaging. First, an introduction to the problem was provided, followed by an extensive literature review on the background of the study. Based on the previous investigations in applying transfer learning to medical image classification tasks, the viability of benchmarking a pre-trained model for the classification of myocardial ischemia was discussed. Later, the process of experiment design for the thesis, as well as the challenges of implementing a myriad of models into one unified code, was explained. Finally, In the last chapter, the results of the experiment were demonstrated.

The transfer learning approach proved to be feasible for the classification of ischemia. Some pre-trained models, such as VGG19 performed superior to other pre-trained networks as well as a custom CNN. Besides, with reference to the evaluation metrics, both training a CNN from scratch and transfer learning marginally performed at the human level performance. Yet, it failed to outperform clinical interpretations in specific metrics such as F1-score.

An advantage of employing transfer learning in image classification tasks is the effortlessness of its implementation. Developing CNN models from the ground up can benefit the performance due to the specific architecture design for the available dataset. On the flip side, the process of developing and training CNNs adds up to the time and resources. Alternatively, with transfer learning, a CNN model can be downloaded and tuned for a classification task in a matter of minutes. By using pre-trained models, the computation power for intensive base model training is not required, leaving more space and accessibility for researchers to implement DL even on low-end hardware. Therefore, transfer learning, not only because of the potential higher performance but also due to the ease of access, can be employed instead of developing a custom CNN.

According to the experiment results, in a medical image classification task, it is challenging to choose a pre-trained model as a predictive model to aid the end-point clinical decision. This is because the unfavorable domain similarity between the source

data in transfer learning models and medical image datasets compromises the performance of pre-trained models. Therefore, model selection is a hindrance in employing transfer learning for medical image classification tasks. We have shown that an ensemble of pre-trained models that performs at a clinical interpretation level is a good candidate to use instead of a single model. Accordingly, one way to tackle the challenge of selecting the appropriate architecture for a medical image classification task is using ensemble modeling.

Although it was previously emphasized that DL, in general, does not aim to replace doctors with regard to diagnosis purposes, once again, it was demonstrated that CNN models are not alternative systems to clinical interpretation. Rather, it can be concluded that doctors may possibly use DL methods for second confirmations or partially automate tasks in clinics. Especially with transfer learning's plug-and-play fashion, high-resolution images can be passed through automated classification models for predictive analysis in any stage of disease management.

In this thesis, all research questions were addressed, and the research objectives were accomplished. However, based on the results of this work, more questions are raised. One of the interesting topics to cover in the future is the effect of image augmentation techniques on transfer learning performance during the dataset preparation. Additionally, the effect of domain similarity in transfer learning for medical image classification is an interesting topic to be addressed.

6 Acknowledgments

The completion of this thesis would not be possible without the support and efforts of my supervisors, Dr. Jarmo Teuho and Professor Riku Klén, who provided me the opportunity of conducting my thesis in their research group. A debt of gratitude is also owed to Turku PET Centre and Turku University Hospital (TYKS) for funding this thesis.

Plus, I would like to thank the biomedical imaging master's degree program for helping me during the difficult time of the pandemic and for funding my studies throughout the whole master's degree program.

Finally, I would like to thank my family, who believed in me and supported me in every step of my journey to complete my studies in Finland.

7 References

1. Groot OQ, Bongers MER, Ogink PT, Senders JT, Karhade A v, Bramer JAM, et al. Systematic Review Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? A Systematic Review. Clinical orthopaedics and related research. 2020 Dec;478(12):2751.
2. Brzezicki MA, Bridger NE, Kobetić MD, Ostrowski M, Grabowski W, Gill SS, Neumann S. Artificial intelligence outperforms human students in conducting neurosurgical audits. Clinical Neurology and Neurosurgery. 2020 May 1;192:105732.
3. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ digital medicine. 2018 Mar 14;1(1):1-4.
4. Hahn HK. *Morphological Volumetry: Theory, Concepts, and Application to Quantitative Medical Imaging* (Doctoral dissertation, Universität Bremen).2005.
5. Seifert R, Weber M, Kocakavuk E, Rischpler C, Kersting D. Artificial intelligence and machine learning in nuclear medicine: future perspectives. InSeminars in nuclear medicine 2021 Mar 1 (Vol. 51, No. 2, pp. 170-177). WB Saunders.
6. Okrainec K, Banerjee DK, Eisenberg MJ. Coronary artery disease in the developing world. American heart journal. 2004 Jul 1;148(1):7-15.
7. Shimokawa H, Yasuda S. Myocardial ischemia: current concepts and future perspectives. Journal of cardiology. 2008 Oct 1;52(2):67-78.
8. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low-and middle-income countries. Current problems in cardiology. 2010 Feb 1;35(2):72-115.
9. Packard RR, Huang SC, Dahlbom M, Czernin J, Maddahi J. Absolute quantitation of myocardial blood flow in human subjects with or without myocardial ischemia using dynamic flurpiridaz F 18 PET. Journal of Nuclear Medicine. 2014 Sep 1;55(9):1438-44.
10. Juarez-Orozco LE, Knol RJ, Sanchez-Catasus CA, Martinez-Manzanera O, Van der Zant FM, Knuuti J. Machine learning in the integration of simple variables for identifying patients with myocardial ischemia. Journal of Nuclear Cardiology. 2020 Feb;27(1):147-55.
11. Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data. 2019 Dec;6(1):1-8.
12. Juarez-Orozco LE, Martinez-Manzanera O, Storti AE, Knuuti J. Machine learning in the evaluation of myocardial ischemia through nuclear cardiology. Current Cardiovascular Imaging Reports. 2019 Feb;12(2):1-8.
13. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine. 2001 Aug 1;23(1):89-109.
14. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.
15. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging. 2016 Feb 11;35(5):1285-98.
16. Yamashita R, Nishio M, Do RK, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights into imaging. 2018 Aug;9(4):611-29.

17. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In2017 international conference on engineering and technology (ICET) 2017 Aug 21 (pp. 1-6). Ieee.

18. France FH. Ethics and biomedical information. International journal of medical informatics. 1998 Mar 1;49(1):111-5.

19. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: International conference on artificial neural networks. 2018. p. 270–9.

20. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big data. 2016 Dec;3(1):1-40.

21. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q. A comprehensive survey on transfer learning. Proceedings of the IEEE. 2020 Jul 7;109(1):43-76.

22. Shaha M, Pawar M. Transfer learning for image classification. In2018 second international conference on electronics, communication and aerospace technology (ICECA) 2018 Mar 29 (pp. 656-660). IEEE.

23. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems. 2019;32.

24. Kaur T, Gandhi TK. Deep convolutional neural networks with transfer learning for automated brain image classification. Machine Vision and Applications. 2020 Mar;31(3):1-6.

25. Ardalan Z, Subbian V. Transfer Learning Approaches for Neuroimaging Analysis: A Scoping Review. Frontiers in Artificial Intelligence. 2022;5.

26. Berkaya SK, Sivrikoz IA, Gunal S. Classification models for SPECT myocardial perfusion imaging. Computers in Biology and Medicine. 2020 Aug 1;123:103893.

27. Wang J, Qiao L, Lv H, Lv Z. Deep Transfer Learning-based Multi-modal Digital Twins for Enhancement and Diagnostic Analysis of Brain MRI Image. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2022 Apr 19.

28. Ni YC, Tseng FP, Pai MC, Hsiao IT, Lin KJ, Lin ZK, Lin WB, Chiu PY, Hung GU, Chang CC, Chang YT. Detection of Alzheimer's disease using ECD SPECT images by transfer learning from FDG PET. Annals of Nuclear Medicine. 2021 Aug;35(8):889-99.

29. Kim DH, Wit H, Thurston M. Artificial intelligence in the diagnosis of Parkinson's disease from ioflupane-123 single-photon emission computed tomography dopamine transporter scans using transfer learning. Nuclear medicine communications. 2018 Oct 1;39(10):887-93.

30. Ghosh N, Rimoldi OE, Beanlands RS, Camici PG. Assessment of myocardial ischaemia and viability: role of positron emission tomography. European heart journal. 2010 Dec 1;31(24):2984-95.

31. Juarez-Orozco LE, Martinez-Manzanera O, Storti AE, Knuuti J. Machine learning in the evaluation of myocardial ischemia through nuclear cardiology. Current Cardiovascular Imaging Reports. 2019 Feb;12(2):1-8.

32. Gomez J, Doukky R. Artificial intelligence in nuclear cardiology. Journal of Nuclear Medicine. 2019 Aug 1;60(8):1042-3.

33. Dobrucki LW, Sinusas AJ. PET and SPECT in cardiovascular molecular imaging. Nature Reviews Cardiology. 2010 Jan;7(1):38-47.

34. Institute of Medicine (U.S.). Committee on Social Security Cardiovascular Disability Criteria., Institute of Medicine (U.S.). Board on the Health of Select Populations. Cardiovascular disability : updating the Social Security listings. National Academies Press; 2010. 328 p.

35. Arsanjani R, Xu Y, Dey D, Fish M, Dorbala S, Hayes S, Berman D, Germano G, Slomka P. Improved accuracy of myocardial perfusion SPECT for the detection of coronary artery disease using a support vector machine algorithm. Journal of Nuclear Medicine. 2013 Apr 1;54(4):549-55.

36. Juarez-Orozco LE, Martinez-Manzanera O, van der Zant FM, Knol RJ, Knuuti J. Deep learning in quantitative PET myocardial perfusion imaging: a study on cardiovascular event prediction. Cardiovascular Imaging. 2020 Jan 1;13(1_Part_1):180-2.

37. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics. 2007 Jun 1;31(4-5):198-211.

38. Shiraishi J, Li Q, Appelbaum D, Doi K. Computer-aided diagnosis and artificial intelligence in clinical imaging. InSeminars in nuclear medicine 2011 Nov 1 (Vol. 41, No. 6, pp. 449-462). WB Saunders.

39. Talo M, Baloglu UB, Yıldırım Ö, Acharya UR. Application of deep transfer learning for automated brain abnormality classification using MR images. Cognitive Systems Research. 2019 May 1;54:176-88.

40. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015 May;521(7553):436-44.

41. Shimokawa H, Yasuda S. Myocardial ischemia: current concepts and future perspectives. Journal of cardiology. 2008 Oct 1;52(2):67-78.

42. Stenström I, Maaniitty T, Uusitalo V, Pietilä M, Ukkonen H, Kajander S, Mäki M, Bax JJ, Knuuti J, Saraste A. Frequency and angiographic characteristics of coronary microvascular dysfunction in stable angina: a hybrid imaging study. European Heart Journal-Cardiovascular Imaging. 2017 Nov 1;18(11):1206-13.

43. Tarkin JM, Ćorović A, Wall C, Gopalan D, Rudd JH. Positron emission tomography imaging in cardiovascular disease. Heart. 2020 Nov 1;106(22):1712-8.

44. Slart RH, Williams MC, Juarez-Orozco LE, Rischpler C, Dweck MR, Glaudemans AW, Gimelli A, Georgoulias P, Gheysens O, Gaemperli O, Habib G. Position paper of the EACVI and EANM on artificial intelligence applications in multimodality cardiovascular imaging using SPECT/CT, PET/CT, and cardiac CT. European journal of nuclear medicine and molecular imaging. 2021 May;48(5):1399-413.

45. Khalil MM, editor. Basic science of PET imaging. Cham: Springer International Publishing; 2017.

46. Chen K, Miller EJ, Sadeghi MM. PET-based imaging of ischemic heart disease. PET clinics. 2019 Apr 1;14(2):211-21.

47. Berger aA. How does it work?: Positron emission tomography. BMJ: British Medical Journal. 2003 Jun 28;326(7404):1449.

48. Maddahi J, Packard mRR. Cardiac PET perfusion tracers: current status and future directions. InSeminars in nuclear medicine 2014 Sep 1 (Vol. 44, No. 5, pp. 333-343). WB Saunders.

49. Nakazato R, Berman DS, Alexanderson E, Slomka P. Myocardial perfusion imaging with PET. Imaging in medicine. 2013 Feb 1;5(1):35.

50. Choi H. Deep learning in nuclear medicine and molecular imaging: current perspectives and future directions. Nuclear medicine and molecular imaging. 2018 Apr;52(2):109-18.

51. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. Annals of translational medicine. 2020 Jun;8(11).

52. Nwadiugwu MC. Neural Networks, Artificial Intelligence and the Computational Brain. arXiv preprint arXiv:2101.08635. 2020 Dec 25.

53. Lorente Ò, Riera I, Rana A. Image classification with classic and deep learning techniques. arXiv preprint arXiv:2105.04895. 2021 May 11.

54. Romero M, Interian Y, Solberg T, Valdes G. Targeted transfer learning to improve performance in small medical physics datasets. Medical Physics. 2020 Dec;47(12):6246-56.

55. Turing AM. Computing machinery and intelligence. InParsing the turing test 2009 (pp. 23-65). Springer, Dordrecht.

56. Andresen SL. John McCarthy: father of AI. IEEE Intelligent Systems. 2002 Sep;17(5):84-5.

57. Ofir N, Nebel JC. Classic versus deep learning approaches to address computer vision challenges. arXiv preprint arXiv:2101.09744. 2021 Jan 24.

58. Ofir N, Nebel JC. Classic versus deep learning approaches to address computer vision challenges. arXiv preprint arXiv:2101.09744. 2021 Jan 24.

59. Carbonell JG, Michalski RS, Mitchell TM. An overview of machine learning. Machine learning. 1983 Jan 1:3-23.

60. Coiera EW. AI in medicine: Overview and challenges. InIEE Colloquium on Artificial Intelligence in Medicine (Digest No: 1996-031) 1996 Feb 19 (pp. 1-1). IET.

61. Simeone O. A very brief introduction to machine learning with applications to communication systems. IEEE Transactions on Cognitive Communications and Networking. 2018 Nov 21;4(4):648-64.

62. Giger ML. Machine learning in medical imaging. Journal of the American College of Radiology. 2018 Mar 1;15(3):512-20.

63. Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D, Chen CM. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Scientific reports. 2016 Apr 15;6(1):1-3.

64. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics. 1943 Dec;5(4):115-33.

65. Dechter R. Learning While Searching in Constraint-Satisfaction-Problems. In: Proceedings of the Fifth AAAI National Conference on Artificial Intelligence. AAAI Press; 1986. p. 178–83. (AAAI'86)

66. Aizenberg I, Moraga C. Multilayer feedforward neural network based on multi-valued neurons (MLMVN) and a backpropagation learning algorithm. Soft Computing. 2007 Jan;11(2):169-83.

67. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science. 2021 Nov;2(6):1-20.

68. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. Annals of translational medicine. 2020 Jun;8(11).

69. Siddique F, Sakib S, Siddique MA. Handwritten digit recognition using convolutional neural network in Python with tensorflow and observe the variation of accuracies for various hidden layers.

70. Gavali P, Banu JS. Deep convolutional neural network for image classification on CUDA platform. InDeep learning and parallel computing environment for bioengineering systems 2019 Jan 1 (pp. 99-122). Academic Press.

71. Sharma N, Jain V, Mishra A. An analysis of convolutional neural networks for image classification. Procedia computer science. 2018 Jan 1;132:377-84.

72. Alzubaidi L, Al-Amidie M, Al-Asadi A, Humaidi AJ, Al-Shamma O, Fadhel MA, Zhang J, Santamaría J, Duan Y. Novel transfer learning approach for medical imaging with limited labeled data. Cancers. 2021 Jan;13(7):1590.

73. Krishna ST, Kalluri HK. Deep learning and transfer learning approaches for image classification. International Journal of Recent Technology and Engineering (IJRTE). 2019 Feb;7(5S4):427-32.

74. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In2009 IEEE conference on computer vision and pattern recognition 2009 Jun 20 (pp. 248-255). Ieee.

75. Tan M, Le Q. Efficientnetv2: Smaller models and faster training. InInternational Conference on Machine Learning 2021 Jul 1 (pp. 10096-10106). PMLR.

76. Wang G, Kikuchi Y, Yi J, Zou Q, Zhou R, Guo X. Transfer Learning for Retinal Vascular Disease Detection: A Pilot Study with Diabetic Retinopathy and Retinopathy of Prematurity. arXiv preprint arXiv:2201.01250. 2022 Jan 4.

77. Papandrianos N, Papageorgiou E. Automatic Diagnosis of Coronary Artery Disease in SPECT Myocardial Perfusion Imaging Employing Deep Learning. Applied Sciences. 2021 Jan;11(14):6362.

78. Rahaman MM, Li C, Yao Y, Kulwa F, Rahman MA, Wang Q, Qi S, Kong F, Zhu X, Zhao X. Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. Journal of X-ray Science and Technology. 2020 Jan 1;28(5):821-39.

79. Stenström I, Maaniitty T, Uusitalo V, Pietilä M, Ukkonen H, Kajander S, Mäki M, Bax JJ, Knuuti J, Saraste A. Frequency and angiographic characteristics of coronary microvascular dysfunction in stable angina: a hybrid imaging study. European Heart Journal-Cardiovascular Imaging. 2017 Nov 1;18(11):1206-13.

80. Dietterich T. Overfitting and undercomputing in machine learning. ACM computing surveys (CSUR). 1995 Sep 1;27(3):326-7.

81. Sagi O, Rokach L. Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery 8, 4 (2018), e1249.

82. Thambawita V, Strümke I, Hicks SA, Halvorsen P, Parasa S, Riegler MA. Impact of Image Resolution on Deep Learning Performance in Endoscopy Image Classification: An Experimental Study Using a Large Dataset of Endoscopic Images. Diagnostics. 2021 Dec;11(12):2183.

83. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. Radiology: Artificial Intelligence. 2020 Jan 22;2(1):e190015.

84. Kannojia SP, Jaiswal G. Effects of varying resolution on performance of CNN based image classification: An experimental study. Int. J. Comput. Sci. Eng. 2018 Sep;6(9):451-6.