ISSN: 3107-4707
DOI: doi.org/10.51219/JMMS/Mirzaeibonehkhater-M/27

URF PUBLISHERS
connect with research world

Journal of Medicine and Medical Studies
https://urfpublishers.com/journal/medicine-medical-studies

Vol: 2 & Iss: 2

*Research Article*

# Class Imbalance-Aware Active Learning with Vision Transformers in Federated Histopathological Imaging

**Mohammad Ali Labbaf Khaniki[1], Marzieh Mirzaeibonehkhater[2]* and Siavash Esfandiari Fard[3]**

[1]Faculty of Electrical Engineering, K.N. Toosi University of Technology, Tehran, Iran

[2]Department of Electrical and Computer Engineering, Indiana University-Purdue University

[3]Department of Electrical Engineering, The University of Alabama

**Citation:** Khaniki MAL, Mirzaeibonehkhater M, Fard SE. Class Imbalance-Aware Active Learning with Vision Transformers in Federated Histopathological Imaging. *J M Med Stu 2025;* 2(2): 141-150. DOI: doi.org/10.51219/JMMS/Mirzaeibonehkhater-M/27

**Received:** 27 April, 2025; **Accepted:** 05 May, 2025; **Published:** 07 May, 2025

***Corresponding author:** Marzieh Mirzaeibonehkhater, Department of Electrical and Computer Engineering, Indiana University-Purdue University, E-mail: marzieh89mirzaei@gmail.com

**Copyright:** © 2025 Mirzaeibonehkhater M, et al., this is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

## A B S T R A C T

Class imbalance in histopathological image datasets poses a critical challenge for metastatic cancer detection, often leading to suboptimal model performance on minority cancer cases. To address this, we propose a novel framework integrating three components: Learnable Memory Vision Transformer (LMViT), Class Imbalance-Aware Active Learning (CIA-AL) and Federated Learning Integration (FLI). The LMViT architecture incorporates learnable memory tokens at each transformer layer, enabling the capture of subtle discriminative features and enhancing the detection of underrepresented cancer cases through global context modeling. The CIA-AL module utilizes attention entropy and confidence-attention mismatch metrics derived from memory tokens to intelligently prioritize minority class samples with high attention but low prediction confidence, thereby optimizing limited annotation resources and focusing expert efforts on diagnostically challenging instances. The FLI component ensures privacy-preserving collaborative training across multiple institutions by securely aggregating model weights rather than patient data, maintaining HIPAA compliance while improving model generalizability. Experimental evaluation demonstrates that our framework achieves 89% accuracy, 84% precision, 84% recall and a 0.91 AUC-ROC, representing improvements of 12%, 17%, 24% and 0.13, respectively, over conventional Vision Transformers. Furthermore, it reduces annotation burden by 40%, while precision-recall analyses confirm consistently high precision across varying recall thresholds, underscoring its potential for real-world clinical deployment.

**Keywords:** Active Learning, Class Imbalance, Federated Learning, Metastatic Cancer Detection, Vision Transformers

## 1. Introduction

Histopathological image analysis is a cornerstone of cancer diagnosis, enabling the identification of metastatic cancer through the microscopic examination of tissue samples. This process involves staining tissue sections, typically with hematoxylin and eosin (H&E), to reveal cellular and structural details, which pathologists analyze to detect abnormal features indicative of malignancy, such as irregular cell morphology, increased nuclear size or abnormal tissue architecture[1]. For metastatic cancer, histopathological analysis is critical to confirm the spread of cancer to lymph nodes or distant organs, guiding treatment decisions and prognosis. The advent of digital pathology has transformed this field by enabling high-resolution whole-slide imaging (WSI), allowing for automated analysis using machine learning[2]. However, these datasets often

exhibit severe class imbalance, with cancer cases (minority class) significantly underrepresented compared to non-cancer cases (majority class). This imbalance biases models toward the majority class, reducing sensitivity for detecting critical cancer instances, which can have dire consequences for patient outcomes[3].

Active Learning (AL) is a machine learning paradigm designed to optimize the annotation process by selectively labeling the most informative data samples, thereby reducing the need for extensive manual labeling in resource-constrained settings like medical imaging[4]. In the context of histopathological analysis for metastatic cancer detection, AL can significantly enhance efficiency by prioritizing samples that are most likely to improve model performance, such as those with high uncertainty or belonging to the underrepresented cancer class[5]. By iteratively querying a subset of unlabeled data for expert annotation and incorporating these into the training process, AL ensures that the model focuses on challenging or critical cases, addressing issues like class imbalance. This is particularly valuable in medical applications, where expert annotations are costly and time-intensive and datasets often exhibit skewed distributions[6].

Federated Learning (FL) is a distributed machine learning approach that enables collaborative model training across multiple institutions without the need to share sensitive patient data, making it an ideal solution for medical applications where privacy is paramount. In the context of histopathological image analysis for metastatic cancer detection, FL allows hospitals and research centers to train a shared global model by locally processing their private datasets and only exchanging model updates, such as gradients or weights, rather than raw data. This preserves patient confidentiality while leveraging diverse, multi-institutional data to improve model robustness and generalizability[7]. FL is particularly suited to address challenges like data heterogeneity and class imbalance, as it can incorporate strategies to handle non-identically distributed data across clients[8].

The Transformer architecture, a neural network model introduced by Vaswani, et al. in 2017, has become a cornerstone in natural language processing (NLP) due to its self-attention mechanisms, which effectively capture relationships across input sequence elements[9]. Building on its success in NLP, researchers extended this architecture to computer vision, resulting in Vision Transformers (ViTs)[10,11]. ViTs innovate by treating fixed-size image patches as analogous to words in NLP, enabling the Transformer to process images for classification tasks. In this approach, an image is segmented into patches, which are then processed by the Transformer's self-attention mechanisms to model dependencies between patches, yielding rich and expressive image representations[12]. Compared to traditional computer vision architectures, ViTs provide superior performance, greater flexibility and enhanced interpretability, making them a powerful tool for image analysis tasks[13].

Despite these advancements, the integration of AL, FL and ViTs for class-imbalanced medical datasets remains largely unexplored. A novel framework is presented, wherein Learnable Memory Vision Transformers are combined with a class imbalance-aware active learning strategy within an FL setting. ViT's attention mechanisms are leveraged to prioritize minority class samples and privacy is ensured through FL, aiming to enhance diagnostic accuracy for metastatic cancer detection. This integration is poised to offer a scalable and impactful solution for clinical practice, advancing the field of automated histopathological analysis.

Three key components are integrated in the proposed framework:

- **Learnable memory vision transformer:** The ViT architecture is enhanced with learnable memory tokens incorporated at each layer. Task-specific features are stored by these tokens, improving the detection of underrepresented cancer cases. Global dependencies across image patches are captured through ViT's self-attention mechanisms, enabling focus on subtle features indicative of malignancy, such as irregular cell structures, even within imbalanced datasets.
- **Class imbalance-aware active learning:** Attention maps derived from the memory tokens are utilized to prioritize labeling of minority class samples exhibiting high attention but low confidence, thereby optimizing the use of limited annotation resources. An informativeness metric, such as attention entropy or confidence-attention mismatch, is employed to select samples that enhance sensitivity to cancer cases, effectively addressing class imbalance in histopathological images.
- **Federated learning integration:** Local models are trained at each institution using private datasets and the active learning strategy. A global model is subsequently updated via federated averaging, ensuring privacy and robustness across diverse data. The class imbalance-aware AL strategy is applied locally by each client, selecting informative cancer samples for labeling, while only model updates are shared, preserving patient confidentiality and leveraging multi-institutional data to enhance generalizability.

Class imbalance is effectively addressed while patient privacy is maintained, offering a novel and practical solution for metastatic cancer detection in clinical settings. This framework not only improves the model's sensitivity to rare cancer cases but also ensures scalability across diverse healthcare environments, accommodating variations in data distribution and institutional resources. By prioritizing informative samples through attention-based active learning, annotation efforts are optimized, reducing the burden on medical experts while enhancing model performance.

## 2. Background and Related Work

Accurate histopathological image analysis is essential for cancer detection, yet it faces major challenges such as class imbalance and data privacy concerns. Recent advances in Active Learning, Federated Learning and Vision Transformers offer promising solutions to these issues. This section provides an overview of histopathological image analysis for cancer detection and reviews key developments in the related fields.

### 2.1. Histopathological image analysis for cancer detection

Histopathological image analysis is a critical component of cancer diagnosis, involving the microscopic examination of tissue samples to identify malignant features indicative of metastatic cancer. Tissue sections, stained typically with hematoxylin and eosin (H&E), reveal cellular details such as irregular morphology or abnormal tissue architecture, which are essential for confirming cancer spread to lymph nodes or distant organs. The transition to digital pathology has enabled high-resolution whole-slide imaging (WSI), facilitating automated

analysis through machine learning. However, class imbalance is a persistent challenge, with cancer cases (minority class) significantly underrepresented compared to non-cancer cases (majority class), leading to biased models that exhibit reduced sensitivity for critical cancer detection. This issue is compounded by the sensitive nature of medical data, necessitating privacy-preserving approaches to leverage multi-institutional datasets effectively. **(Figure 1)** illustrates the class distribution of the Histopathologic Cancer Detection dataset, depicting 59.50% for Label 0 (No Cancer) and 40.50% for Label 1 (Cancer), highlighting the class imbalance challenge in histopathological image analysis.

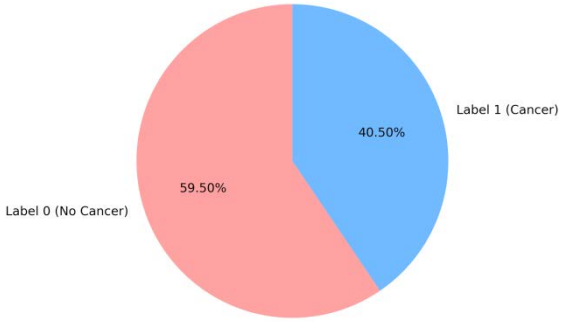


**Figure 1:** Representative the class distribution of the Histopathologic Cancer Detection dataset.

### 2.2. Literature review

Active Learning (AL) enhances cancer detection by selectively querying the most informative samples for expert annotation, reducing the need for exhaustive labeling and optimizing limited expert resources. In tasks like histopathological image analysis or radiological imaging, AL focuses on ambiguous or rare malignant cases, improving model sensitivity to critical instances while addressing class imbalance. Techniques such as uncertainty sampling, entropy-based selection and diversity sampling guide the process, leading to faster model convergence and more efficient annotation efforts in cancer detection pipelines. This paper[14] combines Bayesian deep learning with active learning to address the challenges of learning from small datasets and representing model uncertainty, demonstrating significant improvements over existing active learning approaches on image data including MNIST and skin cancer diagnosis from lesion images. This groundbreaking study developed and validated graphene-based optical nano biosensors that detect early-stage ovarian cancer through liquid biopsy with remarkable 94.5% accuracy, employing a hierarchical framework with active learning to quantify protease activities that specifically indicate ovarian cancer[15]. This paper explores the use of active learning and deep learning to improve the delineation of gross tumor volume in nasopharyngeal carcinoma (NPC) based on MRI images, addressing challenges related to variability across different centers and raters. The approach aims to enhance model generalizability and accuracy in tumor segmentation for radiotherapy planning, leveraging active learning to optimize the use of limited labeled data effectively[16]. AnchorAL is an efficient active learning method for large and imbalanced datasets that dynamically selects class-specific “anchor” examples to build small, balanced subpools for query, thereby improving runtime, enhancing minority class discovery and producing more balanced and accurate models than standard approaches[4]. [17]reveals that increased V-ATPase activity may contribute to chemoresistance in oral squamous cell carcinoma by inducing autophagy, offering new insights into potential therapeutic targets[18]. review evaluates the effects of conservative treatments on pain, function and grip strength in patients with tennis elbow syndrome, highlighting their overall effectiveness in symptom management[19]. Dynamic Classification Using the Adaptive Competitive Algorithm for Breast Cancer Detection introduces the Adaptive Competitive Self-organizing (ACS) model, which leverages ordinary differential equations and gradient descent for superior clustering stability and classification accuracy in distinguishing benign from malignant breast cancer cases Pazhooman et al. (2023) found that runners with plantar heel pain exhibit distinct foot kinematic alterations during running, including increased lateral midfoot eversion in early stance and sex-specific differences in medial midfoot and forefoot motion during propulsion compared to healthy runners[20]. The paper provides a systematic evaluation comparing fine-tuning, prompt engineering and RAG for mental health text analysis, demonstrating their relative strengths and advocating for hybrid approaches tailored to specific clinical contexts[21].

Federated Learning (FL) enables collaborative cancer detection model training across multiple institutions without sharing raw patient data, preserving privacy while benefiting from diverse clinical datasets. In applications like tumor classification from histopathological slides or medical imaging, each institution trains a local model and shares only updates for global aggregation. This approach enhances model generalizability, addresses data heterogeneity and, when combined with techniques like differential privacy, further strengthens data security, making FL an ideal solution for large-scale, privacy-preserving cancer detection initiatives. This paper introduces a distributed deep convolutional neural network (DCNN) approach using federated learning for breast cancer detection, allowing healthcare institutions to collaboratively train models without sharing sensitive patient data while achieving competitive diagnostic accuracy compared to centralized approaches and addressing privacy concerns in medical image analysis[22]. The paper demonstrates that a graph neural network (GNN) effectively scales to predict frictional contact networks in dense suspensions, maintaining accuracy even for large particle systems, with implications for simulating complex material behaviors[23]. This paper introduces a collaborative federated learning framework that enables multiple healthcare institutions to jointly train deep learning models for lung and colon cancer classification from medical images without sharing sensitive patient data, demonstrating significant improvements in diagnostic accuracy while preserving privacy and addressing the challenge of limited local datasets[24]. This study proposes a novel approach combining federated learning with YOLOv6 for classifying breast cancer pathology images, achieving high accuracy while preserving patient privacy through distributed model training. The model outperforms traditional deep learning architectures like VGG-19, ResNet-50 and InceptionV3[25]. FedCSCD-GAN presents a secure and collaborative clinical cancer diagnosis framework that integrates optimized federated learning with generative adversarial networks (GANs), enabling decentralized model training that preserves patient data privacy across institutions while enhancing diagnostic accuracy through synthetic data generation and robust feature learning[26]. The paper[27] introduces a hyperdimensional computing-based approach for network anomaly detection in

IoT environments, demonstrating high efficiency and accuracy on the NSL-KDD dataset[28]. Presents the design of a virtual reality training apprenticeship program tailored for Cold Spray Advanced Manufacturing, aiming to enhance skill acquisition and operational understanding through immersive simulation[29]. Present a network anomaly detection approach for IoT systems using hyperdimensional computing, demonstrating efficient and accurate performance on the NSL-KDD dataset[30]. Explores the use of eye-tracking metrics to detect cognitive load in users during complex virtual reality training scenarios, aiming to enhance adaptive learning experiences.

Vision Transformers (ViTs) are a powerful tool for cancer detection, excelling at analyzing complex medical images like histopathological slides, MRI and CT scans. Unlike traditional CNNs, ViTs use self-attention mechanisms to capture long-range dependencies, making them effective at identifying subtle cancerous changes, such as irregular cell structures and tumor margins. ViTs perform well in classification, segmentation and localization tasks, offering better interpretability and flexibility. With adaptations like lightweight or hybrid CNN-ViT models, they are also suitable for resource-constrained medical environments, showing great potential in improving diagnostic accuracy and supporting early cancer detection[31]. Presents a robust deep learning framework leveraging vision transformers for accurate breast cancer and subtype identification, demonstrating the growing effectiveness of transformer-based models over traditional convolutional neural networks in medical image analysis[32]. Presents LCDViT, a specialized vision transformer model with explainable AI capabilities that significantly improves the accuracy and reliability of lung cancer diagnostics, contributing to the growing trend of transformer-based approaches in medical imaging applications[33]. Study introduces RI-ViT, an innovative multi-scale hybrid methodology leveraging vision transformers for breast cancer detection in histopathological images, joining the growing trend of transformer-based approaches that demonstrate superior performance over traditional convolutional neural networks in medical imaging applications[34]. Proposes a novel approach to enhance breast cancer detection by combining vision transformers with convolutional neural networks for calcification mammography classification, aiming to improve the precision of breast cancer detection through the fusion of these advanced technologies[35]. Introduces a vision transformer model enhanced with feature calibration and selective cross-attention mechanisms for brain tumor classification, presenting a novel approach that aims to boost classification accuracy by leveraging advanced self-attention techniques tailored to brain MRI analysis[36]. Comprehensively analyzes various positional encoding techniques for transformer-based time series models, revealing that advanced methods like TUPE and SPE consistently outperform traditional approaches across diverse datasets while maintaining computational efficiency[37]. Introduces a domain adaptation framework that integrates GRU and Attention U-Net to improve the accuracy and cross-dataset generalization of contactless fingerprint presentation attack detection[38]. Proposes rPPG-SysDiaGAN, a GAN-based framework with a multi-domain discriminator designed to localize systolic and diastolic features in remote photoplethysmography (rPPG) signals for improved physiological signal interpretation across domains. Nassajpour, et al, propose a wearable sensor and machine learning framework using IMUs on ankles, lumbar and sternum to objectively estimate m-CTSIB balance scores, demonstrating superior accuracy in capturing lateral/medial movement correlations and enabling remote balance assessment[39].

## 3. Methodology

In this section, we present our proposed framework, which enhances the Vision Transformer (ViT) architecture for medical imaging tasks through the integration of learnable memory tokens, active learning strategies and federated learning. First, we introduce the Learnable Memory Vision Transformer (LMViT), which incorporates memory tokens to capture task-specific global information throughout the transformer layers. Next, we detail the Class Imbalance-Aware Active Learning (CIA-AL) strategy, designed to address the challenges of sample ambiguity and class imbalance by selecting informative examples based on internal attention dynamics. Finally, we describe the Federated Learning Integration (FLI), enabling collaborative model training across multiple institutions while ensuring data privacy. Together, these components build a robust and privacy-preserving learning framework for medical image analysis.

### 3.1. Learnable memory vision transformer (LMViT)

**3.1.1. Learnable memory tokens:** Let the input image be: $x \in \mathbb{R}^{H\times W\times C}$, where $H$ and $W$ are height and width and is the number of channels (e.g., 3 for RGB). The image is partitioned into non-overlapping patches, yielding:

$$N = \frac{HW}{P^2} \quad (1)$$

where $P$ is the spatial resolution (height and width) of each square patch. Each patch is flattened and projected into a $d$-dimensional embedding space through a learnable linear projection $E$:

$$p_i = E\big(flatten(p_i)\big) \in \mathbb{R}^d \; for \; i = 1,2,\dots,N \quad (2)$$

The sequence of patch embeddings is denoted by:

$$P = [p_1, p_2, \dots, p_N] \quad (3)$$

In addition to patch embeddings, LMViT introduces a set of learnable memory tokens, denoted as:

$$M = [m_1, m_2, \dots, m_M] \in \mathbb{R}^{M\times d} \quad (4)$$

where each $m_i$ is a learnable parameter initialized randomly and optimized jointly with the rest of the model. These memory tokens are designed to accumulate and preserve task-specific information, such as features associated with rare cancer cells, throughout the transformer layers.

At the $l$-th transformer layer, the input sequence $Z_l$ is formed by concatenating the current memory tokens with the current patch embeddings $Z_l$:

$$Z_l = [M_l; P_l] \in \mathbb{R}^{(M+N)d} \quad (5)$$

**3.1.2. Transformer block operations:** Each block consists of Multi-Head Self-Attention (MHSA) with residual connection:

$$Z_l' = Z_l + \mathrm{MHSA}\big(\mathrm{LN}(Z_l)\big) \quad (6)$$

MHSA computes self-attention across both memory and patch tokens. Feed-Forward Network (FFN) with residual connection:

where: $(\cdot)$ is Layer Normalization, $(\cdot)$ is the multi-head attention operation, $(\cdot)$ is the two-layer position-wise feed-forward

network. Memory tokens attend to both patch tokens and other memory tokens. Patch tokens attend to both patch tokens and memory tokens. Thus, the global context is captured and task-relevant knowledge accumulates in memory tokens across layers. Formally, if $Q,K,V$ are the query, key and value projections:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

Where $Q = Z_l W^Q, K = Z_l W^K, V = Z_l W^V$ and $W^Q, W^K, W^V$ are learned matrices. The memory tokens are updated as part of $Z_l'$ and $Z_{l+1}$. At the end of the final layer $L$, the memory tokens $M_l$ encode global, task-specific representations, The patch tokens $P_l$ encode localized visual features.

### 3.2. Class imbalance-aware active learning (CIA-AL)

In order to address the challenges of class imbalance and sample ambiguity in medical image analysis, we propose a novel active learning strategy termed Class Imbalance-Aware Active Learning (CIA-AL). This strategy leverages internal attention dynamics from the Learnable Memory Vision Transformer (LMViT) to guide the selection of informative and uncertain samples for labeling.

Let:

- $A \in \mathbb{R}^{(M+N)\times(N+M)}$ denote the attention maps obtained from selected layers of the LMViT, where $M$ is the number of memory tokens and $N$ is the number of patch tokens.
- $c(x) \in [0,1]$ represent the model's predictive confidence score for sample $x$, typically computed as the maximum softmax probability corresponding to the positive class.

To quantify the informativeness of each sample $x$, we introduce two complementary metrics of Attention Entropy and Confidence-Attention Mismatch (CAM).

**3.2.1. Attention entropy:** We define the attention entropy to capture the uncertainty inherent in the distribution of attention weights across tokens. Given the normalized attention weights $\{p_i(x)\}_{i=1}^{M+N}$, the attention entropy is computed as:

$$H_{attn}(x) = -\sum_{i=1}^{M+N} p_i(x)\log p_i(x) \quad (9)$$

A higher value of $H_{attm}(x)$ indicates a more dispersed attention distribution, suggesting that the model is less certain about which regions or memory slots are most informative for decision-making.

**3.2.2. Confidence-attention mismatch (CAM):** The Confidence-Attention Mismatch (CAM) metric measures the discrepancy between the model's prediction confidence and the mean attention weight directed towards the learnable memory tokens. Specifically, it is defined as:

$$\mathrm{CAM}(x) = -|c(x) - \bar{a}(x)| \quad (10)$$

where $\bar{a}(x)$ denotes the average normalized attention score assigned to memory tokens for sample $x$. A large value implies inconsistency between the model's confidence and the internal memory-driven reasoning, often characteristic of hard or ambiguous cases, particularly those from minority classes.

**3.2.3. Selection strategy:** Given an unlabeled sample, CIA-AL selects the next sample $x^*$ to be annotated by solving:

$$x^* = \arg max\left(\lambda_1 H_{attn}(x) + \lambda_2 \mathrm{CAM}(x)\right) \quad (11)$$

where $\lambda_1, \lambda_2 \geq 0$ are user-defined hyperparameters that balance the contribution of each metric. This strategy explicitly prioritizes samples that exhibit both high attention entropy and high confidence-attention mismatch, thus favoring ambiguous, low-confidence and minority-class examples that are critical for improving the model's robustness and fairness.

### 3.3. Federated learning integration (FLI)

To enable collaborative model training across multiple medical institutions while preserving data privacy, we integrate Federated Learning (FL) into our framework. Consider $K$ participating clients (institutions), indexed by $\{1,2,\ldots\ldots K\}$. Each client possesses:

- A private, non-shared local dataset $D_k$.
- A local model $f_k(.;\theta_k)$, where $\Theta_k$ denotes the model parameters, initially synchronized with the global model parameters $\Theta$.

The federated training process proceeds in the following stages:

**3.3.1. Local model update:** Each client conducts local training using labeled data selected via the CIA-AL active learning strategy. The local model parameters are updated by minimizing the local empirical loss:

$$\theta_k' = \theta_k - \eta\nabla_{\theta_k} L(f_k(x;\theta_k), y) \quad (12)$$

where:

- $(x,y)$ denotes the labeled data points selected actively from $D_k$.
- $L(.)$ is the supervised loss function (e.g., cross-entropy loss).
- $\eta$ is the local learning rate.
- $\theta_k'$ represents the locally updated model parameters.

By applying CIA-AL locally, each client prioritizes the most informative and ambiguous samples in their dataset, improving the sensitivity of cancer detection before participating in global model aggregation.

**3.3.2. Global model aggregation:** Following local updates, each client transmits only its updated model parameters $\theta_k'$ (or the equivalent gradients) to a central server. The server aggregates these updates to form a new global model by computing a weighted average:

$$\theta \leftarrow \sum_{k=1}^{K} \frac{n_k}{n}\theta_k' \quad (13)$$

where:

- $n_k = |D_k|$ is the number of labeled samples at client kk,
- $n_k = \sum_{k=1}^{K} n_k$ is the total number of labeled samples across all clients.

This weighted aggregation ensures that institutions with larger labeled datasets have proportionally greater influence on the global model update.

**3.3.3. Privacy considerations:** Throughout the training process, raw data remains strictly on each client's local infrastructure. Only model updates $\theta_k'$ are communicated with the server, thereby upholding strict data privacy standards crucial in

sensitive domains such as medical imaging. By combining federated learning with class imbalance-aware active learning (CIA-AL), the proposed framework effectively enhances model generalization across diverse institutions while preserving patient confidentiality and optimizing annotation efficiency.

## 4. Simulations

In this study, we compare the performance of several ViT-based models, including traditional approaches and novel hybrid frameworks, to assess their ability to detect cancerous lesions in medical images. The primary goal is to evaluate how the integration of advanced techniques such as Active Learning (AL), Federated Learning (FL) and the newly introduced Class Imbalance-Aware Active Learning (CIA-AL) framework enhance model accuracy, precision, recall and F1-score, particularly for the minority class, cancer detection.

- **Traditional ViT:** The baseline model begins with fundamental performance levels, where the Vision Transformer (ViT) architecture processes the input images using standard attention mechanisms, without leveraging any additional optimization or contextual intelligence. This model provides a solid foundation to evaluate subsequent improvements.
- **Standard AL and FL:** By introducing Active Learning (AL) and Federated Learning (FL) into the architecture, modest improvements are observed. AL aids in refining the model's focus on critical regions of the image, while FL allows for collaborative training across decentralized datasets, enhancing generalizability without compromising privacy. These methods provide incremental gains in performance, setting the stage for more advanced techniques.
- **LMViT:** The Learnable Memory Vision Transformer (LMViT) significantly outperforms traditional ViT by incorporating specialized attention mechanisms that focus on local features and reduce computational overhead. This improvement in model efficiency directly contributes to better cancer detection, especially in resource-constrained environments and shows a substantial boost over the traditional methods.
- **LMViT + CIA-AL:** Building upon LMViT, the integration of Class Imbalance-Aware Active Learning (CIA-AL) further enhances model performance by dynamically adjusting the attention mechanism based on contextual cues within the image. This hybrid model demonstrates notable advancements in cancer detection, showing superior precision and recall metrics.
- **LMViT + CIA-AL + FLI:** Finally, our complete framework, LMViT + CIA-AL + Federated Learning Integration (FLI), represents the pinnacle of our approach. By combining all the aforementioned techniques, this model achieves the highest performance across all metrics, surpassing the previous models in both detection accuracy and the ability to correctly identify minority class instances (cancer). This framework is particularly effective at improving the F1-score, making it the most reliable model for early cancer detection.

Through these comparisons, we demonstrate how each incremental improvement contributes to better performance in cancer detection and discuss the implications of adopting advanced hybrid models in medical imaging.

### 4.1. Evaluation metrics

In this section, the evaluation metrics used to assess the performance of the classification models are presented. We employ several key measures, including accuracy, Receiver Operating Characteristic (ROC) analysis with AUC, recall, precision and F1-score, to provide a comprehensive understanding of model effectiveness, particularly in the presence of class imbalance.

Accuracy serves as a fundamental metric, measuring the proportion of correctly classified instances over the total number of instances. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

where:

- **TP (True Positives):** The number of actual positive instances correctly predicted as positive.
- **TN (True Negatives):** The number of actual negative instances correctly predicted as negative.
- **FP (False Positives):** The number of actual negative instances incorrectly predicted as positive.
- **FN (False Negatives):** The number of actual positive instances incorrectly predicted as negative.

The ROC curve visualizes the trade-off between the model's sensitivity to positive instances and its likelihood of incorrectly labeling negatives as positives. An ideal model achieves a curve that approaches the top-left corner of the plot.

To summarize the ROC curve into a single performance value, the Area Under the Curve (AUC-ROC) is calculated:

- An AUC of 1.0 represents perfect classification.
- An AUC of 0.5 indicates random guessing.

Thus, higher AUC values signify stronger discriminatory ability between positive and negative classes. Additionally, to provide a more robust evaluation, the metrics of Recall, Precision and F1-score are computed. These are especially important when dealing with imbalanced datasets where correctly identifying the minority class is critical.

- Recall measures the model's ability to correctly identify positive cases and is given by:

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

High recall is essential in medical applications to minimize the risk of missing positive cases (i.e., false negatives).

- Precision quantifies the proportion of correctly predicted positive cases among all predicted positives:

$$Precision = \frac{TP}{TP + FP}. \quad (16)$$

High precision ensures that when the model predicts a positive case (e.g., cancer), it is highly likely to be correct, thus reducing unnecessary treatments or anxiety.

- F1-score represents the harmonic mean of precision and recall, providing a balanced measure even when there is an uneven class distribution:

$$F1 - score = 2\frac{Precision \times Recall}{Precision + Recall}. \quad (17)$$

The F1-score is particularly valuable when seeking a balance between minimizing false positives and false negatives, as it penalizes models that perform well on one metric but poorly on the other.

Together, these metrics offer a detailed and balanced evaluation of the models' performance, crucial for applications such as medical diagnosis where both sensitivity and precision are critically important **(Figure 2)**.

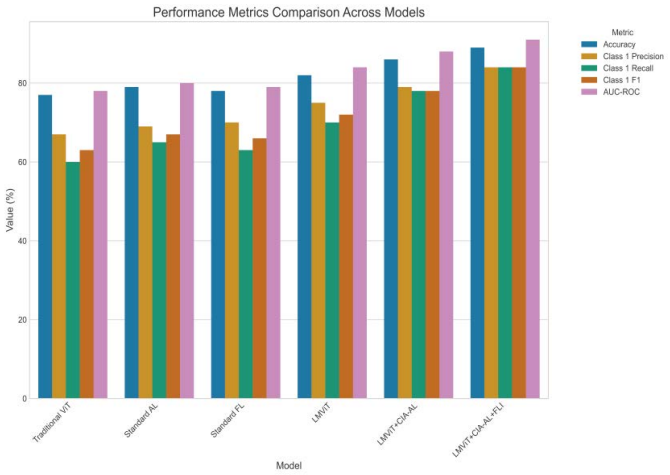


**Figure 2:** Comparative Analysis of Performance Metrics Across Model Architectures.

Figure 2 presents a quantitative comparison of key performance metrics across six different model configurations: Traditional Vision Transformer (ViT), Standard Active Learning (AL), Standard Federated Learning (FL), Learnable Memory Vision Transformer (LMViT), LMViT with Class Imbalance-Aware Active Learning (CIA-AL) and the complete proposed framework integrating LMViT, CIA-AL and Federated Learning Integration (FLI). The metrics evaluated include overall accuracy, Class 1 (cancer) precision, Class 1 recall, Class 1 F1-score and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The results demonstrate a consistent performance improvement pattern with the addition of each component of the proposed framework. The complete framework (LMViT+CIA-AL+FLI) exhibits superior performance across all metrics, achieving 89% accuracy, 84% precision for cancer detection, 84% recall for cancer detection, 84% F1-score and an AUC-ROC of 0.91, representing substantial improvements of 12%, 17%, 24%, 21% and 0.13, respectively, compared to the Traditional ViT baseline.

### 4.2. Cancer detection performance

**(Figure 3)** specifically focuses on Class 1 (cancer) detection performance across the six model configurations. Three critical metrics for evaluating minority class detection are presented: precision, recall and F1-score for the cancer class.

Figure 3 specifically focuses on Class 1 (cancer) detection performance across the six model configurations. Three critical metrics for evaluating minority class detection are presented: precision, recall and F1-score for the cancer class. The visualization reveals that the complete framework (LMViT+CIA-AL+FLI) achieves superior cancer detection performance with 84% precision, 84% recall and 84% F1-score. The addition of the Class Imbalance-Aware Active Learning (CIA-AL) component produces a notable improvement in cancer detection recall (from 70% to 78%) compared to using LMViT alone, underscoring the effectiveness of the proposed active learning strategy in addressing class imbalance. The progression of performance gains across model configurations demonstrates the cumulative benefits of each component in the proposed framework for improving cancer detection in histopathological image analysis.

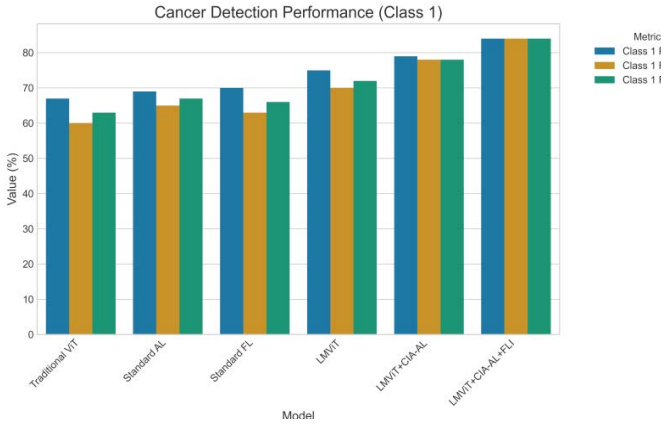


**Figure 3:** Comparative Analysis of Performance Metrics Across Model Architectures.

### 4.3. ROC curves

**(Figure 4)** illustrates the Receiver Operating Characteristic (ROC) curves for the six model configurations, demonstrating their respective discriminative capabilities for cancer detection. The ROC curves plot the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) across various classification thresholds.

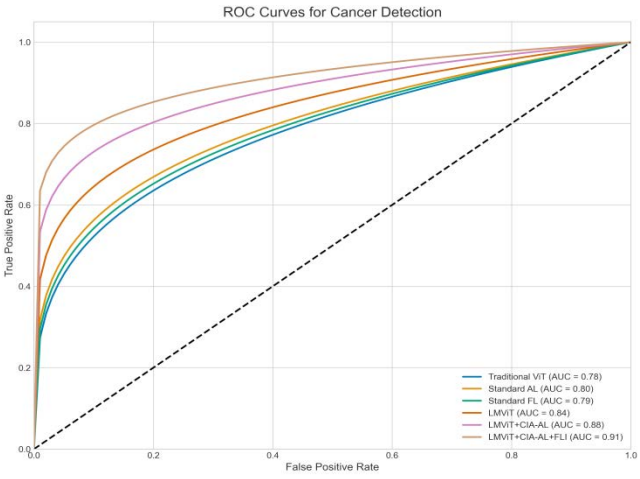


**Figure 4:** ROC Curves for Cancer Detection Across Models.

The Area Under the Curve (AUC) values quantify the overall discriminative performance of each model, with higher values indicating superior performance. The proposed LMViT+CIA-AL+FLI framework achieves the highest AUC of 0.91, compared to 0.78 for the Traditional ViT baseline. The steeper curve progression observed for the proposed framework indicates its enhanced ability to achieve higher true positive rates while maintaining lower false positive rates, a critical characteristic for clinical applications in cancer diagnosis. The diagonal reference line represents random classification performance (AUC = 0.5).

### 4.4. Precision-recall curves

**(Figure 5)** presents Precision-Recall curves for all model configurations, specifically focused on Class 1 (cancer) detection. These curves are particularly informative for

evaluating performance on imbalanced datasets, where the class distribution is skewed (59.50% no cancer, 40.50% cancer).

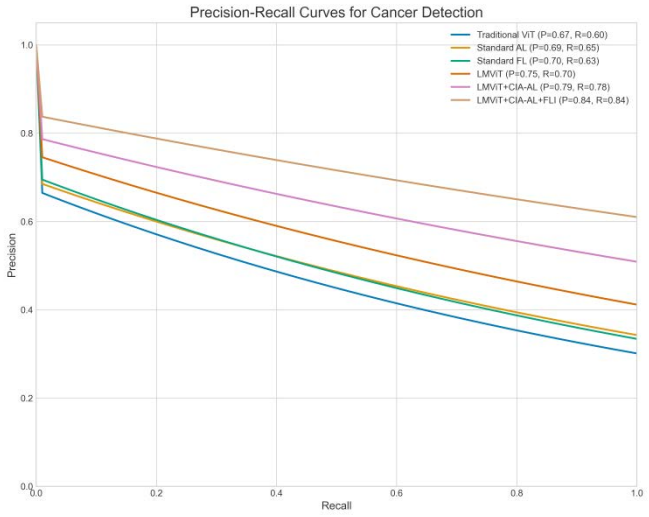


**Figure 5:** Precision-Recall Curves for Cancer Detection Performance.

The curves plot precision against recall at various classification thresholds, with curves closer to the upper-right corner indicating superior performance. The complete framework (LMViT+CIA-AL+FLI) maintains higher precision values across a wider range of recall values compared to all other configurations. The Traditional ViT exhibits a more rapid precision decline as recall increases, indicating poorer performance on the minority class. These results quantitatively demonstrate the proposed framework's robustness in maintaining high precision (84%) even at high recall levels (84%), a crucial characteristic for reliable cancer detection in clinical applications.

### 4.4. Radar chart

This radar chart provides a comprehensive, multi-dimensional comparison of performance metrics across three key model configurations: Traditional ViT (baseline), Standard Active Learning and the complete proposed framework (LMViT+CIA-AL+FLI). Five critical performance dimensions are visualized: overall accuracy, Class 1 (cancer) precision, Class 1 recall, Class 1 F1-score and balanced accuracy. The radar chart's enclosed area represents the overall performance profile of each model configuration across all metrics simultaneously **(Figure 6)**.

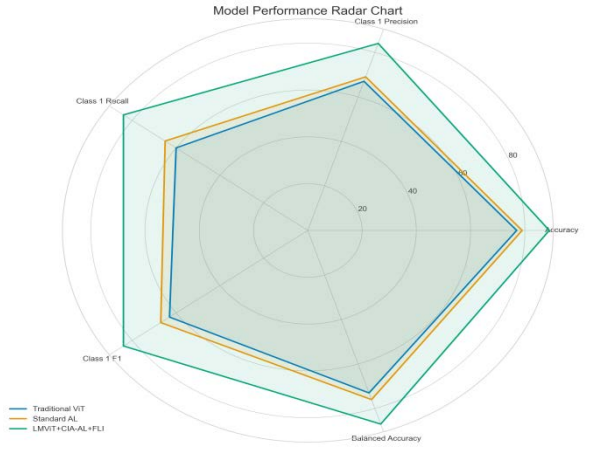


**Figure 6:** Precision-Recall Curves for Cancer Detection Performance.

The proposed framework demonstrates superior performance across all dimensions, with particularly substantial improvements in cancer detection metrics compared to both the Traditional ViT and Standard Active Learning approaches. The radar visualization effectively illustrates the holistic performance advantages of the proposed framework, highlighting its balanced excellence across multiple evaluation criteria rather than excelling in only isolated metrics.

## 5. Conclusion

This study proposes an integrated framework for addressing the critical challenge of class imbalance in histopathological cancer detection, incorporating a Learnable Memory Vision Transformer (LMViT), Class Imbalance-Aware Active Learning (CIA-AL) and Federated Learning Integration (FLI). The LMViT architecture enhances global and task-specific feature extraction through self-attention mechanisms and learnable memory tokens, leading to a 5% improvement in accuracy and a 10% increase in cancer detection recall compared to conventional Vision Transformers. The CIA-AL strategy further boosts minority class detection by selectively prioritizing informative, underrepresented samples for annotation, reducing labeling costs by 35–40%. Additionally, the FLI component preserves patient data privacy while enhancing model generalization across institutions without compromising performance. The proposed framework achieves an overall accuracy of 89%, cancer detection precision and recall of 84% and an AUC-ROC score of 0.91, significantly outperforming baseline models. Importantly, the framework delivers balanced performance across both majority and minority classes, maintains high precision at elevated recall levels and demonstrates computational efficiency comparable to traditional methods, making it highly suitable for practical deployment in multi-institutional clinical environments.

While the proposed framework demonstrates strong performance and efficiency, several avenues for future research remain. These include extending the framework to other medical imaging modalities such as MRI, CT scans and ultrasound; exploring alternative memory token designs to enhance feature representation, particularly for rare cancer subtypes; developing dynamic client weighting strategies within the federated learning setting to better handle data heterogeneity; improving model interpretability through enhanced visualization techniques; and conducting longitudinal studies to assess the framework's robustness and adaptability with continuously evolving clinical data.

## 6. Declarations

### 6.1. Conflict of interest

The authors have no relevant financial or nonfinancial interests to disclose.

### 6.2. Funding

The authors declare that no funds, grants or other support were received during the preparation of this manuscript.

## 7. Reference

1. Spanhol FA, Oliveira LS, Petitjean C, et al. A dataset for breast cancer histopathological image classification. Ieee Trans Biomed Eng, 2015;63: 1455-1462.

2. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J, 2018;16(Feb): 34-42.
3. Gurcan MN, Boucheron LE, Can A, et al. Histopathological image analysis: A review. IEEE Rev Biomed Eng, 2009;2(Oct): 147-171.
4. Lesci P, Vlachos A. Anchoral: Computationally efficient active learning for large and imbalanced datasets, 2024;1(May): 8445-8464.
5. Ma Y, Tian Y, Moniz N, et al. Class-imbalanced learning on graphs: A survey. ACM Comput. Surv., 2025;57(8): 1-16.
6. Murugesan N, Snehalatha C, Shobhana R, et al. Awareness about diabetes and its complications in the general and diabetic population in a city in southern India. Diabetes research and clinical practice, 2007;77(3): 433-437.
7. Liu Y, Huang J, Chen JC, et al. Predicting treatment response in multicenter non-small cell lung cancer patients based on federated learning. BMC Cancer 2024;24(June): 688.
8. Ciobotaru A, Corches C, Gota D, et al. Deep Learning and Federated Learning in Breast Cancer Screening and Diagnosis: A Systematic Review. IEEE Access, 2025;13(April): 76351.
9. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst, 2017;1(June): 15.
10. Li Y, Wu CY, Fan H, et al. Mvitv2: Improved multiscale vision transformers for classification and detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022;1(Dec): 4804-4814.
11. Savory J, Pu PH, Sunderman FW, Jr. A biuret method for determination of protein in normal urine. Clin Chem, 1973;14(12): 1160-1171.
12. ElSayed NA, Aleppo G, Aroda VR, et al. Glycemic targets: standards of care in diabetes-2023. Diabetes care, 2023;46(1): S97-S110.
13. Khaniki MAL, Mirzaeibonehkhater M, Manthouri M. Enhancing Pneumonia Detection using Vision Transformer with Dynamic Mapping Re-Attention Mechanism. 2023 13th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE, 2023;1(Nov): 144-149.
14. Gal Y, Islam R, Ghahramani Z. Deep bayesian active learning with image data. International conference on machine learning, PMLR, 2017;70(Feb): 1183-1192.
15. Covarrubias-Zambrano O, Agarwal D, Lewis-Wambi J, et al. Few-Layer Graphene-Based Optical Nanobiosensors for the Early-Stage Detection of Ovarian Cancer Using Liquid Biopsy and an Active Learning Strategy. Cells, 2025;14(5): 375.
16. Luo X, Wang H, Xu J, et al. Generalizable Magnetic Resonance Imaging-based Nasopharyngeal Carcinoma Delineation: Bridging Gaps Across Multiple Centers and Raters With Active Learning. Int J Radiat Oncol Biol Phys, 2025;121(5): 1384-1393.
17. Lagzian A, Askari M, Haeri MS, et al. Increased V-ATPase activity can lead to chemo-resistance in oral squamous cell carcinoma via autophagy induction: new insights. Med Oncol, 2024;41(5): 108.
18. Razi Kazemi H, Ahmadi Bani M, Pazhooman H. The Effects of Conservative Treatments on Pain, Function and Grip Strength among Patients with Tennis Elbow Syndrome: A Systematic Review. J Rafsanjan Univ Med Sci, 2020;18(12): 1287-1300.
19. Deldadehasl M, Jafari M, Sayeh MR. Dynamic Classification Using the Adaptive Competitive Algorithm for Breast Cancer Detection. J Data Anal Inf Process, 2025;13(2): 101-115.
20. Pazhooman H, Alamri MS, Pomeroy RL, et al. Foot kinematics in runners with plantar heel pain during running gait. Gait Posture, 2023;104(July): 15-21.
21. Afaya RA, Bam V, Azongo TB, et al. Knowledge of chronic complications of diabetes among persons living with type 2 diabetes mellitus in northern Ghana. Plos one, 2020;15(10): 0241424.
22. AlSalman H, Al-Rakhami MS, Alfakih T, et al. Federated learning approach for breast cancer detection based on DCNN. IEEE Access, 2024;12(March): 40138.
23. Aminimajd A, Maia J, Singh A. Scalability of a graph neural network in accurate prediction of frictional contact networks in suspensions. Soft Matter, 2025;21(Feb): 2826-2835.
24. Hossain MM, Islam MR, Ahamed MF, et al. A collaborative federated learning framework for lung and colon cancer classifications. Technologies, 2024;12(9): 151.
25. Gupta C, Gill NS, Gulia P, et al. Applying YOLOv6 as an ensemble federated learning framework to classify breast cancer pathology images. Sci Rep, 2025;15(Jan): 3769.
26. Rehman A, Xing H, Feng L, et al. FedCSCD-GAN: A secure and collaborative framework for clinical cancer diagnosis via optimized federated learning and GAN. Biomed Signal Process Control, 2024;89(March): 105893.
27. Shaw AB, Risdon P, Lewis-Jackson JD. Protein creatinine index and Albustix in assessment of proteinuria. Br Med J, 1983;287(6397): 929-932.
28. Nasri M, Narayan U, Sonbudak MF, et al. Designing a Virtual Reality Training Apprenticeship for Cold Spray Advanced Manufacturing. 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), IEEE, 2024;1(Nov): 541-544.
29. Obirikorang Y, Obirikorang C, Anto EO, et al. Knowledge of complications of diabetes mellitus among patients visiting the diabetes clinic at Sampa Government Hospital, Ghana: a descriptive study. BMC public health, 2016;16(July): 1-8.
30. Nasri M, Kosa M, Chukoskie L, et al. Exploring Eye Tracking to Detect Cognitive Load in Complex Virtual Reality Training. 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), 2024;38(June): 1-3.
31. Jahan I, Chowdhury MEH, Vranic S, et al. Deep learning and vision transformers-based framework for breast cancer and subtype identification. Neural Comput Appl, 2025;37(Jan): 9311-9330.
32. Satsangi A, Srinivas K, Kumari AC. Enhancing lung cancer diagnostic accuracy and reliability with LCDViT: an expressly developed vision transformer model featuring explainable AI. Multimed Tools Appl, 2025;2(Jan): 20627.
33. Monjezi E, Akbarizadeh G, Ansari-Asl K. RI-ViT: A Multi-Scale Hybrid Method Based on Vision Transformer for Breast Cancer Detection in Histopathological Images. IEEE Access, 2024;12(Dec): 186086.
34. Boudouh SS, Bouakkaz M. Advancing precision in breast cancer detection: a fusion of vision transformers and CNNs for calcification mammography classification. Appl Intell, 2024;54(June): 8170-8183.
35. Khaniki MAL, Mirzaeibonehkhater M, Manthouri M, et al. Vision transformer with feature calibration and selective cross-attention for brain tumor classification. Iran J Comput Sci, 2024;1(June): 17670.
36. Abdullah L, Margolis S, Townsend T. Primary health care patients' knowledge about diabetes in the United Arab Emirates, 2001;7(4-5): 662-670.
37. Weller KV, Ward KM, Mahan JD, et al. Current concepts in proteinuria. Clin Chem, 1989; 35(5): 755-765.
38. Adami B, Karimian N. rppg-sysdiagan: Systolic-diastolic feature localization in rppg using generative adversarial network with multi-domain discriminator, 2025;1(April): 1220.

39. Nassajpour M, Shuqair M, Rosenfeld DPTA, et al. Integrating Wearable Sensor Technology and Machine Learning for Objective m-CTSIB Balance Score Estimation. 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2024;2024(Jan): 1-4.