From Scratch to Large Pre-Trained Models: a Comparative Study for Medical Image Classification

Giovanni Lonia[1,*], Davide Ciraolo[1], Fabrizio Celesti[2], Maria Fazio[1] and Antonio Celesti[1]

[1]Department of Mathematics and Computer Sciences, Physical Sciences and Earth Sciences, University of Messina, Messina, Italy
[2]Department of Medical, Surgical and Neuro Sciences, University of Siena, Siena, Italy

Abstract

Thanks to the notable developments in Deep Learning (DL) techniques and the availability of High-performance computing (HPC) resources, medical image analysis has advanced significantly in recent years. Convolutional Neural Networks (CNNs) have been the standard model for image classification for the past ten years, and they have demonstrated outstanding effectiveness in various medical applications. However, the emergence of Vision Transformers (ViTs) has challenged the supremacy of CNNs. This paper aims to investigate the potential of ViTs in healthcare by comparing their efficacy with conventional CNN models. The first step in our comparative study is analyzing the key benefits of the two models: CNNs are generally good for extracting features from images using convolutional operations. On the other hand, ViTs use self-attention processes for recognizing long-range relationships, useful to handle intricate patterns in images. After this comparison, we evaluate the behavior of both the technologies from-scratch and large pre-trained models on both a consumer laptop using a MacBook Pro, and a Cloud HPC Infrastructure as a Service (IaaS) using an Azure Virtual Machine (VM), pointing out the variations in their performances and shedding light on the suitability of Transfer Learning (TL) in healthcare.

Keywords

Deep Learning, Computer Vision, Artificial Intelligence, Machine Learning, Transfer Learning, HPC.

1. Introduction

In the past decade, computer vision has undergone a remarkable transformation, largely driven by two key Deep Learning (DL) approaches: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The literature on CNNs and ViTs in healthcare reflects a dynamic and rapidly evolving field, where both these approaches have significantly contributed to medical image analysis and diagnostics, albeit with distinct strengths and limitations [1]. CNNs have long been the cornerstone of medical image analysis due to their ability to learn hierarchical features from imaging data. Their architecture, composed of convolutional, pooling, and fully connected layers, is particularly well-suited for tasks such as disease detection, organ segmentation, and anomaly classification [2]. Despite their success, CNNs face limitations in capturing global contextual information due to their localized receptive fields. This shortcoming has paved the way for the adoption of VTs, which leverage self-attention mechanisms to model long-range dependencies in images. ViTs, originally developed for natural language processing, have been adapted for medical imaging with promising results. Specifically, ViTs often outperform CNNs in classification tasks, particularly when trained on large datasets and fine-tuned for specific medical applications [3].

The comparative analysis between CNNs and ViTs reveals a nuanced landscape. While CNNs remain highly effective for tasks requiring local feature extraction, ViTs excel in scenarios demanding holistic image understanding. This research compares these two approaches, examining their designs, performances, and transfer learning capabilities. Our study focuses on delineating the differences

ITADATA-WS 2025: The 4th Italian Conference on Big Data and Data Science – Workshops, September 9–11, 2025, Turin, Italy

[*]Corresponding author.

[†]All authors contributed equally.

glonia@unime.it (G. Lonia); davciraolo@unime.it (D. Ciraolo); fabrizio.celesti@unisi.it (F. Celesti); mfazio@unime.it (M. Fazio); acelesti@unime.it (A. Celesti)

0009-0000-1824-0123 (G. Lonia); 0009-0000-0982-5453 (D. Ciraolo); 0000-0003-1629-0922 (F. Celesti); 0000-0003-3574-1848 (M. Fazio); 0000-0001-9003-6194 (A. Celesti)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-15

between CNNs and ViTs, and assessing their performance, adaptability, and potential, especially considering their usage in healthcare applications. To perform our comparison, we considered two different test beds, i.e., a consumer laptop consisting of a Macbook Pro, and a Cloud HPC Infrastructure as a Service (IaaS) consisting of an Azure Virtual Machine (VM). This allowed us to highlight the limitations of consumer systems compared to HPC infrastructures for training large models. A Cloud-based HPC IaaS provides scalability, cost-efficiency, and flexibility, eliminating the need for significant upfront hardware investments while delivering on-demand, high-performance computing power. Nowadays, HPC environments play a crucial role in executing these DL models, given the computational intensity required to leverage their full potential in real-world scenarios. However, they present business challenges that can limit their usage.

The reminder of this paper is organised as follows. Section 2 briefly analyses the state of the art in medical image classification. In Section 3, we discuss materials and methods adopted in our research work. Section 4 presents a comprehensive performance assessment between CNN and ViT models considering both from-scratch and pre-trained approaches. Section 5 concludes the paper also providing light on the future.

2. Related Work

In recent years, the use of DL models, such as CNNs and ViTs, has significantly advanced disease recognition using medical images. Medical photographs represent 90% of the data in digital medicine applications [4], making the use of DL techniques increasingly prevalent in healthcare.

CNNs have become foundational tools in healthcare AI due to their ability to automatically extract hierarchical features from complex medical data, particularly imaging modalities such as MRI, CT, X-rays, and histopathology slides. [2] traces the evolution of CNN architectures from early models like LeNet to more advanced ones such as ResNet and EfficientNet. The paper emphasizes how CNNs have been successfully applied to critical diagnostic tasks, including cancer detection, Alzheimer's disease diagnosis, and brain tumor identification. It also discusses the architectural innovations that have improved performance and efficiency in medical imaging tasks, while highlighting challenges such as overfitting, data imbalance, and the need for interpretability in clinical settings. Another significant contribution is [5], which focuses on the application of CNNs in MRI image analysis. This study provides a detailed taxonomy of CNN-based approaches for tasks such as image pre-processing, segmentation, and classification. It also explores the integration of CNNs with large-scale retrieval systems to enhance the efficiency of medical image processing. The authors underscore the importance of domain-specific adaptations and the role of CNNs in improving diagnostic accuracy and workflow automation. [6] offers a comparative perspective by evaluating CNNs alongside Vision Transformers (ViTs) in medical image analysis. While the paper primarily focuses on the comparative performance of these models, it reaffirms the robustness and computational efficiency of CNNs, especially in tasks involving limited data or requiring real-time inference. CNNs are noted for their strong inductive biases, such as locality and translation invariance, which make them particularly effective in medical imaging contexts where annotated data is often scarce. [7] presents an enhanced CNN model for the early detection and classification of ophthalmic diseases. The model was trained on retinal images and demonstrated high accuracy in identifying conditions such as diabetic retinopathy and glaucoma. The authors emphasize the importance of early diagnosis in preventing vision loss and show how CNNs can automate and improve diagnostic workflows in ophthalmology. These studies collectively demonstrate that CNNs have become indispensable in healthcare AI, offering scalable, accurate, and efficient solutions for a wide range of diagnostic and prognostic tasks. However, they also point to ongoing challenges, including the need for explainability, generalization across diverse patient populations, and integration into clinical workflows.

In the last few years, ViTs have gained popularity in healthcare and other fields as a potential alternative to CNNs. Authors in [8] developed a multi-feature ViT model for classifying infant cry audio to diagnose neonatal conditions such as sepsis and respiratory distress syndrome. The model achieved 99% classification accuracy and incorporated explainable AI techniques like LIME and LRP to enhance interpretability. This study highlights the versatility of ViTs beyond imaging, extending their utility to audio-based diagnostics in pediatrics. In the domain of biometric healthcare, [9] applied ViTs to vein biometric recognition. Their work involved fine-tuning pre-trained ViTs on various vascular datasets, achieving high identification rates across multiple modalities such as finger, palm, and wrist veins. This study underscores the effectiveness of ViTs in handling limited data scenarios through transfer learning. These studies collectively demonstrate that ViTs are not only capable of matching but often surpassing CNNs in healthcare applications, particularly when global context and multimodal integration are essential.

Transformers have a large capacity for learning as they can recognize and learn global relationships in images. However, they might have limited generalization issues since they do not typically explain local correlation in images. To utilize both local and global image representations, ViTs have started combining the convolution operation with the self-attention mechanism. These designs, also known as hybrid vision transformers, have shown impressive performance in vision applications [10]. ViT models are also employed in circumstances where annotated data is scarce. Lagunas et al. in [11] demonstrate how implementing semi-supervised learning techniques, ViTs outperform CNNs.

3. Materials and Method

In this Section, we introduce the specific architectures that define CNN and ViT models. The key design principles of CNNs, such as their hierarchical convolution and pooling layers, and ViTs, characterized by their self-attention mechanisms, significantly impact how these models extract and process features from input images.

3.1. Convolutional Neural Network

CNN is a popular type of neural network that is specifically designed for handling matrix-structured inputs, such as images [12]. Figure 1 depicts our CNN architecture. A CNN architecture comprises three main parts: feature extraction, classification and probability distribution. First, convolutional and pooling layers extract important features from the input image, reducing its dimensionality while preserving key information. The resulting feature maps are then flattened into 1D vectors. These vectors are passed into a fully connected neural network, which classifies the image by estimating the likelihood of each class. Finally, the softmax activation function converts these outputs into probabilities, selecting the class with the highest probability as the predicted label.

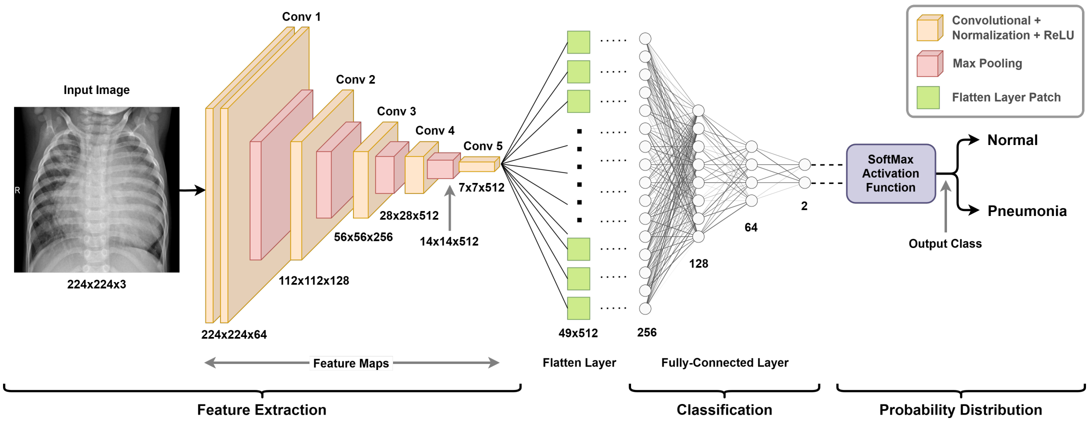
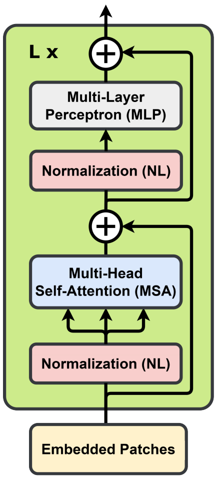


Figure 1: Architecture of our CNN.

3.2. Vision Transformer

ViT is a modern neural network architecture, built upon the well-known Transformer model [13], developed for Natural Language Processing (NLP) applications. ViT has shown remarkable results in image classification tasks, thanks to its capability of capturing long-range dependencies and spatial hierarchies in images, making it a highly effective model. Figure 3 shows our ViT implementation for this case study. It is based on the most famous architecture for image classification [14]. Its workflow can be divided into four main steps: input patching and linear embedding, positional embedding, transformer encoding and classification (commonly with an Multi-Layer Perceptron (MLP) head). The input image is divided into patches, which are flattened and linearly embedded to reduce dimensionality. Positional embeddings are added to each patch to capture spatial relationships within the image. The combined embeddings are processed by a transformer encoder, which uses self-attention mechanisms to learn global dependencies and spatial hierarchies. Finally, the output of the encoder is passed through a Multi-Layer Perceptron head, which classifies the image, often using a softmax function to produce a probability distribution over classes. The core of this architecture is the transformer encoder (shown in Figure 2), which is composed of three main components: Normalization Layer (NL), Multi-Head Self-Attention (MSA) and MLP. The NL is used at two points in the encoder: first to normalize embedded patches (including positional embeddings), and then to normalize the output of the MSA before feeding it to the MLP. Normalization helps standardize data for comparability. The MSA is a key component of transformers, where Query (Q), Key (K), and Value (V) vectors are derived from input patches. The attention mechanism involves calculating attention scores by taking the dot product of Q and K, scaling the results, and applying softmax to generate a probability distribution. The attention output is computed by weighting the V vectors based on this distribution. Multiple attention heads are used in parallel, and their outputs are concatenated and linearly projected to capture various patterns between patches. Finally, the MLP processes the sum of the MSA outputs and residual from the previous normalization layer, normalizes the result, and introduces non-linearity to help the model capture complex spatial patterns in the data.



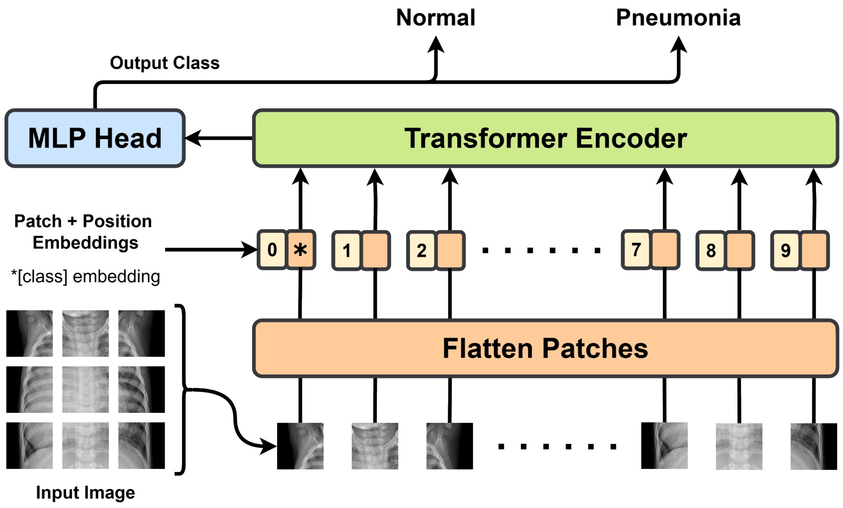


Figure 2: Transformer Encoder. **Figure 3:** Architecture of our Vision Transformer implementation.

3.3. Dataset

The dataset used for this study is the *Chest X-Ray Images (Pneumonia)*[1]. It is made up of three directories (i.e., train, test, val), each of which includes a subdirectory for each image category. In total 5,863 JPEG X-ray images were divided into two groups: Normal (25%) and Pneumonia (75%). There is a significant

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

difference in the number of samples, indicating an unbalanced dataset. It is common to encounter data imbalance in healthcare datasets as investigations often focus on cases more likely to detect or confirm a particular health issue. This leads to certain medical conditions being underrepresented in the dataset as compared to others. However, our study's focus is not on achieving the highest prediction accuracy, but rather on measuring the difference in performance between common "traditional" learning and transfer learning techniques using different test beds.

3.4. Pre-Trained Models

To assess the performance of the Transfer Learning technique we selected one of the well-known pre-trained models for both CNN and ViT family of models.

3.4.1. CNN

ResNet is a deep neural network architecture that introduced skip connections with residual learning to address vanishing gradient issues, which enables the training of extremely deep networks [15]. The introduction of residual blocks is ResNet's primary innovation in CNNs. In CNNs, layers are typically stacked one on top of the other with the expectation that each layer will capture a specific input data transformation. However, as the network becomes deeper, training becomes more challenging. Common problems include the degradation problem, which leads to increased training error with the addition of more layers, and the vanishing gradient problem. To address these issues, ResNet introduces shortcut connections, where the input and output of a layer are combined in a residual block. This allows the layer to learn the residual or the difference between the input and output. By utilizing this method, the model can more effectively optimize the learning process and successfully train very deep networks. In particular, we considered the ResNet50 model available from HuggingFace;

3.4.2. ViT

In literature, there are different variants of ViTs. Most of them are available from Hugging Face, Torchvision and Google Research. All pre-trained models are based on the architecture derived by An Image is Worth 16x16 Words [14] article, some of them are detailed in the table 1.

Table 1
Details of Vision Transformer Model Variants.

Model	Layers	Hidden size (D)	MLP size	Heads	Params
ViT-Tiny	12	192	768	3	5.5M
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M

For this work, we decided to consider the ViT_Base_16 as pre-trained model as it is one of the most used.

3.5. Our Custom Models

3.5.1. CNN

The CNN model used in this study follows a standardized architecture to better analyze and evaluate its performance. We implemented the model using PyTorch, a well-known Python library for building neural networks. Using a well-established framework makes it easier to understand the model's structure, making it more transparent and comparable. Our CNN_scratch, represented in Figure 1 has three blocks for feature extraction and one block for classification. Each of the three feature extraction blocks contains two pairs of convolutional and normalization layers, followed by a 2D max pooling layer with a pool size of 2x2. The convolutional layers have a kernel size of 3x3 and use the Relu activation function, with kernel sizes of 32, 64, and 128 for the three blocks, respectively. The classifier, or fully connected layer, processes the flattened output from the previous layers to capture the data's patterns. It includes a dense input layer of 128 neurons and the Relu activation function, followed by a dropout layer with a dropout rate of 0.2. Finally, there is a dense output layer with one neuron for binary classification, using the Sigmoid activation function.

3.5.2. ViT

To implement the ViT, we began with open-source code and made modifications to certain parts and parameters to tailor the model to our needs. We utilized the PyTorch framework, similar to how we did for the CNN, but for this model, we had to create custom functions to execute all the required operations. To begin with, we designed the 'get_patches' function to implement image patching. It requires a list of images (such as a batch or a dataset) and the number of patches to extract. Only square patches from square images are generated, which are then returned as flattened patches. The positional embedding has been implemented using sine and cosine functions. This function takes as input the sequence length, which is the length of the flattened patches, and the number of hidden dimensions. The output is a 2D tensor that contains the positional embeddings of related patches over the hidden connections for each row (patch). One of the most complicated components of ViT is the MSA. Since the purpose of this work is not to explain in detail the model and the associated mathematical relationships, it is possible to delve deeper into [16, 17]. However, we want to point out that the linear mappings of *Q, K, V* vectors are parallelised as many times as the number of hidden dimensions and for each of them, the attention score is evaluated. At the end of this process, the scores of all the dimensions are concatenated. Then we realized the encoder, as shown in Figure 2, by using normalization layers and linear layers to shape the MLP. The last part of ViT is the classification head, which is an MLP classifier that maps the encoder's outputs to the probabilistic distribution of classes. This enables class prediction for the input samples. Finally, we defined our ViT_scratch model with the following hyperparameters: 4 layers, 56 for the hidden size dimension "D", 224 as MLP size and 4 heads for a total of 155,682 parameters.

4. Experiments

In this Section, we describe a comprehensive evaluation of the performance of both CNNs and ViTs, including pre-trained and from-scratch models. To emphasize the variations in the models' performances, we conducted two sets of experiments, one for each test bed, to perform a comparison based on several factors the number of parameters, accuracy, loss, training and inference time. In both experiment sets, we kept the number of epochs fixed at 30 and used a batch size of 16. The tests were initially performed on an Apple MacBook Pro 15 featuring an i7-4770HQ quadcore CPU @ 2.2Ghz and 16Gb of RAM DDR3 and then on an Azure Virtual Machine (VM) with the Windows 10 operating system, equipped with an 8 cores CPU @ 2.5GHz, 56 GB of RAM DDR4 and a Nvidia Tesla T4 Graphics Processing Unit (GPU) with 16 GB of dedicated RAM GDDR6. As for the software side, we used the Nvidia CUDA driver v12.1 to employ the GPU as the processing unit (only for the VM) and Python v3.12.2 as the programming language, with the following libraries: PyTorch v2.2.1 (to implement our custom models), Timm v0.9.16 (to download pre-trained models), Scikit-learn v1.4.1, Numpy v1.26.4 and Pandas v2.2.1. To ensure an impartial and unbiased evaluation process, we set all seed parameters to the same value for all experiments.

Table 2
Considered models comparison.

Model	Type	Params	Learning Mode	Accuracy
From-scratch CNN	CNN	17,6M	FST	73.8%
From-scratch ViT	ViT	155k	FST	70.5%
Pre-trained ResNet	CNN	23,5M	TL	85.9%
Pre-trained ViT Base 16	ViT	86M	TL	76.3%

4.1. Results

We decided to compare both from-scratch and pre-trained models to perform a global comparative analysis. All the models were trained on the considered dataset (discussed in Section 3.3), with fixed epochs and batch size. ResNet50 was pre-trained on the ImageNet-1k [18] dataset (including 1 million images, 1,000 classes), while the ViT model was pre-trained on the ImageNet-21k dataset (14 million images, 21,843 classes) and fine-tuned on the ImageNet-1k. Thus, starting from these models we performed a fine-tuning by training them on the dataset considered in this study. Figures 4 and 5 show comparative graphs of the results obtained respectively considering the test accuracy and loss.

Table 2 summarizes the accuracy results achieved for all experiments. We denoted with FST (From-scratch training) the models we implemented and trained from scratch, and with TL (Transfer Learning) the pre-trained models we fine-tuned. As we can observe from Table 2 and Figure 4, the ResNet stands out as the most accurate model, which highlights the effectiveness of pre-trained models and the transfer learning approach. In addition, considering the test loss in Figure 5, we can see that the ViT_scratch achieved the lowest loss, suggesting that it is more prone to better generalization. However, looking at the CNN_scratch and ViT_base, we can see that they exhibit divergent trends, indicating that the models are overfitting and are unlikely to improve further as the number of epochs increases. The same doesn't apply for the ViT_scratch which appears to have a slowly decreasing trend and the ResNet which, although doesn't find a plateau, looks like having a stable mean value.

Finally, Figures 6 and 7 show the results obtained in terms of training and inference times respectively. In the experiments conducted on the MacBook, the training times for all models were significantly higher than those on the Azure VM, with a ratio of approximately 14:1. This discrepancy in performance is expected due to the limitations of the hardware in handling large computational workloads. Interestingly, while most models followed this pattern, the ViT model trained from scratch, despite being the smallest in terms of parameters, exhibited a unique behavior. On the HPC system, both training and inference times were similar to those of much larger models with significantly more parameters. This suggests that on HPC infrastructure, the efficiency gains from parallelization may add an overhead, which becomes significant when training small ViT models.

5. Conclusion

In this work, we explored various alternative models in the realm of computer vision for healthcare, comparing their effectiveness, efficiency, and performance in different computational environments. The potential of computer vision in healthcare is immense, with applications ranging from medical imaging to diagnostics and patient monitoring. Our exploration included comparing the training process of these models on a consumer laptop and a Cloud-based HPC system. We conducted several experiments to highlight the differences in parameter numbers and processing times of different models. Through our experiments, we identified that Transfer Learning is particularly useful in domains like healthcare,

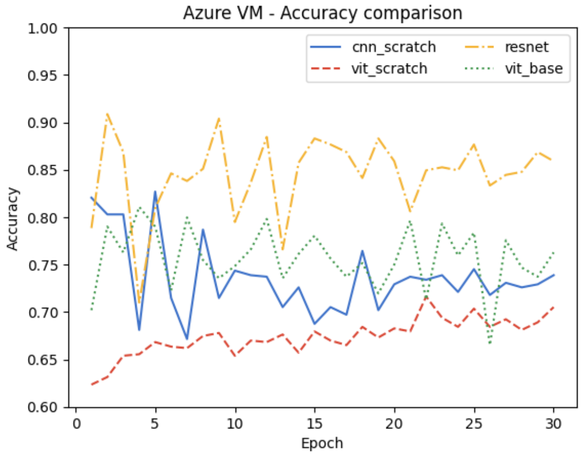


Figure 4: Test Accuracy.

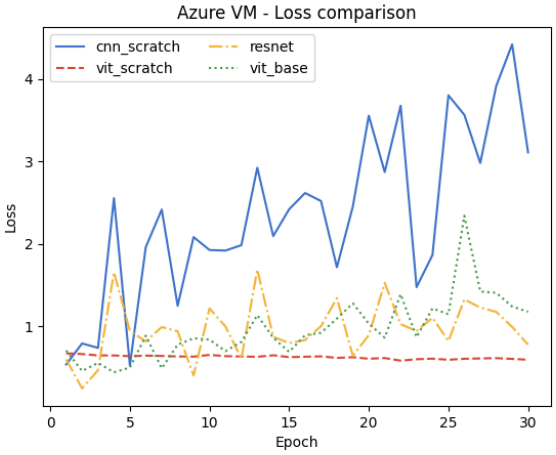


Figure 5: Test Loss.

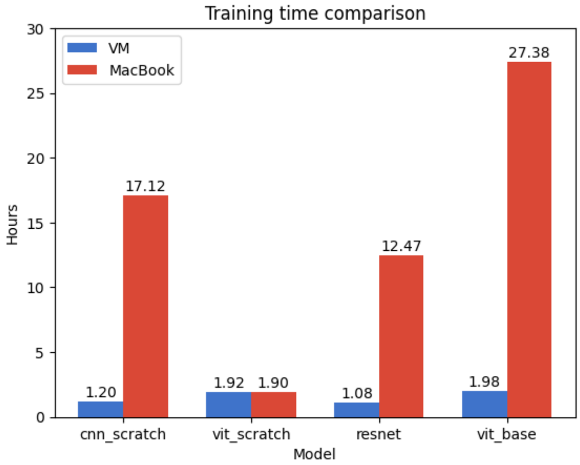


Figure 6: Training Times.

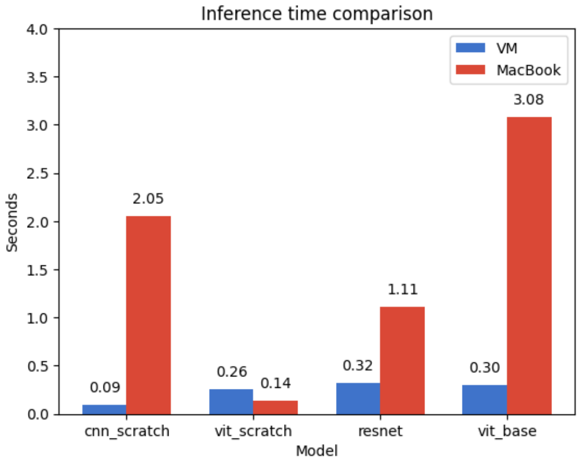


Figure 7: Inference Times.

where data availability is scarce. Utilizing a cloud-based HPC environment significantly accelerates the development process of large deep learning models by providing the necessary computational power and scalability. After reviewing the results, many potential directions for future research have emerged. We will investigate more optimized ViT implementations, add other models to this comparison, and conduct further analysis considering other hyperparameters, devices and HPC services. Furthermore, we will explore alternative approaches such as federated learning to ensure clinical data privacy and parallel processing towards the perspective of classifying Big Medical Data in a collaborative federated healthcare scenario using HPC services.

Acknowledgments

This work has been partially supported by European Union (NextGeneration EU), through the MUR-PNRR project SAMOTHRACE (ECS00000022), the Italian Ministry of Health, Piano Operativo Salute (POS) Trajectory 2 through the project "Rete eHealth: AI e strumenti ICT Innovativi orientati alla Diagnostica Digitale (rAIdD)" (code T2-AN-04 - CUP J43C22000380001), and the Italian PRIN 2022 project Tele-Rehabilitaiton as a Service (TRaaS) (code PRIN_202294473C_001 - CUP J53D23007060006). Giovanni Lonia is a PhD student enrolled in the National PhD in Artificial Intelligence, XL cycle, course on AI for society, organised by Università degli Studi di Pisa. Davide Ciraolo is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVIII cycle, course on Health and life sciences, organised by Università Campus Bio-Medico di Roma.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 (https://openai.com/it-IT/index/gpt-4/) and Grammarly (https://www.grammarly.com/ai) in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Z. Zhang, H. Wu, H. Zhao, Y. Shi, J. Wang, H. Bai, B. Sun, A novel deep learning model for medical image segmentation with convolutional neural network and transformer, Interdisciplinary Sciences: Computational Life Sciences 15 (2023) 663–677.

[2] A. Lin, B. Su, Y. Ning, L. Zhang, Y. He, Convolutional neural networks in medical imaging: A review, in: Advances in Swarm Intelligence (ICSI 2024), volume 14789 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 419–430.

[3] S. Aburass, O. Dorgham, J. A. Shaqsi, M. A. Rumman, O. Al-Kadi, Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases, Journal of Imaging Informatics in Medicine 38 (2025) 1–25.

[4] K. Al-hammuri, F. Gebali, A. Kanan, I. T. Chelvan, Vision transformer architecture and applications in digital health: a tutorial and survey, Visual Computing for Industry, Biomedicine, and Art 6 (2023) 14. URL: https://doi.org/10.1186/s42492-023-00140-9. doi:10.1186/s42492-023-00140-9.

[5] S. S. Kshatri, D. Singh, Convolutional neural network in medical image analysis: A review, Archives of Computational Methods in Engineering 30 (2023) 2793–2810.

[6] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, et al., Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review, Journal of Medical Systems 48 (2024).

[7] M. S. Alam, M. R. Islam, M. M. Hasan, Enhanced convolutional neural networks for early detection and classification of ophthalmic diseases, in: 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), IEEE, 2023.

[8] A. Hasasneh, S. Masri, C. Tadj, Unveiling hidden patterns in infant cry audio: A multi-feature vision transformer approach with explainable ai, IEEE Access (2025).

[9] R. Garcia-Martin, R. Sanchez-Reillo, Vision transformers for vein biometric recognition, IEEE Access (2023).

[10] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, U. Farooq, A survey of the vision transformers and their cnn-transformer based variants, Artificial Intelligence Review 56 (2023) 2917–2970. URL: http://dx.doi.org/10.1007/s10462-023-10595-0. doi:10.1007/s10462-023-10595-0.

[11] M. Lagunas, B. Impata, V. Martinez, V. Fernandez, C. Georgakis, S. Braun, F. Bertrand, Transfer learning for fine-grained classification using semi-supervised learning and visual transformers, 2023. arXiv:2305.10018.

[12] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel, Handwritten digit recognition with a back-propagation network, Advances in neural information processing systems 2 (1989).

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. arXiv:2010.11929.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.

[16] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014. arXiv:1409.0473.

[17] Y. Bazi, L. Bashmal, M. Al Rahhal, R. Dayil, N. Ajlan, Vision transformers for remote sensing image classification, Remote Sensing 13 (2021) 516. doi:10.3390/rs13030516.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.