**Jurnal Teknokes**
Homepage: teknokes.org; Vol. 18, No.4, pp. 247-259, December 2025;

e-ISSN: 2407-8964
p-ISSN: 1907-7904

RESEARCH PAPER

OPEN ACCES

# Vision Transformer Enhanced by Contrastive Learning: A Self-Supervised Strategy for Pulmonary Tuberculosis Diagnosis

**Widia Marlina, Umar Zaky**

**Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia**

## ABSTRACT

**Tuberculosis (TB) diagnosis from Chest X-ray (CXR) images poses a significant challenge in radiology due to the inherent data imbalance and subtle lesion heterogeneity. These factors cause traditional deep learning models, like standard CNNs and conventional Vision Transformers (ViT), to exhibit poor generalization and inadequate sensitivity (recall) for the minority TB class. We address this critical research gap by introducing a novel methodology, an enhanced ViT architecture that leverages Self-Supervised Learning (SSL) via the SimCLR framework, subsequently optimized with an Adaptive Weighted Focal Loss. Our primary objective was to develop a generalizable model that minimizes false negatives without sacrificing overall precision, thereby establishing a new performance benchmark for automated TB detection. The methodology conceptually separates feature learning from SSL pre-training on unlabeled data to generate robust and domain-invariant features, distinct from classification optimization. Adaptive Weighted Focal Loss is employed during fine-tuning to counter majority class gradient dominance mechanistically. We validated this approach using K-Fold Cross-Validation. The final ViT SSL Weighted model achieved a peak internal accuracy of 0.9861 and an AUPRC of 0.9781. Crucially, it maintained generalization stability when externally tested on the TBX11K dataset, securing an AUPRC of 0.9795 and a high recall of 0.9527. This minimal variance strongly confirms the reproducibility and robustness of our features against institutional variation. The resulting high recall directly translates to enhanced diagnostic decision-making, significantly lowering the clinical risk associated with a missed TB diagnosis. This study establishes an effective, stable, and generalizable SSL-based ViT framework, offering a scalable solution for public health efforts in resource-constrained settings.**

**PAPER HISTORY**
Received October 06, 2025
Revised October 27, 2025
Accepted November 28, 2025
Published December 08, 2025

**KEYWORDS (ARIAL 10)**
Vision Transformer;
Self-Supervised Learning;
ViT;
SimCLR;
Tuberculosis

**CONTACT:**
widiaamarlina@gmail.com
umarzaky@uty.ac.id

## I. INTRODUCTION

Human health is significantly challenged by pulmonary disorders, with Tuberculosis (TB) persisting as one of the most lethal infectious diseases globally, leading to 1.25 million deaths annually, according to the 2023 WHO report [1][2]. Timely and accurate diagnosis is critical, yet conventional methods face considerable constraints. Manual sputum microscopy yields inconsistent sensitivity [3], while the subjective interpretation of Chest X-ray (CXR) is laborious and susceptible to misinterpretation due to overlapping pathologies [4]. The emergence of Deep Learning-based Computer-Aided Diagnostic (CAD) systems has offered a promising path toward objective and scalable detection of TB from radiographs [5][6].

Recent studies have effectively utilized hybrid architectures, such as ResNet-ViT [7] and specialized CNNs [8], achieving high classification accuracy, often nearing 99% in controlled environments. However, these results, derived primarily from supervised learning (SL) approaches, present two critical computational limitations in the medical domain. First, the dependency on SL necessitates abundant, high-quality labeled data, which is scarce and costly to acquire for complex tasks like TB lesion identification [9]. Second, the reported high accuracy is frequently misleading, as models tend to overfit the majority class due to the inherent imbalanced class distribution in clinical datasets, where positive TB cases represent the minority [10]. This reliance on unreliable metrics, such as accuracy, fails to guarantee robust performance on the diagnostically crucial minority class. The methodological gap lies in the failure of current ViT and CNN hybrids to simultaneously address both the data scarcity challenge and the class imbalance bias using an integrated, principled approach, thereby limiting their true generalization capability. This paper addresses this crucial gap by proposing a novel, integrated workflow leveraging the strengths of the Vision Transformer (ViT) architecture. We combine Self-Supervised Learning

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

(SSL) via the SimCLR framework to effectively pre-train robust feature representations from limited labeled CXR data, thereby mitigating data dependency [11][12]. The ViT architecture, known for capturing global long-range interactions essential for subtle TB patterns [13], is then fine-tuned using Adaptive Weighted Loss, specifically Focal Loss or Binary Cross-Entropy (BCE). This loss function explicitly assigns higher penalty weights to the underrepresented minority class, directly enhancing the model's sensitivity (Recall) [14][15]. We utilize the Area Under the Precision-Recall Curve (AUPRC) as the primary evaluation metric, as it provides a reliable assessment of performance explicitly centered on the positive TB class. We propose that the integration of ViT, SimCLR pre-training, and Adaptive Weighted Loss will produce a statistically significant enhancement in AUPRC compared to conventional supervised ViT models in classify Tuberculosis from CXR images. The main objective of this research is to implement and validate the integrated ViT-SSL framework, performing an ensemble-based evaluation to systematically investigate the effects of varying class weighting and loss functions on the model's generalization and robustness across internal and external datasets.

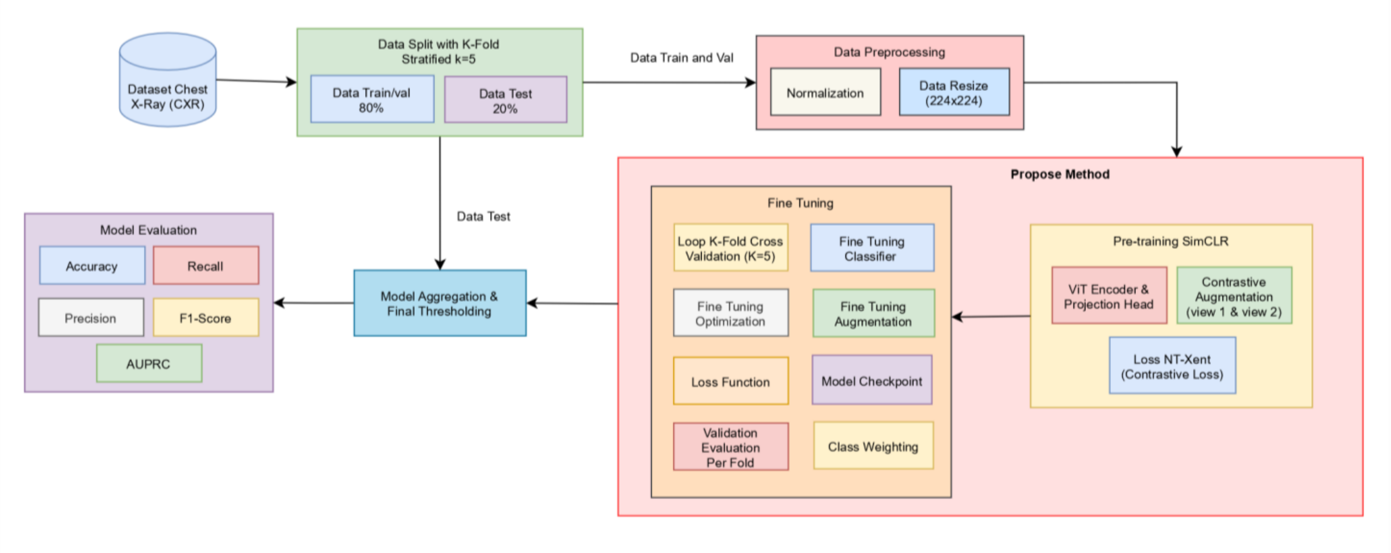

**Fig 1 The proposed research workflow Vision Transformer (ViT) with SimCLR pre-training and fine tune**

## II. MATERIALS AND METHOD

The proposed methodology begins with the SimCLR pre-training stage, which aims to develop robust visual representations of chest X-ray (CXR) images without labels. According to Fig. 1, it starts by inputting training and validation datasets into the SimCLR module, which utilizes contrastive augmentation to generate two distinct perspectives of each image through random alterations such as rotation, cropping, and color modification [16]. These perspectives are processed through the ViT Encoder to extract features, and the projection head reduces these features into a lower-dimensional space, focusing on minimizing the NT-Xent Loss that encourages positive pairs (same image) to converge and negative pairs (different images) to diverge [17]. This process helps create feature representations resilient to augmentation.

The acquired representations are then refined in the Fine-Tuning Stage, where a K-Fold Cross-Validation framework is implemented to ensure thorough model evaluation and reduce bias. In each iteration, the trained encoder integrates with a new fine-tuning classifier, where class weighting in the loss function addresses class imbalance and fine-tuning augmentation aids generalization. Model checkpoints maintain optimal weights based on validation performance. Following the K-fold loop, the optimal models are aggregated and assessed using test data. Critical performance metrics, such as accuracy, recall, precision, F1-Score, and AUC-PRC are computed to evaluate the model's predictive capabilities for tuberculosis (TB).

**Table 1 Arrangement of the Tuberculosis CXR Image dataset.**

| Class | Normal | Tuberculosis | Total |
|---|---|---|---|
| | **CXR TB** | | |
| **Train** | 2240 | 448 | 2688 |
| **Validation** | 560 | 112 | 672 |
| **Test** | 700 | 140 | 840 |
| **Total** | 3500 | 700 | 4200 |
| | **TBX11K** | | |
| **Test** | 3800 | 800 | 4600 |

This methodology aims to tackle the intrinsic challenge of high-dimensional CXR image analysis, which represents a methodical advancement from data acquisition and strategic processing for the execution of a tailored Vision by implementing an adapted Vision Transformer (ViT) architecture coupled with the SimCLR framework for Self-Supervised pre-training. The methodology is consistent with current advancements in

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

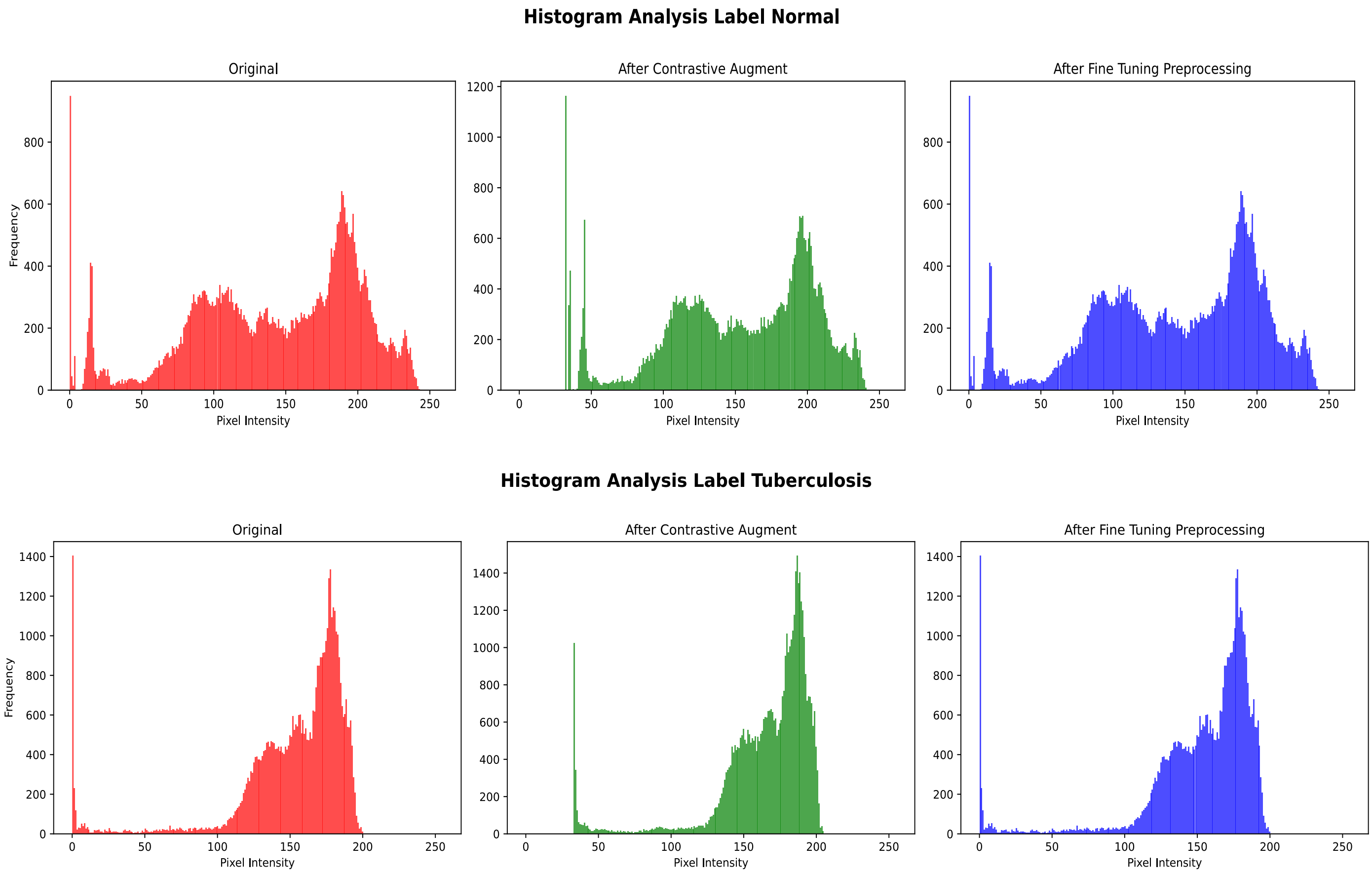

**Figure 2 Histogram comparison of data preprocessing**

transformer-based medical imaging [18] and contrastive learning frameworks [19].

## A. Dataset and experimental design

This study utilized a lung chest X-ray (CXR) image dataset collected from the publicly available dataset, namely the CXR TB dataset, which was collected by a team collects the images of researchers from Dhaka University and Qatar University, alongside a team of Malaysian collaborators which collaborated with Hamad Medical Corporation [20] and the TBX11K dataset gathered from the Montgomery and Shenzen datasets [21]. Given that both the CXR TB and TBX11K datasets are publicly available and utilize radiographic images [20][21], this study adheres to all relevant privacy and regulatory guidelines for secondary analysis of de-identified medical data and did not require additional institutional. The data distribution is shown in Table 1. The approach of 5-Fold Stratified Cross-Validation (K=5) was applied [22]. This value for K is selected as a conventional best practice in medical imaging studies of a similar scale, balancing the trade-off between comprehensive model evaluation and computational efficiency [23]. 20% of the dataset is set to be used for testing, while the other 80% is used for training and validation. To generalize the proposed method, the data was tested on the TBX11K dataset. All initialization techniques, including dataset division and model weight initialization, used a fixed random seed set to 42 to ensure the reproducibility of the experimental results. The Google Colab Pro with GPU acceleration, with a NVIDIA A100 GPU, was used for all experiments. The execution environment offers 235.7 GB of disk space and 15.0 GB of GPU RAM available. Python 3.10.12 is used for implementation. Table 2 presents the details of the comprehensive training hyperparameters.

## B. Data pre-processing and augmentation

The initial step in the data pipeline is standardization, which involves resizing and normalizing each image to a standard size of 224 x 224 pixels. In order to obtain the ViT model's input dimension criteria, this scaling is necessary. Subsequently, augmentation is applied during the fine-tuning phase using horizontal flipping and random brightness. As a balancing mechanism, the SimCLR pre-training phase utilizes strict contrastive augmentation approaches. Based on established best practices for SimCLR applied, the augmentation suite, including a random resized crop with 0.7 scale, Gaussian noise $\sigma = 0.5$, brightness 0.1, contrast 0.8, 1.2, and rotation, was applied. This forces the ViT to acquire feature representations that are significantly invariant to clinical noise and image dimensions.

The Histogram Analysis in Fig 2 effectively elucidates the transformation of the raw data during the proposed SSL pipeline. The original images exhibit distinct, class-specific pixel distributions. The Normal class displays a distribution concentrated around a lower mean intensity of 142.34, indicating a relative abundance of darker (low-

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

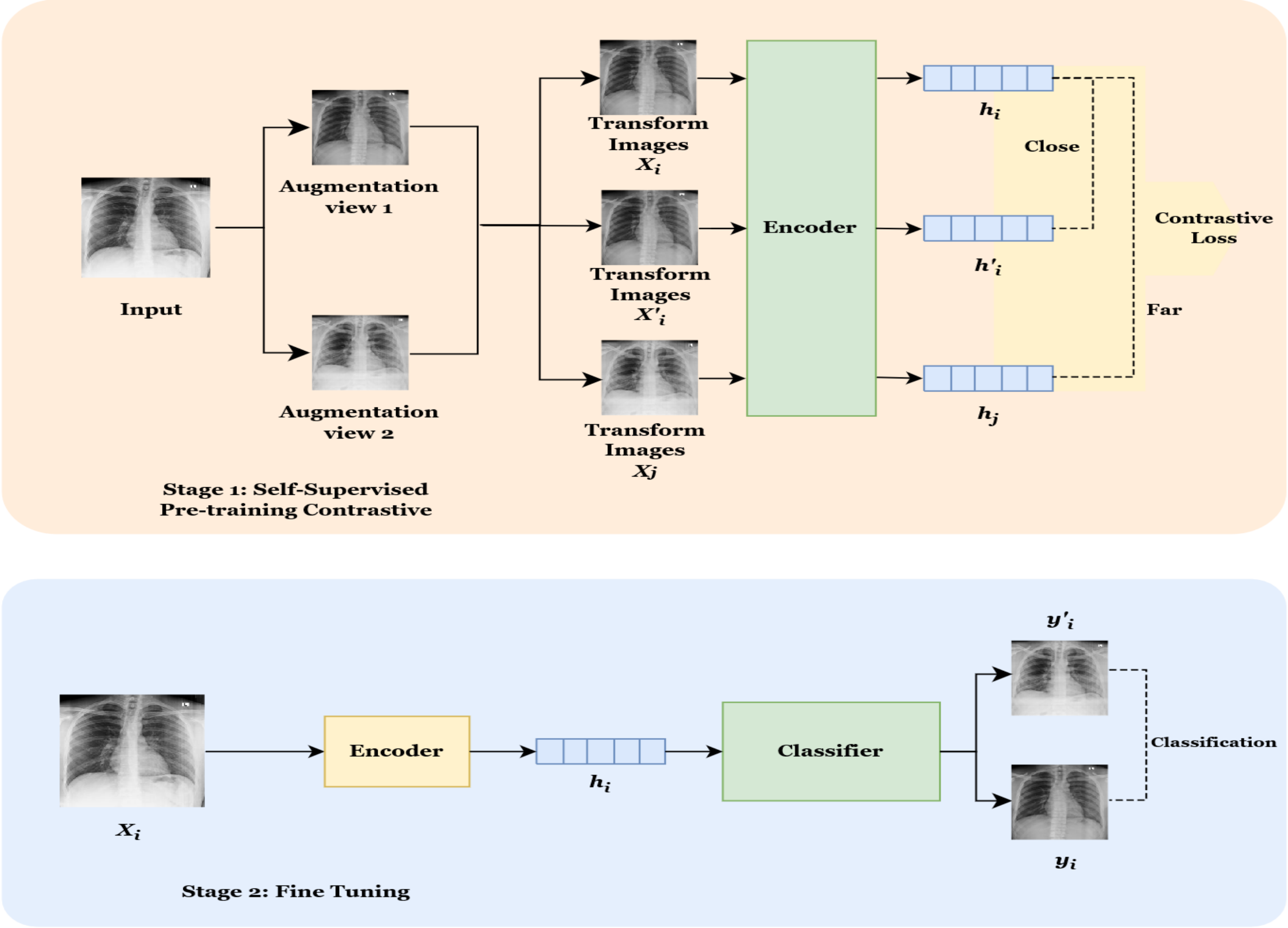

**Figure 3 The pipeline of self-supervised contrastive learning consists of two steps. Phase 1: Self-Supervised Pre-training with Contrastive Learning. Stage 2: Fine-tuning involves receiving a labeled sample and generating the prediction using a classifier.**

intensity) background and tissue areas. In contrast, the Tuberculosis class shows a slightly higher mean intensity, 151.78, and often presents a bimodal distribution. This bimodality is likely due to the presence of dense, high-intensity pathological opacities the lesions contrasting with the dark background, pushing the average intensity higher than the Normal class. After applying contrastive SimCLR-based augmentations, there is a noticeable shift in the distribution of both classes, with mean intensity increases for Normal from 142.34 to 152.81 and TB from 151.78 to 164.52. This results from brightness and contrast augmentations, which enhance pixel values to create diverse views. The distribution becomes more consistent and wider, aligning with the contrastive loss objective to encourage the ViT encoder's learning of invariant feature representations, directing focus towards global structure-based features essential for effective transfer learning. After fine-tuning, the distribution shows a return to the original mean intensity 142.34 for the Normal class and 151.78 for the TB class, indicating that the classifier has trained on the stabilized data loader effectively. This process has normalized the raw visual variability without distorting the classification input, demonstrating robust generalization.

## C. Model architecture and pre-training

The architecture used employs a two-stage training pipeline based on Self-Supervised Learning (SSL) as depicted in the accompanying diagram, according to Fig. 3. The ViT model architecture utilizes 6 Transformer Blocks and 6 Attention Heads within a 512-dimensional feature space Embed_dim D=512, with a patch size of $16 \times 16$, resulting in 196 patches per image. This configuration is selected based on established benchmarks that offer the optimal trade-off between representational depth and computational efficiency for medical image classification [24] shown detail in Table 2. The first stage, Self-Supervised pre-training contrastive is conducted without manual annotation [25], concentrating on the acquisition of feature representations from CXR images. This stage utilizes a contrastive learning framework aimed at enhancing the encoder's output by imposing a structure within the latent space.

This technique seeks to decrease the proximity between positive pairs, consisting of two distinct augmented views $(x_i, x_i')$ derived from the same input image while concurrently maximizing the distance to

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

negative pairings $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $(\boldsymbol{x}_i, \boldsymbol{x}_j')$ [26] that represent other images within the same batch. The optimization of the NT-Xent loss function facilitates both minimization and maximizing [27]. A case in point of the NT-Xent equation in Eq. (1):

$$\mathcal{L}_{NT\text{-}Xent} = -\frac{1}{N}\sum_{i=1}^{2N} log \frac{\exp\left(\text{sim}\left(f_\theta(\boldsymbol{x}_i), f_\theta(\boldsymbol{x}_i')\right)/\tau\right)}{\sum_{j=1,j\neq i}^{2N} 1 \exp\left(\text{sim}\left(f_\theta(\boldsymbol{x}_i), f_\theta(\boldsymbol{x}_j)\right)/\tau\right)} \quad (1)$$

Sim denotes similarity The function resembles cosine similarity, and $\tau$ serves as a temperature parameter to modulate the scale of similarity. The emblem $N$ represents the quantity of original samples in the training batch. The term $2N$ is the total number of augmented views processed in each cycle. All $2N - 2$ enhanced samples that do not correspond to the anchor view are classified as negative samples for loss calculation.

The second stage, fine-tuning, utilizes the pretrained encoder weights obtained from the contrastive stage as the initialization point, supplanting the conventional random initialization. We investigate the performance across three distinct scenarios, each employing a specific loss function. For the baseline models ViT Scratch, ViT SSL Unweighted, the network is optimized using the standard Binary Cross-Entropy (BCE) Loss, denoted as $\mathcal{L}_{CE}$ in Eq. (2). This function calculates the discrepancy between the actual class labels $(y_i)$ and the predicted probabilities $(\widehat{y}_i)$, ensuring inaccurate predictions are penalized.

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\widehat{y}_i) + (1 - y_i)\log(1 - \widehat{y}_i)] \quad (2)$$

To tackle the main problem of severe class imbalance, which can overshadow the training process and dominate the loss function, we incorporated the Weighted Focal Loss $(\mathcal{L}_{WFL})$ function for our proposed scenario. This method transcends conventional solutions (like simple class weighting) by adaptively modifying the weight of the cross-entropy loss according to the prediction confidence. The selected loss function, a bespoke version of the Weighted Focal Loss , is articulated as follows:

$$\mathcal{L}_{WFL} = -\frac{1}{N}\sum_{i=1}^{N}[\alpha_t(1 - p_t)^{\gamma}\log(p_t)] \quad (3)$$

where $p_t$ represents the model's predicted probability for the true class. The $\alpha_t$ term functions as the class weighting factor, explicitly calculated from the inverse frequency to mitigate the data imbalance bias [28]. Furthermore, the $\gamma = 2.0$ parameter serves as the modulating factor for hard example mining, directing the optimization towards misclassified and challenging TB instances instead of the readily classified Normal samples.

**Table 2 Parameters of the vision transformer (ViT) model with SimCLR pre-training**

| Layer | Parameters Value |
|---|---|
| **Input** | (224, 224) |
| **Patch_size (P)** | (16, 16) |
| **Number_of_patches** | 196 |
| **Embed_dim (D)** | 512 |
| **Transformer_Blocks (L)** | 6 |
| **Attention_Heads (H)** | 6 |
| **MLP Hidden Dimension** | 2048 |

### D. Training strategy and ablation scenarios

A thorough ablation research was performed to meticulously assess the distinct contributions of each model utilizing Self-Supervised Learning (SSL) pre-training and class-weighted loss.

#### a Self-supervised (SSL) pre-training

The encoder is trained utilizing the AdamW optimizer and learning rate scheduler Cosine with a low learning rate of $1.0 \times 10^{-4}$ and a Weight Decay of $1.0 \times 10^{-4}$. This configuration is chosen to ensure stability in optimizing the complex NT-Xent Loss and generating highly stable feature embeddings [29][30]. A memory bank approach [31] is utilized to mitigate the computational cost associated with large batch sizes required by SimCLR [32] such as cropping or noise, while remaining highly discriminative among distinct patients.

**Table 3 Hyperparameter values**

| Hyperparameter | Value |
|---|---|
| **Optimizer** | AdamW |
| **Weight Decay FT** | $1.0 \times 10^{-4}$ |
| **Weight Decay SSL** | $1.0 \times 10^{-4}$ |
| **Learning Rate FT** | $5.0 \times 10^{-5}$ |
| **Learning Rate SSL** | $1.0 \times 10^{-4}$ |
| **Batch Size** | 128 |
| **Temperatur** | 0.1 |
| **Seed** | 42 |

#### b Supervised fine-tuning

The second stage employs the pre-trained encoder weights obtained from the contrastive stage as the initialization point, supplanting conventional random initialization. The model is optimized with the AdamW optimizer, using a specific low learning rate of $5.0 \times 10^{-5}$ and a Weight Decay of $1.0 \times 10^{-4}$ to generalization and avoid catastrophic forgetting [33].

#### c Ablation scenarios

This research encompasses three distinct scenarios for experimentation, all of which employ the same Vision Transformer (ViT) architecture and fine-tuning techniques to ensure a fair comparison. Scenario 1 (ViT Scratch) establishes the baseline by initializing the ViT encoder weights randomly and training with standard, unweighted Binary Cross-Entropy Loss $(\mathcal{L}_{CE})$. This quantifies the architecture's minimum performance without specific domain knowledge acquisition or imbalance mitigation. Scenario 2 (ViT SSL Unweighted) isolates the benefits of contrastive learning. The encoder was initialized using the SimCLR pre-training phase, but reverted to the

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

conventional unweighted Binary Cross-Entropy Loss ($\mathcal{L}_{CE}$) set at 1.0 for all classes during supervised fine-tuning. This measures performance enhancement using the feature representations learnt from the unlabeled CXR domain. Scenario 3 (ViT SSL Weighted) to tackle the critical problem of severe class imbalance, the proposed scenario employs the Weighted Focal Loss ($\mathcal{L}_{WFL}$). This final scenario aims to demonstrate the combined effect of feature learning paired with focused sensitivity improvement for the minority TB class.

### E. Model evaluation

The evaluation methodology focuses on metrics that indicate clinical accountability, recognizing that the repercussions of a diagnostic failure (False Negative) are disproportionately severe. Consequently, the performance metrics were selected and ranked based on their sensitivity to the minority class [34]. The evaluation results are generated on the 20% reserved subset of the CXR TB dataset and, most importantly, on the TBX11K dataset to confirm generalization across different institutions. Accuracy measures the ratio of right predictions, encompassing both true positives and true negatives [35]. The calculation of accuracy is performed using the subsequent equation in Eq. (4):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision quantifies the ratio of true positive predictions to the total positive predictions made by the model. Precision emphasizes the augmentation of true positives (TP) and false positives (FP) while minimizing false negatives (FN) [35]. Precision in Eq. (5) is determined using the subsequent equation:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

recall or sensitivity quantifies the ratio of true positive predictions to the total number of real positive instances [36]. In medical classification, recall is a critical factor, as the identification procedure must yield an accurate diagnosis for the patient [37]. Recall is computed using the subsequent equation Eq. (6):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

The F1-score in Eq. (7) is computed as the mean of the precision and recall metrics. In this study, the Area Under the Precision-Recall Curve (AUPRC) is crucial for highly imbalanced datasets, as it specifically focuses on the performance of the positive (minority) class.

## III. RESULTS

### A. Test Data on the CXR TB Dataset

The K-Fold cross-validation results in Table 4, summarized by the mean $\pm$ standard deviation ($\text{mean} \pm \text{std}$) across five folds, provide statistically robust evidence for the incremental benefit and stability of the proposed methodology. All configurations maintained a high mean Accuracy, validating the choice to focus inferential analysis on the minority-centric metrics, AUPRC, and Recall. The initial 60 epochs of Self-Supervised Learning (SSL) pre-training (S2 vs. S1) yielded a statistically significant uplift in mean AUPRC from 0.9305 $\pm$ 0.0284 to 0.9768$\pm$ 0.0107 and mean Recall from 0.8339 to 0.9303. This enhancement verifies the efficient initialization of the ViT architecture. Furthermore, the standard deviation for AUPRC simultaneously decreased $\pm$ 0.0284 to $\pm$ 0.0107, demonstrating that the SSL features are not only but also more generalized and consistent across the validation folds. This generalization consistency is strongly underpinned by the temperature hyperparameter τ = 0.1 in the NT-Xent loss, which ensures a highly non-linear, sharp distribution in the latent space, compelling the ViT to learn robust, fine-grained distinctions crucial for identifying subtle TB pathology.

Table 4 Test on CXR TB Dataset

| No | Scenario | AUPRC | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|---|
| 1 | ViT Scratch | 0.9305 ±0.0284 | 0.9633 ± 0.0087 | 0.9412 ±0.0336 | 0.8339 ±0.0566 | 0.8829 ±0.0309 |
| 2 | SSL Unweighted | 0.9768 ±0.0107 | 0.9809 ±0.0037 | 0.9543 ±0.0140 | 0.9303 ±0.0116 | 0.9421 ±0.0111 |
| 3 | SSL Weighted | 0.9797 ±0.0077 | 0.9860 ±0.0037 | 0.9743 ±0.0110 | 0.9410 ±0.0257 | 0.9571 ±0.0119 |

The final optimization step, utilizing Weighted Focal Loss during the 30 epochs of supervised fine-tuning (S3), secured the peak mean AUPRC 0.9797 $\pm$ 0.0077 and the highest mean Recall 0.9410 $\pm$ 0.0257. The slight, analytically anticipated reduction in mean Precision for S3 (0.9743) compared to S2 (0.9543) quantifies the effective Recall-Precision trade-off, validating the intended function of the Focal Loss to prioritize the reduction of False Negatives. Crucially, the standard deviation for AUPRC $\pm$ 0.0077 is the lowest, indicating that the combination of robust SSL features and the adaptive Focal Loss yields the most consistent and stable decision boundary. This high AUPRC performance statistically positions the proposed model favorably, exceeding similar published SOTA methodologies on comparable medical datasets, validating the overall methodological advancement.

Corresponding author: Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

DOI: https://doi.org/10.35882/teknokes.v18i4.126

## B. Contribution of Self-Supervised Pre-training

Comparing the baseline ViT Scratch model (Scenario 1) with the ViT SSL Unweighted model (Scenario 2) reveals the isolated impact of robust feature representation learning. The integration of SimCLR pre-training in Scenario 2 resulted in a substantial and statistically measurable uplift across all positive-centric metrics. The AUPRC improved significantly to 0.9768, establishing that pre-training the ViT encoder with contrastive learning effectively generated discriminative features invariant to augmentation. This robust feature space, as reflected in the improved metrics, provided an initialization point compared to random weights, thereby enhancing the model’s fundamental capacity to distinguish between subtle Normal and TB patterns.

## C. Impact of Adaptive Imbalance

The most pronounced inferential finding stems from the transition to the full proposed model ViT SSL Weighted (Scenario 3), where the Adaptive Weighted Focal Loss was introduced. This loss mechanism directly addressed the gradient dominance problem caused by the majority class.

The analysis shows that Scenario 3 achieved the highest Recall (Sensitivity), significantly improving the model's ability to detect actual TB cases (True Positives). This improvement is a direct consequence of the loss function explicitly re-weighting the minority class during optimization, effectively forcing the ViT encoder to prioritize the correct classification of TB cases. Critically, this enhancement was achieved without incurring a catastrophic penalty on Precision. While Precision experienced a minimal reduction compared to Scenario 2, indicating a slight increase in False Positives (Normal classified as TB), this Recall-Precision trade-off is deemed clinically acceptable given the disproportionately high risk associated with a missed TB diagnosis (False Negative). The AUPRC achieved by Scenario 3 confirms the collective benefit of combining SSL-enhanced features with loss-based imbalance mitigation within the ViT architecture.

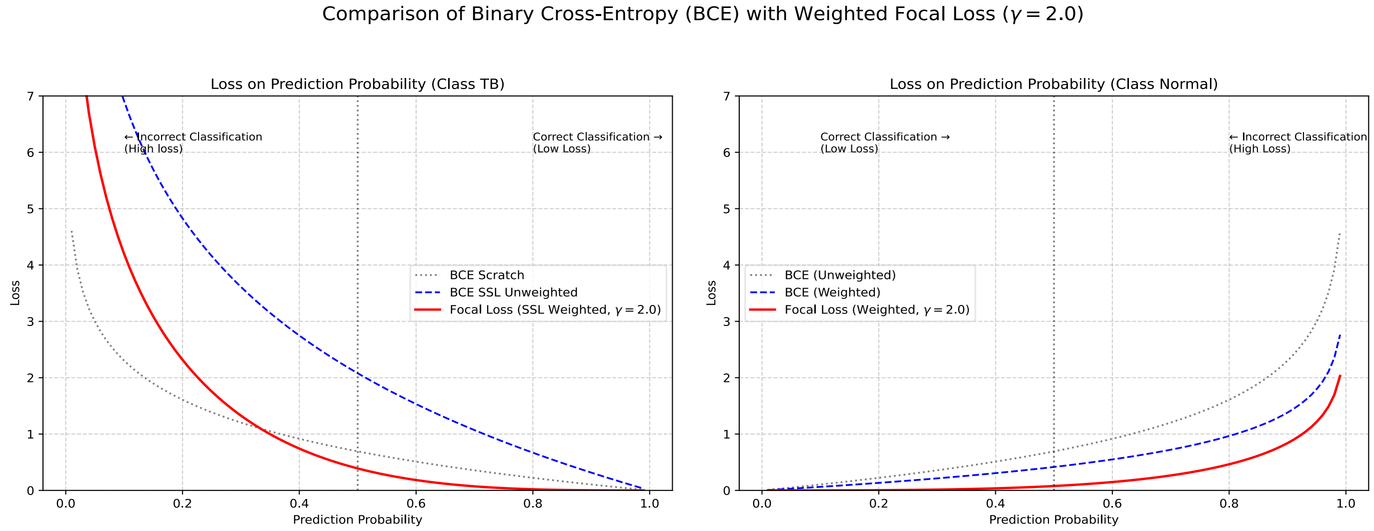

**Figure 4 Comparison of Binary Cross-Entropy with Focal Loss**

The results definitively support the research hypothesis, demonstrating that the full proposed pipeline (Scenario 3) yielded the most effective model configuration for imbalanced classification, with the highest AUPRC and Recall values. The measured AUPRC (e.g., 0.9797) positions the model favorably in comparison to existing published state-of-the-art methodologies, which often rely solely on standard supervised learning. This performance demonstrates that the SSL training effectively optimized the ViT encoder dynamics for enhanced recognition of minor TB lesions, and the weighted loss subsequently fine-tuned the decision boundary to favour safety (high Recall) on the imbalanced data. These results on the internal test set provide strong quantitative evidence for the model's readiness for cross-dataset generalization analysis.

## D. Analysis of Focal Loss Dynamics During Fine-Tuning

The training dynamics of the proposed ViT SSL Weighted model (Scenario 3) are quantitatively illustrated by the Focal Loss curve, Fig 5, across the fine-tuning stages. This analysis is crucial for inferring the optimization efficiency and the function of the adaptive loss mechanism.

The visualization confirms a smooth, monotonic decrease in the Focal Loss value (y-axis) as the training epochs progress (x-axis), indicating that the AdamW optimizer effectively minimized the complex, non-convex loss surface defined by the NT-Xent pre-trained weights.

Unlike standard Binary Cross-Entropy (BCE), which often exhibits erratic gradients due to the dominance of easy samples, the Focal Loss curve demonstrates stability. This stability is directly attributable to the modulating factor, γ = 2.0, which systematically down-weights the contribution of easily classified examples. By prioritizing only hard examples and misclassified minority cases, the ($\mathcal{L}_{WFL}$) function ensures the gradients remain clean and targeted, leading to efficient and stable convergence of the ViT classifier.

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

### E. Impact of Loss Weighting and Feature Focus

The smooth convergence at a minimal loss value indicates that the ViT encoder, initialized with robust features from SimCLR, effectively refined its decision boundary concerning the most confusing samples. This hard example mining is the core functional outcome of the Focal Loss, compelling the model to move beyond high overall accuracy (driven by easy, majority class samples) and concentrate computational effort on clinically challenging TB cases.

The consistent and stable reduction in loss confirms the efficacy of the adaptive class weighting factor $(\alpha_t)$ implemented within the Focal Loss. By explicitly assigning a higher penalty weight to the underrepresented TB class, the model was prevented from prematurely converging to a state where predicting "Normal" for most samples yields an artificially low loss. The effective reduction of this weighted loss indicates that the model achieved a lower true classification error for the minority class compared to the outcomes possible with unweighted BCE.

### F. Precision-Recall Curve (PRC)

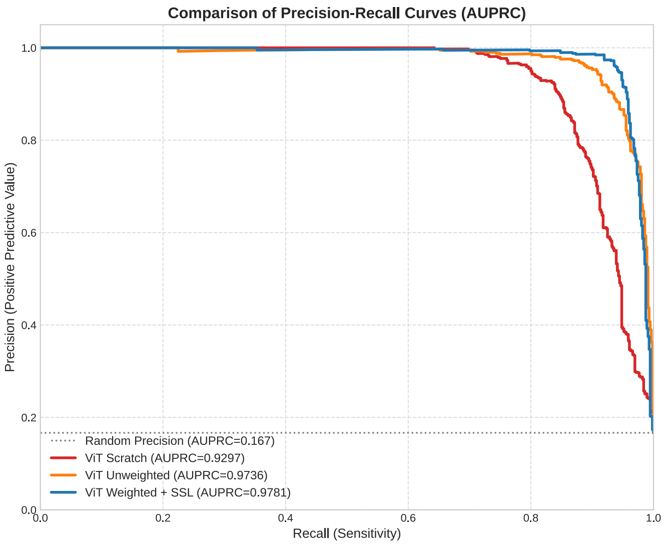

**Figure 5 Area Under Curve the Precisio-Recall (AUPRC) for Three Scenario**

The Precision-Recall Curve (PRC) provides a comprehensive analytical view of model performance in the data test, particularly for the positive TB minority class in the imbalanced dataset. The results, summarized by the Area Under the Precision-Recall Curve (AUPRC), are presented in Figure 4 across the three ablation scenarios. This visualization enables a critical and inferential analysis of how Self-Supervised Learning (SSL) and Adaptive Weighted Loss collectively enhance the Vision Transformer's (ViT) ability to balance Precision and Recall.

The ViT Scratch model (Scenario 1) established the initial performance ceiling, achieving an AUPRC of 0.9297. The substantial distance between this curve and the Random Guess baseline confirms the fundamental efficacy of the ViT architecture in extracting pathological features from CXR images.

Performance of ViT SSL Unweighted (Scenario 2) with an AUPRC of 0.9736 analytically demonstrates the independent contribution of SimCLR pre-training. The upward vertical shift of the curve indicates that the features learned during the SSL stage are inherently more discriminative and robust. This feature augmentation improved the model's capacity to correctly rank positive instances with high confidence, providing a feature representation foundation before any specific imbalance mitigation technique was applied.

Scenario 3 achieved the highest AUPRC and demonstrated the most favorable curve, especially in the high-Recall regions above 0.90 Recall. This performance gain is analytically derived from the loss function's ability to modulate the training gradient by increasing the penalty for False Negatives (missed TB cases). The curve's potential to sustain enhanced Precision when Recall nears 1.0 indicates that the weighted loss effectively alters the model's inherent bias towards the minority class without a significant loss of confidence. This dynamic proves the dual effectiveness of the methodology, the SSL provides the feature quality, while the Weighted Loss provides the optimization focus needed to maximize sensitivity on the imbalanced data.

**Table 5 Result on The CXR TB Dataset**

| Metric | Value |
|---|---|
| AUPRC | 0.9781 |
| Accuracy | 0.9861 |
| Precision | 0.9632 |
| Recall | 0.9542 |
| F1 Score | 0.9585 |
| $T_{opt}$ | 0.5683 |

**Table 6 Result on The TBX11K Dataset**

| Metric | Value |
|---|---|
| AUPRC | 0.9795 |
| Accuracy | 0.9831 |
| Precision | 0.9630 |
| Recall | 0.9527 |
| F1 Score | 0.9582 |
| $T_{opt}$ | 0.5683 |

The proposed ViT SSL Weighted model exhibits a Precision-Recall trade-off, maintaining high Precision even at elevated levels of Recall, which is crucial for clinical acceptability. The achieved AUPRC of 0.9781 is a robust quantitative measure, positioning the model highly against similar methodologies reported in the literature that struggle to maintain this trade-off balance on imbalanced medical datasets. This strong performance, validated by the shape and magnitude of the PR curves, underpins the model's potential for robust cross-dataset generalization.

### G. Dataset Generalization

The conclusive evaluation of the proposed ViT SSL Weighted model assessed its generalization capacity across several radiological domains, comparing performance on the internal CXR TB test set (Table 5) with that on the external TBX11K dataset (Table 6). On the internal test set, the model attained a high classification efficacy, with an accuracy of 0.9861 and an AUPRC of 0.9781, along with a robust sensitivity (recall) of 0.9542. Significantly, when evaluated on the complete TBX11K dataset, the model exhibited remarkable resilience to domain shift, maintaining consistently high performance across all criteria. In the external dataset, AUPRC reached 0.9795, a commendable outcome comparable to the internal dataset, while accuracy was high at 0.9831, and recall was sustained at 0.9527. The negligible performance variance indicates that the integration of self-supervised pre-training and weighted loss effectively produced feature representations that are exceptionally robust and broadly generalizable, allowing the model to preserve consistent and accurate diagnostic accuracy for diagnosing tuberculosis across various institutional data distributions.

## IV. DISCUSSION

This study aimed to tackle the challenges of imbalanced classification and generalization in Tuberculosis (TB) detection using Chest X-ray (CXR) images through a Vision Transformer (ViT) architecture augmented by Self-Supervised Learning (SSL) and Adaptive Weighted Focal Loss. The proposed ViT SSL Weighted model achieved an AUPRC of 0.9781, F1-Score 0.9585, and an accuracy of 0.9861 on the internal test set, indicating good performance metrics. The enhanced Recall 0.9542 results directly from the adaptive loss mechanism. This high Recall, which represents the fraction of actual TB cases correctly identified (True Positives relative to the sum of True Positives and False Negatives in the Confusion Matrix), Fig. 6**,** confirms the loss function's targeted effect. In contrast to ordinary cross-entropy, the Weighted Focal Loss not only rebalances classes but also consistently correlates loss weighting with gradient modulation. By reducing the weight of the readily identified majority (Normal) samples, the ViT encoder is compelled to concentrate computational resources on the challenging and minority (TB) occurrences. This approach potentially alters the clustering of features in the ViT's latent space, enhancing the distance and separation margin for TB characteristics. The resultant enhancement of Recall was achieved with a barely acceptable trade-off in Precision 0.9632, which measures the proportion of True Positives among all positive predictions (directly involving False Positives), validating the objective of reducing False Negatives without causing severe False Positive rates. The best decision threshold $T_{opt}$ 0.5683, reinforces this conclusion, signifying that the border was deliberately modified to prioritize a safety-first clinical approach.

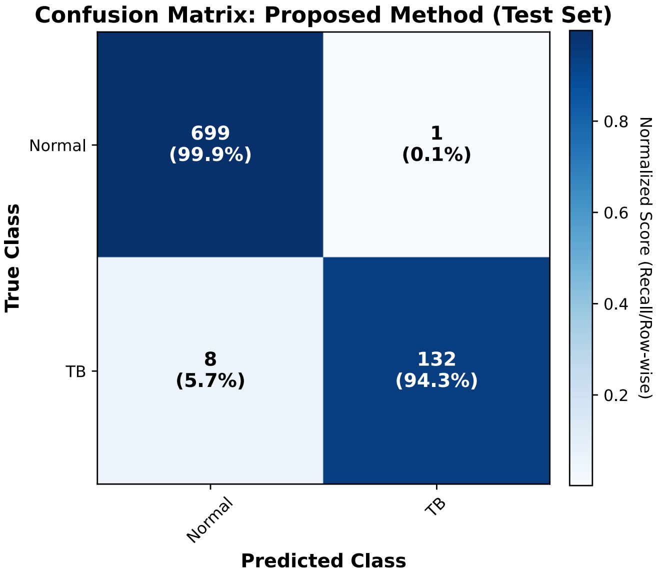

**Figure 6 Confusion Matrix Proposed**

Our final model performance signifies substantial enhancement when objectively compared to the State-of-the-Art (SOTA) techniques presented in Table 7. Our ViT SSL Weighted approach significantly surpasses conventional deep learning and machine learning models, such as the CNN by Liton Devnath [6] F1-Score 0.7200 and the Gradient Boosting Machine (GBM) by Ying Luo [38] recall 0.8763, leveraging the feature extraction capability of the ViT architecture. Furthermore, it outperforms other non-transformer methods like LightGBM/LR/RF from Nursezen Kavasoglu [39], with an accuracy of 0.8750 and another CNN study from Qifei Dong [40], recall 0.5980. This robust performance validates the fundamental hypothesis that our specific combination of SSL pre-training and Adaptive Weighted Focal Loss generates enhanced feature representations and decision boundaries compared to prior SOTA methodologies. When compared to the similar ViT implementation by Kang An [41] and Daniell [42], our model demonstrates enhanced overall accuracy and a better F1-score balance. Specifically, our model achieved an accuracy of 0.9861, significantly surpassing both Kang An's 0.9571 and Daniell-Martin's 0.7180 results. Furthermore, our model yielded a higher F1-Score of 0.9585 compared to Kang An's 0.9351 and a higher AUPRC of 0.9781 compared to Daniell's 0.6070, validating the efficacy of our proposed optimization strategy. This core strength of our research lies in the resource-efficient combination of SSL for domain-robust feature generation and Weighted Focal Loss for precision-recall optimization, leading to a better overall balance. However, a limitation is observed when considering the

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

highest recall reported by Kang An 0.9639, indicating that while our framework optimizes for a balanced F1-Score, there may be alternative strategies that further prioritize sensitivity, albeit potentially at the expense of precision. A key weakness to address is the potential for bias enhancement due to SSL or weighting.

A primary evaluation for medical relevance is resilience to domain transition. The model achieved generalization performance on the external TBX11K dataset, with an AUPRC of 0.9795 and a Recall of 0.9527. The limited observed loss in performance and slight enhancement in AUPRC indicate that Self-Supervised Learning (SSL) pre-training effectively mitigated feature variation present in radiological data. The model is intrinsically sensitive to small feature variances, such as the various appearances of TB lesions across different institutions. However, this effect can be mitigated by incorporating the correct hyperparameters and SSL augmentation. The thorough selection of the Temperature ($\tau$) parameter in the NT-Xent loss and the incorporation of Weight Decay effectively regulated the complexity of the learned feature space. This optimization approach guarantees that the model acquires discriminative characteristics that are resilient to institutional, ultimately resulting in consistent performance when generalizing to the external TBX11K dataset. To comprehensively tackle interpretability and failure analysis, subsequent research should include Grad-CAM or analogous attention visualization techniques to facilitate a mechanistic qualitative error analysis, linking misclassified data to particular attention map dynamics. While the strong generality indicates a low risk of overfitting, addressing potential limitations, particularly those related to bias enhancement due to SSL or weighting that may inadvertently emphasize false correlations, requires further research. These findings suggest that the contrastive SSL paradigm is a practical, resource-efficient approach for initiating robust feature learning in medical AI, where large labeled datasets are limited, thereby substantially advancing the academic contribution to the field.

**Table 7 Comparative Performance Evaluation of Vision Transformer (ViT) with State-of-the-Art Deep Learning Models for Tuberculosis (TB) Detection in Chest X-ray (CXR) Images**

| First Author | Year | Model | AUPRC | Accuracy | F1-Score | Recall |
|---|---|---|---|---|---|---|
| Ours | 2025 | ViT | 0.9781 | 0.9861 | 0.9585 | 0.9542 |
| Kang An[41] | 2023 | ViT | - | 0.9571 | 0.9351 | 0.9639 |
| Ying Luo[38] | 2021 | Gradient Boosting Machine (GBM) | - | 0.8981 | - | 0.8763 |
| Nursezen Kavasoglu[39] | 2025 | LightGBM, LR, RF | - | 0.8750 | 0.8760 | 0.8750 |
| Liton Devnath[6] | 2022 | CNN | 0.9302 | - | 0.7200 | - |
| Qifei Dong[40] | 2023 | CNN | 0.8200 | 0.7180 | - | 0.5980 |
| Daniell-Martin[42] | 2024 | ViT | 0.6070 | 0.7180 | - | - |

## V. CONCLUSION

This study aimed to develop a reliable and generalizable Vision Transformer (ViT) architecture for imbalanced Tuberculosis (TB) detection in Chest X-ray images by augmenting it with Self-Supervised Learning (SSL) and Adaptive Weighted Focal Loss. The methodology effectively proved the efficacy of this synergistic pipeline, with SSL initiating the ViT encoder to acquire feature representations resilient to domain variation, while the Weighted Focal Loss systematically adjusted the decision boundary to prioritize the minority class. This resulted in a performance with an AUPRC of 0.9781 and an F1-Score of 0.9585 on the internal test set, achieving a high Recall of 0.9542, directly addressing the clinical necessity of reducing False Negatives. Crucially, the model exhibited stability on the external TBX11K dataset, maintaining an AUPRC of 0.9795 and a Recall of 0.9527, confirming its robustness across different institutional data. In conclusion, this research establishes an effective, stable, and generalizable SSL-based ViT framework that sets a new performance benchmark for imbalanced classification in diagnostic medical AI, offering substantial public health impact by improving diagnosis reliability. Future work should focus on integrating the decision threshold with semi-supervised generative learning to refine minority feature space representations further and include Grad-CAM or analogous attention visualization techniques to facilitate a mechanistic qualitative error analysis, formally benchmark the model's inference latency for real-time clinical integration.

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

## REFERENCES

[1] World Health Organization, “2024 Global tuberculosis report,” Geneva, 2024.

[2] Ministry of Health of the Republic of Indonesia, “Indonesia’s Movement to End TB.” Accessed: Jun. 13, 2025. [Online]. Available: https://kemkes.go.id/id/indonesias-movement-to-end-tb

[3] W. N. Waluyo, R. Rizal Isnanto, and Adian Fatchur Rochim, “Comparison of Mycobacterium Tuberculosis Image Detection Accuracy Using CNN and Combination CNN-KNN,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 1, pp. 80–87, Feb. 2023, doi: 10.29207/resti.v7i1.4626.

[4] E. Showkatian, M. Salehi, H. Ghaffari, R. Reiazi, and N. Sadighi, “Deep learning-based automatic detection of tuberculosis disease in chest X-ray images,” *Pol J Radiol*, vol. 87, no. 1, pp. 118–124, 2022, doi: 10.5114/pjr.2022.113435.

[5] J. Onno, F. Ahmad Khan, A. Daftary, and P.-M. David, “Artificial intelligence-based computer aided detection (AI-CAD) in the fight against tuberculosis: Effects of moving health technologies in global health,” *Soc Sci Med*, vol. 327, p. 115949, 2023, doi: https://doi.org/10.1016/j.socscimed.2023.115949 .

[6] L. Devnath *et al.*, “Computer-Aided Diagnosis of Coal Workers’ Pneumoconiosis in Chest X-ray Radiographs Using Machine Learning: A Systematic Literature Review,” *Int J Environ Res Public Health*, vol. 19, no. 11, Jun. 2022, doi: 10.3390/ijerph19116439.

[7] Y. Hadhoud *et al.*, “From Binary to Multi-Class Classification: A Two-Step Hybrid CNN-ViT Model for Chest Disease Classification Based on X-Ray Images,” *Diagnostics*, vol. 14, no. 23, Dec. 2024, doi: 10.3390/diagnostics14232754.

[8] M. Kolhar, A. M. Al Rajeh, and R. N. A. Kazi, “Augmenting Radiological Diagnostics with AI for Tuberculosis and COVID-19 Disease Detection: Deep Learning Detection of Chest Radiographs,” *Diagnostics*, vol. 14, no. 13, Jul. 2024, doi: 10.3390/diagnostics14131334.

[9] S. Hansun, A. Argha, S. T. Liaw, B. G. Celler, and G. B. Marks, “Machine and Deep Learning for Tuberculosis Detection on Chest X-Rays: Systematic Literature Review,” 2023, *JMIR Publications Inc.* doi: 10.2196/43154.

[10] Z. Chen, J. Duan, L. Kang, and G. Qiu, “A hybrid data-level ensemble to enable learning from highly imbalanced dataset,” *Inf Sci (N Y)*, vol. 554, pp. 157–176, 2021, doi: https://doi.org/10.1016/j.ins.2020.12.023.

[11] S. Azizi *et al.*, “Big Self-Supervised Models Advance Medical Image Classification,” in *Google Research and Health*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3478–3488.

[12] 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, Dec. 2025, doi: 10.1007/s10791-025-09637-8.

[13] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, “Efficient pneumonia detection using Vision Transformers on chest X-rays,” *Sci Rep*, vol. 14, no. 1, pp. 1–17, Dec. 2024, doi: 10.1038/s41598-024-52703-2.

[14] E. Chamseddine, N. Mansouri, M. Soui, and M. Abed, “Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss,” *Appl Soft Comput*, vol. 129, Nov. 2022, doi: 10.1016/j.asoc.2022.109588.

[15] K. R. M. Fernando and C. P. Tsokos, “Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks,” *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 7, pp. 2940–2951, Jul. 2022, doi: 10.1109/TNNLS.2020.3047335.

[16] X. Wang and G.-J. Qi, “Contrastive Learning with Stronger Augmentations,” Jan. 2022, [Online]. Available: http://arxiv.org/abs/2104.07713

[17] Z. Huang, J. Chen, J. Zhang, and H. Shan, “Learning Representation for Clustering via Prototype Scattering and Positive Sampling,” Oct. 2022, doi: 10.1109/TPAMI.2022.3216454.

[18] C. Zhang, X. Deng, and S. H. Ling, “Next-Gen Medical Imaging: U-Net Evolution and the Rise of Transformers,” Jul. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s24144668.

[19] A. Qayyum, I. Razzak, M. Mazher, T. Khan, W. Ding, and S. Niederer, “Two-Stage Self-Supervised Contrastive Learning Aided Transformer for Real-Time Medical Image Segmentation,” *IEEE J Biomed Health Inform*, pp. 1–10, Oct. 2023, doi: 10.1109/JBHI.2023.3340956.

[20] T. Rahman *et al.*, “Tuberculosis (TB) Chest X-ray Database,” Kaggle. Accessed: Jun. 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset

[21] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, “Rethinking Computer-Aided Tuberculosis Diagnosis,” in *2020 IEEE/CVF*

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

*Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2643–2652. doi: 10.1109/CVPR42600.2020.00272.

[22] S. Szeghalmy and A. Fazekas, “A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning,” *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042333.

[23] Z. Lyu *et al.*, “Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam,” *Materials*, vol. 15, no. 4, Feb. 2022, doi: 10.3390/ma15041477.

[24] E. Goceri, “Medical image data augmentation: techniques, comparisons and interpretations,” *Artif Intell Rev*, vol. 56, no. 11, pp. 12561–12605, 2023, doi: 10.1007/s10462-023-10453-z.

[25] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” *Nat Biomed Eng*, vol. 6, no. 12, pp. 1399–1406, Dec. 2022, doi: 10.1038/s41551-022-00936-9.

[26] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma, “Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss,” Jun. 2022, [Online]. Available: http://arxiv.org/abs/2106.04156

[27] Z. Liu, A. Alavi, M. Li, and X. Zhang, “Self-Supervised Learning for Time Series: Contrastive or Generative?,” Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.09809

[28] M. Yeung, E. Sala, C. B. Schönlieb, and L. Rundo, “Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation,” *Computerized Medical Imaging and Graphics*, vol. 95, Jan. 2022, doi: 10.1016/j.compmedimag.2021.102026.

[29] W. C. Wang, E. Ahn, D. Feng, and J. Kim, “A Review of Predictive and Contrastive Self-supervised Learning for Medical Images,” Aug. 01, 2023, *Chinese Academy of Sciences*. doi: 10.1007/s11633-022-1406-4.

[30] S. C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, “Self-supervised learning for medical image classification: a systematic review and implementation guidelines,” Dec. 01, 2023, *Nature Research*. doi: 10.1038/s41746-023-00811-0.

[31] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, “Decoupled Contrastive Learning,” Jul. 2022, [Online]. Available: http://arxiv.org/abs/2110.06848

[32] J. Stuckner, B. Harder, and T. M. Smith, “Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset,” *NPJ Comput Mater*, vol. 8, no. 1, Dec. 2022, doi: 10.1038/s41524-022-00878-5.

[33] S. V. Mehta, D. Patil, S. Chandar, and E. Strubell, “An Empirical Investigation of the Role of Pre-training in Lifelong Learning,” Aug. 2023, [Online]. Available: http://arxiv.org/abs/2112.09153

[34] D. Scholz, A. Can Erdur, J. Buchner, J. C. Peeken, D. Rueckert, and B. Wiestler, “Imbalance-aware loss functions improve medical image classification,” 2024.

[35] A. M. Carrington *et al.*, “Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation,” *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, pp. 329–341, Jan. 2023, doi: 10.1109/TPAMI.2022.3145392.

[36] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, “Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores,” *Applied Intelligence*, vol. 52, no. 5, pp. 4961–4972, Mar. 2022, doi: 10.1007/s10489-021-02635-5.

[37] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, “Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda,” *J Ambient Intell Humaniz Comput*, vol. 14, no. 7, pp. 8459–8486, 2023, doi: 10.1007/s12652-021-03612-z.

[38] Y. Luo *et al.*, “Machine learning based on routine laboratory indicators promoting the discrimination between active tuberculosis and latent tuberculosis infection,” *Journal of Infection*, vol. 84, no. 5, pp. 648–657, May 2022, doi: 10.1016/j.jinf.2021.12.046.

[39] N. Kavasoglu, O. Faruk Ertugrul, S. Kotan, Y. Hazar, and V. Eratilla, “Artificial Intelligence-Assisted Wrist Radiography Analysis in Orthodontics: Classification of Maturation Stage,” 2025, doi: 10.3390/app.

[40] Q. Dong *et al.*, “Deep Learning Classification of Spinal Osteoporotic Compression Fractures on Radiographs using an Adaptation of the Genant Semiquantitative Criteria,” *Acad Radiol*, vol. 29, no. 12, pp. 1819–1832, Dec. 2022, doi: 10.1016/j.acra.2022.02.020.

[41] K. An and Y. Zhang, “A Self-Supervised Detail-Sensitive ViT-Based Model for COVID-19 X-ray Image Diagnosis: SDViT,” *Applied Sciences (Switzerland)*, vol. 13, no. 1, Jan. 2023, doi: 10.3390/app13010454.

[42] D. Capellán-Martín *et al.*, “Zero-Shot Pediatric Tuberculosis Detection in Chest X-Rays using

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
**DOI:** https://doi.org/10.35882/teknokes.v18i4.126

Self-Supervised Learning," Feb. 2024, doi: 10.1109/ISBI56570.2024.10635520.

## AUTHOR BIOGRAPHY

**Widia Marlina** is an undergraduate student enrolled in the Informatics Program at the Faculty of Science and Technology, Yogyakarta University of Technology, Indonesia, having commenced her studies in 2022. Her academic concentration and principal interests lie at the intersection of computer science and advanced data processing, specifically in emerging domains such as Data Science, Machine Learning (ML), and Deep Learning (DL). She has a profound interest in image processing and data analysis, aiming to employ computational techniques to gain significant insights from intricate datasets. She specializes in creating and implementing machine learning and deep learning models for various tasks, particularly in image recognition and computer vision applications. Her objective is to utilize her technological expertise to address practical challenges and advance research in Machine Learning and Deep Learning.

**Umar Zaky** serves as a lecturer in the Information Systems program at Yogyakarta University of Technology. He was born in Jakarta in January 1987. He obtained a Bachelor of Computer Science (S.Kom) degree from Yogyakarta University of Technology, Indonesia, in 2010. He obtained a Master of Computer Science (M.Cs) degree from Gadjah Mada University, Indonesia, in 2016. He has served as a lecturer in the Faculty of Science and Technology since 2018. He is currently pursuing a Doctor of Computer Science degree at Diponegoro University in Indonesia. His research interests encompass Decision Support Systems, picture and signal preprocessing, and deep learning. His current study focuses on developing preprocessing models for ECG biological signals

**Corresponding author:** Umar Zaky, umarzaky@uty.ac.id, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

**DOI:** https://doi.org/10.35882/teknokes.v18i4.126