Vim-Med: a Vision Mamba-based Model for Pathology Classification in X-Ray Images *

Gregory J. Pitthan[1], Lucas B. V. Cordova[1], Tatiana T. Schein[1], Eduardo L. Silva [1], Gustavo A. Dutra[2], Gustavo P. Almeida[1], Stephanie L. Brião[1], Paulo L. J. Drews-Jr[1]

[1]Universidade Federal do Rio Grande, Rio Grande, RS, Brasil
[2]Universidade Federal de Santa Maria, Santa Maria, RS, Brasil

{gustavo.dutra@acad.ufsm.br ,gustavo.pereira.furg, paulodrews}@furg.br

Abstract. *There is a need to improve medical diagnostics in identifying rare diseases and analyzing unbalanced image data. This work presents Vim-Med, an adaptation of the Vision Mamba (Vim) architecture for pathology classification in X-ray images. To evaluate the model, a comparison was made with other Mamba models and Transformer architectures. The results show that in the Chest X-Ray dataset, Vim-Med achieved the best F1-score with* 0.888*. In the NIH CRX8 dataset, Vim-Med excelled at handling rare classes (Macro-F1 of* 0.192*). Vim-Med achieved the highest inference speed, corresponding to 125 FPS, and achieved a reduction of more than 50% in training time. Thus, the Vim-Med model is efficient in classifying pathologies in X-ray images.*

1. Introduction

Deep learning has significantly advanced medical image analysis by improving diagnostic accuracy and automating image interpretation. However, challenges persist, such as handling high-resolution data, limited labeled datasets, and complex object recognition. Vision Transformers (ViTs) [Dosovitskiy and et al. 2021], adapted from natural language processing [Vaswani et al. 2017], capture long-range dependencies using self-attention and have achieved strong performance in medical imaging, but they demand large datasets and are computationally expensive [Liu et al. 2021]. Recently, State Space Models (SSMs), particularly Mamba networks [Gu and Dao 2024], have emerged as an efficient alternative. SSMs capture long-range dependencies with linear complexity, making them well-suited for high-resolution medical images. Architectures like Vision Mamba (Vim) [Zhu et al. 2024] and Mamba-UNet [Wang et al. 2024] have already demonstrated performance competitive with Transformers but with greater efficiency. Motivated by this, we adapt the Vim [Zhu et al. 2024] architecture to medical imaging, defining the resulting model as **Vim-Med**, tailored for pathology classification in chest X-rays. Unlike ViT-based approaches, Vim-Med leverages Mamba's efficiency for medical imaging tasks. We evaluated it on the NIH chest x-ray and pneumonia-normal chest x-ray datasets using precision, recall, F_1-score, and accuracy. Our contributions include the adaptation and evaluation of Vim [Zhu et al. 2024] model customized for medical

*The authors acknowledge the financial support of CNPq, FINEP, FAURG, and FAPERGS. This research was supported by the Human Resources Program of the National Agency of Petroleum, Natural Gas, and Biofuels (PRH/ANP–PRH22.1/FURG). We also acknowledge the support of the São Paulo Research Foundation (FAPESP), Brazil, under grant number 2024/10523-5.

imaging tasks, referred to as Vim-Med, an analysis of the sensitivity-specificity trade-off, an assessment of its performance under class imbalance, and the use of training optimizations, such as mixed precision and gradient accumulation, to enhance model performance.

2. Methodology for Vim-Med

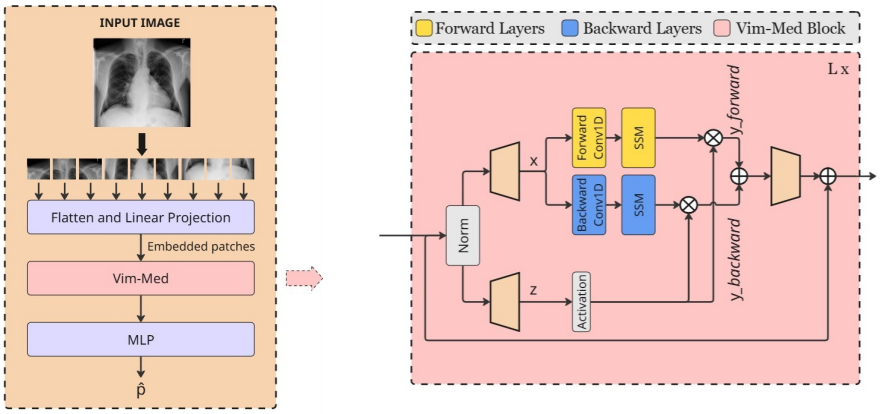
This work evaluates Mamba-based architectures for medical image classification, focusing on traditional performance metrics and the clinically important trade-off between sensitivity and specificity. We also analyze the Vim-Med ability to handle class imbalance, a common challenge in medical datasets. This section outlines the formulation of the Vim-Med proposal. The following state-space model represents Mamba networks:

$$h_t = \overline{\mathbf{A}} h_{t-1} + \overline{\mathbf{B}} \mathbf{x}_t \tag{1a}$$

$$y_t = \overline{\mathbf{C}} h_t, \tag{1b}$$

where $\overline{\mathbf{A}} \in \mathbb{R}^{N \times N}$, $\overline{\mathbf{B}} \in \mathbb{R}^{N \times 1}$, and $\overline{\mathbf{C}} \in \mathbb{R}^{1 \times N}$ are discrete parameters obtained via Zero-Order Hold (ZOH), defined as $\overline{\mathbf{A}} = e^{\Delta \mathbf{A}}$, $\overline{\mathbf{B}} = (\Delta \mathbf{A})^{-1} \left(e^{\Delta \mathbf{A}} - \mathbf{I}\right) \Delta \mathbf{B}$, and $\overline{\mathbf{C}} = \mathbf{C}$, for Δ denoting the timescale factor and N the state dimension. The main idea of Mamba is to process an input sequence of tokens $\{\mathbf{x}_t\}_{t=1}^{N}$, for each $\mathbf{x}_t \in \mathbb{R}^D$, mapping them to outputs y_t via an implicit latent state h_t, using $\mathbf{B}$, $\mathbf{C}$, and Δ as linear projections of the input $\mathbf{x}_t$ into N, parameterized to select relevant information. Mamba networks [Zhu et al. 2024], map a sequence of inputs $\mathbf{x}_t$ to outputs y_t through a latent state h_t. The core innovation of Mamba is its selective mechanism, allowing the model parameters to be input-dependent. This enables Mamba to capture long-range dependencies with linear complexity, serving as an alternative to Transformers. Vim [Zhu et al. 2024] is a recent architecture that utilizes a bidirectional state-space based on Mamba (see Fig. 1), aiming to enhance inference speed and memory efficiency, particularly for long input sequences. First, similar to ViT, Vim transforms the input image $\mathbf{t}$ into flattened 2D patches, projects each patch linearly into a D-dimensional embedding $\{\mathbf{t}_p^i\}_{i=1}^{N}$, and adds positional encodings. Additionally, a special classification token is used to summarize the entire patch sequence. In Fig. 1b, the token sequence $\mathbf{t}_p^i$ passes through a normalization layer, followed by a linear projection to obtain $\mathbf{x}$ and $\mathbf{z}$. For forward and backward directions, a 1D convolution is applied to $\mathbf{x}$ to produce an intermediate representation $\mathbf{x}_0'$, which is then linearly projected to $\overline{\mathbf{B}}$, $\overline{\mathbf{C}}$, and Δ. The outputs y_{forward} and y_{backward} are then computed using the State Space model (SSM), gated by $\mathbf{z}$, and summed to produce the final output token sequence. By default, the number of blocks L is set to 24 and the SSM dimension N to 16. Finally, the output classification token is normalized and passed through a MLP head to produce the final prediction $\hat{p}$. The Vim-Med approach is based on Vim [Zhu et al. 2024] and uses the bidirectional layers to capture global visual context in a data-dependent manner.

The characteristics of each dataset guided the choice of loss functions. For binary classification tasks, such as distinguishing between pneumonia and normal chest x-ray images, we employed the Binary Cross-Entropy Loss Function $\mathcal{L}_{\text{BCE}}$ [Goodfellow et al. 2016]. For the NIH Chest X-ray dataset (NIH CRX8), a multi-label classification problem where multiple thoracic pathologies may co-occur, we adopted the function $\mathcal{L}_{\text{BCE}}$ with Logits Loss. This function applies the sigmoid internally and computes the binary cross-entropy per class. For labels C, with ground truth vector



(a) General Model **(b)** Vim-Med block

Figure 1. Vim-Med overview [Zhu et al. 2024]. (a) Model, from input patches to final prediction. (b) Vim-Med block with processing (Forward/Backward).

$\mathbf{y} \in \{0,1\}^C$ and predicted logits $\mathbf{z} \in \mathbb{R}^C$, the loss is:

$$\mathcal{L}_{\text{BCEwL}} = -\sum_{i=1}^{C} \left[y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i)) \right]. \tag{2}$$

Model performance was assessed using four standard metrics: *accuracy*, *precision*, *recall*, and *F_1-score*. To provide a robust evaluation, especially for unbalanced datasets, we calculated these metrics using two averaging schemes: *micro-averaging*, which aggregates results across all samples, and *macro-averaging*, which calculates the metric independently for each class and then averages the results across classes. This dual approach offers insight into both overall and class-specific performance. To improve efficiency and stability during training, we applied mixed precision and gradient accumulation. Mixed precision leveraged FP16 operations with automatic loss scaling, reducing memory usage and accelerating computation. Gradient accumulation simulated larger batch sizes by aggregating gradients over k mini-batches before each optimizer update, enhancing convergence without exceeding GPU limits.

3. Experimental Results

Two public datasets were used: the multi-label **NIH Chest X-ray** and the binary classification **Pneumonia-Normal Chest X-ray**. All images were preprocessed to a size of 224×224 pixels, normalized, and augmented. Inference efficiency of ViT, Mamba-UNet, and Vim-Med was evaluated on an RTX 4060 GPU setup. As shown in Table 2, Vim-Med achieved the lowest latency on both datasets, followed by Mamba-UNet, highlighting the efficiency of Mamba-based designs. Table 1 details per-class results on NIH CRX8, where Vim-Med consistently obtained the highest recall, particularly for underrepresented classes such as *Mass* and *Nodule*. This sensitivity to rare conditions is clinically relevant, while Mamba-UNet tended to achieve higher Precision, indicating a trade-off between sensitivity and specificity. Aggregated results are summarized in Table 2. On NIH CRX8, Mamba-UNet reached the best frequency-weighted performance with Micro-F1 of **0.396**, while Vim-Med excelled under class imbalance with a Macro-F1 of **0.192** and Recall of **0.672**. ViT provided balanced but lower results across metrics. On the Chest

Table 1. Class-wise comparison of model performance on the NIH CRX8.

Class	Sample	Mamba UNet			Vim-Med (Ours)			Vision Transformer		
		F1-Score ↑	Precision ↑	Recall ↑	F1-Score ↑	Precision ↑	Recall ↑	F1-Score ↑	Precision ↑	Recall ↑
No Finding	60361	**0.5579**	**0.3888**	0.9875	0.5562	0.3853	**1.0000**	0.4562	0.3253	0.8930
Atelectasis	11559	0.2661	**0.2330**	0.3102	**0.2867**	0.1777	**0.7411**	0.2136	0.1424	0.7220
Cardiomegaly	10776	0.0302	**0.3091**	0.0159	**0.2024**	0.2130	**0.1927**	**0.0775**	0.0773	**0.0496**
Effusion	13317	**0.4071**	**0.3256**	0.5432	0.3980	0.2590	**0.8585**	0.2947	0.2473	0.8160
Infiltration	19894	**0.4315**	**0.3326**	0.6139	0.3979	0.2513	**0.9553**	0.3150	0.2511	0.7542
Mass	5782	0.0554	**0.1545**	0.0338	**0.1630**	0.1139	**0.2860**	0.0323	0.1074	0.1808
Nodule	6331	0.0012	0.1250	0.0006	**0.1003**	0.0875	**0.1177**	0.0827	0.1032	0.0690
Pneumothorax	5302	0.0140	**0.3393**	0.0071	**0.2377**	0.2481	**0.2281**	0.2254	0.2233	0.1917
Consolidation	4667	0.0464	0.1332	0.0281	**0.2021**	**0.1335**	**0.4154**	0.1906	0.1325	0.3399
Edema	2303	0.1275	**0.1375**	0.1189	**0.1891**	0.1224	**0.4151**	0.1732	0.1182	0.3243
Emphysema	2516	0.0000	0.0000	0.0000	**0.0565**	**0.0860**	**0.0421**	**0.0131**	**0.0615**	**0.0073**
Pleural Thickening	3385	0.0051	0.0732	0.0026	**0.0828**	**0.0883**	**0.0779**	**0.0591**	**0.0998**	**0.0420**

X-Ray dataset, Vim-Med clearly outperformed both baselines, leading across all metrics. Overall, Vim-Med demonstrates superior Recall and robust performance on both datasets, suggesting its potential for improving diagnostic sensitivity in medical imaging tasks.

Table 2. Metric results and inference speed (FPS) on NIH CRX8 and Chest X-Ray.

Model	NIH CRX8					Chest X-Ray				
	Micro-F1↑	Macro-F1↑	Micro-Pr↑	Micro-Re↑	FPS↑	Accuracy↑	F1-Score↑	Precision↑	Recall↑	FPS↑
Mamba-UNet	**0.396**	0.129	**0.345**	0.467	83.3	0.849	0.844	0.856	0.849	100.0
Vision Transformer	0.372	0.173	**0.261**	0.648	71.4	0.785	0.762	0.824	0.785	76.9
Vim-Med (Ours)	0.368	**0.192**	0.253	**0.672**	**111.1**	**0.889**	**0.888**	**0.891**	**0.889**	**125.0**

4. Conclusion

This study compares our proposed model, Vim-Med, with Mamba-UNet and ViT for medical image classification on the chest X-ray and NIH CRX8 datasets. The results show that Vim-Med outperforms in handling unbalanced data and detecting rare conditions, achieving the best macro-F_1 (0.192) and recall (0.672), while also being the fastest, with a frame rate of 125 FPS. Mamba-UNet excelled in precision, obtaining the highest micro-F_1 (0.396) and micro-precision (0.345), though with a lower recall, evidencing the trade-off between sensitivity and specificity. Future directions include data balance, cross-validation, and evaluation in other medical imaging modalities.

References

Dosovitskiy, A. and et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gu, A. and Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Liu et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF ICCV*.

Vaswani et al. (2017). Attention is all you need. *Advances in NeurIPS*, 30.

Wang, Z. et al. (2024). Mamba-unet: Unet-like pure visual mamba for medical image segmentation.

Zhu et al. (2024). Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st ICML*.