TRANSFER LEARNING AND KNOWLEDGE DISTILLATION for CHEST X-RAY CLASSIFICATION

Master Degree in Artificial Intelligence

Course: Computer Vision and Deep Learning

Academic Year: 2024–2025

Author Name: Gebre Haftom Desbele
Matriculation number:VR526525

Abstract

This project uses two powerful and efficient deep learning approaches for chest X-ray classification. A combination of Transfer Learning and Knowledge Distillation to create a high-performing and lightweight model. The approach for the classification of chest X-ray images, specifically for the diagnosis of pneumonia. A baseline Convolutional Neural Network (CNN) is first established to benchmark performance. Subsequently, transfer learning is applied using a pretrained ResNet-18 model, which is fine-tuned and also used as a fixed feature extractor. The project then investigates knowledge distillation by training a smaller student model (ResNet-18) under the guidance of a more powerful teacher model (ResNet-50). The goal is to transfer the knowledge from the larger network to the smaller one, thereby enabling the student to achieve comparable performance with fewer parameters and faster inference times. The **Final Student Model** achieved an accuracy of 0.9806, a precision of 0.9851, a recall of 0.9882, and an F1-score of 0.9866. The results demonstrate that the proposed knowledge distillation approach with transfer learning provides an effective solution for medical image classification, balancing high accuracy with computational efficiency, which is crucial for real-world clinical applications with limited resources.

Contents

List of Tables

List of Figures

Chapter 1

Motivation and Rationale

Introduction

Deep learning for computer vision has enabled significant advancements in image processing and machine learning applications. The data-intensive nature of deep learning makes neural networks trained with large datasets capable of learning rich representations that often generalize well. However, smaller networks trained on smaller datasets are considered to learn fewer representations, which implies that they do not generalize well. [1], [2]. Similarly, Another major challenge for deep learning solutions is the availability of sufficiently labeled data, which is very expensive, time-consuming, and difficult in the case of medical diagnosis [3], [4].

Chest X-Rays (CXRs) are a frequent diagnostic technique in medicine. They provide information on many lung and cardiovascular disorders, such as pneumonia, lung cancer, and tuberculosis. A radiologist interprets the X-ray to look for these abnormal findings, along with other signs like air bronchograms, to help guide treatment. Since deep learning has been developed, there have been more attempts to automate the interpretation of CXR images in an effort to assist radiologists with diagnosis and therapy [5]. However, there are several obstacles to using deep learning models for CXR diagnosis. These include the requirement for sizable annotated datasets, computing power, and model interpretability.

Since advanced computing and massive data are not accessible to everyone, we can pre-train a network and reuse it to fine-tune smaller data multiple times when needed with **Transfer Learning**. It could be a data-efficient way to achieve faster training and inference on problems with limited available resources effectively. Similarly, a large previously trained network, called the **teacher network**, can also be used to teach another network, called the **student network**, by emulating the teacher through **Knowledge Distillation** [6]. This helps a smaller student network achieve the higher performance of a larger teacher network. This process works across many domains, modalities, or problems already like image classification [6], audio recognition, image captioning etc[7]. This addresses the problem of requiring large computation by providing a way to combat the requirement to train large networks every time we need them. This project uses knowledge distillation with transfer learning to obtain higher performance and small model over transfer learning for Chest X-ray classification.

1.1 Motivation

Convolutional Neural Networks (CNNs) were utilized as the primary method for classifying chest X-rays using Deep Learning (DL). CNNs had the capability to extract important features from medical images, which made them highly effective for classification tasks [8], [9]. Furthermore, this approach demonstrated potential efficacy in the detection and diagnosis of other critical medical conditions, such as lung cancer. However, this approach required a substantial amount of data, resulted in large models, and demanded high computational power [6]. Within the domain of DL, producing smaller models with good performance remained a challenge. **Knowledge Distillation (KD)** addressed this by transferring insights from a more complex model (teacher) to a more compact model (student), thereby enabling the student to achieve enhanced performance [6]. With **transfer learning**, we can reuse the representation of a model trained on a related dataset and improve performance while training on a limited dataset. This means we are getting the best of both worlds, that is big model-like performance (near to big model) with a limited in-domain dataset.

1.2 Problem Statement

The problem is developing an effective framework that combines transfer learning and knowledge distillation to improve the accuracy, efficiency, and scalability of chest X-ray classification, ensuring they are both clinically useful and computationally feasible.

- **Type of Problem:** Binary Image Classification.
- **Target Variable:** `Pneumonia_Label` (`0 = Normal`, `1 = Pneumonia`).

1.3 Objectives

The main objective of this study is to develop and evaluate deep learning approaches for chest X-ray image classification using convolutional neural networks (CNNs), transfer learning, and knowledge distillation. The goal is to improve classification accuracy, efficiency, and deployability of models in real-world medical applications.

1.3.1 Specific Objectives

1. To implement a baseline CNN model with two convolutional layers and two fully connected layers in order to establish a performance benchmark for chest X-ray classification.
2. To apply transfer learning using ResNet-18 by:
   (a) Fine-tuning the convolutional network to adapt pretrained features to chest X-ray images.
   (b) Using the convolutional network as a fixed feature extractor for efficient feature representation.

3. To investigate knowledge distillation by training a student model (ResNet-18) under the supervision of a stronger teacher model (ResNet-50), with the aim of achieving comparable accuracy while reducing computational complexity.

4. To compare performance metrics (accuracy, precision, recall, F1-score, AUC, number of parameters, and inference time) across the three experimental setups.

5. To identify the most effective approach that balances high classification accuracy with computational efficiency, making the model more suitable for deployment in real-world medical environments with limited resources.

1.4 Scope

The focus of the project is on the use of transfer learning and knowledge distillation techniques for the classification of pneumonia in chest X-ray images. The project utilizes the dataset by Paul Mooney, *Chest X-Ray Images (Pneumonia)*, available on Kaggle [10], which contains 5,863 frontal-view chest X-ray images from pediatric patients. The images are divided into two classes: **Normal** and **Pneumonia**.

The scope of the study includes:

- Using convolutional neural networks (CNNs) for feature extraction and classification of chest X-ray images.
- Applying transfer learning from pretrained networks (e.g., ResNet-18 and ResNet-50) to improve model performance on this dataset.
- Implementing knowledge distillation to compress a high-capacity teacher model into a smaller, more efficient student model while maintaining high classification accuracy.
- Evaluating AI models based on standard metrics such as accuracy, precision, recall and F1-score,

The scope of this Project is limited to the features and images provided in the *Chest X-Ray Images (Pneumonia)* dataset by Paul Mooney, available on Kaggle [10]. This dataset includes frontal-view chest X-ray images categorized as either normal or pneumonia-affected. Despite its focus on a specific set of images, the dataset provides a solid foundation for experimenting with AI techniques such as convolutional neural networks, transfer learning, and knowledge distillation. The insights and models developed in this study can be generalized to other chest X-ray datasets and similar medical imaging tasks, contributing to more accurate and efficient diagnostic solutions in clinical settings.

1.5 Significance of the Project

This project aims to enhance the decision-making process for healthcare professionals by providing an accurate and efficient diagnostic tool for chest X-ray analysis. By combining transfer learning and knowledge distillation, the model enables:

- Accurate and faster diagnosis of pneumonia.

- Development of lightweight models suitable for deployment on low-resource devices in clinical settings.
- Reduction of computational costs and energy consumption associated with large-scale deep learning models.

The integration of deep learning into medical diagnostics supports more informed and strategic clinical decisions, contributing to better patient outcomes and more efficient use of medical resources.

1.6 Paper Organization

The rest of the paper is organized as follows. Section 2 presents the state of the art of the existing technologies, Section 3 introduces the Methodology, including the proposed methodology, dataset, and models. Section 4 introduces metrics, experimental setup, and analyzes the results. Eventually, Section 5, conclusion, and final recommendation.

Chapter 2

State of the Art

Image classification in the medical field has become increasingly important, enabling more accurate and faster diagnosis. The state of the art reflects a shift from traditional image processing techniques to a more advanced application of deep learning (DL), particularly Convolutional Neural Networks (CNNs).

2.1 Transfer Learning

Early approaches to medical image classification relied on handcrafted features and classical machine learning models. These methods, although interpretable, were often limited in capturing complex, high-level features from medical images. The advent of deep learning, especially CNNs, revolutionized this field. However, training deep networks from scratch requires massive, well-annotated datasets, which are scarce and expensive in medical domains [8]

Transfer learning has emerged as a powerful solution to this problem. By using models pretrained on large-scale datasets like ImageNet, the learned features can be transferred and fine-tuned for a specific medical task, such as chest X-ray classification. This approach has shown remarkable success, achieving high accuracy with significantly less data and computational resources [9].

2.2 Convolutional Neural Networks (CNNs)

a. ResNet Architectures

Deep CNN architectures, such as ResNet (Residual Network), are widely used for image classification due to their ability to mitigate the vanishing gradient problem in deep networks. The ResNet family, including ResNet-18, ResNet-34, and ResNet-50, has become a standard for computer vision tasks **he2016deep**. ResNet-18, a more compact version, offers a good balance between performance and computational cost, making it an excellent candidate for a student model in knowledge distillation. ResNet-50, with its greater depth and capacity, serves as an ideal teacher model due to its ability to learn more complex feature representations.

b. Knowledge Distillation (KD)

Knowledge Distillation is a model compression technique where a smaller, more efficient student model is trained to mimic the output of a larger, more powerful teacher model*. This is particularly useful when the goal is to deploy a compact model without sacrificing performance. The student model learns not only from the hard labels (e.g., 0 or 1) but also from the soft labels (probability distributions) generated by the teacher, which contain richer information about the data [6].

In medical imaging, where deploying models on embedded devices or at the edge is often necessary, KD can significantly reduce the model size and inference time, making it a viable solution for real-time diagnostic support [5].

Chapter 3

Methodology

3.1 Workflow

The Workflow of the system is summarized as follows:

1. **Chest X-ray Dataset:** The process begins with the collection of raw chest X-ray images from the Kaggle dataset.
2. **Data Preprocessing:** This step handles image resizing, normalization, and makes it ready for the model.
3. **Split Train/Test:** The dataset is divided into training, validation, and testing sets to evaluate the generalization of the models.
4. **Build Models:** Three different deep learning approaches are implemented and trained:
   (a) Baseline CNN from scratch.
   (b) Transfer learning with ResNet-18 (Fine-tuning and fixed feature extractor).
   (c) Knowledge distillation with ResNet-50 (Teacher) and ResNet-18 (Student).
5. **Evaluation:** The trained models are evaluated using performance metrics to determine their effectiveness.
6. **Final Model Selection:** The best-performing and most efficient model is selected for potential deployment.

This workflow ensures a systematic approach to solving the classification task with optimal model selection and reliable results.

3.2 Dataset

The dataset used in this project was obtained from Kaggle, titled *Chest X-Ray Images (Pneumonia)* by Paul Mooney [10]. It contains 5,863 frontal-view chest X-ray images from pediatric patients, categorized into two classes: **Normal** and **Pneumonia**. The dataset is pre-divided into training, validation, and test sets.

- **Training Set:** 5,216 images.

- **Test Set:** 624 images.

The images are in JPEG format and have varying resolutions. All images are resized to a uniform dimension 224x224 pixels to be used as input for the models.

Figure 3.1: Sample data visualization from the Chest X-Ray Images (Pneumonia) dataset.

3.3 Models

3.3.1 Baseline CNN

A custom CNN model is developed from scratch to serve as a performance benchmark. This simple network consists of:

- Two convolutional layers with ReLU activation and max-pooling.
- A flattening layer.
- Two fully connected (dense) layers with dropout for regularization.
- An output layer with a sigmoid activation function for binary classification.

This model helps assess the difficulty of the task and provides a baseline to compare the effectiveness of transfer learning and knowledge distillation.

3.3.2 Transfer Learning with ResNet-18

ResNet-18, a deep residual network pre-trained on the large ImageNet dataset, is used for transfer learning. Two approaches are investigated:

- **Fixed Feature Extractor:** The convolutional base of ResNet-18 is frozen, and only the final classification layers are trained on the chest X-ray dataset. This is computationally efficient and effective when the new dataset is small.
- **Fine-tuning:** The entire ResNet-18 network is fine-tuned, allowing the pretrained weights to be slightly adjusted to better suit the specific features of chest X-ray images. This often yields higher accuracy but requires more computational resources.

3.3.3 Knowledge Distillation

Knowledge Distillation is implemented to transfer the knowledge from a powerful teacher model o a smaller student model[6].

- **Teacher Model:** **ResNet-50**, a deeper and more complex network pretrained on ImageNet, serves as the teacher. It is first fine-tuned on the chest X-ray dataset to achieve high performance.
- **Student Model:** ResNet-18 serves as the student. It is trained to mimic both the hard labels and the soft probabilities (logits) from the teacher's output. The loss function is a weighted sum of the standard cross-entropy loss and a distillation loss, which is a Kullback-Leibler (KL) divergence between the student's and teacher's softmax outputs.

This technique involves the student model replicating the behavior and predictions of the teacher model. Figure 3.2 illustrates how this technique is performed. The teacher model, which already possesses knowledge, is trained alongside the student model on the input data. Both models are trained up to the last fully connected layer to produce raw logits. These raw logits are processed through the softmax activation function, which includes a temperature parameter. The formula for the softmax with temperature is shown below:

$$P(x) = \frac{\exp(z_c(x)/T)}{\sum_{c' \in C} \exp(z_{c'}(x)/T)}$$

where $z_c(x)$ is the model that produces logits for class c and T is the temperature parameter that produces a softer probability distribution over classes. The result of this process is soft labels and soft predictions. On the other hand, the student model also performs a standard softmax activation function with a temperature of 1 to produce standard soft predictions. Then, The Kullback-Leibler (KL) Divergence between the soft labels and soft predictions is used to calculate the Distillation Loss. Additionally, the Student Loss is determined by applying standard Cross-Entropy between the conventional soft labels and the hard labels. Both losses are combined with a weighting factor alpha (α). The formulas for KL Divergence, Cross-Entropy Loss, and Total Loss are presented below:

$$\text{KL Divergence} = \sum_{x \in X} P_t(x) \log\left(\frac{P_t(x)}{P_s(x)}\right)$$

$$\text{Cross-Entropy Loss (Student Loss)} = -\sum_{x \in X} y(x) \log(P_s(x))$$

$$\text{Total Loss} = (1 - \alpha) \times (\text{KL}) + (\alpha) \times (\text{Cross-Entropy Loss})$$

Where $P_t(x)$ is the probability distribution over the classes predicted by the teacher model for a particular input x, $P_s(x)$ is the probability distribution over the classes predicted by the student model for the same input x, and α is a weighting factor that balances the contributions of the Student and Distillation Loss. This Total Loss is used to update the weights of the student model, enabling it to acquire knowledge from the teacher model. During this process, the teacher model's weights remain frozen, preventing further learning.

3.4 Architecture

The project's methodology is based on the principles outlined in the paper *On effects of Knowledge Distillation on Transfer Learning*. The core idea is to train a smaller, more efficient "student" model to replicate the performance of a larger, more powerful "teacher" model. This approach aims to achieve high accuracy with a more lightweight network, making it suitable for deployment in resource-constrained environments.

The proposed system follows a structured pipeline to perform chest X-ray classification. The system is composed of several sequential steps, as illustrated in Figure 3.2.

This setup aims to train the compact ResNet-18 model to achieve a performance level close to the larger ResNet-50, making it a highly efficient solution for deployment.

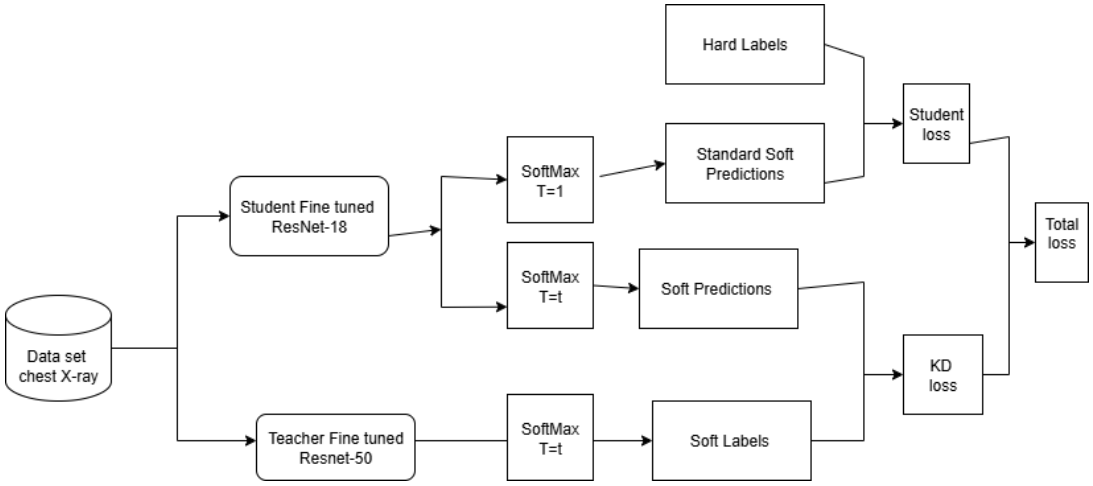


Figure 3.2: Data Flow Diagram of the Proposed System

Chapter 4

Experiment setup and Results

Experiment Setup

All models were implemented using the PyTorch deep learning framework.

Evaluation Metrics

The models are evaluated using several standard classification metrics to provide a comprehensive assessment of their performance:

Accuracy: The proportion of correctly classified instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision: The proportion of positive identifications that were actually correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall: The proportion of actual positives that were identified correctly.

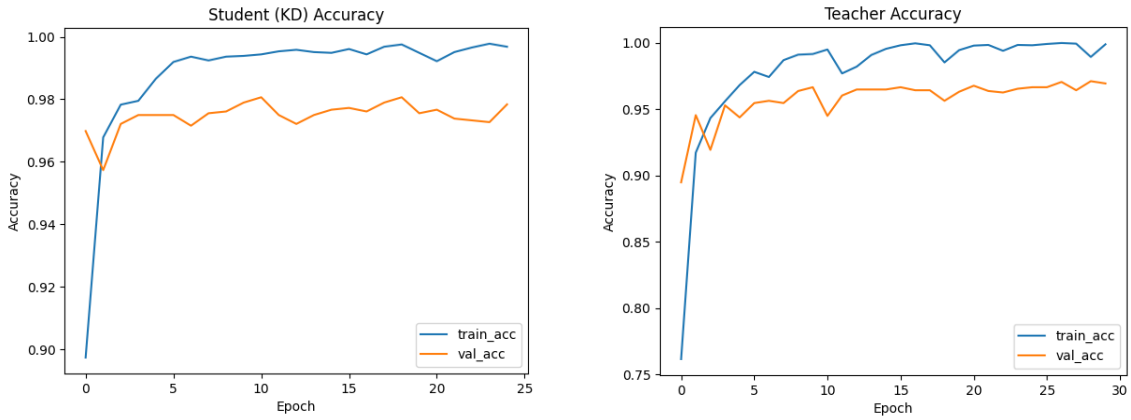
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-Score: The harmonic mean of precision and recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Results

The graphs show the model performance of teacher model and student model.



Confusion Matrix of Models

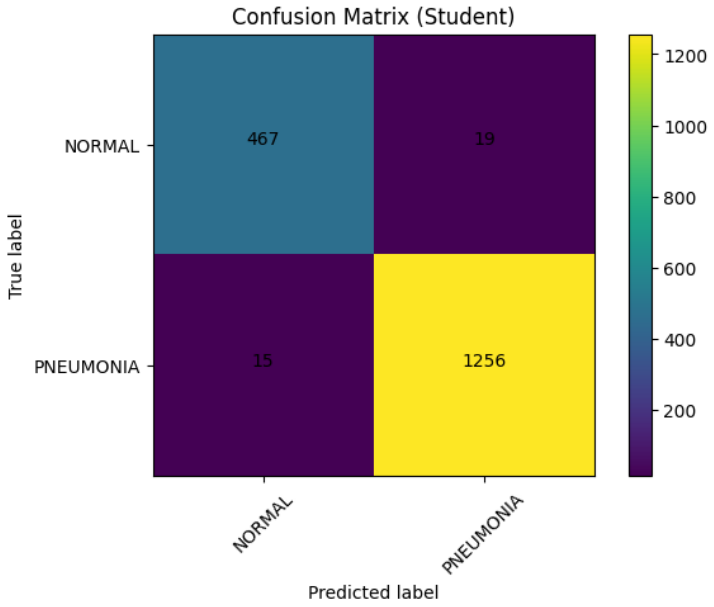


Figure 4.1: Confusion matrix of Final Student Model

Final Student Model Performance

The Final Student Model, a product of knowledge distillation, shows a very high recall of approximately 0.9882 (1256+15/1256). This indicates its excellent ability to correctly identify positive cases (pneumonia), meaning it minimizes false negatives (only 15 cases were missed). This is particularly important in medical diagnosis to ensure that actual pneumonia cases are not overlooked.

The model's precision is also very high, at approximately 0.9851 (1256+19/1256), which is very close to its recall. This suggests that when the model predicts pneumonia, it is correct a large majority of the time, with a very low number of false positives (only 19 normal cases were incorrectly classified as pneumonia).

The F1-score provides a balanced view of the model's performance on these two metrics, and it is also very high at approximately 0.9866. This high F1-score confirms the model's overall strong and balanced performance.

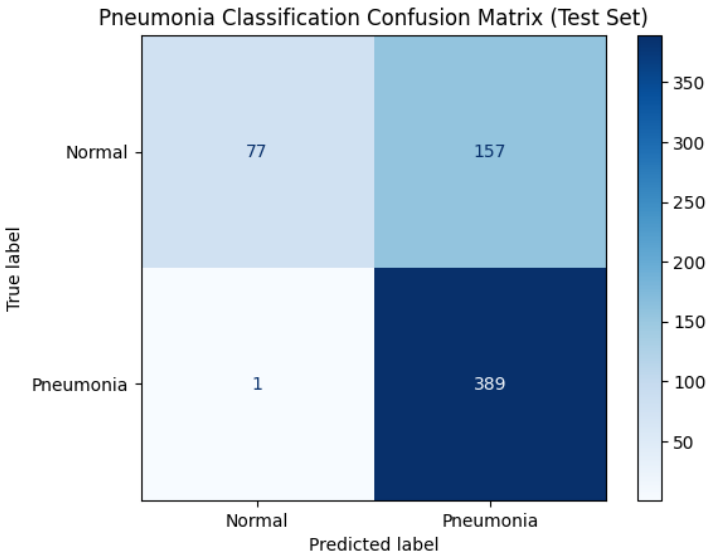


Figure 4.2: Confusion matrix of Baseline CNN

Model Performance Summary

The table below shows the training accuracy results in models.

Models	Accuracy
Baseline CNN	0.747
Transfer Learning (ResNet-18 Fine-tuned)	0.825
Transfer Learning (ResNet-18 Fixed Extractor)	0.875
Final Student Model	**0.9806**

Table 4.1: Performance comparison of model using Accuracy.

The Transfer Learning (ResNet-18 Fixed Extractor) model performed very competitively, with an accuracy of 0.875.

The Knowledge Distillation (ResNet-18 Student) model achieved an accuracy of 0.98, While slightly lower than the ResNet-50 teacher and fine-tuned ResNet-18, the student model significantly outperformed the baseline CNN and the fixed feature extractor. This confirms that knowledge distillation successfully transferred a large portion of the teacher's performance to the smaller student model, which has fewer parameters and faster inference time.

The Baseline CNN had the lowest performance across all metrics, highlighting the challenges of training a deep network from scratch on a limited dataset.

Chapter 5

Conclusion

This project implemented and evaluated various deep learning approaches for the classification of chest X-ray images. The results demonstrate that transfer learning is highly effective, enabling high accuracy even with a relatively small dataset. The Knowledge Distillation approach, using a powerful ResNet-50 teacher to guide a smaller ResNet-18 student. The student model achieved a performance close to the teacher's, validating the hypothesis that knowledge can be effectively transferred to a more compact network. This makes the distilled model a practical solution for deployment in resource-constrained medical environments.

Recommendations

For future work, we recommend:

- Exploring different teacher-student network architectures for knowledge distillation.
- Testing the result with different possible hyperparameter.
- Deploying the final model on a mobile or edge device to test its real-world performance in a clinical setting.

Bibliography

[1] A. Brutzkus and A. Globerson, "Why do larger models generalize better? a theoretical perspective via the xor problem," in *International conference on machine learning*, PMLR, 2019, pp. 822–830.

[2] H. Mhaskar, Q. Liao, and T. Poggio, "When and why are deep networks better than shallow ones?" In *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[3] W. Zhang, H. Wang, Z. Lai, and C. Hou, "Constrained contrastive representation: Classification on chest x-rays with limited data," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE Computer Society, 2021, pp. 1–6.

[4] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, U. Naseem, and Y. Gu, "Robust and explainable framework to address data scarcity in diagnostic imaging," *arXiv preprint arXiv:2407.06566*, 2024.

[5] J. Kishore, A. Jain, K. Krishna Koushika, P. K. Mishra, S. Karanwal, and S. Solanki, "Enhancing medical diagnosis on chest x-rays: Knowledge distillation from self-supervised based model to compressed student model," *Discover Computing*, vol. 28, no. 1, p. 118, 2025.

[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[7] S. Thapa, *On Effects of Knowledge Distillation on Transfer Learning*. New Mexico Institute of Mining and Technology, 2022.

[8] B. Pardamean, T. W. Cenggoro, R. Rahutomo, A. Budiarto, and E. K. Karuppiah, "Transfer learning from chest x-ray pre-trained convolutional neural network for learning mammogram data," *Procedia Computer Science*, vol. 135, pp. 400–407, 2018.

[9] A. Susanto, T. W. Cenggoro, B. Pardamean, *et al.*, "Transfer-learning-aware neuroevolution for diseases detection in chest x-ray images," *arXiv preprint arXiv:2004.07136*, 2020.

[10] P. Mooney, *Chest x-ray images (pneumonia)*, 2018. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia.