# MULTI-CLASS PNEUMONIA CLASSIFICATION ON X-RAY IMAGES USING TRANSFER LEARNING AND SELF-SUPERVISED FINE-TUNING

*Sebnem Demirtas (76813), Mete Erdogan (69666)*

Koc University, Istanbul
Department of Computer Engineering

## ABSTRACT

Detecting pneumonia through X-Ray images is crucial especially for early diagnosis. The application of Machine Learning and Deep Learning techniques in this domain not only eases the workload of doctors but also enhances diagnostic accuracy. Therefore, leveraging these advanced techniques holds significant value and numerous studies have addressed this issue using various approaches. In this work, we focus on not only detecting the illness, but also classifying given X-Ray images as either viral or bacterial pneumonia or healthy, offering better treatment for patients. For this purpose, we experiment with ResNet models in both Transfer Learning and from scratch training scenarios. Furthermore, a Self-Supervised Transformer model DINO ViT-B/16 is fine-tuned on pneumonia classification with a linear classifier. Most importantly, we experimented with the Extreme Learning Machines (ELM) aided with Principle Component Analysis (PCA) to train a linear classifier on the extracted feature vectors. This approach yields successful generalization with a training speed equivalent to one epoch fine-tuning with gradient descent. [1]

***Index Terms***— Pneumonia Classification, X-Ray, Transfer Learning, Self-Supervised Learning, Extreme Learning Machines

## 1. INTRODUCTION

Pneumonia, a common respiratory infection that affects people of all ages, represents a major global healthcare challenge. Early and accurate detection is essential for effective treatment and the prevention of serious complications. Furthermore, classifying the cause of the disease as viral or bacterial infection further eases the job of medical personnel to propose more effective treatment for the patients [1]. Non-invasive imaging methods, such as X-Ray imaging, have become crucial diagnostic tools for assessing pneumonia. With the advent of large-scale datasets and advances in deep learning, transfer learning and self-supervised learning techniques have gained prominence. These techniques leverage the capabilities of pre-trained Convolutional Neural Network (CNN) such as ResNet [2] and Vision Transformer (ViT) [3] models to enhance pneumonia detection.

In this paper, we explore the utilization of the ResNet model [2] across different training paradigms, including from-scratch training, transfer learning, and fine-tuning the last layer. Additionally, we investigate the application of self-supervised learning-based Transformer models, such as DINO [4], for pneumonia classification by fine-tuning a linear classifier. Finally, we employ Extreme Learning Machines (ELM), which offer a least-squares solution for training linear classifiers, together with the Singular Value Decomposition (SVD)-based Principal Component Analysis (PCA) for feature elimination. This approach, applied to both the fine-tuning of ResNet and DINO models, provides a fast and robust alternative to gradient-based methods and demonstrates strong generalization performance in our experiments.

## 2. RELATED WORK

Transfer learning has emerged as a fundamental method in machine learning, allowing pre-trained models to be applied to new, often related tasks with less data. Transfer learning makes use of the information gathered from big datasets and challenging tasks to help models generalize more effectively and converge more quickly when applied to new tasks. This greatly minimizes the requirement for a lot of labeled data and computer power. In particular, ResNet models, which were first presented by He et al. [2], transformed deep learning by using residual connections to solve the vanishing gradient problem. ResNet has shown impressive results in applications involving image detection and classification. Pre-trained versions on massively parallel datasets, like ImageNet, have been widely used as a foundation for a variety of downstream tasks, demonstrating the model's versatility and resilience in transfer learning situations. For instance Chouhan et al. [5] utilized a variety of transfer learned CNN models in an ensembled fashion to achieve good performence in pneumonia detection.

Furthermore, self-supervised learning has proven to be an effective strategy, especially in situations when obtaining labeled data is costly or difficult. Through the utilization of self-supervised learning techniques, models can acquire valuable representations without requiring substantial human

[1]This work is done as the project of the COMP448 - "Medical Image Analysis" course in Koç University.

annotation, by capitalizing on the underlying structure and patterns contained in unlabeled data. Fields such as computer vision and natural language processing, have effectively used this paradigm [6]. In particular, DINO (self-DIstillation with NO labeling) transformer models have attracted a lot of interest in the area of vision tasks [4]. With DINO, a student network gains knowledge from a teacher network using a self-distillation framework. The two networks are initiated the same but change on their own during training. This method yields outstanding results on many picture identification and segmentation tasks by allowing the model to collect complex visual elements and representations without labeled supervision. Transformer architectures are used in DINO models to further improve their ability to represent complex structures and extracting global features inside images. This makes DINO models ideal for a wide range of computer vision self-supervised learning applications.

Extreme Learning Machines (ELM) [7] are known for its fast training speed and strong generalization. Unlike traditional neural networks that optimize all parameters using gradient-based methods, ELMs fix and randomly assign input weights and biases, learning only the output weights via a simple least-squares solution. This avoids iterative optimization strategies, which leads to noticeably quicker training times. In our work, rather than randomly initializing the weights and biases of the first layers of a network, we use the features extracted by either the transfer learning and self-supervised learning based models, and train a linear classifier. ELMs are valuable for rapid training and deployment in real-time and low-resource applications, offering competitive performance in tasks like feature learning, regression, and classification. For instance, Nahiduzzaman et al. [8] used ELMs with PCA for feature elimination, employing a generic CNN for feature extraction and fine-tuning a linear layer for binary and multi-class pneumonia classification.

## 3. METHODOLOGY

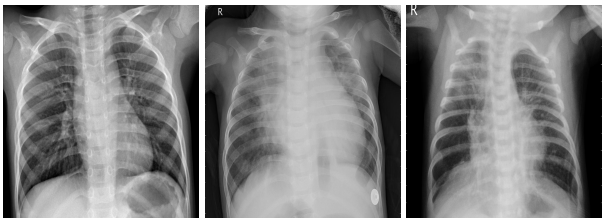
### 3.1. Dataset Specifications

The Chest X-Ray images of pneumonia were sourced from the Guangzhou Women and Children's Medical Centre in Guangzhou [9] and are also publicly accessible on Kaggle [2]. This dataset contains a fixed train and test splits of X-ray images with resolutions ranging from 400 to 2,000 pixels. The images are categorized into three groups: normal, bacterial pneumonia, and viral pneumonia; where number of images are given as in table 1 for the train and test splits. Example instances for different classes of the dataset can be seen in figure 1.

[2]https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data

| Type | Train Set | Test Set |
|---|---|---|
| Normal | 1341 | 234 |
| Bacterial | 2530 | 242 |
| Viral | 1345 | 148 |
| Total | 5216 | 624 |

**Table 1**: Number of X-Ray Images in the Dataset



(a) (b) (c)

**Fig. 1**: Sample images of (a) normal, (b) bacterial pneumonia, and (c) viral pneumonia from the dataset.

### 3.2. Dataset Preparation and Preprocessing

For the preprocessing of the images, we started by converting resizing the greyscale images to 224x224 pixels to ensure uniform input dimensions. Then, we applied mean and standard deviation normalization by calculating training data statistics and performed adaptive histogram equalization (CLAHE) to enhance the model performance and reliability. In later steps, we use the training portion of the dataset to perform k-fold cross validation where we apply data augmentation to the training set of each fold for balancing the class distribution. Also, while giving the greyscale inputs to the models we duplicate the image tensors to 3-channels for model compatibility.

### 3.3. Pretraining and Fine Tuning Using Gradient Descent Algorithm

By using the gradient descent algorithm, we train the following models on our pneumonia classification dataset: 1) ResNet18 trained from scratch, 2) ResNet18 where all parameters trained after transfer learning initialization, 3) Resnet18 where the last linear layer is fine-tuned after transfer learning initialization and 4) DINO ViT-B/16 model fine-tuned using a linear classifier. The results of these models are presented in 3.

### 3.4. Training Specifications for Gradient Descent Based Methods

In the training of each of these models, we used the Adam [10] optimizer with a weight decay for regularization. We performed a grid-search to find the optimal values for the batch size, learning rate and the weight decay rate and performed k-fold cross validation with $k = 5$ on each grid-search combination. As we have a fixed test set, for each k-fold we divided

| Model | Epoch | Optimizer | Learning Rate | Weight Decay | Batch Size | Loss | Scheduler |
|---|---|---|---|---|---|---|---|
| ResNet From Scratch | 30 | Adam | 0.0001 | 1e-05 | 16 | Cross Entropy | StepLR |
| ResNet Transfer Learn | 30 | Adam | 0.0005 | 1e-05 | 32 | Cross Entropy | StepLR |
| ResNet Fine-Tuning | 30 | Adam | 0.001 | 1e-05 | 16 | Cross Entropy | StepLR |
| DINO Fine-Tuning | 30 | Adam | 0.001 | 1e-04 | 16 | Cross Entropy | StepLR |

**Table 2**: Training details of gradient descent based models. The parameter values are found with a grid-search on learning rate, weight decay and batch size, and validated with 5-fold cross validation.

the training set into 80% (training) and 20% (validation) portions and reported the test set results of the model with the best validation F1-score. Then, as the data has imbalanced class distribution, after each train-validation split in each fold, we applied data augmentation to the training portion including random horizontal flip, random rotation (10 degrees) and gaussian blur. The training parameters calculated after grid-search and 5-fold cross validation are presented in table 2.

### 3.5. Fine-Tuning by Extreme Learning Machines (ELM)

Extreme Learning Machines (ELMs) [7] offer a fast and efficient method for training linear classifiers, particularly when utilized on top of features extracted from self-supervised models. Unlike traditional neural networks that rely on iterative backpropagation, ELMs employ a straightforward closed-form solution to learn the output weights, significantly reducing the computational complexity and training time.

#### *3.5.1. Pseudo-Inverse Based ELM Formulation*

In this approach, a model is first employed to extract meaningful features from the input data. Let $\mathbf{X}$ denote the matrix of these extracted features, where $\mathbf{X} \in \mathbb{R}^{N\times d}$ represents $N$ samples with $d$ features each. These features serve as the input to the ELM, which is used to train a linear classifier.
The target matrix $\mathbf{T}$, containing the target vectors $\mathbf{t}_i$ for all samples, is constructed as follows together with the output linear layer weight matrix $\boldsymbol{\beta}$:

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ \vdots \\ \mathbf{t}_N^T \end{bmatrix}_{N\times m} \qquad \beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_d^T \end{bmatrix}_{d\times m} \tag{1}$$

The relationship between the feature matrix $\mathbf{X}$ and the target matrix $\mathbf{T}$ through the output weight matrix $\boldsymbol{\beta}$ is established by the linear system:

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{T} \tag{2}$$

To determine the output weight matrix $\boldsymbol{\beta}$, the Moore-Penrose generalized inverse of $\mathbf{X}$ is used:

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^{\dagger}\mathbf{T} \tag{3}$$

Here, $\mathbf{X}^{\dagger}$ denotes the Moore-Penrose pseudoinverse of $\mathbf{X}$. This approach minimizes the least squares error between the predicted outputs and the actual targets, providing an efficient and effective solution for training the linear classifier. By using the powerful feature extraction capabilities of self-supervised or transfer learning models; and the simplicity of ELMs, this method facilitates rapid training and robust performance, making it a valuable technique for various linear classifier fine-tuning tasks.

#### *3.5.2. Feature Extraction Using Principal Component Analysis (PCA)*

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms data into a new coordinate system where the greatest variances by any projection of the data come to lie on the first coordinates (called principal components). The Singular Value Decomposition (SVD) is a powerful method to perform PCA.

First, we center the data matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$ by subtracting the mean of each feature:

$$\mathbf{X}_{\text{centered}} = \mathbf{X} - \mathbf{1}_N\mathbf{m}^T$$

where $\mathbf{1}_N$ is a column vector of ones of length $N$ and $\mathbf{m}$ is a column vector of the means of each feature. Next, we perform Singular Value Decomposition (SVD) on the centered data matrix:

$$\mathbf{X}_{\text{centered}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

Here, $\mathbf{U}$ is an $N\times N$ orthogonal matrix, $\boldsymbol{\Sigma}$ is an $N\times d$ rectangular diagonal matrix with non-negative real numbers on the diagonal (the singular values), and $\mathbf{V}$ is a $d \times d$ orthogonal matrix. Then, principal components are given by the columns of $\mathbf{V}$ and the principal component scores are given by the projection of the centered data onto these principal components:

$$\mathbf{Z} = \mathbf{X}_{\text{centered}}\mathbf{V}$$

For dimensionality reduction, we select the first $k$ principal components corresponding to the largest singular values. Let $\mathbf{V}_k$ be the $d\times k$ matrix containing the first $k$ columns of $\mathbf{V}$, then the reduced data matrix is:

| Model | Train | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|
| | F1-Score (%) | Accuracy (%) | Recall (%) | Precision (%) | F1-Score (%) | Accuracy (%) | Recall (%) | Precision (%) |
| **M1: ResNet18** | <u>96.17</u> | <u>97.51</u> | <u>95.99</u> | <u>96.36</u> | 69.74 | 81.20 | 71.20 | 74.71 |
| **M2: ResNet18 TL** | **97.34** | **98.22** | **97.31** | **97.38** | 73.54 | 83.76 | 74.65 | 77.56 |
| **M3: ResNet18 FT** | 78.63 | 86.12 | 79.13 | 78.22 | <u>77.82</u> | <u>86.00</u> | <u>79.35</u> | <u>79.37</u> |
| **M4: ResNet18 ELM** | 83.19 | 89.12 | 82.57 | 84.05 | 74.69 | 84.29 | 76.53 | 78.68 |
| **M5: ResNet50 ELM** | 66.85 | 80.99 | 67.56 | 70.68 | 65.69 | 79.06 | 65.23 | 70.70 |
| **M6: DINO FT** | 77.99 | 86.45 | 77.87 | 79.10 | 71.37 | 81.84 | 71.59 | 74.09 |
| **M7: DINO ELM** | 83.59 | 89.61 | 83.40 | 84.22 | **84.27** | **90.28** | **84.49** | **85.49** |

**Table 3**: Average performance metrics for our models on the train and test datasets for 3-class pneumonia classification. The evaluated models are ResNet18 train from scratch (M1), ResNet18 with transfer learning (TL) initialization (M2), ResNet18 fine-tuned (FT) last layer with gradient descent (M3), ResNet18 with Extreme Learning Machine (ELM) (M4), ResNet50 with ELM (M5), DINO fine-tuned linear classifier with gradient descent (M6), and DINO with ELM (M7). The best-performing metrics are highlighted in bold, and the second-best metrics are underlined.

$$\mathbf{Z}_k = \mathbf{X}_{\text{centered}}\mathbf{V}_k$$

where $\mathbf{Z}_k$ is the transformed data in the reduced $k$-dimensional space. Thus, SVD-based PCA allows us to effectively reduce the dimensionality of the data while preserving as much variance as possible. This operation offers two main advantages for our resulting classifier. First, it serves as a form of regularization by emphasizing the most significant feature dimensions, focusing on the largest singular values. Second, when the training data is limited and the samples are highly correlated, the feature matrix $\mathbf{X}$ can have a high condition number. Multiplying the feature matrix by the orthogonal $\mathbf{V_k}$ matrix projects the features onto an orthogonal basis, reducing the singularity of the feature matrix and improving numerical stability.

### 3.6. Training Specifications for Extreme Learning Machine (ELM) Based Methods

Using our ELM and PCA based fine-tuning method, we train a linear classifier on top of the following models: 1) ResNet18, 2) ResNet50 and 3) DINO ViT-B/16. We again perform grid-search and k-fold cross validation with $k = 5$ to find the best PCA parameter to reduce the feature dimension. The results for the ELM based models are presented in table 3.

## 4. RESULTS

The results of our experiments shown in the table 3 demonstrate a clear distinction in performance between different models and methodologies for multi-class pneumonia classification using X-ray images. Among the evaluated models, the DINO ELM (M7) model, which combines a self-supervised transformer with Extreme Learning Machines, outperformed all other models on the test set, achieving an F1-Score of 84.27% and an accuracy of 90.28%. This indicates superior generalization capability of self-supervised models and ELMs that have robust performance in unseen data scenarios. Furthermore, ResNet18 FT (M3), which involves fine-tuning the last layer, demonstrated the second-best test performance with an F1-Score of 77.82%, an accuracy of 86.00%.

The effectiveness of DINO ELM suggests that self-supervised learning can leverage large-scale, unlabeled data to capture intricate feature representations that transfer learning models may miss. While ResNet18 ELM (M4) and ResNet50 ELM (M5) also utilized ELMs, their performance was inferior compared to DINO ELM, with M5 performing particularly poorly, indicating that increased model complexity without appropriate fine-tuning can lead to suboptimal results. Lastly, the ResNet18 model trained from scratch (M1) and Transfer Learning initialized ResNet18 (M2) exhibited the two best training set performance, but could not generalize well to the test set with a notably lower performance with an F1-Score of 69.74% and 73.54% respectively.

To further validate the performance differences between models, we used the McNemar-Bowker Test, which is a generalization of the regular McNemar's Test for multi-class, a statistical method for comparing paired predictions. The McNemar-Bowker test examines the disagreements between predictions of multiple classifiers, using a contingency table based on their predictions on the test dataset. The test statistic follows a chi-squared distribution, and the resulting p-value indicates if the prediction difference is statistically significant. For example, comparing ResNet18 (M1) and DINO ELM (M7) showed significant differences, with a p-value less than 0.05, indicating the difference in model predictions. This validation ensures our performance metrics reflect true differences. Applying McNemar-Bowker test across all model pairs created a matrix of p-values as in table 4. These results confirm significant differences, particularly between DINO ELM and other

models. Furthermore, this highlights the potential of combining self-supervised learning with Extreme Learning Machines to improve model generalization and classification accuracy in medical image analysis, offering a promising approach for developing efficient diagnostic tools, especially in data-scarce environments.

| Model | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
|---|---|---|---|---|---|---|---|
| M1 | - | 0.000 | 0.013 | 0.000 | 0.000 | 0.000 | 0.000 |
| M2 | 0.000 | - | 0.082 | 0.000 | 0.000 | 0.000 | 0.000 |
| M3 | 0.013 | 0.082 | - | 0.000 | 0.005 | 0.000 | 0.001 |
| M4 | 0.000 | 0.000 | 0.000 | - | 0.117 | 0.000 | 0.074 |
| M5 | 0.000 | 0.000 | 0.005 | 0.117 | - | 0.000 | 0.001 |
| M6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | - | 0.000 |
| M7 | 0.000 | 0.000 | 0.001 | 0.074 | 0.001 | 0.000 | - |

**Table 4**: Pairwise McNemar-Bowker Test [11] p-values between model outputs of the models presented in table 3. This table summarizes the p-values from the McNemar-Bowker test for pairwise comparisons between the outputs of seven different classifiers (M1 to M7). Significant p-values (typically less than 0.05) indicate that the performance difference between the paired classifiers is statistically significant.

## 5. CONCLUSION

In this paper, we demonstrated that combining self-supervised learning with Extreme Learning Machines (ELMs) significantly enhances multi-class pneumonia classification using X-ray images. The DINO ELM model achieved the highest performance, with an F1-Score of 84.27% and an accuracy of 90.28%, showing their superior generalization capabilities compared to traditional transfer learning based fine-tuning methods. While transfer learning approaches like ResNet models initialized with ImageNet parameters, provided strong initial performance, they were outperformed in generalizing to test data compared to self-supervised transformer models. Our findings underline the effectiveness of utilizing self-supervised models and the rapid training capabilities of ELMs in medical imaging tasks, highlighting a promising approach for developing efficient and accurate diagnostic tools in healthcare.

## 6. REFERENCES


[1] Jithin Thomas, Aiste Pociute, Rimantas Kevalas, Mantas Malinauskas, and Lina Jankauskaite, "Blood biomarkers differentiating viral versus bacterial pneumonia aetiology: a literature review," *Italian journal of pediatrics*, vol. 46, pp. 1–10, 2020.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[5] Vikash Chouhan, Sanjay Kumar Singh, Aditya Khamparia, Deepak Gupta, Prayag Tiwari, Catarina Moreira, Robertas Damaševičius, and Victor Hugo C De Albuquerque, "A novel transfer learning based approach for pneumonia detection in chest x-ray images," *Applied Sciences*, vol. 10, no. 2, pp. 559, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[8] Md Nahiduzzaman, Md Omaer Faruq Goni, Md Shamim Anower, Md Robiul Islam, Mominul Ahsan, Julfikar Haider, Saravanakumar Gurusamy, Rakibul Hassan, and Md Rakibul Islam, "A novel method for multivariant pneumonia classification based on hybrid cnn-pca based feature extraction using extreme learning machine with cxr images," *IEEE Access*, vol. 9, pp. 147512–147526, 2021.

[9] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al., "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley data*, vol. 2, no. 2, pp. 651, 2018.

[10] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] Albert H Bowker, "A test for symmetry in contingency tables," *Journal of the american statistical association*, vol. 43, no. 244, pp. 572–574, 1948.