

TRANSFORMER-BASED OTITIS MEDIA CLASSIFICATION: A COMPARATIVE STUDY OF ViT, DeiT, AND PVT ARCHITECTURES IN OTOSCOPIC IMAGE ANALYSIS

Sedat ORENC 

Batman University, Electrical Electronics Engineering Department, Batman, Türkiye,
sedat.orenc@batman.edu.tr

KEYWORDS

Image processing
Deep learning
Transformer models
Vision transformer
Data-efficient image transformer
Pyramid vision transformer

ARTICLE INFO

Research Article

DOI:

[10.17678/beuscitech.1920012](https://doi.org/10.17678/beuscitech.1920012)

Received March 31, 2026

Revised May 11, 2026

Accepted June 3, 2026

Year 2026

Volume 16

Issue 1

Pages 53-73



ABSTRACT

Otitis media (OM) is one of the most prevalent middle ear diseases globally. It represents a major clinical and public health burden. It imposes significant health and socio-economic challenges, particularly among pediatric populations. Conventional diagnostic tools, such as otoscopy and tympanometry, are constrained by inherent subjectivity and operator dependency. Furthermore, they suffer from low specificity. This underscores the critical need for reliable, scalable, and automated diagnostic solutions. Recent advances in deep learning have improved image-based diagnosis; however, traditional Convolutional Neural Networks (CNNs) remain constrained by their inability to capture long-range dependencies essential for distinguishing subtle tympanic membrane pathologies. To address these limitations, this study systematically benchmarks three transformer-based architectures: Vision Transformer (ViT), Data-Efficient Image Transformer (DeiT), and Pyramid Vision Transformer (PvT). The objective is to automate the classification of otoscopic images. Classification targets include Acute Otitis Media (AOM), Chronic Otitis Media (COM), and normal cases. A balanced dataset of 1,800 images was curated and augmented. This ensured a fair evaluation under standardized training conditions. The experimental results demonstrate that ViT performs well for AOM, achieving an accuracy and F1-score of 0.98. However, its performance declines for chronic and normal cases. In contrast, DeiT produces the most consistent results across all categories, achieving near-perfect accuracy of 1.00 for acute OM cases and 0.96 for chronic and normal cases. PvT also demonstrates strong performance, achieving an accuracy of 1.00 for OM and 0.99 for normal. These findings demonstrate the superior robustness and clinical potential of DeiT and PvT compared to ViT, suggesting their suitability for real-world applications. Beyond delivering a reproducible benchmarking framework, this work contributes toward bridging the gap between algorithmic innovation and clinical translation. This provides a way of achieving reliable, interpretable and scalable AI-assisted diagnosis of otitis media.

1. INTRODUCTION

Otitis media, which encompasses both acute and chronic forms, is one of the most prevalent middle ear diseases worldwide, particularly affecting children [1]. The disease poses a major clinical and public health burden, as untreated or recurrent OM can lead to severe complications, including conductive hearing loss, language development delays, impaired academic performance, and even intracranial infections [2]. Beyond the medical consequences, OM imposes a substantial socio-economic impact, with healthcare expenditures and productivity losses disproportionately affecting low- and middle-income countries [3]. AOM is a common childhood disease, affecting about two-thirds of children under 3 years old, with one-third experiencing recurrent episodes (>3 in 6 months or >4 in 1 year). In Europe, the incidence of AOM in outpatient care is 256 episodes per 1000 person-years, with a peak of 22.2 episodes per 100 children in Italy. Acute otitis media (AOM) is the main cause of antibiotic prescriptions in children and carries significant economic costs, ranging from €332 per episode in the Netherlands to €752 in the United Kingdom. These epidemiological facts and figures highlight the urgent need for accurate, accessible, and scalable diagnostic solutions [4].

Conventional diagnostic methods for OM include otoscopy and tympanometry. In more advanced cases, computed tomography (CT) or magnetic resonance imaging (MRI) can be used [5]. Although otoscopy is simple and cost-effective, it is inherently limited by operator dependency and subjectivity. Even experienced otolaryngologists may misinterpret subtle pathological findings such as mild tympanic membrane bulging, early effusion, or minute perforations [6]. Tympanometry provides objective information on middle ear function, yet its diagnostic specificity is limited in differentiating between acute and chronic conditions. Advanced imaging modalities, while offering detailed structural information, are costly, time-consuming, and unsuitable for routine clinical screening [7]. Consequently, the lack of standardized, objective, and reliable diagnostic practices can lead to both over- and under-diagnosis of OM, often resulting in delayed treatment or unnecessary interventions. These diagnostic challenges highlight the potential of computational image analysis and artificial intelligence to enhance clinical decision-making processes [8].

With the rapid development of artificial intelligence (AI), [9] deep learning-based medical image analysis has emerged as a promising solution for OM diagnosis. CNNs have been widely applied to otoscopic image classification, and it achieved remarkable success in feature extraction and disease recognition [10]. However, CNNs are limited by their small receptive fields, which prevent them from modelling long-range dependencies and global structural patterns. These are crucial for identifying subtle pathological changes in the tympanic membrane [11]. To overcome these limitations, recent studies have increasingly adopted transformer-based architectures. By leveraging self-attention mechanisms, these models can capture both local texture details and global contextual information simultaneously.

Despite their success, CNNs have intrinsic limitations [12]. Their receptive fields are inherently local, restricting the ability to capture long-range dependencies and global structural relationships within images [13]. This limitation is particularly critical in OM diagnosis, where distinguishing between acute inflammation, chronic perforations, and normal anatomical variations requires both fine-grained texture analysis and holistic structural understanding. Moreover, CNNs often require large-scale datasets to achieve generalizable performance, posing challenges in medical imaging domains where data availability is limited [14]. These shortcomings have motivated the exploration of novel architectures that can overcome the local bias of convolutional models.

Transformer-based architectures, initially introduced in the field of natural language processing, have recently gained traction in computer vision due to their self-attention mechanisms [15]. Unlike CNNs, transformers can simultaneously model local and global dependencies, capturing both texture-level and structural-level features. The ViT marked a breakthrough by demonstrating that, when trained on sufficient data, transformers can rival or surpass CNNs in image classification. Subsequent variants, such as the DeiT and the PvT, have further optimized training efficiency, reduced computational cost, and adapted transformer architectures to medical imaging applications. Their ability to integrate multi-modal information, such as combining otoscopic images with clinical metadata, further enhances diagnostic robustness and clinical interpretability [16].

However, there are still several critical gaps. Firstly, there is a lack of comparative benchmarks for different transformer families, particularly within standardized, balanced otoscopic datasets. Without systematic evaluations, it is difficult to determine the relative strengths and limitations of each architecture in practical diagnostic scenarios. Secondly, many existing studies rely on small or imbalanced datasets, limiting reproducibility and generalizability. Thirdly, there is a lack of standardized experimental protocols, including consistent preprocessing pipelines, augmentation strategies, and evaluation metrics, which hinders fair comparison across studies. Finally, although recent works emphasize model accuracy, relatively few studies address clinical interpretability and real-world deployment, both of which are essential for translating research findings into clinical practice.

The present study seeks to address these gaps by providing a systematic evaluation of three representative transformer architectures for automated OM classification. A carefully curated and balanced dataset of otoscopic images was used. This dataset included cases of acute and chronic otitis media, as well as normal cases. Uniform preprocessing and augmentation were applied to ensure consistency. All models were trained under identical hyperparameter settings, enabling fair comparison across architectures. Performance was assessed using a comprehensive suite of metrics, including accuracy, precision, recall, and F1-score, to capture both overall predictive performance and sensitivity to misclassification. By establishing a reproducible benchmarking framework, this work not only advances the state of the art in OM classification but also provides practical insights into the diagnostic potential and clinical applicability of transformer-based models. Ultimately, this study helps to bridge the gap between algorithmic innovation and real-world medical practice by offering a way to make AI-assisted diagnosis of otitis media more reliable, interpretable, and scalable.

Several promising areas of research could further enhance the role of AI in OM diagnosis. First, future studies should focus on the integration of multimodal data, including otoscopic images, audiometry, patient history, and laboratory findings, to improve diagnostic robustness and reduce ambiguity. Second, larger and more diverse datasets collected across multiple populations and clinical settings are necessary to ensure generalizability and fairness, particularly given the global burden of OM. Third, explainable AI (XAI) approaches must be incorporated into transformer-based pipelines to improve interpretability and foster trust among clinicians, making AI-assisted systems more acceptable for routine clinical use. Finally, lightweight transformer architectures and hardware-efficient optimizations should be explored to enable deployment in resource-limited environments, such as rural clinics or mobile health units. By addressing these future directions, researchers can move toward building comprehensive, interpretable, and globally deployable AI frameworks that not only improve diagnostic accuracy but also contribute to equitable and sustainable healthcare solutions. The main contributions of this study can be summarized as follows:

- This work provides one of the comprehensive benchmarking studies of three representative transformer models, ViT, DeiT, and PvT, for otitis media classification using otoscopic images.
- A carefully constructed dataset was employed, including acute otitis media, chronic otitis media, and normal cases, ensuring class balance and standardized preprocessing to eliminate biases frequently encountered in prior OM studies.
- All models were trained and tested under identical hyperparameters, augmentation techniques, and evaluation protocols, enabling fair performance comparison and addressing the lack of standardization in existing OM literature.
- This study moves beyond accuracy metrics to address the translational gap between deep learning research and its clinical implementation by establishing a reproducible benchmarking framework that emphasises interpretability and deployment considerations.

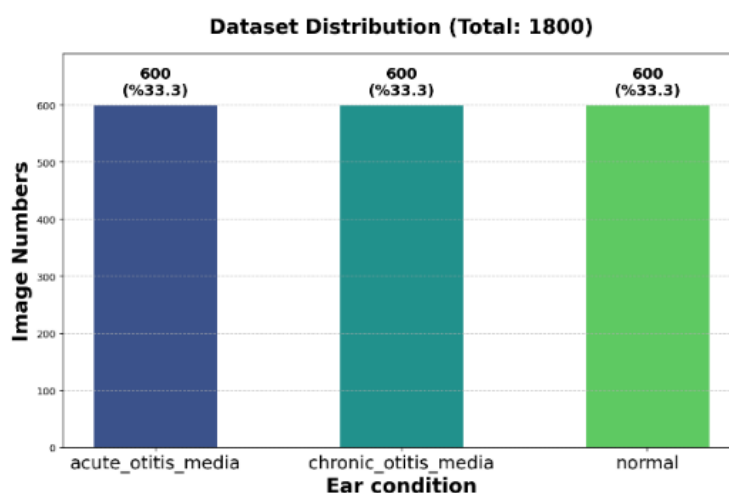


Figure 1. The number of data points distributed across the otitis media and healthy control classes.

The dataset consists of 1,800 images evenly distributed across three diagnostic categories: acute otitis media, chronic otitis media, and normal controls, as it is shown in Figure 1. This balanced distribution is advantageous for deep learning model development, as it mitigates class imbalance and reduces the risk of bias toward any particular category. The bar chart visually confirms this uniform distribution, with the x-axis labeled ear condition to reflect the clinical nature of the analyzed classes.

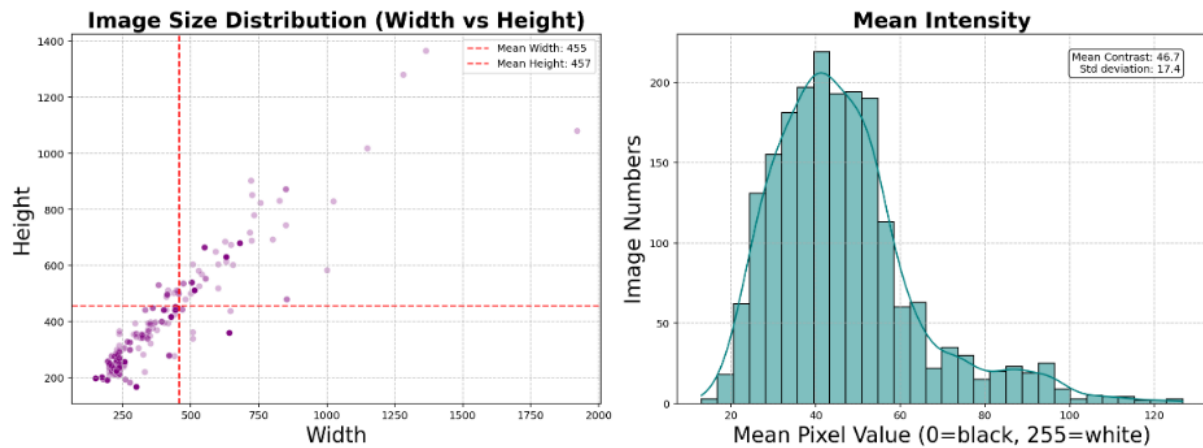


Figure 2. The number of data points distributed across the otitis media and healthy control classes

The visual analysis reveals substantial variability in both the spatial dimensions and intensity characteristics of the dataset images in Figure 2. The image size distribution scatter plot indicates a positive correlation between image width and height, with most samples clustered within a lower-resolution range of approximately 200–500 pixels. Nevertheless, several outliers exhibit considerably larger dimensions, reaching up to approximately 1900 pixels in width and 1400 pixels in height.

The red dashed lines indicate the dataset's average dimensions, corresponding to a mean width of 455 pixels and a mean height of 457 pixels. These values suggest that the aspect ratio of most images is square or nearly square.

The mean intensity histogram illustrates the distribution of pixel brightness values and displays a right-skewed pattern. The majority of images have mean pixel intensities between 30 and 50 on a scale ranging from 0 to 255, indicating predominantly dark images. The distribution has a mean intensity of 46.7 and a standard deviation of 17.4. This suggests relatively consistent lighting conditions overall, alongside a notable variation in contrast levels that must be addressed during deep learning model training.

2. RELATED WORKS

The classification of OM using medical imaging has rapidly evolved from traditional CNN-based methods to more advanced transformer-driven architectures. These models leverage self-attention mechanisms to capture local textures and global structural features in otoscopic images, addressing limitations in earlier frameworks for distinguishing between normal ears, AOM, and COM. We classify prior studies into three subclasses below, based on the transformer variants employed: ViT, DeiT, and PvT. Although ViT has been adopted more widely in OM diagnostics, DeiT and PvT applications are still in their infancy. However, they show great potential due to their efficiency and hierarchical designs, despite the lack of domain-specific studies.

2.1. Vision Transformer-Based Approaches

Recently, ViT architectures have gained significant attention in OM diagnosis due to their ability to capture global dependencies and structural variations in tympanic membrane images. Across the literature, ViT-based models consistently demonstrate superior performance compared to conventional CNNs, although challenges remain in terms of dataset scale, heterogeneity, and real-world generalizability. Cömert et al. (2024) developed a computerized otoscopy model that extracts deep features from ViT, followed by grid search optimization and support vector machine (SVM) classification, achieving an accuracy of 99.37% in multi-class OM detection across normal, AOM, and COM categories. This hybrid method emphasizes cost reduction and robustness on public datasets like the Ear Imagery dataset with 880 images [17]. Similarly, Demircan et al. (2025) proposed a ViT-based feature extractor combined with machine learning classifiers, including k-nearest neighbors (kNN), SVM, and random forest, attaining 99.00% accuracy with kNN on a dataset of 880 eardrum images divided into CSOM, earwax, myringosclerosis, and normal classes. The approach focuses on preprocessing images to 420×380 pixels and providing second-opinion support for ENT specialists to minimize human error [18]. Deva et al. (2025) introduced a feature fusion framework utilizing ViT alongside symptom encoding to integrate otoscopic images with clinical metadata, surpassing CNN baselines in accuracy for classifying normal, AOM, and COM cases from the Kaggle "Otoscope Data" dataset. This multimodal strategy employs data augmentation and adaptive learning to handle image variability and sparse annotations, promoting cost-effective healthcare access [19]. Khan

et al. (2024) presented the Sumotosima framework, a transformer-based pipeline combining triplet and cross-entropy losses for classification, followed by summarization using Knowledge Enhanced Multimodal BART on a custom OCASD dataset of 500 images across five categories, yielding 98.03% accuracy. The model prioritizes resource efficiency and patient-friendly summaries, addressing both diagnostic and interpretive needs [20]. In a complementary convolutional deep learning study, Dubois et al. (2024) validated a smartphone-based system employing an Inception-v2 model on 41,664 otoscopic images across 11 middle-ear conditions, achieving expert-level sensitivity (99.0%) and specificity (95.2%) for binary normal vs. abnormal classification, with near-perfect detection of wax plugs (sensitivity 100%, specificity 97.7%). This mobile-deployable tool underscores the translational potential of DL for routine OM screening, offering a benchmark for ViT's efficiency in similar resource-constrained settings, though with variable performance across finer-grained pathologies [21].

2.2. Data-Efficient Image Transformer-Based Approaches

While ViT-based studies have already demonstrated strong performance in OM classification tasks, the use of DeiT has not yet been directly explored in this domain. To the best of our knowledge, there is currently no published study applying DeiT specifically to otoscopic image classification for OM diagnosis. Existing DeiT-related works in medical imaging have focused on other modalities, such as MRI, histopathology, and multimodal fusion tasks, highlighting the model's strength in sample efficiency, robustness to noisy inputs, and integration with clinical metadata. However, the absence of DeiT applications in OM presents a clear research gap. By leveraging DeiT's data-efficient training paradigm and distillation-based learning strategy, this paper aims to be the first to classify otitis media using DeiT models, thereby extending the transformer-based diagnostic framework to a new clinical imaging application.

2.3. Pyramid Vision Transformer-Based Approaches

To date, no published study has applied PVT directly for OM diagnosis. Existing OM literature is dominated by CNN-based and ViT-based approaches, which, while effective, are limited by either their local receptive fields or their lack of hierarchical multi-scale representations. In contrast, PVT has the potential to overcome these limitations by jointly modeling global and local patterns of the tympanic membrane, offering a more robust feature extraction pipeline. This research, therefore, positions itself as the first work to investigate PVT architectures for otitis media classification. By exploiting the hierarchical representation capability of PVT, our study aims to demonstrate improved diagnostic accuracy and generalization on otoscopic datasets compared with existing transformers and CNN baselines. This fills a clear gap in the literature and extends transformer-based diagnostic frameworks to a previously unexplored model family in the OM domain.

3. MATERIALS

3.1. Dataset Description

The dataset employed in this study consists of 1,800 otoscopic images, evenly distributed across three diagnostic categories: acute otitis media (600 images), chronic otitis media (600 images), and normal healthy cases (600 images). All images are provided in JPEG format with a standardised resolution of 500 x 500 pixels to ensure uniformity for use in deep learning applications. The dataset is balanced across classes, which minimizes bias during model training and evaluation. AOM cases are characterized by signs of infection such as redness, bulging of the tympanic membrane, and fluid accumulation. COM cases present with long-term pathological features, including tympanic membrane perforation and persistent discharge, often associated with conductive hearing loss. Normal cases depict a healthy ear, represented by a pearly-grey, translucent tympanic membrane without infection, fluid buildup, or structural abnormalities. This curated dataset was specifically constructed to support the development and evaluation of automated otitis media classification models. A subset of the dataset is illustrated in Figure 3.

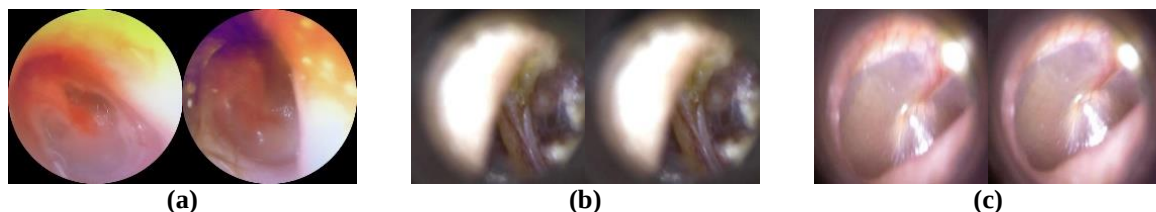


Figure 3. A subset of the Ear Image dataset: (a) Chronic otitis media, (b) Acute otitis media, (c) Normal.

Table 1. The distribution of the samples over classes after data augmentation.

Class	Training set	Validation set	Test set	Total samples
Chronic otitis media	4700	1000	1000	6700
Acute otitis media	4700	1000	1000	6700
Normal	4700	1000	1000	6700

Table 1 summarizes the final sample distribution across training, validation, and test subsets following the data preparation process. Initially, the original dataset was divided into training, validation, and testing sets using an 80%, 10%, and 10% split ratio prior to augmentation to ensure strict subset independence and eliminate the risk of information leakage. Subsequently, data augmentation was employed to increase sample diversity and improve model robustness, particularly for the training process. After augmentation, each diagnostic category, namely acute otitis media, chronic otitis media, and normal cases, contained 4,700 training images, 1,000 validation images, and 1,000 test images. This balanced class distribution minimizes class imbalance, strengthens model generalization, and provides a reliable framework for evaluating classification performance on unseen data.

3.2. Experimental Setup

All training experiments were conducted on an NVIDIA® GeForce RTX 3060 Ti GPU utilizing the CUDA® parallel computing framework, under a Windows 10 operating system. The hardware setup consisted of a PowerBoost desktop equipped with an Intel® Core™ i5-10400F processor running at 2.90 GHz and supported by 16 GB of RAM.

4. TRANSFORMER ARCHITECTURES

The methodology of this study employs three transformer-based architectures. These are the vision transformer, the data-efficient image transformer, and the pyramid vision transformer. They are used to systematically evaluate their performance on the prepared image datasets.

4.1. Vision Transformers

The Vision Transformer architecture, which is illustrated in Figure 4, was pioneered by Dosovitskiy et al. in 2020 [22], marked a significant shift in computer vision by directly applying the standard Transformer model, initially developed for Natural Language Processing (NLP) tasks, to image recognition. This approach demonstrated that reliance on CNNs is not strictly necessary for achieving state-of-the-art results in image classification, particularly when the model is pre-trained on large datasets [23]. Transformers are self-attention-based architectures renowned for their capability to individually weight the importance of each part of the input data and capture long-range dependencies and spatial correlations, which often exceed the non-local modeling capabilities of traditional CNN structures [20], [24].

The methodology begins by transforming the two-dimensional image input into a one-dimensional sequence, which is the native input format for the transformer encoder [20]. An input image $x \in \mathbf{R}^{H \times W \times C}$ where H and W are height and width, and C is the number of channels, is partitioned into fixed-size, non-overlapping 2D patches, such as 16×16 pixels, which are then flattened and linearly projected into a sequence of patch embeddings $x_p \in \mathbf{R}^{N \times (P^2 \times C)}$ [25]. Here, $N = HW / P^2$ represents the sequence length, or the total number of patches. To facilitate downstream classification, a learned classification token is prepended to this sequence, acting as the aggregated image representation. Crucially, to preserve the inherent spatial structure lost during the flattening process, positional embeddings (E_{pos}) are added to the patch embeddings. If the model is fine-tuned using a higher resolution than that used during pre-training, the pre-trained positional embeddings are typically adjusted via 2D interpolation. The complete input sequence z_0 that is subsequently fed into the transformer encoder, incorporating the classification token, embedded patches, and positional embeddings, is thus defined:

$$z_0 = [x_{class}; Ex_p^1; Ex_p^2; \dots; Ex_p^N] + E_{pos} \quad (1)$$

The core of the ViT architecture consists of a stack of L identical transformer encoder blocks, following the structure introduced by Vaswani et al. in 2017. Each layer, l , processes the input sequence z_{l-1} through two main sub-components: a Multi-Head Self-Attention (MSA) module and a position-wise Multi-Layer Perceptron (MLP) block, often referred to as a Feed-forward Neural Network (FFN). Layer Normalization (LN) is applied

before each block, and residual connections (skip connections) are added after each block to stabilize training [17]. The sequence of operations for transforming the state from layer $l-1$ to l is described by the recurrent equations.

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1} \quad (2)$$

$$z_l = MLP(LN(z'_l)) + z'_l \quad (3)$$

The Multi-Head Self-Attention (MHSA) module is the defining feature, enabling the model to jointly attend to features from multiple representation subspaces and dynamically establish long-range dependencies between all input patches (Z) [24]. For a given input vector Z , the mechanism first projects it into three conceptual vector spaces: Query (Q), Key (K), and Value (V) using corresponding learned weight matrices (W_Q , W_K , W_V). The self-attention function then computes the output by determining the normalized dot-product compatibility between Q and K , scaled by D_k , and applying the resulting weight matrix to V [17].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \quad (4)$$

where $\sqrt{D_k}$ is a scaling factor. MHSA extends this by performing the attention function concurrently h times where h is the number of heads, concatenating the results ($\text{head}_1, \dots, \text{head}_h$), and projecting them through a final linear weight matrix W_O :

$$MSA(Z) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_O \quad (5)$$

where the i -th head output is $\text{head}_i = \text{Attention}(QW_Q^i, KW_K^i, VW_V^i)$.

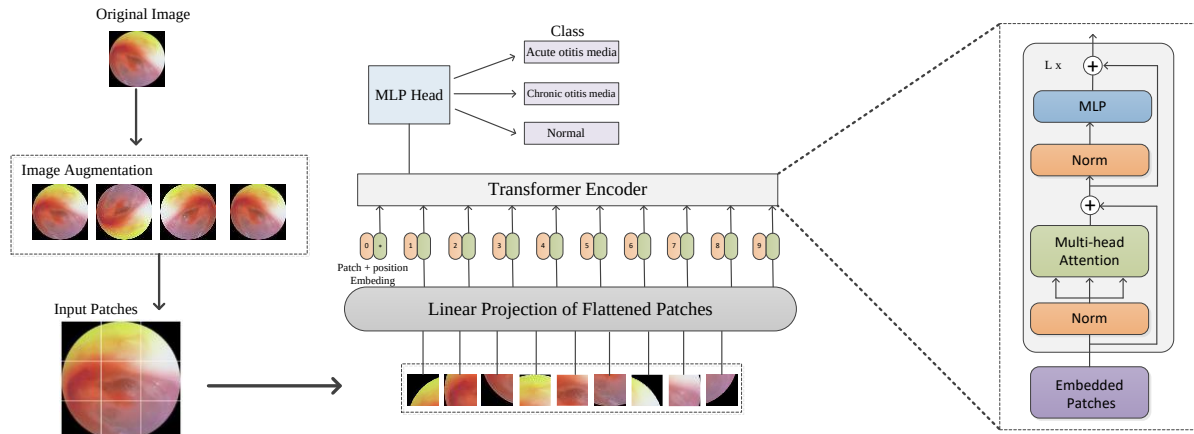


Figure 4. The Architecture of The Vision Transformer Model.

4.2. Data-Efficient Image Transformer

The DeiT architecture was introduced by Touvron and colleagues [26] in their foundational work on "Training data-efficient image transformers & distillation through attention". This model was developed to directly address the key limitation of the original ViT architecture. It depends on large-scale datasets to perform comparably to CNNs [16]. DeiT successfully reframed the scalability debate by demonstrating that competitive vision performance could be attained by pre-training solely on mid-sized datasets like ImageNet. The core innovation enabling this data efficiency is the integration of a teacher-student knowledge distillation paradigm. This transfer learning approach leverages the knowledge captured by a highly efficient teacher model to guide the training of the smaller DeiT student model. DeiT maintains a pure transformer architecture, meaning it achieves visual recognition without incorporating any convolutional layers. This methodology leads to better accuracy than the classic ViT while simultaneously reducing computational complexity and parameter size [24].

The DeiT structure is built upon the ViT encoder framework, utilizing a deep stack of self-attention layers to process image sequences. Initially, the input image is partitioned into non-overlapping patches, typically 16×16 pixels, which are flattened and projected into patch embeddings. The critical input sequence (X) is then formed by

augmenting the patch embeddings (X_{patch}) with positional information (X_{pos}) and two specialized learnable tokens. They are the classification token (t_{cls}) and the unique distillation token (t_{distil}).

$$X = X_{patch} + X_{pos} \quad (6)$$

The core mechanism within each transformer encoder layer is the Multi-Head Self-Attention (MHSA), which calculates the attention score by normalizing the dot-product similarity between the Query (Q) and Key (K) matrices using the softmax function [25]. This operation is central to modeling long-range dependencies across image patches:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)Vz_l = \text{MLP}(\text{LN}(z'_l)) + z'_l \quad (7)$$

where d_k is the dimension of the key vectors. The MHSA mechanism runs h attention operations (heads) in parallel, concatenating their outputs and linearly projecting them via W^O :

$$\text{MultiHead}(Q, K, V) = [\text{Concat}(\text{head}_1, \dots, \text{head}_h)]W^O \quad (8)$$

The popular DeiT-base configuration employs 12 encoder layers, resulting in a total of 144 attention heads, demonstrating significant capacity while requiring only approximately 85 million (M) parameters.

Each of the L encoder layers (z_l) incorporates MHSA and a Feed-Forward Neural Network (FFN) through residual connections and Layer Normalization (LN) to ensure stable gradient flow.

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \quad (9)$$

$$z_l = \text{FFN}(\text{LN}(z'_l)) + z'_l \quad (10)$$

The final output for classification relies on the representation of the classification token ($z_{CLS}[L]$) from the last layer, mapped to class probabilities y via a classification head [17]:

$$y = \text{softmax}(W_{cls} \cdot z_{CLS}[L] + b_{cls}) \quad (11)$$

The defining characteristic of DeiT's efficiency is the integration of knowledge from a pre-trained teacher model Z_t . This process is often performed using hard distillation, where the student model Z_s minimizes a combined loss function that includes standard cross-entropy (LCE) against the true label (y) and LCE against the hard pseudo-label (y_t) provided by the teacher:

$$L_{hard} = \frac{1}{2}L_{CE}(\varphi(Z_s), y) + \frac{1}{2}L_{CE}(\varphi(Z_s), y_t) \quad (12)$$

$$\text{where } y_t = \text{argmax}_c Z_t(c)$$

Here, φ represents the softmax function. This meticulous combination of architectural purity, sophisticated attention mechanics, and an explicit knowledge transfer mechanism allows DeiT to generalize powerfully even when trained with relatively scarce data.

4.3. Pyramid Vision Transformer

The PVT, designed by Wang et al. at Nanjing University, represents a versatile and pure Transformer backbone that was introduced primarily to overcome the limitations of the original ViT architecture, particularly in its applicability to dense prediction tasks such as object detection and semantic segmentation [27]. Unlike ViT, which produces a single-scale, low-resolution output feature map and incurs high computational costs, PVT integrates a hierarchical, progressive shrinking pyramid structure, analogous to traditional CNNs. This design allows PVT to generate essential multi-scale feature maps with strides progressively shrinking from high (4-stride) to low (32-stride), making it plug-and-play compatible with existing dense prediction pipelines like RetinaNet or Mask R-CNN. The PVT architecture is divided into four stages, where each stage consists of a patch embedding layer and L_i transformer encoder layers, allowing it to take fine-grained image patches as input to learn high-resolution representations. This is a task that typically causes standard ViT models to exhaust memory resources. To efficiently manage the computational and memory costs associated with processing these high-resolution

feature maps, PVT introduces the Spatial-Reduction Attention (SRA) layer to replace the traditional MHA mechanism within the encoder block. The SRA achieves resource reduction by applying a spatial reduction operation (SR) to the Key (K) and Value (V) sequences before calculating attention, enabling the maintenance of a global receptive field under limited computation resources. This reduction operation on an input sequence $x \in \mathbb{R}^{H_i W_i \times C_i}$ is defined using a reduction ratio R_i and a linear projection W^S [27]:

$$SR(x) = \text{Norm}(\text{Reshape}(x, R_i)W^S) \quad (13)$$

The subsequent self-attention mechanism, known as SRA, then incorporates this reduction by utilizing the attention function applied to the original Query (Q) and the spatially reduced Key and Value components:

$$SRA(Q, K, V) = \text{Attention}(Q, SR(K), SR(V)) \quad (14)$$

This ability to effectively learn multi-scale, high-resolution features without relying on convolutions has made PVT a strong alternative to CNN backbones. PVT achieves superior performance not by abandoning CNN principles, but by integrating the strengths of transformers with a pyramid structure that CNNs rely on, overcoming key limitations of pure CNNs and ViTs.

4.4. Framework Design

The proposed framework, which is illustrated in Figure 5, illustrates a complete deep learning pipeline for the automated classification of otoscopic images into AOM, COM, and Normal categories. The process begins with image acquisition, followed by preprocessing, where images are resized and augmented to ensure consistency and to improve model generalization. The dataset is then divided into training (80%), validation (10%), and test (10%) subsets. This provides a balanced approach for both model development and unbiased performance evaluation. With identical hyperparameter configurations applied, the three transformer models, such as ViT, DeiT, and PVT, were employed to allow for comparison of their classification capability. Performance is measured by using accuracy, precision, recall, and F1-score. These metrics collectively capture both predictive performance and sensitivity to misclassification. This is an essential consideration in medical image analysis. The classification section of the framework demonstrates the final outputs, where each image is assigned to one of the three classes. This step highlights the clinical relevance of the approach, as an accurate distinction between acute and chronic otitis media is crucial for appropriate treatment planning, while reliable identification of normal cases prevents unnecessary interventions.

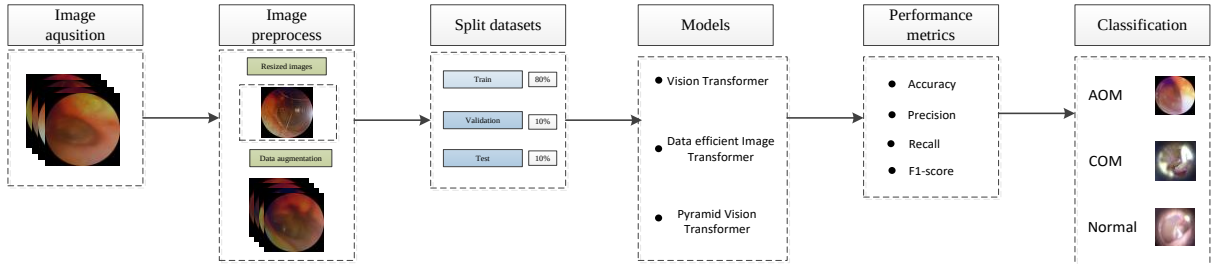


Figure 5. The workflow for transformer-based image classification models.

Line	Procedure: BENCHMARK_OM_CLASSIFIER(D_{raw})
1:	$D_{\text{raw_AOM}}, D_{\text{raw_COM}}, D_{\text{raw_Normal}} \leftarrow \text{Image_Acquisition}(N_{\text{total}}=1800)$
2:	$D_{\text{train}}, D_{\text{val}}, D_{\text{test}} \leftarrow \text{DATA_PREPARATION}(D_{\text{raw}})$
3:	$M_{\text{ViT}(\text{trained})} \leftarrow \text{TRAIN_MODEL}(\text{ViT}, D_{\text{train}}, D_{\text{val}})$
4:	$M_{\text{DeiT}(\text{trained})} \leftarrow \text{TRAIN_MODEL}(\text{DeiT}, D_{\text{train}}, D_{\text{val}})$
5:	$M_{\text{PvT}(\text{trained})} \leftarrow \text{TRAIN_MODEL}(\text{PvT}, D_{\text{train}}, D_{\text{val}})$
6:	$\text{Results}_{\text{ViT}} \leftarrow \text{EVALUATE_PERFORMANCE}(M_{\text{ViT}(\text{trained})}, D_{\text{test}})$
7:	$\text{Results}_{\text{DeiT}} \leftarrow \text{EVALUATE_PERFORMANCE}(M_{\text{DeiT}(\text{trained})}, D_{\text{test}})$
8:	$\text{Results}_{\text{PvT}} \leftarrow \text{EVALUATE_PERFORMANCE}(M_{\text{PvT}(\text{trained})}, D_{\text{test}})$
9:	Output ($\text{Results}_{\text{ViT}}, \text{Results}_{\text{DeiT}}, \text{Results}_{\text{PvT}}$)

Figure 6. Algorithmic pseudo-code representation of the OM classification framework.

4.5. Evaluation Metrics

Accuracy, precision, recall, and F1-score are among the most widely applied metrics for evaluating classification models, yet each captures performance from a different perspective. Accuracy measures the overall proportion of correct predictions out of all predictions, providing a straightforward view of performance. However, it can be misleading when the dataset is imbalanced, as it may overstate the success of a model that simply predicts the majority class. Precision shows the proportion of correctly predicted positive instances out of all instances predicted as positive. This highlights how trustworthy your positive predictions are. It is especially valuable when false positives lead to serious consequences. Recall focuses on how many of the actual positive cases the model successfully identifies. This is crucial in applications where missing positive cases are costly. The F1-score is a single measure that balances precision and recall. It offers a more informative view when datasets are imbalanced or when both false positives and false negatives are significant errors.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

Where TP : True Positive, TN : True Negative, FP : False Positive, FN : False Negative

$$\text{Recall} = \frac{TP}{TP + FN} \quad (16)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (17)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (18)$$

4.6. Training Configuration and Hyperparameter Settings

The hyperparameter configuration designed for the ViT, DeiT, and PvT reflects a consistent and balanced approach for comparative evaluation in demonstrated in Table 2. All models were trained with an input size of 256×256 pixels, which offers a suitable balance between retaining sufficient image detail and maintaining computational efficiency. The training process was conducted for 50 epochs with a batch size of 16. This setting represents a moderate choice that enables stable learning while addressing potential hardware memory limitations. For optimization, the Adam optimizer with a fixed learning rate of 0.001 was employed. This choice is widely accepted in deep learning studies because it provides adaptive learning-rate adjustment and stable convergence. The models were trained using the categorical cross-entropy loss function, which is the standard criterion for multi-class classification tasks. ReLU was used as the activation function, supporting efficient training by reducing vanishing gradient problems and promoting sparse activations.

Table 2. Hyperparameter configuration for transformer-based models.

Models	ImageSize	Epochs	HyperParameters				Activation Function
			Batch Size	Optimizer	Learning Rate	Loss Function	
Vision Transformer	256x256	50	16	Adam	0.001	CategoricalCrossentropy	ReLU
Data-efficient Image Transformer	256x256	50	16	Adam	0.001	CategoricalCrossentropy	ReLU
Pyramid Vision Transformer	256x256	50	16	Adam	0.001	CategoricalCrossentropy	ReLU

5. RESULTS AND DISCUSSION

The training performances of the three models are illustrated in Figures 7, 8, and 9, and a comparative analysis highlights their distinct characteristics.

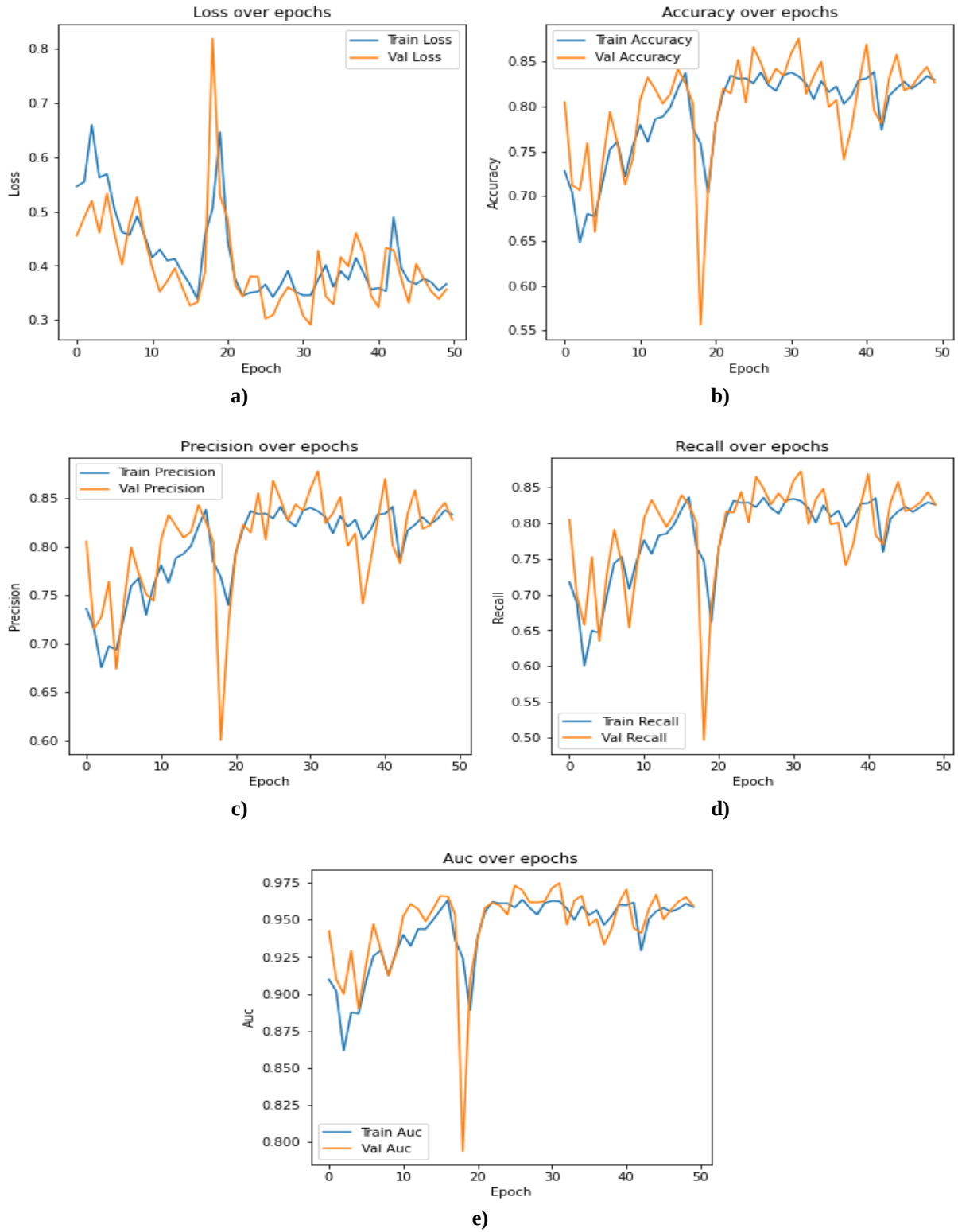


Figure 7. The graphical train performances of the ViT model for **a)** Loss over epochs, **b)** Accuracy over epochs, **c)** Precision over epochs, **d)** Recall over epochs, **e)** AUC over epochs.

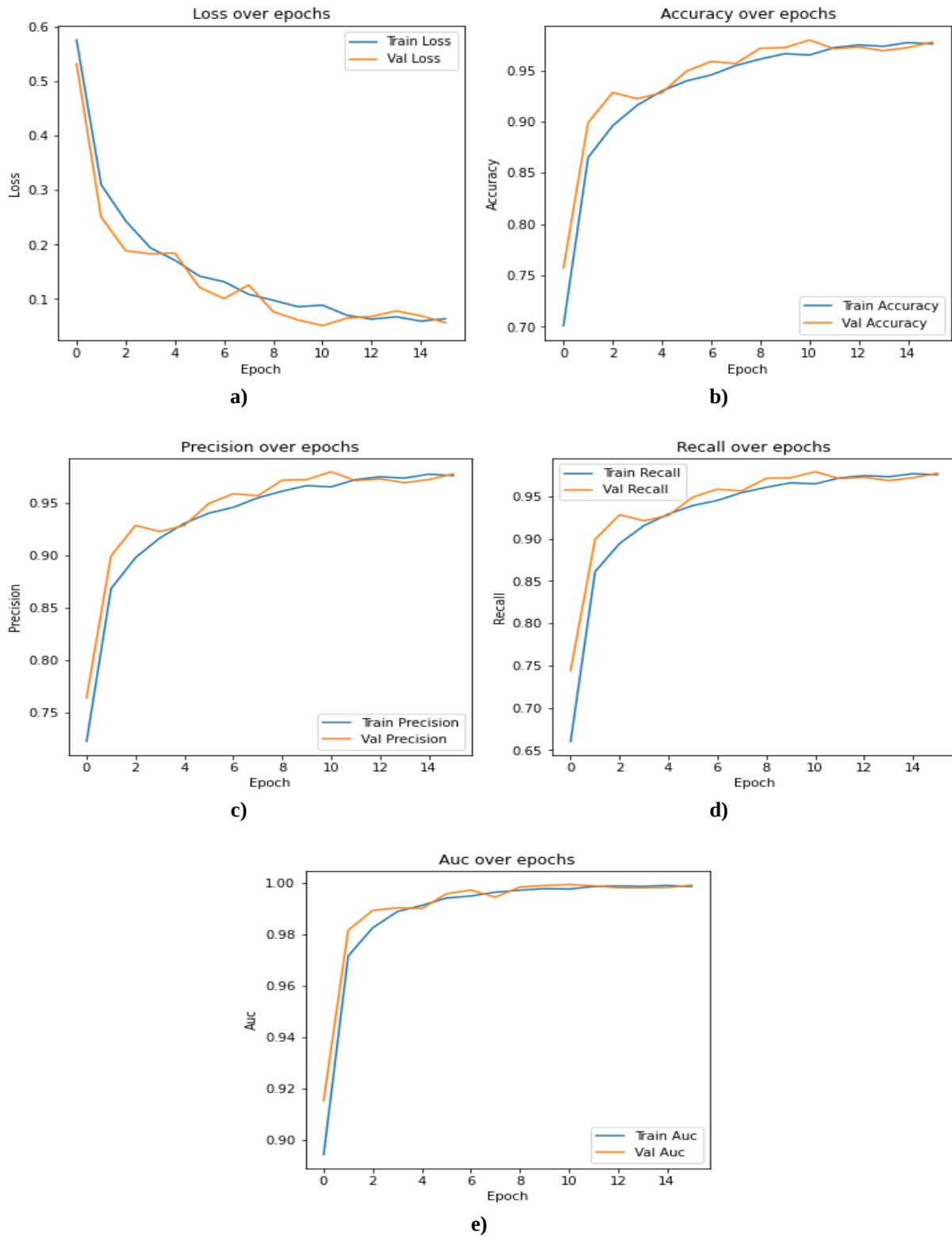


Figure 8. The graphical train performances of the DeiT model for **a)** Loss over epochs, **b)** Accuracy over epochs, **c)** Precision over epochs, **d)** Recall over epochs, **e)** AUC over epochs.

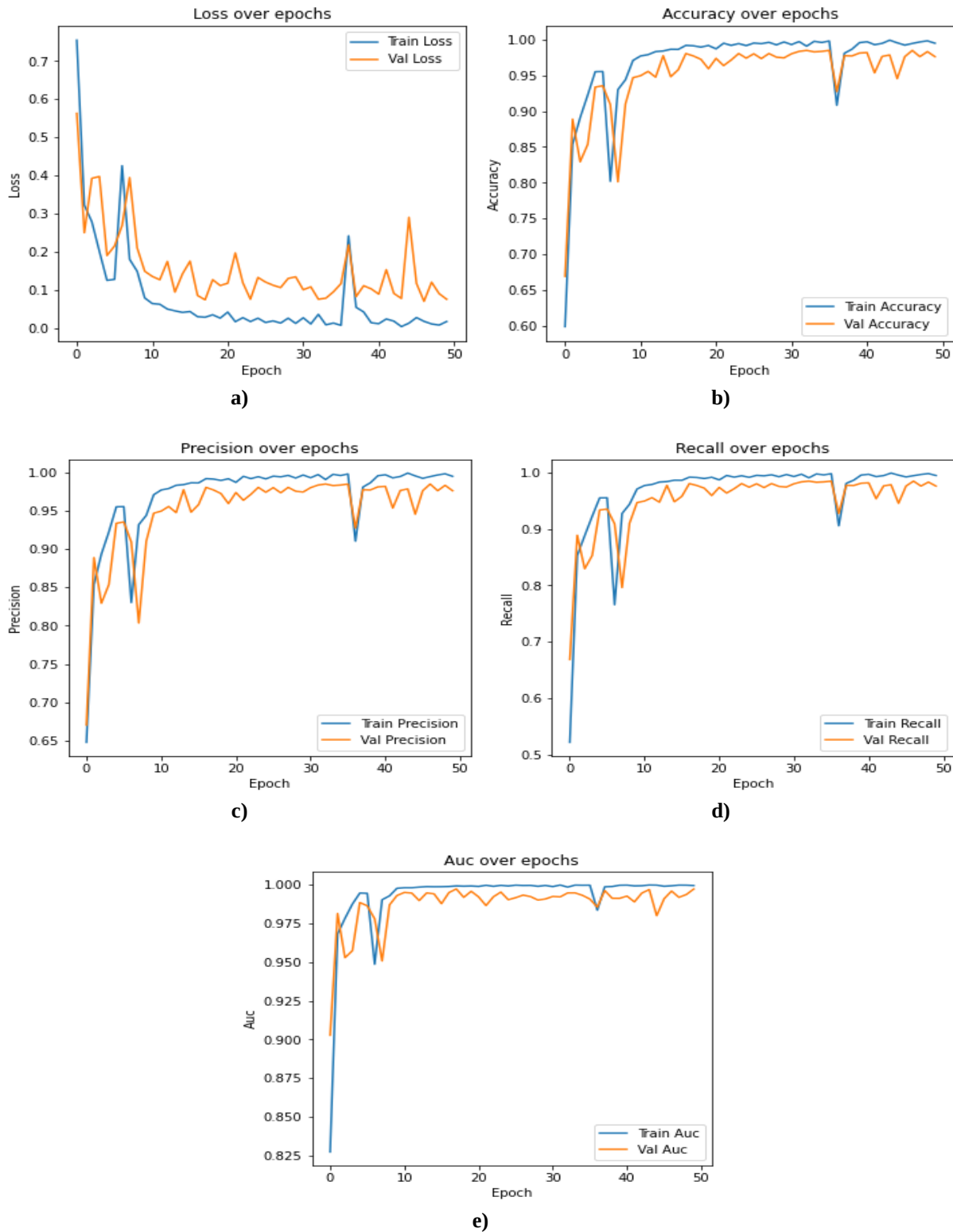


Figure 9. The graphical train performances of the PvT model for **a)** Loss over epochs, **b)** Accuracy over epochs, **c)** Precision over epochs, **d)** Recall over epochs, **e)** AUC over epochs.

As shown in Figures 7, 8, and 9, the training performance of the ViT, DeiT, and PvT models differs significantly in terms of convergence stability and overall classification quality. The ViT model exhibits fluctuating training dynamics across 50 epochs. While the accuracy oscillates between approximately 0.75 and 0.85, the validation accuracy does not follow a smooth trend and shows considerable variation, suggesting unstable generalization. Precision and recall also vary in the range of 0.70–0.90, reflecting inconsistency in the model's ability to correctly identify positive cases. Although the AUC values remain relatively high, around 0.93,

indicating the model's ability to distinguish between classes, the lack of stability in loss and accuracy curves suggests that ViT struggles to achieve robust learning for this task.

In contrast, the DeiT model demonstrates much more stable and efficient training behavior. Both training and validation loss steadily decrease and converge to very low values within approximately 15 epochs, showing rapid optimization. Accuracy improves consistently and reaches nearly 0.98, with training and validation curves closely aligned, indicating minimal overfitting. Precision and recall also stabilize around 0.95–0.98, highlighting the balanced performance of the model in terms of both correct positive predictions and sensitivity. The AUC value approaches 0.99, confirming that DeiT achieves near-perfect separability between classes while maintaining strong generalization capacity. The PvT model also achieves outstanding performance, characterized by rapid convergence and strong stability. Training and validation loss decrease sharply within the first 10 epochs and stabilize at low levels, while accuracy consistently remains in the range of 0.97–0.99 across the training process. Precision and recall values both exceed 0.95 and maintain stable trends, ensuring reliable classification outcomes. The AUC value is almost perfect (≥ 0.99), further indicating the model's excellent discriminative capability. Compared to DeiT, PvT converges slightly more gradually but reaches equally high-performance levels with robust stability throughout training.

The comparison reveals that DeiT and PvT significantly outperform the standard ViT model. While ViT achieves moderate accuracy and reasonable AUC, it suffers from unstable training behavior and inconsistent precision and recall. In contrast, DeiT and PvT not only converge faster and more stably but also achieve consistently high performance across all metrics. Among them, PvT provides the strongest balance of robustness, stability, and predictive power, making it the most effective model for this classification task.

Table 3 summarizes the diagnostic performance of three transformer-based models, ViT, DeiT, and PvT, applied to otitis media classification. The models are evaluated across acute otitis media, chronic otitis media, and normal cases using accuracy, precision, recall, and F1-score.

The results clearly show differences in performance among the models. ViT performs very well in detecting acute otitis media, with high accuracy (0.98) and balanced precision and recall. However, its performance decreases in chronic otitis media (accuracy 0.83, precision 0.71) and normal cases (accuracy 0.84, recall 0.67). This indicates that ViT has limitations in generalising. These limitations apply to all categories. DeiT provides almost perfect results in all conditions. It achieves 1.00 in every metric for acute otitis media and 0.96 for both chronic otitis media and normal cases. These values show that DeiT is highly robust and reliable. PvT also shows excellent results. It reaches 1.00 in acute otitis media and performs strongly in chronic and normal categories. It is especially strong in precision for chronic otitis media (0.99) and recall for normal cases (0.99). When compared to ViT, both DeiT and PvT are more stable, more sensitive, and more balanced between precision and recall. This reduces the chance of misclassification. Finally, DeiT shows the most consistent performance, while PvT is also highly competitive with advantages in some metrics. Both are more suitable for clinical use than ViT.

Table 3. Comparative evaluation of ViT, DeiT, and PvT models in otitis media classification.

Model	Classification	Performance metrics			
		Accuracy	Precision	Recall	F1-score
ViT	Acute otitis media	0.98	0.97	0.98	0.98
	Chronic otitis media	0.83	0.71	0.84	0.77
	Normal	0.84	0.83	0.67	0.74
DeiT	Acute otitis media	1.00	1.00	1.00	1.00
	Chronic otitis media	0.96	0.96	0.96	0.96
	Normal	0.96	0.96	0.96	0.96
PvT	Acute otitis media	1.00	1.00	1.00	1.00
	Chronic otitis media	0.94	0.99	0.94	0.96
	Normal	0.99	0.94	0.99	0.96

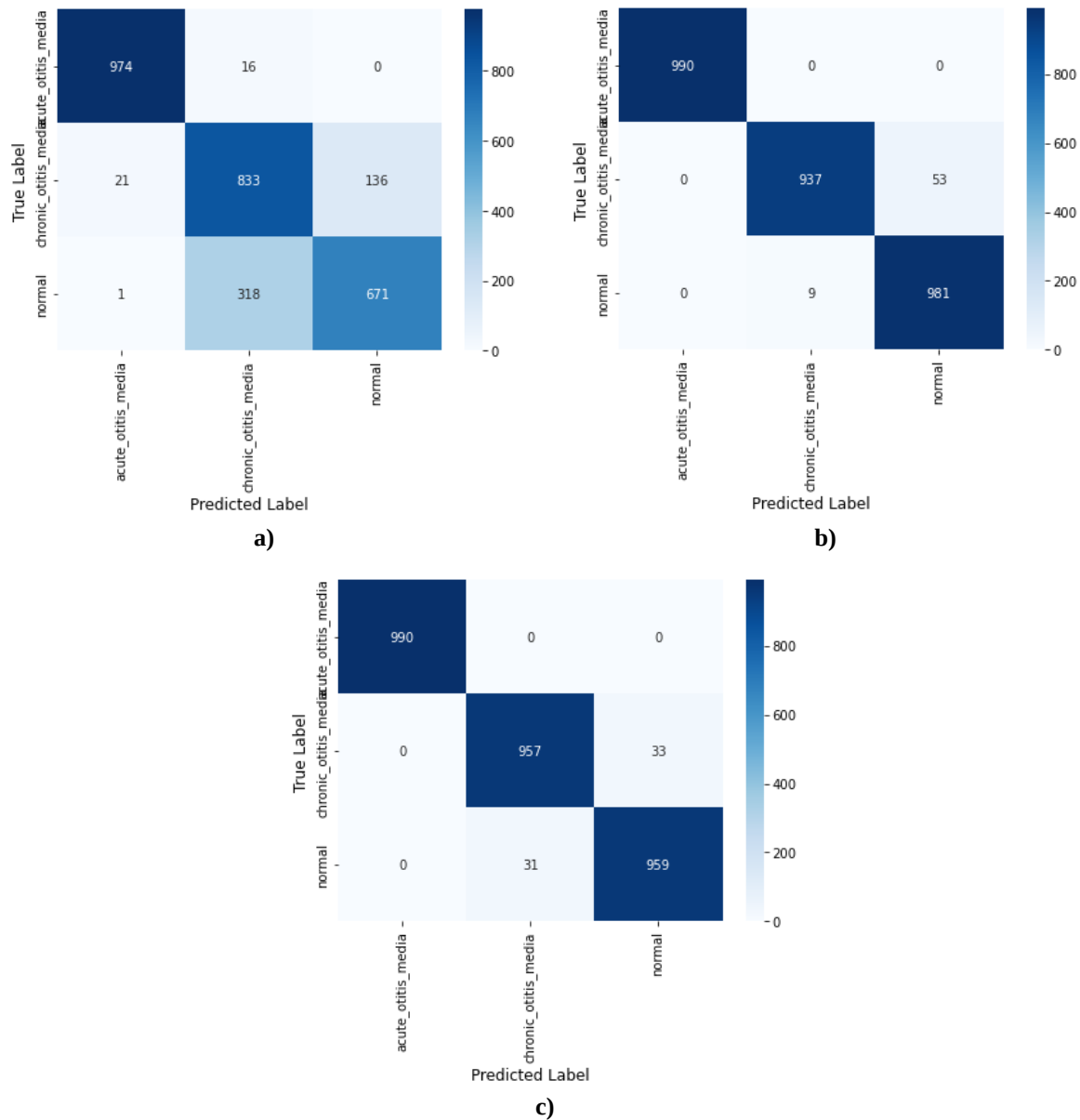


Figure 10. The confusion matrix of the models: **a)** ViT **b)** PvT **c)** DeiT.

The performance of the three transformer-based models, which is illustrated in Figure 10, was rigorously evaluated on a balanced test dataset comprising 2,970 otoscopic images (990 per class: normal, acute otitis media, and chronic otitis media). Confusion matrices provide a detailed visualization of classification outcomes, highlighting true positives along the diagonal and off-diagonal misclassifications. The comparative performance analysis of the ViT, PvT, and DeiT highlights clear differences in classification accuracy across the three otitis media classes. The ViT model achieved strong results for acute otitis media with 974/990 correctly classified cases (accuracy: 98.4%), but its performance degraded on chronic otitis media with 833/990 correct predictions (84.1%) and significant confusion with normal cases, where only 671/990 were correctly classified (67.8%). PvT demonstrated more balanced outcomes, perfectly classifying acute otitis media (990/990, 100%) and achieving high accuracy for chronic otitis media with 937/990 correct predictions (94.6%) and normal cases with 981/990 correctly classified (99.1%). DeiT provided the strongest overall performance, also achieving perfect classification for acute otitis media (990/990, 100%), and surpassing PvT on chronic otitis media with 957/990 correctly identified (96.7%), while slightly trailing PvT for normal cases with 959/990 correct (96.9%). ViT shows limitations due to high misclassification between chronic and normal cases, making it the weakest of the three. PvT offers excellent accuracy, especially for normal class detection, while DeiT delivers the most balanced performance, excelling particularly in chronic otitis media classification. Quantitatively, DeiT achieves the highest robustness, with near-uniform accuracy across all categories, establishing it as the best-performing model for this task.

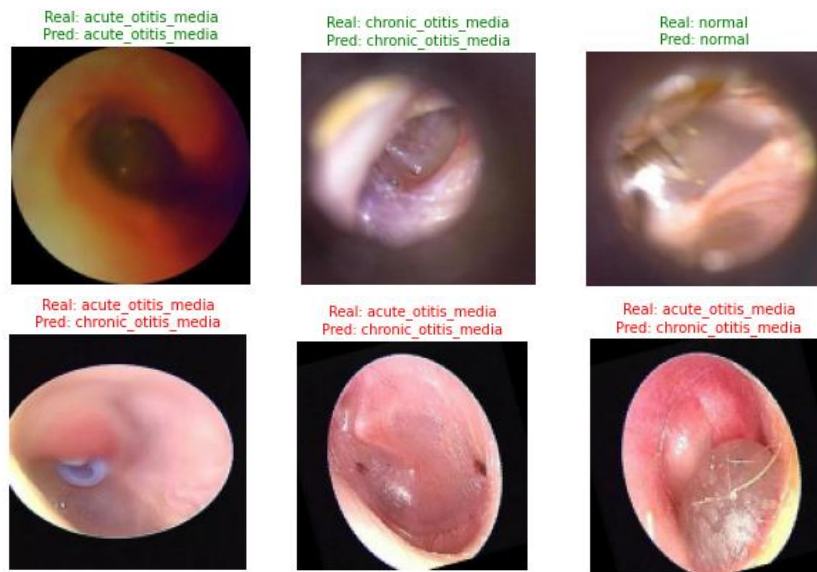


Figure 11. Visual results of the ViT model for otitis media classification, illustrating both true and false predictions.

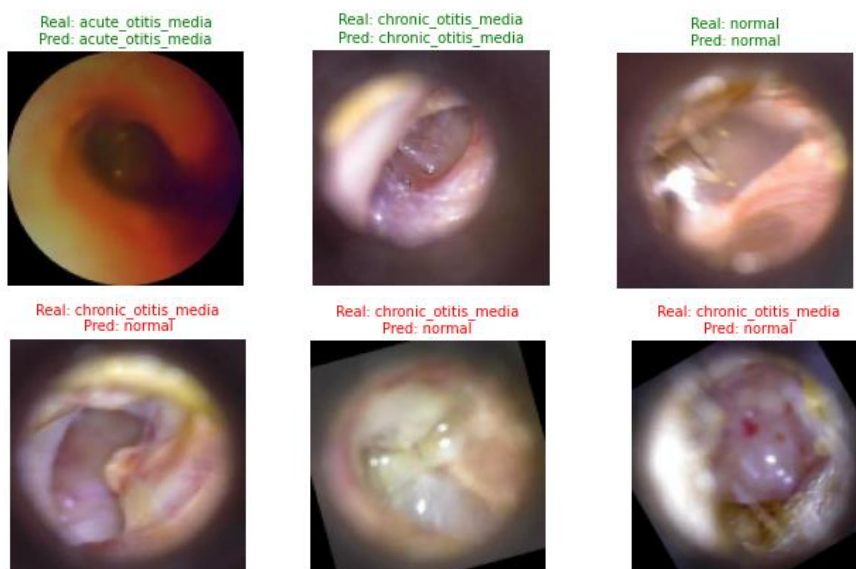


Figure 12. Visual results of the DeiT model for otitis media classification, illustrating both true and false predictions.

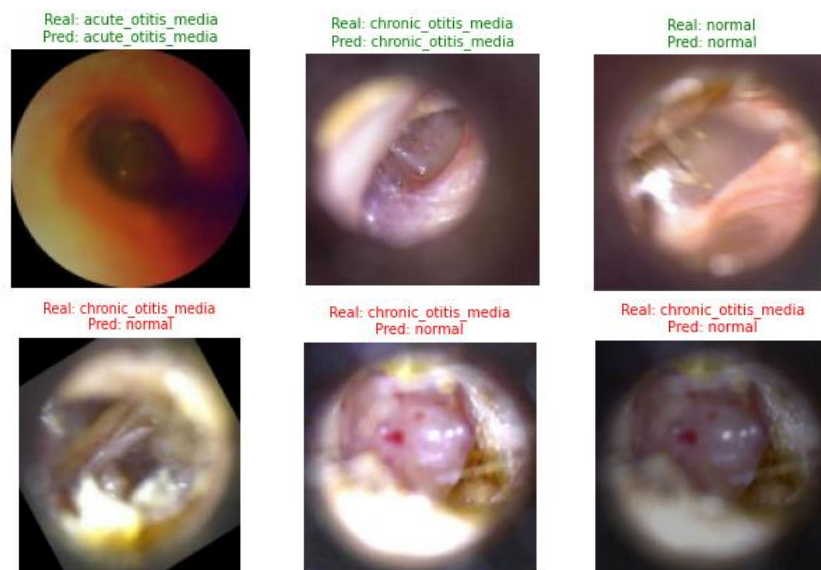


Figure 13. Visual results of the PvT model for otitis media classification, illustrating both true and false predictions.

Figure 11, 12, and 13 present the visual results of otitis media classification using three transformer-based architectures. Each figure illustrates both true positive predictions, where the model correctly identified the condition, and misclassifications, where the model struggled to distinguish between similar categories. For the ViT model, the results demonstrate its capability to correctly classify acute, chronic, and normal cases. However, it also shows notable confusion between acute and chronic otitis media, as observed in multiple false predictions. This suggests that ViT captures the structure of the tympanic membrane but may struggle with subtle inflammatory differences between acute and chronic cases, which often present overlapping visual features. Despite these challenges, ViT still produced several accurate classifications, particularly in normal cases.

The DeiT model exhibits a slightly different error pattern. While DeiT correctly identifies many samples, it shows frequent misclassifications of chronic otitis media as normal. This indicates that DeiT, despite being designed as a more data-efficient variant of ViT, might underrepresent pathological features in certain conditions, leading to higher false negatives. This is particularly concerning in clinical contexts, as mislabeling diseased cases as normal could delay diagnosis and treatment.

The PvT model similarly demonstrates both strengths and limitations. It successfully detects acute and chronic cases in several examples. However, consistent errors occur when differentiating chronic otitis media from normal. These findings highlight the difficulty of chronic case recognition, likely due to heterogeneous and less distinct visual patterns compared to acute inflammation.

5.1. 5-Fold Cross-Validation and Statistical Analysis

To further evaluate the robustness and reliability of the transformer-based models, additional validation analyses were conducted using five-fold cross-validation and the McNemar statistical test. These analyses were performed to assess model stability across different data partitions and to determine whether the observed performance differences between models were statistically significant.

Table 4. 5-Fold Cross-Validation Accuracy Results of ViT, DeiT, and PvT Models.

Model	k=1	k=2	k=3	k=4	k=5
ViT	0.9972	0.9972	1.000	0.9972	1.0000
DeiT	0.9916	0.9972	0.9972	0.9944	0.9916
PvT	0.9694	0.9972	0.9500	0.9666	0.9805

Table 4 presents the five-fold cross-validation accuracy results for the ViT, DeiT, and PvT models. The findings demonstrated consistently high classification performance across all models. ViT achieved the highest and most stable performance, with accuracy values ranging from 0.9972 to 1.0000 across the five folds. Similarly, DeiT exhibited robust performance, with accuracy values varying between 0.9916 and 0.9972. In contrast, PvT demonstrated relatively greater variability and lower accuracy values, ranging from 0.9500 to 0.9972. Despite this variation, PvT maintained strong classification performance overall. These results suggest that ViT and DeiT achieved more stable generalization across different data splits, while PvT exhibited comparatively lower consistency.

Table 5. Pairwise McNemar test results for performance comparison among ViT, DeiT, and PvT models.

Models	McNemar Statistic (χ^2)	p-value	Result
ViT vs DeiT	1.7857	0.0907	No significant difference
ViT vs PvT	46.8167	0.0010	Significant difference
DeiT vs PvT	35.6290	0.0010	Significant difference

To statistically compare model predictions, the McNemar test was performed for pairwise comparisons between the transformer architectures. As shown in Table 5, no statistically significant difference was observed between ViT and DeiT, with a p-value of 0.0907, which exceeded the significance threshold of 0.05. However, statistically significant differences were identified between ViT and PvT and between DeiT and PvT, with p-values below 0.001. These findings indicate that ViT and DeiT exhibited comparable prediction performance, whereas PvT demonstrated significantly different prediction behavior.

Table 6 presents a comparative summary of recent studies employing ViT and its variants for medical image analysis, particularly focusing on otoscopic and dermatological datasets. The table highlights model architectures, datasets, and performance metrics, including accuracy, precision, recall, and F1-score.

According to Table 6, Vision Transformer-based approaches have demonstrated consistently strong performance across various medical imaging domains. Among the listed studies, Comert et al. achieved robust results using different ViT variants (B8, B16, B32) with accuracies ranging from 0.90 to 0.94 on otoscopic images involving four classes (COM, myringosclerosis, earwax plug, and normal). Similarly, Vuran et al. evaluated multiple transformer architectures, including ViT, MAE, DINO, and Swin, on a six-class skin lesion dataset from

Kaggle, where ViT and Swin models reached 0.93 accuracy, demonstrating their generalization ability beyond otoscopic data.

Table 6. Literature comparison of Vision Transformer-Based studies for medical image classification

Study	Model	Accuracy	Precision	Recall	F1-score	Dataset	Number of classes
Comert et al	ViT-B8	0.90	0.90	0.90	0.90	The Otorhinolaryngology Department of the Universidad de Chile Clinical Hospital	4(COM, myringosclerosis, earwax plug, and normal.)
	ViT-B16	0.94	0.94	0.94	0.94		
	ViT-B32	0.93	0.93	0.93	0.93		
Vuran et al	ViT	0.93	0.93	0.94	0.92	Kaggle platform (skin lesion)	6 (Mpox, Chickenpox, Measles, Cowpox, HFMD, Healthy)
	MAE	0.84	0.86	0.84	0.84		
	DINO	0.90	0.91	0.90	0.89		
	Swin	0.93	0.94	0.93	0.93		
Deva et al	Vision Transformer	0.95 (AOM)	0.96 (AOM)	0.89 (AOM)	0.92 (AOM)	Kaggle platform (Otoscope Data)	5(AOM, CSOM, Tympanoskleros, Otitis Externa)
Demircan et al	Vision Transformer	0.97	0.97	0.97	0.97	ENT department of the Clinical Hospital of Universidad de Chile	4 (CMOS, earwax, normal and myringosclerosis)
Habib et al	Vision Transformer	0.92	—	0.86	0.96	848 otoscopic images (Özel Van Akdamar Hospital), 540 otoscopic images (University of Chile Clinical Hospital), 454 images (Nationwide Children’s Hospital)	2 (normal, abnormal)
Our Study	ViT	0.98	0.97	0.98	0.98	Kaggle Platform (Otitis Media)	3 (AOM, COM, Normal)
	DeiT	0.96	1.00	0.96	1.00		
	PvT	0.99	0.99	0.94	0.96		

Deva et al. applied ViT to a five-class otoscope dataset, obtaining 0.95 accuracy, while Demircan et al. reported an even higher accuracy of 0.97 on a four-class otoscopic dataset, confirming the model’s effectiveness in complex ear image classification tasks. Furthermore, Habib et al. combined multi-source otoscopic datasets and achieved 0.92 accuracy, underscoring ViT’s adaptability across heterogeneous medical data. In comparison, our proposed study achieved the highest performance, with 0.98 accuracy, 0.97 precision, 0.98 recall, and 0.98 F1-score using a ViT-based model on the Kaggle otitis media dataset comprising three classes (AOM, COM, and Normal). These results indicate that the proposed model not only outperforms previous transformer-based studies but also achieves superior balance among all evaluation metrics. The comparison emphasizes the powerful representational capacity of transformer-based architectures in disease detection and classification from medical images, particularly in otoscopic analysis.

5.2. Limitations and Gaps

While this study provides one of the first systematic benchmarking analyses of transformer-based architectures for otitis media classification, several limitations should be acknowledged. First, the dataset employed, although carefully balanced and augmented, is relatively modest in scale (1,800 original images expanded through augmentation). This may not fully reflect the heterogeneity of otoscopic presentations encountered in diverse clinical environments. Second, the data originates from a single source, which could limit

the generalizability of the models to different populations, imaging devices, or clinical settings. Third, the study focused exclusively on image-based diagnosis without incorporating multimodal clinical data such as audiometry, patient history, or laboratory findings, which could provide additional diagnostic context. From a methodological perspective, the models were trained and tested under controlled conditions with identical hyperparameters. While this ensured fair benchmarking, it does not capture the potential variability in performance when hyperparameter optimization or transfer learning is applied. Moreover, although evaluation metrics such as accuracy, precision, recall, and F1-score provide quantitative insights, the study did not incorporate explainability frameworks to enhance clinical interpretability and trust. Finally, computational cost and deployment feasibility were not assessed, leaving open questions regarding the suitability of these architectures for real-time diagnosis in resource-limited or mobile healthcare settings. Addressing these gaps will require larger, multi-center datasets, the integration of multimodal clinical information, and the development of lightweight, explainable transformer models. Such advancements are essential to move beyond proof-of-concept benchmarking toward robust, trustworthy, and clinically deployable AI-assisted diagnostic tools for otitis media.

6. CONCLUSION

This study presented a comprehensive benchmarking of transformer-based architectures for the automated diagnosis of otitis media, addressing a critical clinical and public health challenge. By systematically evaluating the ViT, DeiT, and PvT on a carefully curated and balanced otoscopic dataset. The study demonstrated the potential of transformers to overcome the limitations of conventional diagnostic methods and convolutional neural networks. The results clearly indicate that while ViT performs strongly for acute otitis media, its reduced generalizability across chronic and normal cases limits its clinical utility. In contrast, DeiT achieved the most consistent and robust performance across all diagnostic categories, and PvT exhibited complementary strengths in precision and recall, underscoring their promise for reliable clinical deployment. Beyond performance, the study contributes a reproducible and standardized evaluation framework that can serve as a foundation for future research in this domain. Nevertheless, several limitations remain, including the reliance on a single dataset and the absence of multimodal clinical information, which may constrain generalizability to diverse patient populations. Future research should explore larger, multi-center datasets, integration of multimodal features such as audiometric data and patient history, and the incorporation of explainable AI techniques to foster clinical trust and interpretability. Additionally, the development of lightweight transformer models will be essential for enabling deployment in resource-limited healthcare settings. Finally, this work provides new evidence for the feasibility and effectiveness of transformer architectures in otitis media classification and offers a meaningful step toward the integration of advanced AI methods into routine otolaryngology practice.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics.

Artificial Intelligence (AI) Contribution Statement

Artificial intelligence (AI) tools were used solely for language editing and English grammar improvement. The scientific content, study design, data analysis, interpretation of results, and preparation of figures were entirely performed by the authors. The authors reviewed and approved all manuscript content and take full responsibility for its accuracy and integrity.

REFERENCES

- [1] I. A. Mohamed, Z. A. Mohamed, F. Ning, and W. Xin, "The Prevalence and Risk Factors Associated with Otitis Media in Children under Five Years of Age in Mogadishu, Somalia: A Hospital-Based Cross-Sectional Study," *International Journal of Otolaryngology and Head & Neck Surgery*, vol. 12, no. 06, pp. 426–443, 2023, doi: 10.4236/ijohns.2023.126046.
- [2] M. W. You *et al.*, "The Roles of NOD-like Receptors in Innate Immunity in Otitis Media," Feb. 01, 2022, *MDPI*. doi: 10.3390/ijms23042350.
- [3] L. Cupples, T. Y. C. Ching, and S. Hou, "Speech, language, functional communication, psychosocial outcomes and QOL in school-age children with congenital unilateral hearing loss," *Front Pediatr*, vol. 12, Mar. 2024, doi: 10.3389/fped.2024.1282952.

- [4] F. Folino, M. Caruso, P. Bosi, M. Aldè, S. Torretta, and P. Marchisio, "Acute otitis media diagnosis in childhood: still a problem in 2023?," *Ital J Pediatr*, vol. 50, no. 1, Dec. 2024, doi: 10.1186/s13052-024-01588-y.
- [5] J. H. Stephens *et al.*, "A multimodal machine learning algorithm improved diagnostic accuracy for otitis media in a school aged Aboriginal population," *J Biomed Inform*, vol. 164, Apr. 2025, doi: 10.1016/j.jbi.2025.104801.
- [6] R. G. Kashani *et al.*, "Shortwave infrared otoscopy for diagnosis of middle ear effusions: a machine-learning-based approach," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-91736-9.
- [7] C. B. Botelho, R. Tortoriello, S. N. Koch, and J. I. Fernandes, "Clinical aspects of middle ear tympanokeratoma in dogs diagnosed through advanced imaging, otoendoscopy and histopathological evaluation," *Vet Dermatol*, vol. 36, no. 4, pp. 417–423, Aug. 2025, doi: 10.1111/vde.13321.
- [8] D. Muhammad and M. Bendeche, "Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis," Dec. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.csbj.2024.08.005.
- [9] S. Örenç, E. Acar, M. S. Özerdem, S. Şahin, and A. Kaya, "Automatic Identification of Adenoid Hypertrophy via Ensemble Deep Learning Models Employing X-ray Adenoid Images," *Journal of Imaging Informatics in Medicine*, 2025, doi: 10.1007/s10278-025-01423-8.
- [10] S. Orenc, M. S. Ozerdem, E. Acar, and M. Yilmaz, "Automatic segmentation of chest X-ray images via deep-improved various U-Net techniques," *Digit Health*, vol. 11, May 2025, doi: 10.1177/20552076251366855.
- [11] Y. Choi, J. Chae, K. Park, J. Hur, J. Kweon, and J. H. Ahn, "Automated multi-class classification for prediction of tympanic membrane changes with deep learning models," *PLoS One*, vol. 17, no. 10 October, Oct. 2022, doi: 10.1371/journal.pone.0275846.
- [12] A. Younesi, M. Ansari, M. Fazli, A. Ejlali, M. Shafique, and J. Henkel, "A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends," Feb. 2024, [Online]. Available: <http://arxiv.org/abs/2402.15490>
- [13] Y. Lu, S. Wang, B. Wang, X. Zhang, X. Wang, and Y. Zhao, "Enhanced Window-Based Self-Attention with Global and Multi-Scale Representations for Remote Sensing Image Super-Resolution," *Remote Sens (Basel)*, vol. 16, no. 15, Aug. 2024, doi: 10.3390/rs16152837.
- [14] A. W. Salehi *et al.*, "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," Apr. 01, 2023, *MDPI*. doi: 10.3390/su15075930.
- [15] A. Khan *et al.*, "A survey of the vision transformers and their CNN-transformer based variants," *Artif Intell Rev*, vol. 56, pp. 2917–2970, Dec. 2023, doi: 10.1007/s10462-023-10595-0.
- [16] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," Jan. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/technologies13010032.
- [17] Z. Cömert, A. Sbröllini, F. Demircan, and L. Burattini, "Computerized otoscopy image-based artificial intelligence model utilizing deep features provided by vision transformer, grid search optimization, and support vector machine for otitis media diagnosis," *Neural Comput Appl*, vol. 36, no. 36, pp. 23113–23129, Dec. 2024, doi: 10.1007/s00521-024-10457-y.
- [18] F. Demircan, M. Ekinçi, and Z. Cömert, "Enhancing intra-aural disease classification with attention-based deep learning models," *Neural Comput Appl*, vol. 37, no. 9, pp. 6601–6616, Mar. 2025, doi: 10.1007/s00521-025-10990-4.
- [19] B. Ananda Deva, V. Sood, A. Fahim, P. M. Arjun Krishna, and T. Anjali, "Feature Fusion: Enhancing Middle Ear Infection Diagnosis with Vision Transformer and Symptom Encoding," in *Proceedings of 8th International Conference on Inventive Computation Technologies, ICICT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 233–239. doi: 10.1109/ICICT64420.2025.11005139.
- [20] E. A. Khan and A. A. H. Khan, "Sumotosima: A Framework and Dataset for Classifying and Summarizing Otoscopic Images," Aug. 2024, [Online]. Available: <http://arxiv.org/abs/2408.06755>
- [21] C. Dubois *et al.*, "Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images," *NPJ Digit Med*, vol. 7, no. 1, Dec. 2024, doi: 10.1038/s41746-024-01159-9.

- [22] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 2021, [Online]. Available: <http://arxiv.org/abs/2010.11929>
- [23] J. Maurício, I. Domingues, and J. Bernardino, “Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review,” May 01, 2023, *MDPI*. doi: 10.3390/app13095521.
- [24] R. Azad *et al.*, “Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review,” Nov. 2023, [Online]. Available: <http://arxiv.org/abs/2301.03505>
- [25] B. Guo *et al.*, “An enhanced network for extracting tunnel lining defects using transformer encoder and aggregate decoder,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 136, Feb. 2025, doi: 10.1016/j.jag.2024.104259.
- [26] H. Touvron, M. Cord, and H. Jégou, “DeiT III: Revenge of the ViT.”
- [27] M. Asante, G. Adbdul-Salam, B. Nkrumah, and W. K. Adu-Gyamfi, “Data-Efficient Image Transformers for Robust Malware Family Classification,” *Journal of Cyber Security*, vol. 6, no. 1, pp. 131–153, 2024, doi: 10.32604/jcs.2024.053954.
- [28] W. Wang *et al.*, “Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions.” [Online]. Available: <https://github.com/whai362/PVT>