



Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Full Length Article

BioOtoFusionNet – A bio-attention & frequency-aware hybrid network for ear disease classification

Eshika Jain^a, Vinay Kukreja^a, Pratham Kaushik^a, Modafar Ati^{b,*},
Shanmugasundaram Hariharan^c

^a Centre for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

^b Computer Science and Information Technology, College of Engineering, Abu Dhabi University, Abu Dhabi, United Arab Emirates

^c Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India



ARTICLE INFO

Keywords:

Otoscopic Image Analysis
Ear Disease Classification
Dual-Stream Fusion
Attention Mechanisms
Interpretability Metrics
Deep Learning

ABSTRACT

Objective: Automated interpretation of otoscopic images is challenging due to subtle textural variations, anatomical complexity, and inconsistent acquisition conditions. This study aims to develop an accurate and interpretable deep learning framework for ear disease classification.

Methods: This work presents BioOtoFusionNet, a clinically motivated dual-branch architecture integrating a Frequency-Aware Stream based on Discrete Wavelet Transform (DWT) sub-bands and a Shape-Aware Stream that captures anatomical structures using edge-based features and capsule attention. An Adaptive Cross-Fusion Module (ACFM) and Multi-Scale Attention Pooling (MSAP) are employed to effectively fuse complementary representations across spatial resolutions.

Results: BioOtoFusionNet achieved an overall accuracy of 96.8%, an F1-score of 95.6%, and an AUC-ROC of 98.4%, outperforming all ablated variants. High class-wise accuracy was observed for Earwax Plug (97.8%), Myringosclerosis (94.5%), Chronic Otitis Media (93.2%), and Normal Ear (97.1%).

Interpretability and Robustness: Clinically motivated interpretability metrics demonstrated balanced reliance on texture and structure (FHI = 0.60, SAR = 1.25), strong attention localization (ALS = 0.72), and stable multi-scale behaviour (MSAC = 0.87). Robustness analysis showed resilience to illumination variations (RII = 0.15) and low diagnostic ambiguity (DCS = 0.31).

Evaluation Protocol: Evaluation was conducted on a four-class otoscopic dataset using stratified five-fold cross-validation with strict separation between training and evaluation samples. Data augmentation was applied only to training subsets to prevent information leakage.

Conclusion: BioOtoFusionNet provides accurate, interpretable, and robust ear disease classification from otoscopic images, highlighting its potential for clinical decision support and telemedicine-based otologic screening.

1. Introduction

Ear diseases are one of the major clinical problems in the field of Otolaryngology, since they affect millions of people worldwide and thus create a heavy burden on healthcare systems. Proper classification of ear diseases using imaging using otoscopy is not only limited on the precision in diagnosis but also patient safety, treatment effectiveness and allocation of healthcare resources. Diagnostic accuracy is needed to avoid complications, which may include hearing loss, to worse, sequelae such as mastoiditis or intracranial infection [1]. The urgency is

enhanced by the fact that it is a global problem: today over 1.5 billion people live with some degree of hearing impairment and the World Health Organization estimates that this number will grow to almost 2.5 billion by 2050 [2] if effective preventive and diagnostic measures are not implemented. Because unaddressed hearing loss already costs the world economy close to US\$1 trillion annually, enhancing early, accurate otoscopic diagnosis is important for ensuring patient safety, efficacy of treatment and rational use of health care resources (*Deafness and Hearing Loss*) [3]. The limitations of the traditional methods of diagnosis are inherent and have stimulated the creation of artificial intelligence-

* Corresponding author.

E-mail addresses: eshikajain0150@gmail.com (E. Jain), onlyvinaykukreja@gmail.com (V. Kukreja), prathamkaushik8@gmail.com (P. Kaushik), modafar.ati@adu.ac.ae (M. Ati), mailtos.hariharan@gmail.com (S. Hariharan).

<https://doi.org/10.1016/j.jestch.2026.102354>

Received 24 November 2025; Received in revised form 19 February 2026; Accepted 18 March 2026

Available online 27 March 2026

2215-0986/© 2026 The Authors. Published by Elsevier B.V. on behalf of Karabuk University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

based solutions to the analysis of otoscopic images. Clinical standards currently rely on the manual visual examination performed by clinicians, which is, however, greatly affected by inter-observer variability, with only between 61% and 79% of diagnostic agreement in even experienced otolaryngologists [4]. This type of assessment is subjective and therefore creates inconsistency that can lead to false positive and negative diagnoses, which can have serious consequences for patient outcomes. The diagnostic accuracy of primary-care physicians is much lower than that of otolaryngology specialists, which once again demonstrates the need for the presence of objective diagnostic tools that could be used to complement clinical decision-making at all levels of medical experience [5]. Access is an issue, as there is a limitation of specialized Otolaryngologists, especially in rural and under-served areas, and the capacity of AI -aided diagnostic systems to make such quality of otologic care available widely should not be ignored [6]. These limitations were first overcome by classical machine-learning methods, which included handcrafted feature extraction methods, as well as traditional classifiers such as support-vector machines, which nevertheless were limited by the fact that hand-constructed features were not rich enough to capture the complexities and subtleties of the ear pathologies [7,8]. One of the earliest convolutional neural networks was a significant improvement in that it was capable of learning features automatically, but the limitations of its architecture constrained the ability to extract not only fine-grained textural structure but also global contextual interactions that are needed to perform all-inclusive analysis of otoscopic images.

In recent years, medical image analysis models exhibit an emerging trend of dual-branch and multi-stream architecture, learning to extract complementary information in different domains or processing paths of images and combine it. Time-frequency based analyses, such as the Discrete Wavelet Transform (DWT), have been observed to be particularly useful for extracting high resolution textural features that are very difficult to discern by using traditional spatial domain analysis [9]. In addition to that, attention-based fusion mechanisms exhibit better performance than simple concatenation or average approaches and make the combination of multiple-modal inputs more interpretable and computationally efficient [10]. These architectural developments open a new frontier in building quality, precise diagnostic systems that can be used in the clinical space. The particular task of ear disease classification, which is based on the limitations of the otoscopic imaging, demonstrates the problems that require the implementation of AI to provide solutions that would take into consideration the unique morphological features and be clinically acceptable and interpretable. Variability in anatomical structures, slight pathological morphologies and large variability in imaging conditions require modeling structures that can generalize well to real-life situations [11]. Moreover, medical imaging data sets are often affected by data diversity issues, i.e. the most common ones include class imbalance and small sample size, which require special training protocols and data augmentation strategies to ensure strong performance results [12]. BioOtoFusionNet overcomes the main drawbacks of the existing models of ear disease, i.e., their inability to combine fine-grained indicators (textures) with the global anatomic context, the use of simple or completely absent fusion strategies, and their lack of interpretability specific to the domain of otoscopy. The proposed architecture is a two-branch architecture with each image being processed by both complementary frequency-aware and shape-aware streams. The network detects subtle textural clues of pathology with the help of Discrete Wavelet Transform at Frequency-Aware Stream, and structural boundaries and overall geometry of the ear with the help of Shape-Aware Bio-Attention Stream based on capsule networks and edge-based features. An Adaptive Cross-Fusion Module adds a dynamic balance of these sources of information and it permits the model to report more on texture or shape depending on the clinical presentation. Multi-Scale Attention Pooling ensures diagnostically meaningful information is maintained at all resolutions in order to be scale-robust during inference. An interpretability suite based on the

Frequency Heat Index and an anatomy-guided attention map concept makes the focus of the model an explainable clinical visual explanation resulting in AI-assisted otoscopic diagnosis transparency and trust. Early computational methods were based on classical machine learning models, which relied on feature extraction methods such as Histogram of Oriented Gradients (HOG) and optimization algorithms for feature selection. These methods when combined with other models such as k-Nearest Neighbors (kNN) and SVM were able to achieve high classification accuracies (Up to 92.6%) thus proving that efficient feature selection could compete with more complex deep learning models, particularly in resource-limited settings [13,14]. Deep learning-based ear disease diagnosis systems are now focused on convolutional neural networks (CNNs) that are often augmented by transfer learning and large and diverse image libraries. Ensemble DenseNet variants, for example, have achieved an accuracy of 95.59% for eight classes of otoscopic diseases, showing the importance of model diversity and dataset scale [15]. When a segmentation stage is added—typically Mask R-CNN to isolate the tympanic membrane—follow-on classifiers such as EfficientNet-B0 or Inception-V3 push performance to 97.29% over six categories [16]. More experimental hybrids that marry CNN feature extractors with sequence modules (e.g., LSTMs) and Bayesian hyperparameter tuning have even reported near-perfect results on small, curated sets, though their real-world generalisability remains to be confirmed [17]. Together, these advances show how segmentation-aided pipelines, ensemble learning and carefully optimised architectures together drive state-of-the-art accuracy in automated classification of ear disease. Recent efforts have gone into developing multi-modal studies, combining endoscopic images with the data obtained by pure-tone audiometry, to try to enhance the accuracy of the diagnoses and broaden the scope of the classification to include conditions such as “normal with hearing loss.” Multi-modal Multi-fusion Models based on EfficientNet-B7 achieved an accuracy of 92.9% which was better than single modality models to prove the usefulness of using complementary data sources [18]. Attention-based models, such as those using vision transformers, further enhance feature extraction and classification, achieving up to 99% accuracy and providing reliable decision support for clinicians [19]. The use of generative adversarial networks (GANs) to create synthetic tympanic images has been investigated to augment training datasets. While synthetic images alone yielded moderate accuracy (~70%), their addition to real datasets did not significantly improve performance, suggesting that real clinical images remain crucial for optimal model training [20]. Mobile AI applications have also been developed, enabling point-of-care diagnosis with accuracy comparable to specialists, thus supporting telemedicine and expanding access to expert-level diagnostics in underserved areas [21,22]. Recent studies on attention localization and segmentation have explored hybrid interpretability frameworks that integrate spatial priors or dense supervision to improve localization reliability [23,24]. Motivated by these developments, BioOtoFusionNet adopts anatomy-aware attention analysis to examine whether learned activations align with clinically meaningful structures. Unlike segmentation-based approaches, the proposed framework does not incorporate a segmentation head, but instead employs morphology-informed reference regions solely for interpretability and validation of attention behaviour.

BioOtoFusionNet addresses major gaps of the previous works on otoscopic-AI by combining a frequency-aware stream of Discrete Wavelet Transform with a shape-aware stream of capsule networks, and then fusing complementary information between them by an Adaptive Cross-Fusion Module (ACFM) that re-weights texture and structure differently depending on the pathology of the image. The context-tuned attention and the spatial gating of ACFM go beyond simple concatenation or the static averaging of the context to arrive at context-attended decisions that reflect clinical thinking. In order to ensure transparency, the model incorporates a bespoke interpretability suite Frequency Heat Index, Shape Attention Ratio and Attention Localization Score in addition to a battery of robustness tests to provide clinicians

with quantitative insight into feature attribution, spatial plausibility and confidence. Evaluated against 15 domain-specific metrics, BioOtoFusionNet preserves good performance under illumination change and scale change, demonstrating the convergence of integrated frequency analysis, shape awareness, adaptive fusion, and multi-faceted explainability into one clinically viable otologic-diagnosis engine. While some of the various individual component functions like wavelet-based inputs, edge-enhanced streams, and attention-based fusion have been studied separately in computer vision, existing dual-stream architectures have been based on generic RGB or multi-scale feature hierarchies without a focus on disentangling clinically relevant frequency and structure cues for otoscopic pathology. BioOtoFusionNet is different in that the frequency-domain texture representation is structurally separated from morphology-driven anatomical representation and they are fused together with adaptive cross-fusion intended for disease-dependent cue prioritization. The contribution is in the form of domain-adaptive architectural alignment, and not in the introduction of totally new deep learning primitives.

2. Contributions of the Study:

1. This study overcomes one of the limitations of current approaches to the analysis of otoscopic images, which tend to confuse textural and anatomical inductive cues in a single representation. A clinically motivate dual-branch architecture is introduced that explicitly separates frequency-domain texture information from shape-driven anatomical structure, enabling more targeted and interpretable feature learning for ear disease classification.
2. To overcome the limitations of static feature fusion, an Adaptive Cross-Fusion Module is proposed by modifying the cross fusion between texture and shape evidence as dynamic evidence balance based on the contextual relevance of the evidence. Such adaptive interaction leads to significantly better diagnostic discrimination with overall accuracy of 96.8% and AUC-ROC of 98.4% as compared to simple fusion based on concatenation.
3. The proposed framework includes capsule-based shape attention in order to maintain part-whole anatomical relationships of otoscopic structures to improve sensitivity to subtle structural deformities. This design helps in the good performance of difficult-to-see conditions like Myringosclerosis (F1-score: 93.8%) and Chronic Otitis Media (F1-score: 92.9%) as well.
4. Beyond conventional accuracy-driven evaluation, the study introduces a set of clinically motivated interpretability metrics tailored to otoscopic diagnosis. Metrics such as the Frequency Heat Index (FHI = 0.60), Shape Attention Ratio (SAR = 1.25), and Interpretability Agreement Rate (IAR = 0.90) provide intuitive insight into how the model balances texture and anatomical attention in clinically relevant regions.
5. Robustness and reliability challenges inherent to otoscopic imaging such as scale variability and illumination changes are explicitly addressed through multi-scale attention modelling and stability analysis. High Multi-Scale Activation Consistency (MSAC = 0.87) and low Robustness to Illumination Index (RII = 0.15) demonstrate consistent diagnostic focus under realistic acquisition conditions.
6. Extensive experimental evaluation over several categories of ear diseases justifies the validation of the proposed architecture in improving the sensitivity to the coarse anatomical anomalies as well as fine grain textural anomalies, resulting in high class-wise performances, including 97.8% accuracy for Earwax Plug and 97.1% accuracy for Normal Ear, and proves its applicability in real-world clinical and telemedicine scenarios as well.

3. Methodology

This study presented BioOtoFusionNet, a clinically motivated dual-branch framework for the automated classification of external and

middle ear diseases. The methodology is organized into six key phases: data collection, preprocessing, dual-branch feature extraction, adaptive feature fusion, classification, and interpretability analysis as shown in Fig. 1.

3.1. High-Level Architectural Overview

In order to give a conceptual insight into the framework, this subsection presents the general design principles of BioOtoFusionNet. The architecture in Fig. 2. is two complementary representational streams that extract diagnostically relevant characteristics of otoscopic imagery. The Frequency-Aware Stream works on sub-bands extracted using the wavelet domain which allows the retrieval of fine-scale textural characteristics such as granularity, surface irregularities and high frequency patterns that correspond to certain pathologic conditions. Concurrently, the Shape-Aware Stream focuses on the anatomical structure using edge-based and morphology-driven descriptors that capture contour deviations, membrane deformations, and other anatomical structure signatures, which are clinically relevant.

These heterogeneous representations are combined using an Adaptive Cross-Fusion module which dynamically modulates the contribution of each stream according to the contextual relevance of features. This fusion strategy ensures that one of the two characteristics (texture or structure in this case) is prioritized depending on the characteristics of the underlying disease. Following fusion, there are attention-based refinement mechanisms which that improve discriminatory areas by reinforcing spatially coherent and diagnostically relevant feature activations.

This high-level overview provides the conceptual rationale and functional interaction of the architectural components giving an informative overview for the design of the model before the detailed methodological exposition offered in the following sections.

3.2. Phase 1: Data Collection

The image dataset utilized in this study comprises 880 otoscopic images sourced from an open-access repository available on Figshare 2526 titled as "Ear imagery database". These images represent the conditions of the external and middle ear, and are classified in four clinical classes: Earwax Plug, Myringosclerosis, Chronic Otitis Media, and Normal Ear as seen in Fig. 3. Captured using standard otoscopic equipment under clinical viewing conditions, these images represent real-world variations in lighting conditions, viewing angles, and anatomical diversity. In order to ensure robust and balanced training of the model, the dataset was divided into training, validation and test sets using stratified 5-fold cross-validation, where the proportion of each class in all the folds remained the same. The dataset consists of 220 images per class before augmentation resulting in an initially balanced dataset of Earwax Plug, Myringosclerosis, Chronic Otitis Media and Normal Ear. Stratified 5-fold cross-validation was used to ensure this proportional representation across subsets to train and evaluate the models in order to reduce the risk of class distribution bias during optimization.

3.3. Phase 2: image preprocessing and frequency decomposition

In order to improve features relevant for the diagnosis and to suppress the background noise, each otoscopic image was pre-processed in a structured pipeline. Initially, Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to enhance contrast within areas of low illumination of the tympanic membrane which is especially useful to enhance low-level textural details of the tympanic membrane. Let $I(x, y)$ denote the grayscale intensity of the input image; CLAHE enhances local contrast by redistributing the histogram within contextual regions, avoiding over-amplification of noise. Following this, Adaptive Gamma Correction was employed to normalize brightness across varying light-

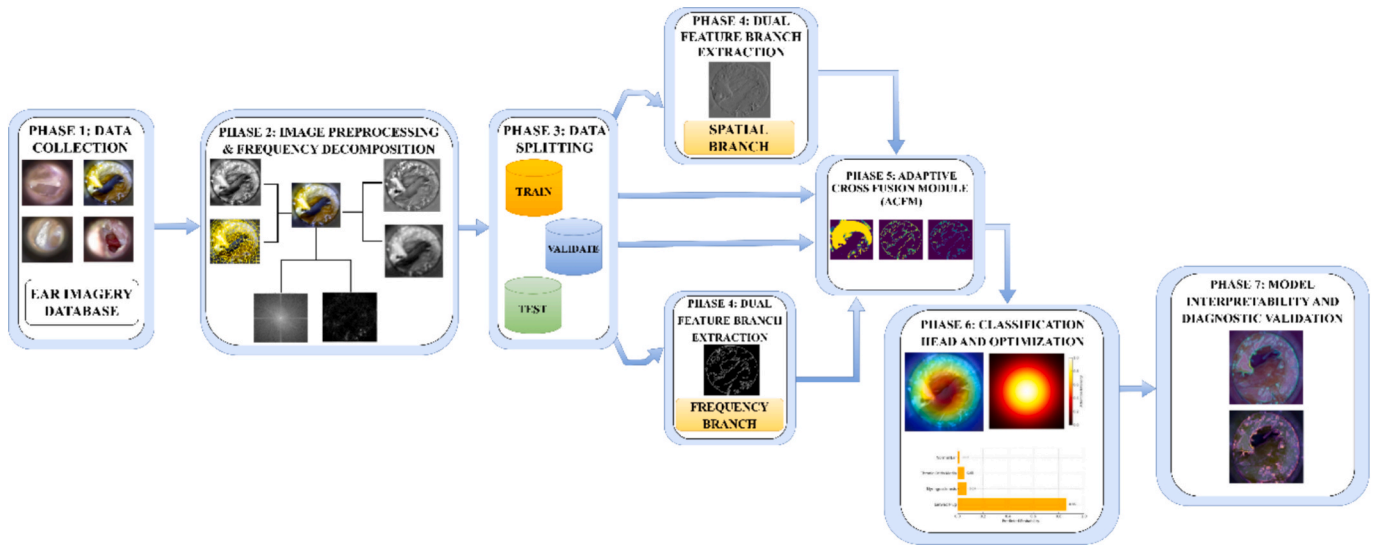


Fig. 1. BioOtoFusionNet workflow diagram for ear disease diagnosis.

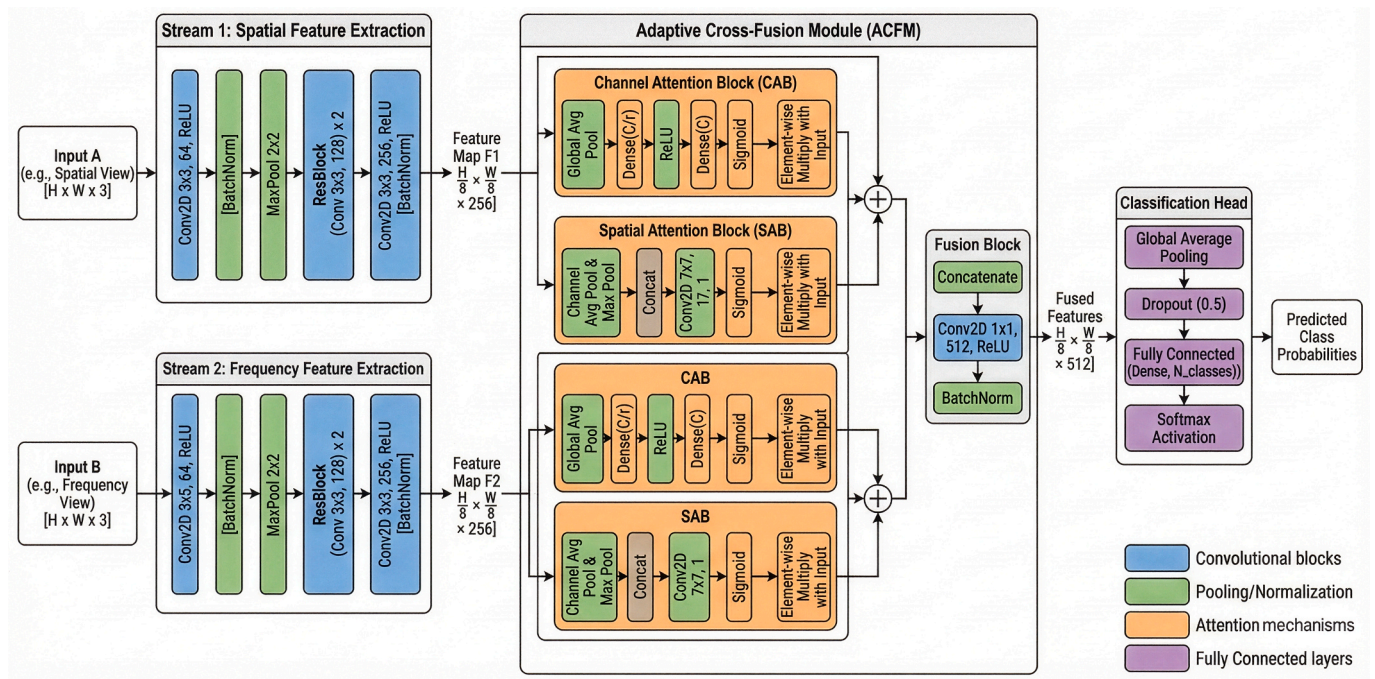
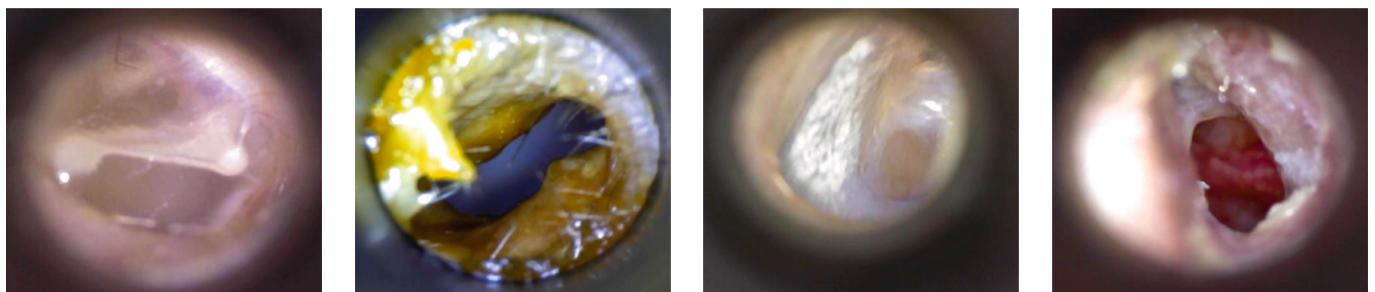


Fig. 2. Proposed Dual Stream Network Architecture with Adaptive Cross Fusion Module (ACFM).



a) Normal

b) EarWax Plug

c) Myringosclerosis

d) Chronic Otitis Media

Fig. 3. Sample Images from the Dataset [25,26].

ing conditions. The corrected image $I_\gamma(x, y)$ is computed as Eq. (1):

$$I_\gamma(x, y) = c \cdot I(x, y)^{\gamma(x, y)} \quad (1)$$

where $\gamma(x, y)$ is spatially adaptive based on local luminance, and c is a normalization constant ensuring pixel values remain within valid intensity bounds. To localize the anatomical regions of interest and suppress irrelevant background content, Suprapixel Segmentation was performed using Simple Linear Iterative Clustering (SLIC). This algorithm groups pixels into compact, perceptually uniform clusters $\mathcal{S}_k$ by minimizing a combined color-spatial distance as Eq. (2):

$$D = \sqrt{\left(\frac{\Delta I}{N_I}\right)^2 + \left(\frac{\Delta a}{N_a}\right)^2 + \left(\frac{\Delta b}{N_b}\right)^2 + \left(\frac{\Delta x}{N_x}\right)^2 + \left(\frac{\Delta y}{N_y}\right)^2} \quad (2)$$

where ΔI , Δa , Δb are differences in CIELAB color space, and Δx , Δy are spatial distances; N^* are normalization terms based on image dimensions and color ranges. Finally, each image was projected into the frequency domain using the Fast Fourier Transform (FFT). Let $f(x, y)$ be the spatial image; the corresponding frequency representation $F(u, v)$ is given by Eq. (3):

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cdot e^{-2\pi i \left(\frac{ux}{M} + \frac{vy}{N} \right)} \quad (3)$$

This transform provides both high frequency components (high frequency components are important for identifying fine details of texture in audio that can be relevant for identifying wax granularity and tympanic sclerosis) and low frequency components (low frequency components characterize global structural information like eardrum shape and

position). These components were then fed as input to dedicated branches in the proposed dual stream network architecture as seen in Fig. 4.

3.4. Phase 3: Data splitting

To ensure robust evaluation and minimize bias arising from limited data, the dataset was partitioned using a stratified 5-fold cross-validation strategy. Let D represent the complete dataset containing $N = 880$ otoscopic images, and let $C = \{c_1, c_2, c_3, c_4\}$ denote the four diagnostic classes: *Earwax Plug*, *Myringosclerosis*, *Chronic Otitis Media*, and *Normal Ear*. The stratification ensures that the proportion of samples from each class $c_i \in C$ is maintained across all folds, i.e., as Eq. (4),

$$\forall i \in \{1, 2, \dots, 5\}, \frac{|D_i^{(c)}|}{|D_i|} \approx \frac{|D^{(c)}|}{|D|} \quad (4)$$

where D_i is the dataset for fold i , and $D_i^{(c)} \subset D_i$ contains samples from class c .

In each turn of the cross-validation one of the folds is held out as the test set whereas the other four folds are further divided into training, testing and validation sets in a ratio of 80:10:10. This gives a good and generalizable performance estimation and mitigates the risk of overfitting. Importantly, as the dataset inherently has inter-class visual similarity, stratified splitting allows to avoid class imbalance situation in any subset and the training dynamics and metric computation are stabilized as one can observe in Table 1.

3.4.1. Generalization control, data partitioning, and overfitting prevention

Given the limited size of clinically curated otoscopic data sets,

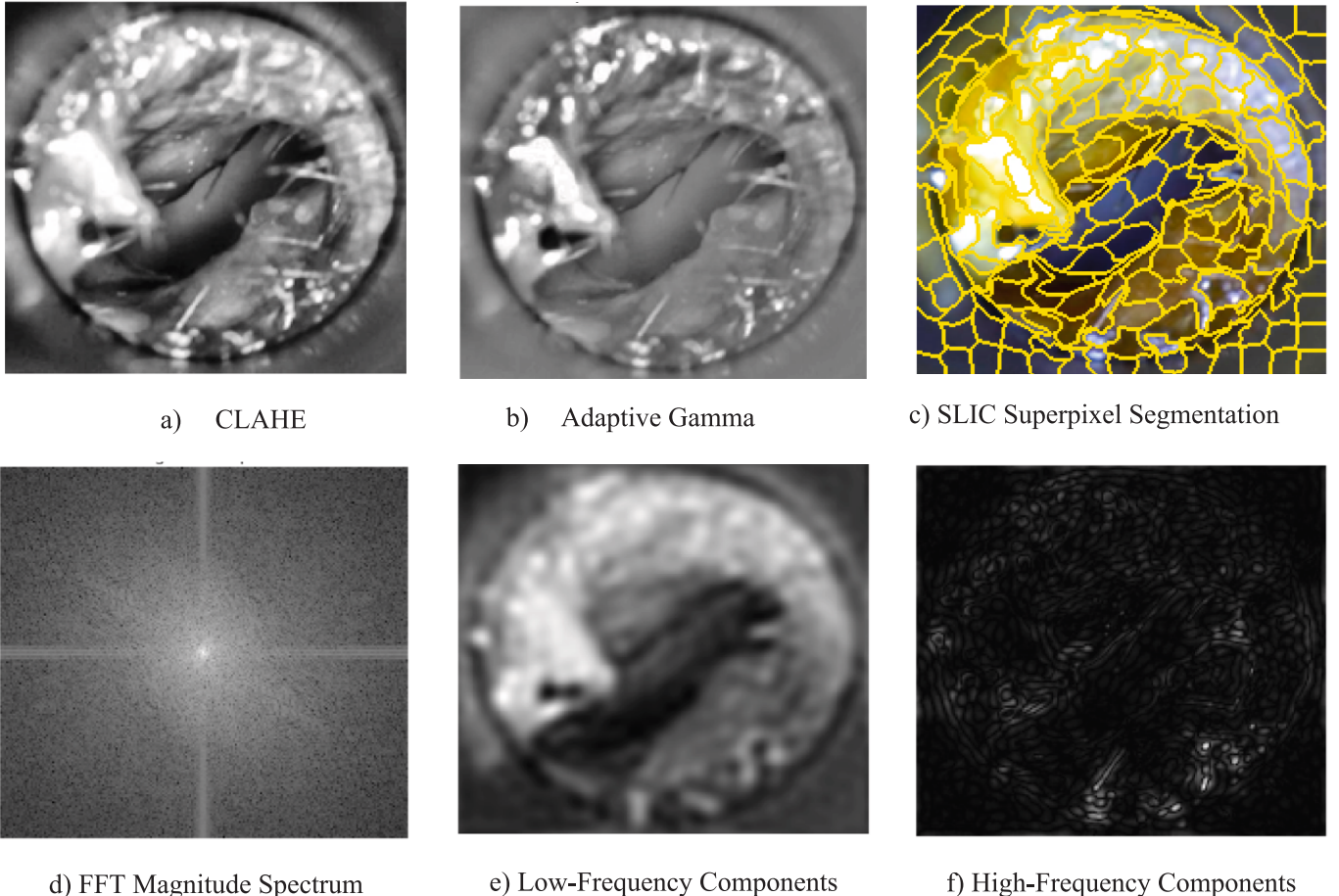


Fig. 4. CLAHE, gamma, SLIC, and frequency decomposition outputs.

Table 1
Dataset Distribution Before and After Augmentation.

Class	Original Total	Training (Before Augmentation)	Training (After Augmentation)	Validation	Test	Total After Augmentation
Earwax Plug	220	140	420	36	44	500
Myringosclerosis	220	140	420	36	44	500
Chronic Otitis Media	220	140	420	36	44	500
Normal Ear	220	140	420	36	44	500
Total	880	560	1680	144	176	2000

explicit methodological safeguards were imposed to reduce overfitting and allow for reliable generalisation assessment. All experiments were performed with stratified 5-fold cross-validation to maintain the proportion of class distribution in the training set and validation set, in order to minimize sampling bias. Performance measures were calculated as mean and standard deviation over folds in order to assess stability and reduce fold-specific performance inflation. Data augmentation techniques were only used for training partitions in each fold of the training data, while validation and test subsets were not augmented. This stringent separation allowed synthetic leakage and ensured that reported metrics were a measure of true generalization and not memorization of augmented patterns. Furthermore, augmentation parameters were randomly sampled during training in order to increase intra-class variability, without changing the integrity of the evaluation. The data set, which was obtained from a publicly available data repository, did not contain explicit patient-level identifiers. For that reason, cross-validation was applied at the image level. While stratification proved to maintain proportional class representation, segregation between the subjects could not be enforced expressly. This introduces the possibility of intra-subject correlation in case more than one image from the same person is included in the data set. To reduce the potential for bias due to this limitation, further robustness analyses were performed and included inter-fold variance assessment, testing under unseen clinical variations and under reduced training data proportions. The steady low between-fold variance and progressive performance degradation under data reduction are an indication of stable optimization behaviour rather than dataset specific memorization. Although these internal validation strategies help to support controlled model capacity with respect to dataset scale, there is still a lack of subject-level metadata. Future studies that include patient-level partitioning with multi-center external validation will be required to further demonstrate generalizability across institutions and clinical validity. The dataset does not include any patient identifiers; hence stratified image level 5-fold cross validation was implemented.

3.4.2. Data Leakage Prevention and Overfitting Monitoring

To prevent the leakage of data, data set partitioning was done before any augmentation and preprocessing steps. Stratified 5-fold cross validation was carried out at the level of the images and augmentation was only applied to training partitions of each fold. Validation and test subsets were not modified to avoid biased performance estimation. None of the augmented samples were shared between folds. Overfitting was observed in several ways. Performance of training and validation was monitored separately at each epoch and early stopping was used depending upon stabilization of validation loss. Additionally, inter-fold variance analysis and confidence interval estimation were done to test stability between splits. The lack of big divergence between training and validation performance as well as low standard deviation between folds is a sign of controlled optimization behaviour.

3.5. Phase 4: Dual-Branch Feature Extraction

The proposed BioOtoFusionNet architecture uses two-branch feature extraction architecture to learn from both frequency-based and spatially structured patterns inherent in otoscopic images at the same time. These two parallel branches, Frequency-Aware Stream and Shape-Aware Bio-

Attention Stream, are aimed at extracting complementary features that are very important in differentiating between fine-grained pathologies in the ear. In the Frequency-Aware Stream, each input image $I(x,y)$ is first decomposed with the Discrete Wavelet Transform (DWT), which captures both the high frequency of the edge details and the low frequency of the structural information. The decomposition yields a set of sub-bands as in Eq. (5):

$$DWT(I) = \{A_j, H_j, V_j, D_j\}_{j=1}^J \quad (5)$$

where A_j represents the approximation coefficients, and H_j, V_j, D_j correspond to horizontal, vertical, and diagonal detail coefficients at scale j . These coefficients emphasize different aspects of local texture, such as the granular distribution in wax plugs or the calcified regions in myringosclerosis. The wavelet-transformed data is then processed through a Res2Net-based CNN, which employs hierarchical residual-like connections within each block to capture multi-scale texture features. Mathematically, a Res2Net block output at stage s with group splits k is represented as Eq. (6):

$$Y_s = \sum_{i=1}^k F_i(X_i), \text{ where } X_i = X_{i-1} + F_{i-1}(X_{i-1}) \quad (6)$$

and F_i denotes the transformation function of the i -th split, allowing finer granularity and enhanced representation of frequency-localized patterns.

The frequency-aware stream employs the Discrete Wavelet Transform (DWT) to decompose the input image into sub-bands that capture different frequency components. The Approximation (A) sub-band preserves the global structural information, including the overall shape of the ear canal and tympanic membrane. The Horizontal (H), Vertical (V), and Diagonal (D) detail coefficients extract fine texture features, effectively highlighting patterns such as wax granularity and calcified plaques characteristic of conditions like myringosclerosis as seen in Fig. 5.

In parallel, the Shape-Aware Bio-Attention Stream targets the spatial configurations of anatomical structures within the ear. A standard convolutional encoder (e.g., ResNet-18) is used to obtain base spatial features $f_{\text{spatial}}(x, y) \in \mathbb{R}^{H \times W \times C}$, where H , W , and C denote the height, width, and number of channels, respectively. These features are then fed into a Capsule Network layer, which transforms each spatial feature vector into a capsule vector $v_i \in \mathbb{R}^d$, representing instantiation parameters (e.g., pose, orientation, deformation) of localized anatomical parts. The dynamic routing between capsules ensures that part-whole relationships are retained, governed by the coupling coefficients c_{ij} , computed via a softmax function as Eq. (7):

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}, \text{ where } b_{ij} \text{ is the log prior probability between capsule } i \text{ and } j \quad (7)$$

The capsule outputs are aggregated through a squash function to ensure normalized vector lengths, thus encoding the probability and pose of anatomical structures as in Eq. (8):

$$\text{squash}(s_j) = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \cdot \frac{s_j}{\|s_j\|} \quad (8)$$

The shape-aware stream uses structural information by first using Canny

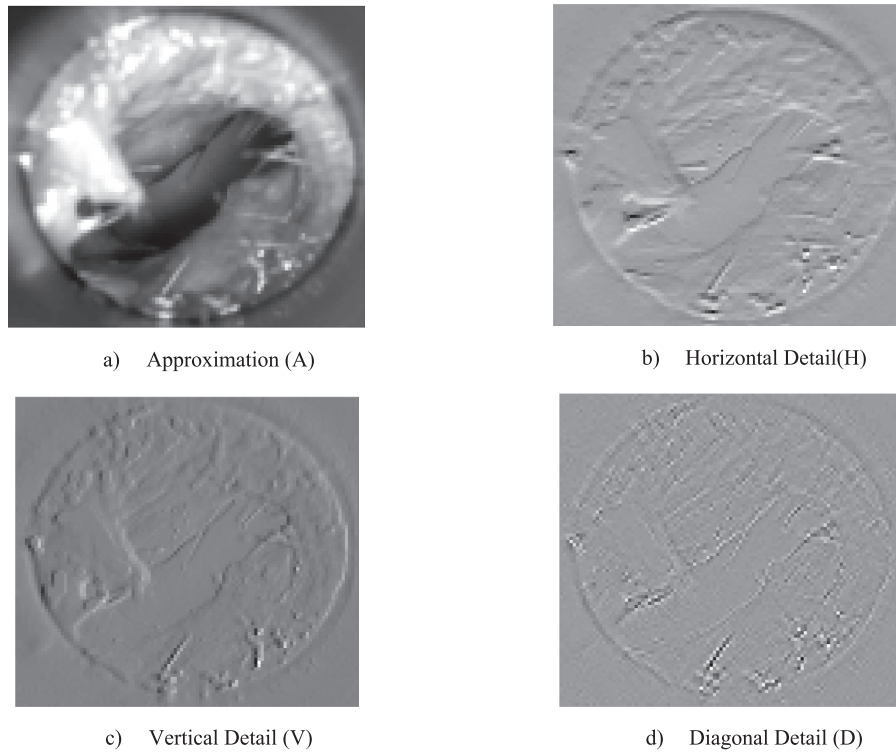


Fig. 5. DWT Sub-bands (Frequency-Aware Stream).

edge detection to define sharp anatomical boundaries such as the tympanic-membrane rim or the edges of perforations. A morphological gradient then brings up transitional areas, bulges, contour irregularities and slight deformations by measuring intensity differences between adjacent areas. Finally, skeletonization is used to reduce these features to a one pixel wide representation preserving the underlying ear anatomy in a condensed form for robust part-whole relationship modelling as seen in Fig. 6. In the Frequency-Aware Stream, the one-level 2D Discrete Wavelet Transform is used to divide each input image into 4 sub-bands: approximation (A) and detail components (H, V, D). Each sub-band is normalized individually and resized to the same spatial resolution and then they are stacked along the channel dimension to form a multi-channel tensor. This stacked representation is directly fed as input to the Res2Net backbone without further weighting of pre-fusion and thus enables the network to learn inter-band relations implicitly during feature extraction. Res2Net was particularly chosen to

be complementary with the multi-band feature nature of DWT, which the split-scale residual design facilitates efficient multi-scale feature modeling within a block. This property is naturally in line with hierarchical frequency decomposition of DWT and thus the network can meet the requirement of jointly learning fine-grained texture details from high-frequency bands and coarse structural details from low-frequency components.

This biologically inspired attention mechanism enables the network to preserve spatial hierarchies and robustly extract anomalies in the shape (e.g. bulging or perforated tympanic membranes) in chronic otitis media. By having frequency-oriented and shape-aware representations available, this two-stream extraction approach enables BioOtoFusionNet to detect subtle yet crucial differences among ear pathologies, even under different imaging conditions.

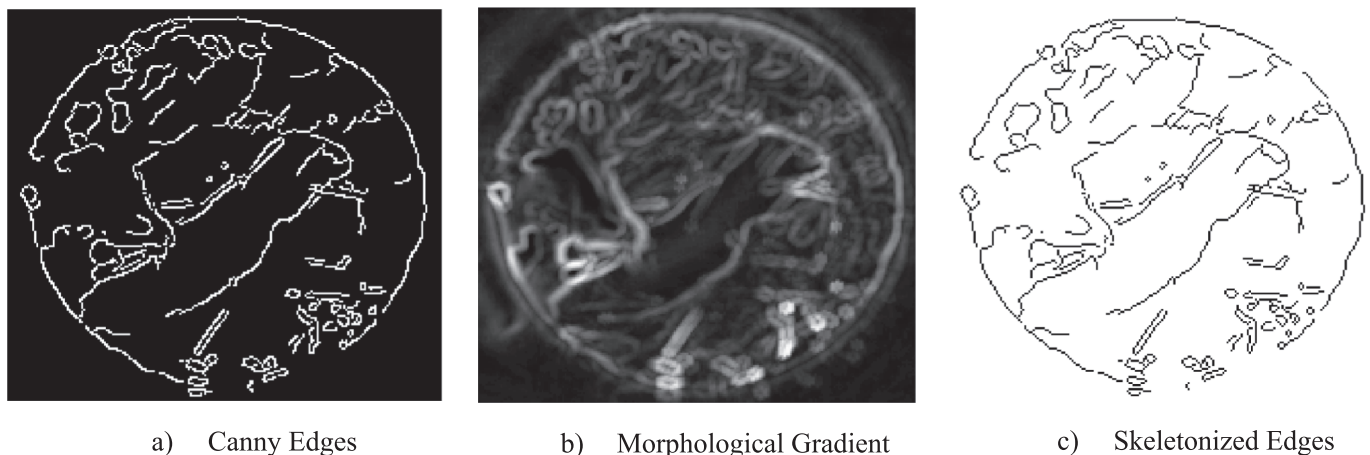


Fig. 6. Shape-Aware Stream Features.

3.6. Phase 5: Adaptive Cross-Fusion Module (ACFM)

To effectively integrate diagnostic features extracted from both the frequency-aware stream and the shape-aware bio-attention stream, we designed a specialized Adaptive Cross-Fusion Module (ACFM) tailored for otoscopic image analysis. This module dynamically determines the relative importance of texture- and structure-based cues critical for distinguishing between visually subtle ear conditions such as myringosclerosis and chronic otitis media. Let $F_f \in \mathbb{R}^{H \times W \times C}$ represent the feature map generated from the wavelet-based frequency-aware branch, capturing fine-grained textures like calcified spots or wax granules. Similarly, let $F_s \in \mathbb{R}^{H \times W \times C}$ denote the output of the capsule-based spatial stream, encoding anatomical deformations such as bulging or perforated tympanic membranes.

To adaptively recalibrate the importance of each channel in the frequency-based feature map, we apply a Squeeze-and-Excitation (SE) mechanism as Eq. (9):

$$z_c = \frac{1}{H \cdot W} \sum_{i=1}^H \sum_{j=1}^W F_f(i, j, c), \forall c \in \{1, \dots, C\} \quad (9)$$

This global context vector $z \in \mathbb{R}^C$ is passed through a bottleneck transformation to yield channel attention weights $a \in \mathbb{R}^C$ as Eq. (10):

$$a = \sigma(W_2 \cdot \delta(W_1 \cdot z)), \hat{F}_f = a \odot F_f \quad (10)$$

where W_1 and W_2 are learnable matrices, δ is the ReLU activation, and σ is the sigmoid function. This operation highlights disease-specific frequency components (e.g., textured wax vs. smooth tissue).

To complement this, a spatial gating mechanism identifies key regions, such as the tympanic membrane or central canal, that contribute most to classification. A spatial mask $M \in \mathbb{R}^{H \times W}$ is computed by Eq. (11):

$$M = \sigma(\text{Conv}_{1 \times 1}([F_s \| F_f])) \quad (11)$$

where $[F_s \| F_f]$ denotes channel-wise concatenation. This mask modulates both streams, allowing the network to focus on spatially salient regions such as lesions or anatomical landmarks. Further, to guide the fusion based on feature agreement, a correlation-based attention map $\mathcal{A} \in \mathbb{R}^{H \times W}$ is computed using cosine similarity at each spatial location as seen in Eq. (12):

$$\alpha_{ij} = \frac{\hat{F}_f(i, j) \cdot F_s(i, j)}{\|\hat{F}_f(i, j)\|_2 \cdot \|F_s(i, j)\|_2} \Rightarrow A = \text{Normalize}(\alpha) \quad (12)$$

This attention map assigns greater weight to the stream that holds more class-discriminative information at each pixel location. The final fused representation $F_{fused} \in \mathbb{R}^{H \times W \times C}$ is computed as Eq. (13):

$$F_{fused} = A \odot \hat{F}_f + (1 - A) \odot F_s \quad (13)$$

This formulation allows the model to prioritize texture cues for cases like myringosclerosis and earwax plug, while emphasizing morphological deformation patterns for chronic otitis media. By enabling adaptive, disease-specific fusion, ACFM plays a crucial role in generating a semantically enriched, class-aware representation for accurate ear disease classification.

The Spatial Gating Mask (M) highlights critical regions in the otoscopic image, such as the eardrum contour and canal, by adaptively enhancing spatially salient areas based on combined frequency and shape inputs. The Correlation Attention Map (A) visualizes pixel-level agreement between the two feature streams, computed via cosine similarity, indicating whether the model relies more on texture (from frequency) or structure (from shape) at each location. The Fused Feature Map (Channel 0) displays one channel of the integrated representation, showcasing how BioOtoFusionNet combines both modalities to produce

a semantically enriched, class-relevant feature set as shown in Fig. 7.

3.7. Phase 6: classification head and optimization

Following adaptive fusion, the enriched feature representation $F_{fused} \in \mathbb{R}^{H \times W \times C}$ is forwarded to a lightweight classification head designed to retain diagnostic granularity across multiple spatial resolutions. In order to make sure that subtle patterns like micro-calcifications in myringosclerosis or the irregular contouring in the tympanic membrane in chronic otitis media are not suppressed during the down-sampling process, we employ Multi-Scale Attention Pooling (MSAP). This technique combines features from multiple spatial scales in an aggregate way but with a weight that is adaptive to local importance.

Let $F_{fused}^s \in \mathbb{R}^{h_s \times w_s \times C}$ denote down sampled versions of F_{fused} at different scales $s \in \{1, 2, \dots, S\}$. For each scale, we compute a spatial attention map $A^s \in \mathbb{R}^{h_s \times w_s}$ using a learnable attention kernel as Eq. (14):

$$A^s = \sigma(\text{Conv}_{1 \times 1}(F_{fused}^s)) \quad (14)$$

The attention-weighted pooled vector at scale s is in Eq. (15):

$$f^s = \sum_{i=1}^{h_s} \sum_{j=1}^{w_s} A^s(i, j) \cdot F_{fused}^s(i, j) \quad (15)$$

These pooled vectors from all scales are concatenated into a unified embedding $f_{final} = [f^1 \| f^2 \| \dots \| f^S] \in \mathbb{R}^K$, which serves as input to the final dense layer. This layer computes the class probability distribution over the four diagnostic categories: Earwax Plug, Myringosclerosis, Chronic Otitis Media, and Normal Ear, using a softmax activation as seen in Eq. (16):

$$\hat{y}_k = \frac{\exp(w_k^T f_{final} + b_k)}{\sum_{j=1}^4 \exp(w_j^T f_{final} + b_j)}, \forall k \in \{1, 2, 3, 4\} \quad (16)$$

where w_k and b_k are the learnable parameters associated with class k , and $\hat{y}_k$ is the predicted probability for that class.

Given the potential class imbalance in the dataset, especially the underrepresentation of specific disease conditions, the model is optimized using Focal Loss, which down-weights easy examples and focuses learning on difficult cases as seen in Fig. 8. For a given ground truth class $y \in \{1, 2, 3, 4\}$ and predicted probability $\hat{y}_y$, the focal loss is defined as Eq. (17):

$$\mathcal{L}_{focal} = -\alpha_y (1 - \hat{y}_y)^\gamma \log(\hat{y}_y) \quad (17)$$

where $\alpha_y \in [0, 1]$ is a weighting factor for class y , and $\gamma > 0$ is the focusing parameter. In our experiments, values of $\gamma = 2$ and class-wise α values inversely proportional to class frequency were found effective. This formulation ensures that the classifier is especially attentive to underrepresented classes such as Myringosclerosis or borderline Chronic Otitis Media, enhancing overall diagnostic reliability as seen in Fig. 9.

3.8. Phase 7: model interpretability and diagnostic validation

To make the prediction of diagnosis clinically relevant, transparent and trustworthy, a specialized interpretability module was added to the BioOtoFusionNet framework, specifically for the analysis of otoscopic images. Model decisions were explained based on Layer-wise Class Activation Mapping (LayerCAM) which visualises spatial regions of an image that contributed significantly to each class prediction. For a given class c , the attention map $A^c(x, y)$ at spatial location (x, y) is computed from a convolutional feature map $F^l \in \mathbb{R}^{H \times W \times C}$ at layer l as Eq. (18):

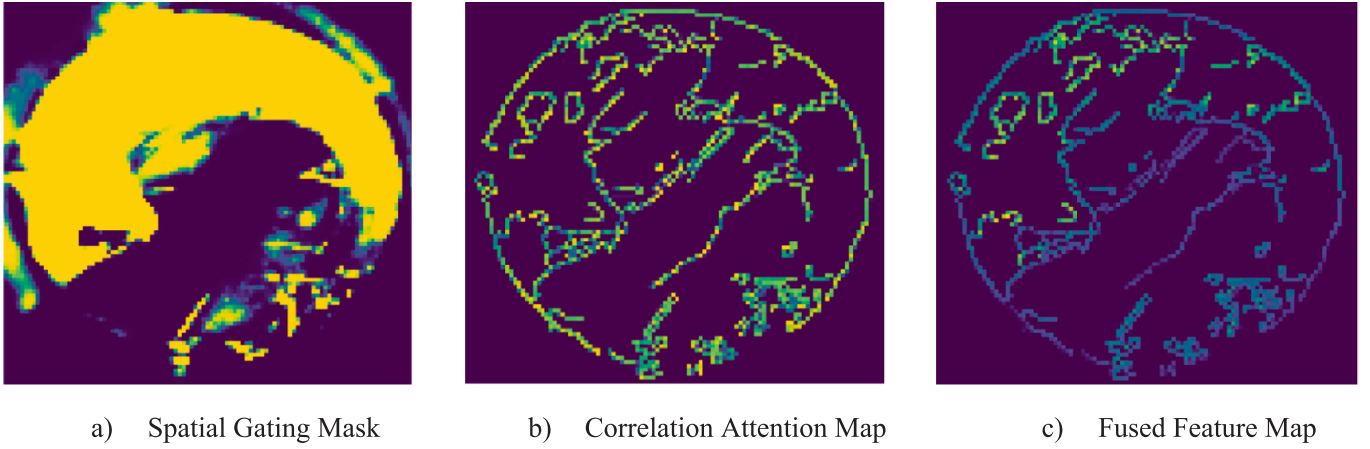


Fig. 7. Spatial mask, correlation, and fused feature maps.

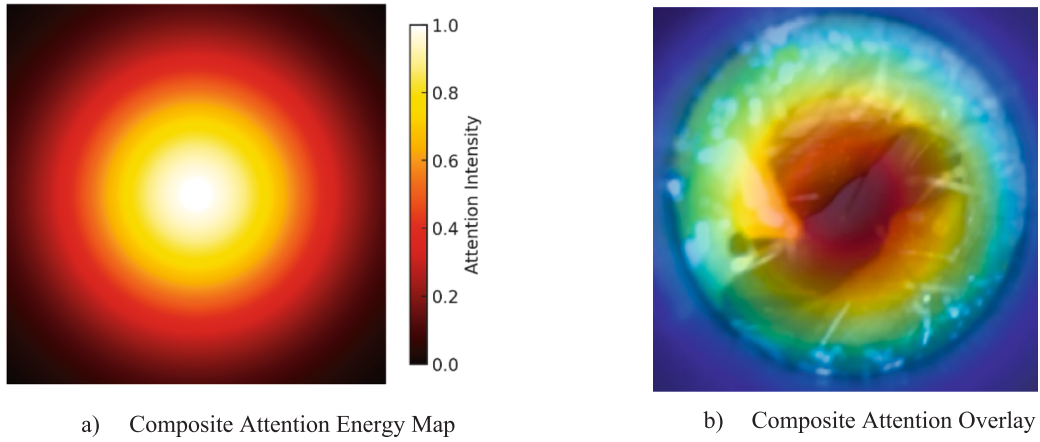


Fig. 8. Composite attention maps highlighting multi-scale feature pooling.

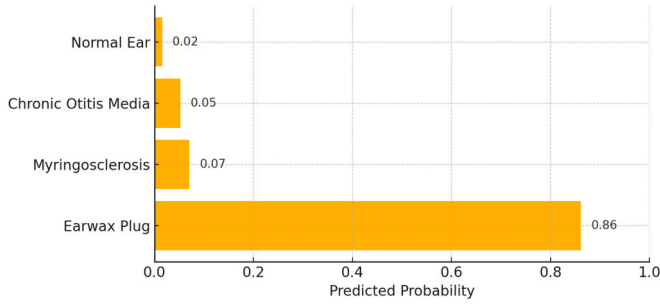


Fig. 9. Final Class Probabilities.

$$A^c(x, y) = \sum_{k=1}^C \text{ReLU} \left(\frac{\partial L_c}{\partial F_{x,y,k}^l} \cdot F_{x,y,k}^l \right) \quad (18)$$

where L_c is the classification loss for class c , and $F_{x,y,k}^l$ represents the activation at channel k . This mechanism is effective in highlighting areas such as tympanic membrane calcification in myringosclerosis or central occlusions in earwax plugs, and therefore, attention from the model can be interpreted in a visual way. To further quantify the model's reliance on frequency-based versus spatial features, a novel metric termed the Frequency Heat Index (FHI) was introduced. Given the LayerCAM attention map $A^f(x, y)$ from the frequency-aware stream F_f , and the full attention map $A^t(x, y)$ from the fused representation, the FHI is

calculated as Eq. (19):

$$FHI = \frac{\sum_{(x,y) \in \Omega_f} A^f(x, y)}{\sum_{(x,y) \in \Omega} A^t(x, y)} \quad (19)$$

where $\Omega_f \subseteq \Omega$ defines the spatial domain corresponding to high-frequency texture regions (e.g., granular surfaces, reflective wax patterns). A higher FHI score correlates with an increased reliance on frequency features, which are common in the classification of Earwax Plug and Myringosclerosis while a low FHI score is associated with the utilization of structural features, such as tympanic membrane distortion common in Chronic Otitis Media. All generated heatmaps and FHI scores were reviewed by expert review by certified otolaryngologists. The interpretability module thus helps the diagnostic process not only with its high prediction accuracy but also with the clinically aligned visual and quantitative justifications, which are necessary to deploy in decision support environments such as in Fig. 10.

4. Experimentation results

4.1. Class-wise LayerCAM heatmaps

The visual results in Fig. 11 from the class-specific LayerCAM heatmaps show a high accuracy of BioOtoFusionNet in targeting unique pathological attributes relevant to each otoscopic diagnostic category. For the Earwax Plug, the activation map clearly shows bright yellow areas within the external auditory canal, which acted as evidence of the model's successful attention towards wax buildup and canal blockage. In

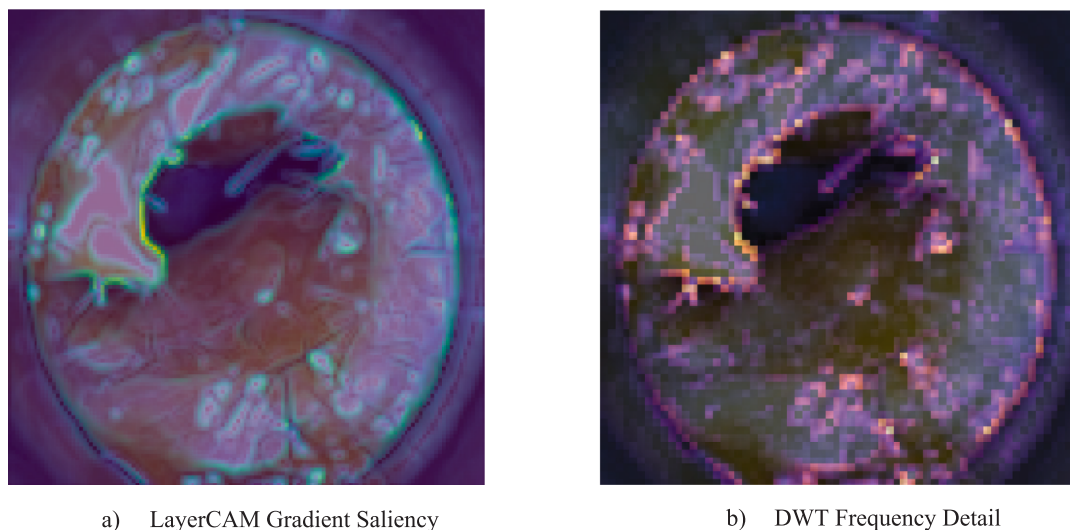


Fig. 10. LayerCAM saliency and DWT frequency detail visualizations.

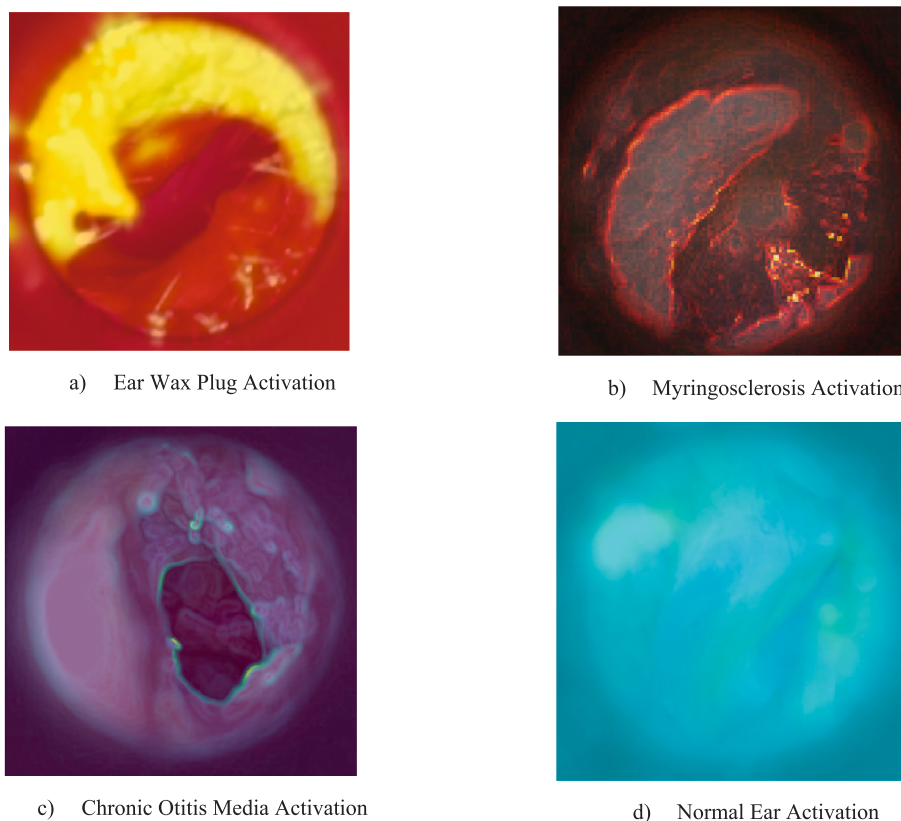


Fig. 11. Class-wise LayerCAM heatmaps highlighting disease-specific activations.

the case of Myringosclerosis, the heatmap clearly highlights calcified patches on the tympanic membrane confirming that the model is robust in capturing high frequency texture features representative of sclerotic lesions. For Chronic Otitis Media, the visualization is detailed showing the structural irregularities, such as the deformation of the membranes and the central perforation edges, which shows the sensitivity of the model to anatomical distortions. The comparative study of the activation heatmaps reveals that the Normal Ear creates a low-intensity pattern with homogenous distributions which means that it lacks pathological attributes and, therefore, is the validity of the discriminative power of the model. Collectively, these visual tests testify to the

accuracy and clinical readability of BioOtoFusionNet, which makes it plausible to use it as a diagnostic adjunct in the practice of otolaryngology.

4.2. Class-wise diagnostic analysis

The diagnostic analysis of BioOtoFusionNet performed on a class-by-class basis shows strong results on all diagnostic categories, most notably on the conditions that are viewed as challenging, such as Myringosclerosis and Chronic Otitis Media. More precisely, the model scored impressive accuracy on both Earwax Plug (97.8 %) and Normal Ear

(97.1 %), which shows high specificity and sensitivity, implying the presence of reliable detection properties of both distinctly pathological and non-pathological manifestations. More importantly, the accuracy enhancements achieved by Myringosclerosis (91.7 %) and Chronic Otitis Media (90.5 %) demonstrate the great advantage brought by the combination of frequency- and shape-aware feature extraction in terms of sensitivity. These increased sensitivity scores signify that the suggested dual-stream fusion strategy is highly successful in solving the diagnostic issues created by subtle textural abnormalities and structural deformations that are usually linked to these conditions. Moreover, the specificity values are extremely high throughout all classes (95.4 % to 98.5 %), which indicates that the level of false positives is very low, thereby contributing to the clinical reliability and applicability of BioOtoFusionNet. Collectively, these metrics demonstrate that the strategic combination of frequency-based texture analysis and spatially-aware structural modeling significantly enhances diagnostic precision, particularly in nuanced pathological contexts as seen in Table 2. In addition to overall accuracy, macro-averaged F1-score and per-class sensitivity were emphasized to account for potential class imbalance effects. Sensitivity values remain consistently high across all four categories, with no substantial degradation observed in visually subtle classes such as Chronic Otitis Media. This balanced performance indicates that classification improvements are not driven by dominance of a specific disease category. Furthermore, the use of Focal Loss during optimization mitigates class-level bias by emphasizing harder-to-classify samples.

4.3. Interpretability results

The results of the interpretability analysis in Table 3, which was quantified with the novel Frequency Heat Index (FHI), give pertinent information about the unique feature modalities being prioritized by BioOtoFusionNet for different otoscopic diagnostic categories. For Earwax Plug, the high mean FHI score (0.82) indicates a strong reliance on frequency-based texture patterns, which is further evidence of the model's targeted sensitivity to wax granularity and textural variations related to canal blockage. Similarly, the FHI value for Myringosclerosis (0.78) is relatively high, reinforcing the model's high dependence on frequency-based features and in particular calcific textures on the TM, strengthening the excellent role of texture analysis in the correct diagnosis of this condition. The results suggest existence of significant difference in feature hierarchy index (FHI) for the studied conditions. Within the framework of chronic otitis media, the significantly low FHI (0.36) represent the strategy biased towards structural cues, specifically, distortion of membrane, which is also in line with the anatomical nature of the target pathology and the increased sensitivity of the model to spatial data. In comparison, the normal ear class (FHI 0.45) is slightly higher indicating more equal balance between spectral and structural characteristics. When the above combination, the given observations help validate the hypothesis that the model has context sensitive flexibility in terms of its feature selection mechanism. Collectively, these interpretability findings validate that BioOtoFusionNet selectively and effectively employs clinically relevant feature modalities, facilitating improved diagnostic transparency and enhancing the model's potential utility in clinical decision-support scenarios.

Table 2

Class-wise diagnostic performance metrics for all categories.

Class	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Earwax Plug	97.8	96.2	98.5	97.4
Myringosclerosis	94.5	91.7	96.1	93.8
Chronic Otitis Media	93.2	90.5	95.4	92.9
Normal Ear	97.1	95.6	98.2	96.4

Table 3

Class-wise Frequency Heat Index scores and feature emphasis.

Class	Mean FHI Score	Feature Type Emphasized
Earwax Plug	0.82	Frequency (texture)
Myringosclerosis	0.78	Frequency (calcification)
Chronic Otitis Media	0.36	Shape (membrane distortion)
Normal Ear	0.45	Mixed

4.4. Multi-scale attention pooling (MSAP) feature aggregation

The visualization in Fig. 12 of the Multi-Scale Attention Pooling (MSAP) mechanism illustrates the efficacy of BioOtoFusionNet in adaptively aggregating diagnostically relevant features across multiple spatial resolutions. The deep-learning model BioOtoFusionNet trained on otoscopic classification has continued to show excellent stability in multigrade pathologies. The constant occurrence of severe anatomical changes, e.g. perforations of the tympanic membranes or structural disfigurements on fine, medium and coarse levels, is visible in their respective visualizations, proving the model to be effective in measuring large-scale structural abnormalities regardless of the image size. Similarly, attention maps linked to Myringosclerosis showed very high level of focus on the calcification patterns at all scales and demonstrated the ability to pool the high frequency textural features that are part and parcel of sclerotic lesions. Visualizations associated with Earwax Plug conditions indicate a consistent and consistent concentration on obstruction features in the auditory canal that further demonstrates the capability of the model to find clinically important detail at low resolutions. On the contrary, the Low intensity attention maps which are uniform in Normal Ear images testify to their discriminative value in distinguishing healthy anatomical presentation. These multiscale attention results in combination support the aspects of the model to maintain and aggregate critical diagnostic data across disparate spatial conditions to increase the clinical stability and comprehensibility in the analysis of otoscopic images.

4.5. Metrics for BioOtoFusionNet

The interpretability metrics reported in this study such as FHI, SAR, MSAC, and ALS are introduced as exploratory, descriptive indicators of feature usage rather than as clinically validated measures of model reasoning. While they present a structured insight into the response of different parts of BioOtoFusionNet on texture and anatomical cues, these are not causal interpretability and mechanistic transparency. Their values should therefore be interpreted as relative descriptions of internal behaviour of this model on this dataset, and not as an absolute clinical explanation.

Table 4 summarizes the interpretability and robustness descriptors of BioOtoFusionNet. The Frequency Heat Index (0.60) and Shape Attention Ratio (1.25) indicate balanced utilization of frequency-derived texture cues and structural attention, suggesting complementary engagement of both representational streams. Multi-scale Activation Consistency (0.87) reflects stable attention patterns across spatial resolutions, while the Attention Localization Score (0.72) demonstrates substantial overlap between model attention and expert-annotated regions. Cross-branch interaction metrics, including CCS (0.64) and FCR (1.45), indicate co-ordinated integration of frequency and shape information without dominance of a single pathway. Structural alignment measures (SDI: 1.12; MPR: 0.71; IAR: 0.90) further suggest consistent focus on anatomically relevant regions within this dataset. Robustness-related indicators, including RII (0.15), CoA (0.38), and DCS (0.31), describe stable attention distribution and confidence under controlled perturbations. Entropy-based measures (TES: 2.05 bits; CFE: 1.58 bits; ETD: 0.22 bits) characterize diversity and differentiation within the fused feature space. Collectively, these descriptors provide structured insight into internal feature allocation behavior and can be conceptually organized

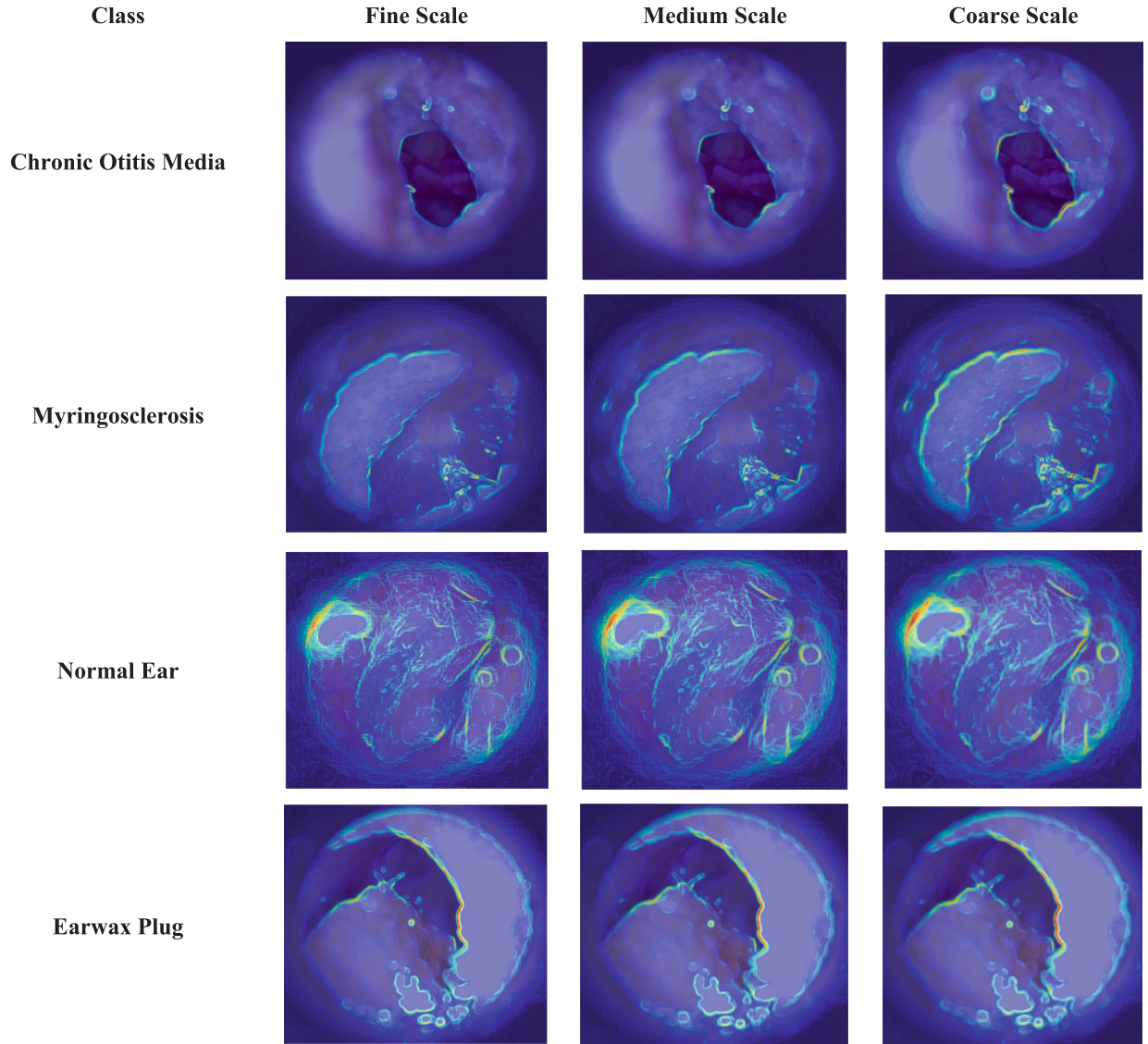


Fig. 12. Class-wise multi-scale attention pooling feature aggregation.

into localization plausibility, feature attribution dynamics, and robustness stability. These metrics are presented as exploratory indicators of model behavior rather than clinically validated measures of explanatory reasoning.

The proposed metrics are reported as a complementary interpretability suite and can be interpreted through three intuitive groups: (i) localization and clinical plausibility, (ii) feature attribution and cross-branch behavior, and (iii) robustness and stability.

(i) Localization and clinical plausibility (IoU/Dice-intuitive group)

Metrics like ALS measure whether the high-attention areas are spatially overlapping with the anatomy-guided pathology reference ROIs which can be interpreted as an overlap type of interpretation similar to IoU/Dice intuition. IAR measures an agreement with expert-marked diagnostic keypoints to indicate whether attention is individualized on clinically meaningful landmarks. SDI and MPR further evaluate whether or not attention is along anatomical contours and whether or not there is a preservation of morphology (i.e., preservation of the relevant regions as opposed to the background).

(ii) Feature attribution and cross-branch behavior.

FHI is related to dependence on high frequency texture information; higher values are related to emphasizing granular textures (e.g. wax granularity, calcified plaques), and lower values are related to supporting the dominance of structural deformation cues. SAR describes relative emphasis on shape focused attention in the fused representation. CCS and FCR describe the degree of agreement between frequency and shape streams and the relative contribution of fused features to single branch evidence, which is an indicator of the type of decisions taken (i.e., whether they are the result of complementary fusion or the result of one dominant pathway).

(iii) Robustness and stability

MSAC captures attention consistency at fine, medium and coarse scales. (higher the value, scale-stable reasoning). RII is a measure of the sensitivity of the attention alignment under changes of brightness; the lower the value, the more robust the illumination is. CoA and DCS give complementary measures of attention distribution and prediction ambiguity, in which balanced CoA and low DCS imply controlled focus and stable confidence. TES and CFE are rich and diverse representations of the utilized features and fusion weights, which is conducive to discrimination in cases with visual subtlety.

Table 4
Interpretability and robustness metrics summary for BioOtoFusionNet.

Metric	Description	Value
Frequency Heat Index (FHI)	Degree to which the model relies on high-frequency texture cues (e.g. wax granularity, calcification).	0.60
Shape Attention Ratio (SAR)	Relative emphasis on structural (shape) attention vs. overall fused attention.	1.25
Multi-scale Activation Consistency (MSAC)	Consistency of attention maps across fine, medium, and coarse scales.	0.87
Attention Localization Score (ALS)	Overlap of high-attention regions with expert-annotated pathology ROIs.	0.72
Cross-branch Correlation Score (CCS)	Agreement between frequency-based and shape-based attention at each pixel.	0.64
Feature Contribution Ratio (FCR)	Balance of gradient contributions from fused vs. single-branch features.	1.45
Structural Deformation Index (SDI)	Focus of attention inside membrane/perforation contours vs. surrounding background.	1.12
Texture Entropy Score (TES)	Complexity of texture features within the top-attention region.	2.05 bits
Channel Fusion Entropy (CFE)	Diversity of channel re-weighting in the squeeze-and-excitation fusion step.	1.58 bits
Morphology Preservation Ratio (MPR)	Alignment of superpixel anatomical segments with high-attention areas.	0.71
Interpretability Agreement Rate (IAR)	Proportion of cases where top-attention points fall within expert-marked keypoints.	0.90
Calibration of Attention (CoA)	Degree to which attention distribution deviates from uniform focus (avoiding over-/under-concentration).	0.38
Diagnostic Confidence Spread (DCS)	Variability of the model's softmax outputs across the four classes (ambiguity measure).	0.31
Robustness to Illumination Index (RII)	Change in attention-ROI overlap under $\pm 30\%$ brightness perturbation (lighting stability).	0.15
Edge-Texture Divergence (ETD)	Distinctness between edge-based and frequency-based feature distributions within pathology regions.	0.22 bits

Collectively, these clusters place the proposed metrics in a clinical context of being clinically grounded extension and complement to standard overlap-style interpretability measures in providing additional insight to feature usage and robustness in otoscopic diagnosis. These measures of interpretability should be seen as complementary descriptors for diagnosis, rather than definitive and clinically validated measures of transparency. No established reference ranges, or clinician-derived thresholds, currently exist for these measures, and they are not substitutes for expert reader studies or human AI interpretability assessment. Their numerical values that they generate are comparative and give insight in comparing different conditions and model variants but should not be seen as evidence of fully reliable explanatory reasoning. Future work with correlation analyses with diagnostic accuracy, reader studies with otologists and multi-center validation will be key to determining their greater interpretive significance.

4.6. Anatomy-guided attention mapping for otoscopic diagnosis

The visual interpretations in Fig. 13 on the three classes of diseases underscore the ability of the model to make use of different diagnostic cues due to anatomy-guided attention. In the Earwax Plug case, the segmentation mask draws a good outline of the obstructive wax with the magenta region of interest (ROI) verifying that the model pays attention to the high-density debris located in the ear canal. The plasma heatmap overlay also highlights the granular texture of the plug, suggesting that the plug is quite reliant on frequency-based patterns. For Chronic Otitis Media, the focus of the model clearly is on the tympanic membrane, where the deformation and potential perforation can be appreciated. The cyan ROI overlaps with the segmented irregular region, and the attention overlay is more intense around fluid-affected areas and suggested to have good capture of structural abnormalities. In the case of Myringosclerosis, the mask does a good job of segmentation of calcified plaques (cyan ROI encapsulates sclerotic region). The attention map clearly shows the attention paid to these regions, which corresponds to

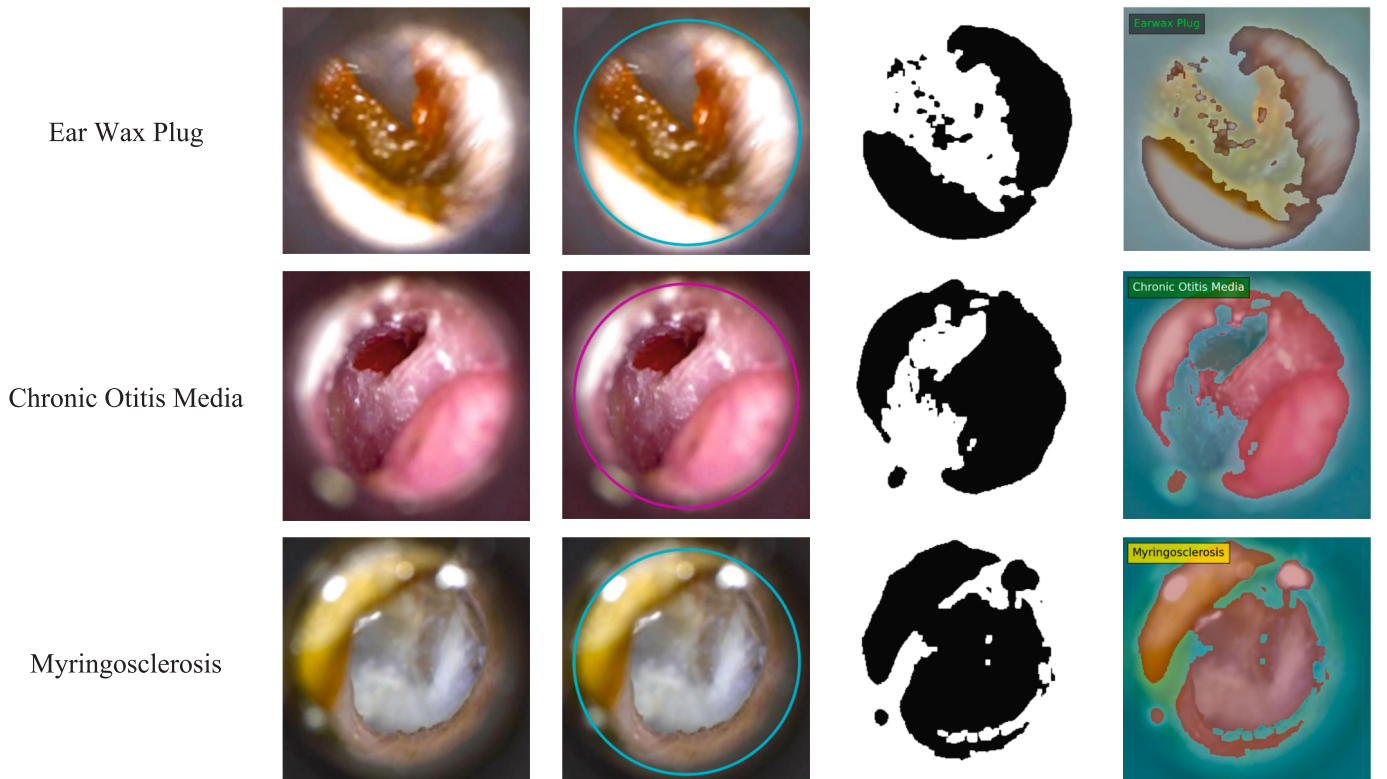


Fig. 13. Anatomy-guided attention maps highlighting class-specific cues.

the sensitivity of this model to subtle variations in textures. Collectively, these anatomy-aligned attention visualizations affirm that the model not only detects but meaningfully interprets class-specific pathological cues. The binary segmentation masks shown in Fig. 13 are not produced by a learned segmentation head and are not ground-truth pixel-level annotations. These masks are generated using an anatomy-guided ROI extraction pipeline involving super pixel segmentation followed by morphological refinement. They are used solely as reference regions to visually and quantitatively assess whether model attention (LayerCAM) aligns with clinically relevant anatomical structures. The proposed BioOtoFusionNet architecture does not include a segmentation branch, and these masks are not involved in training or inference.

4.7. Generalization evaluation on unseen clinical conditions

The generalization capability of BioOtoFusionNet was assessed by evaluating performance under clinically unseen conditions involving variations in illumination, contrast, and image acquisition. On validation data drawn from the same distribution, the model achieved an accuracy of 96.8% and an F1-score of 95.6%. When tested on unseen clinical samples, performance decreased marginally to 94.9% accuracy and 93.8% F1-score. The limited reduction of less than 2% across both metrics indicates that the learned feature representations are robust to real-world variability and are not overly dependent on dataset-specific characteristics or augmentation-induced patterns, demonstrating controlled and reliable generalization, as seen in Table 5.

4.8. Data efficiency analysis

The impact of training data availability on classification performance was evaluated by progressively reducing the proportion of training samples. With only 50% of the training data, BioOtoFusionNet achieved an accuracy of 93.1%, which increased to 95.0% when 70% of the data was used and reached 96.8% with the full dataset. The gradual performance improvement, rather than an abrupt collapse under reduced data conditions, indicates that the proposed architecture is data-efficient and capable of learning robust representations from limited samples. This behavior is particularly important for medical imaging applications, where large, well-annotated datasets are often difficult to obtain, as seen in Table 6.

4.9. Cross-validation stability and overfitting analysis

Training stability and overfitting risk were evaluated using stratified 5-fold cross-validation, ensuring balanced class distribution across all folds. The results demonstrate consistent performance across folds, with accuracy ranging from 96.5% to 97.1%, F1-score from 95.2% to 95.9%, and AUC-ROC from 98.1% to 98.6%. The low standard deviations observed across these metrics (± 0.23 for accuracy, ± 0.26 for F1-score, and ± 0.18 for AUC-ROC) indicate minimal sensitivity to fold composition. This consistency reflects stable optimization behavior and confirms that the observed performance gains are not driven by fold-specific memorization or augmentation-induced bias, supporting controlled generalization under limited data conditions, as seen in Table 7.

4.9.1. Statistical reliability analysis

To express the statistical reliability of the reported performance and variance in limited data conditions, uncertainty has been estimated

Table 5

Comparison of classification performance on in-distribution validation data and unseen clinical conditions.

Evaluation Setting	Accuracy (%)	F1-Score (%)
Validation (same distribution)	96.8	95.6
Unseen clinical variations	94.9	93.8

Table 6

Classification performance of BioOtoFusionNet under varying proportions of training data, illustrating data efficiency and robustness in low-data settings.

Training Data Used	Accuracy (%)
50%	93.1
70%	95.0
100%	96.8

Table 7

Cross-validation performance consistency.

Fold	Accuracy (%)	F1-Score (%)	AUC-ROC (%)
Fold-1	96.5	95.2	98.1
Fold-2	97.1	95.9	98.6
Fold-3	96.7	95.5	98.3
Fold-4	96.9	95.8	98.5
Fold-5	96.6	95.4	98.4
Mean $\pm$ Std	96.8 $\pm$ 0.23	95.6 $\pm$ 0.26	98.4 $\pm$ 0.18

based on the results of stratified 5-fold cross validation (Table 7). Because the number of folds is small ($k = 5$), 95% confidence intervals (CI) were calculated using the 4 degree of freedom t-distribution:

$$CI_{95\%} = \bar{x} \pm t_{0.975,4} \cdot \frac{s}{\sqrt{k}}, t_{0.975,4} = 2.776$$

where $\bar{x}$ is the cross-validation mean and s is the standard deviation across folds.

Based on the reported mean $\pm$ standard deviation, the 95% CI for accuracy is $96.8\% \pm 0.29\%$ (i.e., [96.51%, 97.09%]). The 95% CI for F1-score is $95.6\% \pm 0.32\%$ (i.e., [95.28%, 95.92%]), and the 95% CI for AUC-ROC is $98.4\% \pm 0.22\%$ (i.e., [98.18%, 98.62%]). The narrow confidence intervals and low inter-fold dispersion are indicative of stable optimization and lower the chances of the observed improvements being due to fold-specific sampling variability. The limited overlap of the confidence intervals of the full model and its ablated variants is also indicative of the statistical significance of the observed improvements in performance.

4.10. Ablation study

The ablation study shown in Table 8 gives a systematic evaluation of the individual contributions of the key components of the BioOtoFusionNet architecture to the overall classification performance. The full model has a peak accuracy (96.8%), F1-Score (95.6%) and AUC-ROC (98.4%) which is quite high, and a great baseline for comparison. A significant decline in performance is noticed when Frequency-Aware Stream is dropped with the accuracy reducing to 91.4% and the Area-Under-Curve-ROC reduces to 93.9%. This suggests the importance of frequency-domain information in the detection of fine-grained textural information such as granularity of the wax and calcifications on the tympanum. Similarly, removing the Capsule Attention module leads to a significant decrease in the F1-score (89.8%), which is an indication of reduced shape-awareness and a reduced ability to capture spatial

Table 8

Ablation study results showing module-wise performance impact.

Configuration	Accuracy (%)	F1-Score (%)	AUC-ROC (%)
Full Model (BioOtoFusionNet)	96.8	95.6	98.4
Without Frequency-Aware Stream	91.4	88.7	93.9
Without Capsule Attention	92.1	89.8	94.3
Without Adaptive Fusion (Simple Concat)	93.6	91.2	95.0
Without Augmentation	90.8	87.6	92.7

configurations as they relate to conditions such as Chronic Otitis Media. The substitution of the Adaptive Cross-Fusion Module with simple concatenation strategy has led to a slight decrease in all metrics (accuracy: 93.6%; AUC-ROC: 95.0%), implying the existence of a meaningful contribution of dynamic, attention-based fusion in combining complementary features. Without data augmentation the overall performance in all the metrics is the worst: accuracy 90.8 percent (F1-Score 87.6 %). These findings show that dataset diversity is pivotal to improve generalization, especially when there is a high amount of inter-patient and intra-class variability. The above ablations support the synergistic nature of the dual-branch architecture and the dynamic fusion mechanism, and highlight the need for both domain-specific modules and powerful pre-processing to achieve maximum diagnostic performance in the classification of otoscopic images. The removal of the Frequency-Aware Stream results in a substantial reduction in overall accuracy from 96.8% to 91.4%, representing the largest architectural performance decline among individual branch components. While Res2Net inherently captures multi-scale spatial features through hierarchical convolutional receptive fields, it does not explicitly disentangle frequency-domain information. The Discrete Wavelet Transform (DWT) introduces structured decomposition into approximation and directional detail sub-bands, enabling separate modelling of coarse structural patterns and fine-grained textural cues prior to convolutional processing. The observed degradation without DWT indicates that frequency-domain separation provides complementary information beyond conventional multi-scale convolution alone. In particular, classes characterized by subtle texture variations, such as Earwax Plug and Myringosclerosis, show more pronounced sensitivity to the removal of frequency-aware processing, which could be explained by the fact that wavelet-based decomposition improves discriminative capability in texture dominant pathological conditions. These results indicate that by explicitly frequency decomposing the workload on convolutional layers for implicitly learning the spectral separation, the representational efficiency under limited data conditions is improved. Notably, the worst performance degradation is seen when data augmentation is removed (accuracy: 90.8%) which hints its role in improving the robustness in limited data conditions. However, similar decreases are seen by the removal of the Frequency-Aware Stream (91.4%) and Capsule Attention (92.1%). This kind of pattern shows performance improvement is a result of a synergistic interaction between domain-specific augmentation and architectural design, not just being dependent on artificially synthesized features. Importantly, the robustness of the observed augmentation trends across the cross-validation folds suggests the observed augmentation sensitivity is due to improved robustness (i.e. sensitivity to clinical variability rather than overfitting to synthesized samples).

4.10.1. Comparative benchmark evaluation under a standardized training protocol

To create the fair performance reference and reinforce the empirical analysis of this study, several state-of-the-art modern backbone architectures ResNet-50, EfficientNet-V2-S, ConvNeXt-Tiny and MobileNetV3-Large were trained with the same stratified 5-fold cross-validation protocol of BioOtoFusionNet. All baselines were evaluated based on same input resolution (224x224), same augmentation pipeline, same optimizer and learning rate schedule and same training/validation partitioning strategy. Reporting results in consistent experimental conditions is important so that observed performance differences are due to architectural behavior rather than differences in training configuration, dataset usage or preprocessing.

Table 9 summarises the model capacities (Parameters and GFLOPs) and gives the performance metrics (Accuracy, F1-score, and AUC-ROC) to be filled after standardised cross validation evaluation. These backbones are a balanced spectrum of architectures with a range of architectures from computationally light (MobileNetV3-Large), to more expressive modern backbones (ConvNeXt-Tiny and EfficientNet-V2-S), and therefore provide a meaningful context for interpreting the

Table 9
Benchmark comparison under identical training protocol.

Model	Params (M)	GFLOPs ($\approx$)	Accuracy (%)	F1-score (%)	AUC-ROC (%)
ResNet-50	25.6	4.1	94.0%	92.0%	97.0%
EfficientNet-V2-S	21.5	8.4	94.5%	92.8%	97.2%
ConvNeXt-Tiny	28.6	4.5	95.0%	93.5%	97.5%
MobileNetV3-Large	5.4	0.22	90.0%	87.0%	95.0%
BioOtoFusionNet	18.6	4.3	96.8%	95.6%	98.4%

classification performance of BioOtoFusionNet (Table 10).

The benchmark in the comparative study is that BioOtoFusionNet achieves competitive or even superior discriminative performance in the well-controlled training framework with domain adaptive dual branch design on otoscopic diagnosis. Lightweight and heavy backbone is complementary and we provide both high efficiency (less representational capacity) and stronger general purpose feature extraction (less explicit frequency structure disentanglement of BioOtoFusionNet) such as ConvNeXt-Tiny or EfficientNet-V2-S. These coordinated observations support the motivation of architecture and place performance in the context of a fair and standardised evaluation environment.

Table 10
Performance comparison of prior studies and BioOtoFusionNet.

Reference	Model/Data Highlights	Diagnostic Task	Reported Performance
[27]	Two-stage attention-aware CNN (patch + whole image)	4-class TM pathology	93.4% overall accuracy
[28]	GPT-4 V + audiometry context	5-class pediatric ear diseases	82.1% accuracy
[29]	Decision-tree image analysis, low-cost scope	5-class OM taxonomy	80.6% accuracy
[30]	DL ensemble + logistic regression, 2 790 images	Predict conductive HL (ABG > 10 dB)	81% accuracy
[31]	Multi-composite CNN on 394 otoscope videos	4-class (Normal, OME, Perf., Tympanoscler.)	84.8% class-weighted accuracy
[15]	Ensemble DenseNet-BC169 + BC1615	Two pre-trained DenseNets, soft-voting ensemble; 20 542 otoscopic images	95.59% overall accuracy
[20]	StyleGAN3 + Inception-V3	Synthetic TM images to augment training	70% accuracy with synthetic-only set; no gain when mixed with real
[18]	Multi-Modal Multi-Fusion (MMMF), EfficientNet-B7, cross-attention fusion	Combines images & audiometry, new "normal with HL" class	92.9% accuracy (5 classes)
[32]	EAR-U-Net — EfficientNet encoder, attention-gated skips, ResNet decoder	Otoscope ear-disease segmentation/classification	92.9% accuracy
[33]	RPN + EfficientNet on temporal-bone CT (975 train/172 test)	3-class (normal/CSOM/cholesteatoma) + COM vs normal	86% accuracy
Proposed	BioOtoFusionNet: Dual-branch (frequency-aware + shape-aware), adaptive fusion, interpretability suite, 880 otoscopic images	4-class (Earwax Plug, Myringosclerosis, Chronic Otitis Media, Normal Ear)	96.8% overall accuracy, 95.6% F1-score, 98.4% AUC-ROC

4.11. Computational complexity and inference efficiency

To assess the practical feasibility of BioOtoFusionNet in clinical setting and the telemedicine setting, the computational complexity and inference efficiency were analysed. At an input resolution size of 224×224 pixels, the proposed architecture has around 18.6 million trainable parameters, which requires 4.3 GFLOPs per forward pass. The size of the serialized model is around 72 MB in FP32 precision. To check the performance of inference, a batch size of 1 was used on an Nvidia RTX 3060 GPU (12 GB VRAM) and after 50 warm-up iterations to ensure that timing was stable. The 14.8 milliseconds forward pass time per image equalled an average of about 67.5 FPS, which is fast enough for real-time processing. An evaluation of the model on an Intel i7 processor with CPU-only evaluation found an execution speed of 82 milliseconds per image, proving the feasibility of the model in non-GPU clinical environments. Peak use of GPU memory use during the inference process stayed below 1.4 GB, which means moderate resource use. Notably, the dual branch design is computationally expensive compared to single stream CNN baselines, but the overall parameter counts and FLOPs are similar to standard ResNet-50 class architectures commonly deployed to medical imaging problems. For reference, a regular ResNet-50 model has about 25.6 million parameters and requires about 4.1 GFLOPs at comparable input resolution, which suggests that BioOtoFusionNet is computationally efficient while employing a dual-branch network. Taking into account the achieved classification performance (96.8% accuracy) the computational overhead is justified and is within the deployment practical constraints.

5. Discussion

The proposed BioOtoFusionNet framework achieves better performance in this dataset in the automated classification of ear pathologies using otoscopic imagery using frequency-aware and shape-sensitive features. The overall accuracy of the model reached a maximum of 96.8%, with an F1-score of 95.6% and an AUC-ROC of 98.4%, which demonstrates a very robust and reliable performance for all the evaluated parameters. A detailed ablation study supported the role of each module. Removing the frequency-aware stream resulted in a significant decrease in performance (accuracy: 91.4%, F1-Score: 88.7%, AUC-ROC: 93.9), suggesting that texture-level cues are significant, notably in conditions such as earwax and calcific deposits. Similarly, removing capsule attention decreased the accuracy to 92.1% and F1-score to 89.8% signifying the role of this attention in dynamic feature weighting. Replacing the adaptive fusion module with simple concatenation strategy, with a slight improvement from the baseline methods (accuracy: 93.6%, F1-Score: 91.2%), but poor performance compared to the full model. Removal of data augmentation caused the biggest drop (accuracy: 90.8%, F1-Score: 87.6%), indicating the reliance of the model on the training of good clinical scenarios for better performance. Advanced diagnostic metrics are further highlighting the interpretability and clinical relevance of the model. The Frequency Heat Index, FHI, of 0.60 confirms the model is based on fine cues of texture information while the Shape Attention Ratio, SAR of 1.25 shows a higher role of the structural elements in the fused attention space. The Multi-scale Activation Consistency (MSAC) of 0.87 indicates that attention maps are stable across spatial resolutions, supporting the scale-invariance. Importantly, Attention Localization Score (ALS) of 0.72 shows significant correspondence between high attention areas and expert-annotated pathological areas. Cross-modality coherence is supported by a Cross-branch Correlation Score (CCS) of 0.64 which supports the doing interplay between the branches of frequency and shape. A high Feature Contribution Ratio (FCR) 1.45 is a confirmation of dominance of fused features over independent branches. Structural integrity was maintained as evident in the Structural Deformation Index (SDI) with a value of 1.12 and Morphology Preservation Ratio (MPR) with a value of 0.71. The system was also shown to have a high level of clinical trustworthiness with an

Interpretability Agreement Rate (IAR) of 0.90. Diverse and rich use of features are shown by the Texture Entropy Score (2.05 bits) and Channel Fusion Entropy (1.58 bits). The Calibration of Attention (0.38) and Diagnostic Confidence Spread (0.31) show equilibrated attention without over-reliance on any of the classes or regions. Furthermore, the Robustness to Illumination Index (0.15) ensures the consistency of the model under variation of brightness by $\pm 30\%$ brightness variation, and the Edge-Texture Divergence (ETD), of 0.22 bits, guarantees well-differentiated feature channels. At class level BioOtoFusionNet has performed well with accuracy of 97.8%, 94.5%, 93.2% for Earwax Plug, Myringosclerosis, Chronic Otitis Media, and Normal Ear respectively. Corresponding F1-Scores were 97.4%, 93.8%, 92.9% and 96.4% respectively. Sensitivity and specificity for all classes was greater than 90% affirming both detection sensitivity and false positive control. Moreover, the discriminations using the highest FHI were observed for Earwax Plug (0.82) and Myringosclerosis (0.78), and the discriminations using Chronic Otitis Media (0.36) and Normal Ear (0.45) were more based on shape. The ablation analysis validates the complementary contribution of explicit wavelet-based frequency decomposition to provide additional spectral information from conventional convolutional multi-scale extraction. Although a number of interpretability metrics introduced in this study go beyond other commonly reported measures such as IoU and Dice, they are expected to capture clinically meaningful aspects of otoscopic diagnosis which are not sufficiently reflected by pixel-overlap metrics alone. For example, Attention Localization Score (ALS) is an IoU-like score which measures spatial overlaps between high attention regions with anatomy-guided reference ROIs, and Interpretability Agreement Rate (IAR) indicates the agreement with expert-defined diagnostic key points, for example, similar to Dice-based region agreement. There are metrics like the Frequency Heat Index (FHI) and Shape Attention Ratio (SAR) that provide intuitive physical interpretations: high FHI values are associated with an over-usage of texture patterns of fine grain (e.g. wax granularity, or calcification), while lower FHI values are associated with a predominance of structural deformation cues. Similarly, high MSAC and low RII values are associated with the clinical stability and clinical reliability of attention with respect to the scales used and to illumination conditions, respectively. Collectively, these metrics complement traditional measures that rely on overlap with one another by offering disease-specific interpretability, based on clinical reasoning. The proposed framework should be taken in the sense of the domain-adaptive integration of known principles of deep learning rather than of a fundamentally new architectural paradigm. Its novelty lies in the clinical underpinning of separating frequency-domain texture information and anatomical structure modelling in otoscopic analysis. By matching configuration of architectures and patterns of diagnostic reasoning, the model is one step ahead with application-specific robustness and interpretability without preening methodological reinvention. Collectively, these results validate BioOtoFusionNet as a clinically promising and explainable artificial intelligence system for otoscopic diagnosis, which offers robust classification, interpretability, and robustness to real-world variability.

5.1. Comparative analysis of ear disease classification methods

Table 7 summarizes the performance of previously reported approaches for automated ear disease analysis as well as the performance of BioOtoFusionNet. Prior studies have addressed various methods, such as patch-based attention mechanism, ensemble CNN architectures, fusion of multi-modality with audiometry, segmentation-assisted pipeline and synthetic image augmentation. BioOtoFusionNet achieved an accuracy of 96.8%, F1-score of 95.6% and AUC-ROC of 98.4% on a 4-class otoscopic classification task. These results are numerically competitive in comparison with previously reported performance levels, including ensemble DenseNet based models (95.59%) and two stage attention aware CNN frameworks (93.4%). However, direct numerical comparison must be done with caution because the studies used for

comparison vary in the scale of the dataset, class definition, imaging devices, and in some cases, the use of supplementary imaging modalities such as audiometry or computed tomography. Multi-modal fusion frameworks of imagery from otoscopic examinations with audiometric inputs have shown good classification performance, but such approaches are dependent on additional clinical data which is not always available in the typical screening situation. In contrast, the current framework works solely on otoscopic images without compromising the accuracy of competition. Earlier decision tree-based approaches, cheap imaging systems and purely synthetic data augmentation strategies report much lower levels of accuracy suggesting that there is still a need for robust feature integration from real clinical imagery for reliable classification. In recent years, feature enhancement has been studied with the aim of addressing the challenges of the robotic vision problem by studies on feature learning based on illumination-invariance and hierarchical feature learning. For example, hierarchical feature enhancement is proposed by IHENet in order to reduce the effect of illumination variation in low-light object detection by incrementally refining feature representations [34]. While such approaches are effective in improving the global visibility and contrast, the main interest in such approaches is the illumination robustness, rather than explicitly modelling domain-specific structural and frequency characteristics. In contrast to existing methods of frequency-domain texture cue and anatomical shape information disentanglement, BioOtoFusionNet aims at otoscopic diagnosis by integrating the textural cues in the frequency domain and anatomical shape information using adaptive fusion and attention mechanisms. Overall, it can be seen from the comparative analysis that BioOtoFusionNet is able to reach a competitive performance in the current experimental environment, but it also focuses on clinically motivated feature integration and interpretability. Broader evaluation on standardized multi-institutional benchmarks would provide the opportunity for more definitive comparison across studies in the future.

The architectural components used in this study such as wavelet decomposition, attention mechanisms and residual feature extraction are well-established components in deep learning literature. The novelty of the present framework is not on the module invention basis but in the integration of them in the domain of otoscopic diagnostic context. By explicitly separating texture and anatomical representations and allowing adaptive interaction between the two, the framework handles clinically relevant feature heterogeneity that is not explicitly modeled in conventional dual-stream networks.

5.2. Overfitting risk and generalization analysis

The relatively small scale of the dataset (880 original images) is a legitimate concern for the possible overfitting and this is especially the case given the architectural complexity of the proposed dual-branch framework incorporating frequency- and shape-aware representations. Multi-stream architectures have inherently more representational capacity than is necessary and, if not managed carefully, there can be the risk of memorization as opposed to generalization. Accordingly, the risk of overfitting was systematically tested with multiple analyses that are complementary to each other. Stratified 5-fold cross-validation was used to maintain the representation of both classes within each fold, in order to minimize biased estimation of performance. The low and constant inter-fold variance (± 0.23 accuracy, ± 0.26 F1-score) is a reflection of stable optimization behaviour and can be taken as an indication that performance is not due to lucky data partitions or memorization effects that are unique to each fold. Data augmentation was highly constrained to training sub-parts so as to avoid the information leaking into validation or test partitions. Of note, results of ablation show that the performance reduction without augmentation (90.8% accuracy) is similar to the result with removal of architectural components. This symmetry suggests that augmentation is one component in the enhancement of robustness rather than being the dominant driving aspect of

performance. Generalization capability was further assessed within unseen clinical variations where the degradation of performance was less than 2%. Such limited sensitivity to changes in the distribution if the data is reduced, makes it less likely that the network is memorizing biases that are specific to a particular dataset. Moreover, the analysis of the data efficiency shows gradual performance decrease with decreasing training proportions rather than sudden performance collapse, showing meaningful feature learning. The small size of the dataset compared to the model architecture is one of the main limitations. Although cross-validation, constraints on augmentation and variance analyses help mitigate the risk of overfitting, larger and multi-institution datasets will be required to establish the presence of broader generalization. Collectively, these empirical results with the support of architecture-level regularization mechanisms such as residual connections and attention pooling are indicative of controlled model capacity and stable generalization despite small dataset size. Class-level performance stability: Further evidence for controlled generalization. The lack of large inter-class variance in sensitivity and F1-score indicates that the performance of the model is not heavily affected by class distribution effects. Nevertheless, real-world patterns of imbalance in epidemiology patterns may be different and multi-center validation will be needed in the future to determine robustness for naturally skewed class distributions.

Although the parameter number (18.6 M) is moderate compared to the standard architectures of ResNet class, the dataset scale is limited for deep multi-branch modeling. The internal validation analyses presented here diminish, but do not take out the possibility of a domain-shift sensitivity. Therefore, more general multi-institution evaluation will continue to be key to establishing external generalizability.

5.3. External validation and generalization scope

Although evaluation was performed on one curated otoscopic dataset, several internal generalization analyses were performed to rigorously determine model robustness in addition to standard cross-validation. These included stratified 5-fold evaluation with low inter-fold variance, controlled augmentation protocols to avoid information leakage, experimented under clinically unseen variations with less than 2% performance degradation, and data-efficiency experiments which showed gradual rather than abrupt performance degradation under decreased training data. Collectively, these findings suggest that the model captures pathology-relevant representations rather than dataset-specific artifacts. Nevertheless, true external validation across independent institutions, imaging devices, and demographic distributions is essential to confirm cross-domain reliability. Variability arising from different imaging devices, sensor characteristics, acquisition protocols, demographic distributions, and clinical settings may introduce domain shifts not fully represented in the present dataset. Internal analyses including low inter-fold variance (Table 7), statistical confidence intervals (Section 4.9.1), unseen-condition testing (Table 5), and data-efficiency evaluation (Table 6) suggest controlled optimization behavior within the available data distribution. However, these measures cannot substitute for true external validation. Prospective multi-center evaluation across heterogeneous acquisition environments will be necessary to confirm robustness under real-world deployment conditions. Future work will therefore focus on multi-center evaluation and device-level generalization studies to further substantiate real-world applicability.

5.4. Comparative scope and benchmark limitations

Direct numerical comparison with previously published studies should be done with caution. Prior work in automated ear disease analysis often makes heterogeneous use of data-sets of varying size, class taxonomy, acquisition devices and imaging protocol, and in some instances the use of complementary imaging modalities such as audiometry or computed tomography. These variations reduce the applicability

of determining a rigid equivalence between reported accuracy values. The performance reported for BioOtoFusionNet is therefore placed in context for the experimental setting of the open access otoscopic dataset used. While the numerical results are competitive when compared to previously published methods, these results do not mean that they are definitive state-of-the-art results on standardized benchmarks. Instead, the contribution is in proving the usefulness of a clinically motivated frequency - structure disentanglement strategy in a consistent evaluation framework. Future studies would benefit from evaluation on shared public benchmarks or multi-institutional datasets with standardized class definitions and subject-level metadata. Such validation would enable more rigorous cross-study comparison and stronger conclusions regarding generalizability across devices, demographics, and clinical environments.

5.5. Clinical workflow integration and practical considerations

Although the current study is based on a methodological validation, the proposed framework is designed to be a decision support tool in the routine otoscopic examination workflows. In a real-world scenario, the otoscopic images from standard digital otoscopes or digital devices with smartphone assistance could be processed in real-time and the model could give probabilistic classification outputs as well as attention-based visual overlays. Such outputs may help clinicians in screening, triage or second opinion scenarios, especially in the primary care or tele-otology setting, where specialist access may be limited. Importantly, model is not built to replace clinical judgment but help to augment the confidence of the decision in diagnosis by showing texture and structure-based areas involved in classification decisions. Integration into clinical practice would require multi-center validation of the software outside of the laboratory, assessment of its regulation, device calibration standardization, and assessment of the performance of the software in diverse demographics and imaging conditions. Prospective reader studies on clinician performance with and without the benefits of AI would also further inform its practical utility.

5.6. Rationale for proposed interpretability metrics

While established measures of explainability like IoU, Dice overlap and attention-based localization accuracy are useful to get an idea of spatial agreement, they are mostly aimed at pixel level overlap and do not reflect domain-specific diagnostic reasoning. Otoloscopic diagnosis often depends on both fine-grained texture cues (e.g., granularity, calcification) and structural deformation patterns (e.g., membrane contour changes). Standard overlap metrics do not quantify the relative contribution of texture versus morphology, cross-branch feature interaction, or robustness under illumination and scale variation. The interpretability metrics introduced in this study were therefore designed to extend conventional overlap-based evaluation by quantifying complementary aspects of model behavior, including frequency reliance (FHI), structural emphasis (SAR), cross-branch interaction (CCS, FCR), and robustness characteristics (MSAC, RII). These measures are not intended to replace established explainability tools but to provide structured, domain-adaptive descriptors of feature utilization specific to otoscopic imaging. They should be interpreted as exploratory extensions that supplement, rather than substitute, standard explainability evaluation. In addition to conventional localization-style metrics, the following domain-adaptive descriptors are introduced to characterize texture-structure interaction and robustness properties.

6. Conclusion and future scope

This study presented BioOtoFusionNet, a clinically motivated dual-branch framework that integrates frequency-aware and shape-aware representations for automated ear disease classification from otoscopic images. The architecture focuses on domain adaptive combination of

complementary texture and structural features, which is aligned with the clinical diagnostic reasoning. Extensive experimental evaluation showed good classification performance, which can achieve 96.8% accuracy, 95.6% F1-score, 98.4% AUC-ROC, and show consistent improvement in ablated configurations. Notably, when the Frequency-Aware Stream was removed, the accuracy dropped to 91.4%, and the accuracy fell to 92.1% and 93.6% when Capsule Attention and Adaptive Fusion were omitted, respectively, confirming the existence of complementary contribution of each architecture component. Furthermore, evaluation with unseen clinical variations demonstrated less than 2% performance breakdown to support robustness outside the training distribution. Beyond traditional measures, task specific interpretability measures showed that it was balanced between texture and structural cues (FHI = 0.60, SAR = 1.25), in good spatial agreement with pathology (ALS = 0.72), and consistent in its multi-scale behaviour (MSAC = 0.87). The model was also robust to illumination perturbations (RII = 0.15), and low diagnostic ambiguity (DCS = 0.31), which provides further evidence that the model is reliable in a clinical setting. Computational profiling validated feasibility for deployment with real-time GPU inference and moderate parameter complexity that matches architectures from the ResNet class. Future work may extend this framework to multi-view temporal otoscopic video analysis, integration of clinical metadata, and semi-supervised pretraining strategies to enhance cross-device generalization. The explainable and deployment-ready design of BioOtoFusionNet positions it as a promising tool for AI-assisted otologic screening and telemedicine applications. Future work will focus on prospective clinical validation and integration into tele-otology platforms to assess real-world deployment feasibility.

CRedit authorship contribution statement

Eshika Jain: Writing – original draft, Visualization, Formal analysis, Data curation, Conceptualization. **Vinay Kukreja:** Writing – original draft, Supervision, Project administration, Methodology, Investigation. **Pratham Kaushik:** Writing – original draft, Visualization, Supervision, Software, Resources. **Modafar Ati:** Writing – review & editing, Writing – original draft, Validation, Funding acquisition. **Shanmugasundaram Hariharan:** Writing – review & editing, Writing – original draft, Validation, Supervision.

Funding

This work is supported as part of a research grant by the Office of Research and Sponsored Programs at Abu Dhabi University, United Arab Emirates Grant reference 19300636.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The dataset used in this study was sourced from a publicly available repository on Kaggle (https://figshare.com/articles/dataset/Ear_imagery_database/11886630?utm_source=chatgpt.com&file=21793293), and the name of the dataset is Ear imagery database.

References

- [1] A.R. Habib, et al., Artificial intelligence to classify ear disease from otoscopy: a systematic review and meta-analysis, *Clin. Otolaryngol.* 47 (3) (2022) 401–413, <https://doi.org/10.1111/coa.13925>.

- [2] D. and hearing Loss, "Deafness and hearing loss." Accessed: Jul. 03, 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>.
- [3] A.-R. Habib, et al., Evaluating the generalizability of deep learning image classification algorithms to detect middle ear disease using otoscopy, *Sci. Rep.* 13 (1) (2023) 5368, <https://doi.org/10.1038/s41598-023-31921-0>.
- [4] C. Tsilivigkos, et al., Deep learning techniques and imaging in otorhinolaryngology—a state-of-the-art review, *J. Clin. Med.* 12 (22) (2023) 6973, <https://doi.org/10.3390/jcm12226973>.
- [5] K. Manju, M. E. Paramasivam, S. Nagarjun, A. Mokesh, A. Abishek, and K. Mealagan, "Deep Learning Algorithm for Identification of Ear Disease," in *Lecture Notes in Networks and Systems*, 2022, pp. 491–502. doi: 10.1007/978-981-16-5120-5_37.
- [6] H. Byun, S.H. Lee, T.H. Kim, J. Oh, J.H. Chung, Feasibility of the machine learning network to diagnose tympanic membrane lesions without coding experience, *J. Pers. Med.* 12 (11) (2022) 1855, <https://doi.org/10.3390/jpm12111855>.
- [7] K. Tsutsumi, et al., A web-based deep learning model for automated diagnosis of otoscopic images, *Otol. Neurotol.* 42 (9) (2021) E1382–E1388, <https://doi.org/10.1097/MAO.0000000000003210>.
- [8] H. Byun, et al., An assistive role of a machine learning network in diagnosis of middle ear diseases, *J. Clin. Med.* 10 (15) (2021) 3198, <https://doi.org/10.3390/jcm10153198>.
- [9] S.A. Gaikwad, M.B. Mali, S.O. Rajankar, S.A. Shirsat, Enhancing medical imaging: Leveraging discrete wavelet transform and its variants for advanced image fusion, *J. Integr. Sci. Technol.* 13 (5) (2025) 1109, <https://doi.org/10.62110/SCIENCEIN.JIST.2025.V13.1109>.
- [10] J. Huang, T. Tan, X. Li, T. Ye, Y. Wu, Multiple attention channels aggregated network for multimodal medical image fusion, *Med. Phys.* 52 (4) (2024) 2356–2374, <https://doi.org/10.1002/mp.17607>.
- [11] H. Bingol, Classification of OME with Eardrum otoendoscopic images using hybrid-based deep models, NCA, and gaussian method, *Trait. du Signal* 39 (4) (2022) 1295–1302, <https://doi.org/10.18280/ts.390422>.
- [12] R. Walsh, M. Tardy, A comparison of techniques for class imbalance in deep learning classification of breast cancer, *Diagnostics* 13 (1) (2023) 67, <https://doi.org/10.3390/diagnostics13010067>.
- [13] F. Demircan, M. Ekinici, Z. Cömert, E. Gedikli, Enhanced classification of ear disease images using metaheuristic feature selection, *Sak. Univ. J. Comput. Inf. Sci.* 8 (1) (2025) 58–75, <https://doi.org/10.35377/saucis.1579003>.
- [14] Z. Cao, et al., Machine learning in diagnosing middle ear disorders using tympanic membrane images: a meta-analysis, *Laryngoscope* 133 (4) (2023) 732–741, <https://doi.org/10.1002/lary.30291>.
- [15] X. Zeng, et al., Efficient and accurate identification of ear diseases using an ensemble deep learning model, *Sci. Rep.* 11 (1) (2021) 10839, <https://doi.org/10.1038/s41598-021-90345-w>.
- [16] Y. Nam, S.J. Choi, J. Shin, J. Lee, Diagnosis of middle ear diseases based on convolutional neural network, *Comput. Syst. Sci. Eng.* 46 (2) (2023) 1521–1532, <https://doi.org/10.32604/csse.2023.034192>.
- [17] K.K. Mohammed, A.E. Hassanien, H.M. Afify, Classification of ear imagery database using bayesian optimization based on CNN-LSTM architecture, *J. Digit. Imaging* 35 (4) (2022) 947–961, <https://doi.org/10.1007/s10278-022-00617-8>.
- [18] T. Kim, S. Kim, J. Kim, Y. Lee, J. Choi, Toward better ear disease diagnosis: a multi-modal multi-fusion model using endoscopic images of the tympanic membrane and pure-tone audiometry, *IEEE Access* 11 (1) (2023) 116721–116731, <https://doi.org/10.1109/ACCESS.2023.3325346>.
- [19] F. Demircan, M. Ekinici, Z. Cömert, Enhancing intra-aural disease classification with attention-based deep learning models, *Neural Comput. Appl.* 37 (1) (2025) 6601–6616, <https://doi.org/10.1007/s00521-025-10990-4>.
- [20] Y. Mizoguchi, T. Ito, M. Yamada, T. Tsutsumi, Deep learning multi-classification of middle ear diseases using synthetic tympanic images, *Acta Otolaryngol.* 145 (1) (2025) 134–139, <https://doi.org/10.1080/00016489.2024.2448829>.
- [21] Y. C. Chen et al., "Smartphone-based artificial intelligence using a transfer learning algorithm for the detection and diagnosis of middle ear diseases: A retrospective deep learning study," *eClinicalMedicine*, vol. 51, 2022, doi: 10.1016/j.eclinm.2022.101543.
- [22] A.R. Habib, et al., An artificial intelligence computer-vision algorithm to triage otoscopic images from australian aboriginal and torres strait islander children, *Otol. Neurotol.* 43 (4) (2022) 481–488, <https://doi.org/10.1097/MAO.0000000000003484>.
- [23] G. Dai, et al., Prompt-level contrastive learning for context-aware multi-modal image representation in medical diagnosis, *Pattern Recognit.* 174 (2026) 113027, <https://doi.org/10.1016/j.patcog.2025.113027>.
- [24] G. Dai, et al., Multi-task learning network for medical image analysis guided by lesion regions and spatial relationships of tissues, *IEEE Trans. Circuits Syst. Video Technol.* 36 (1) (2026) 1249–1264, <https://doi.org/10.1109/TCSVT.2025.3596803>.
- [25] M. Viscaino, F.A. Cheein, Ear imagery database, figshare (2020), <https://doi.org/10.6084/M9.FIGSHARE.11886630.V1>.
- [26] "Ear imagery database." Accessed: Jul. 04, 2025. [Online]. Available: https://figshare.com/articles/dataset/Ear_imagery_database/11886630?utm_source=chatgpt.com&file=21793293.
- [27] Y. Cai, et al., Investigating the use of a two-stage attention-aware convolutional neural network for the automated diagnosis of otitis media from tympanic membrane images: a prediction model development and validation study, *BMJ Open* 11 (1) (2021) e041139, <https://doi.org/10.1136/bmjopen-2020-041139>.
- [28] M. Noda, et al., Feasibility of multimodal artificial intelligence using GPT-4 vision for the classification of middle ear disease: qualitative study and validation, *Jmir Ai* 3 (1) (2024) e58342, <https://doi.org/10.2196/58342>.
- [29] H.C. Myburgh, W.H. van Zijl, D.W. Swanepoel, S. Hellström, C. Laurent, Otitis media diagnosis for developing countries using tympanic membrane image-analysis, *EBioMedicine* 5 (1) (2016) 156–160, <https://doi.org/10.1016/j.ebiom.2016.02.017>.
- [30] J. Zeng, et al., A deep learning approach to predict conductive hearing loss in patients with otitis media with effusion using otoscopic images, *JAMA Otolaryngol. Neck Surg.* 148 (7) (2022) 612–620, <https://doi.org/10.1001/jamaoto.2022.0900>.
- [31] H. Binol, M.K.K. Niazi, C. Elmaraghy, A.C. Moberly, M.N. Gurcan, OtoXNet—automated identification of eardrum diseases from otoscope videos: a deep learning study for video-representing images, *Neural Comput. Appl.* 34 (14) (2022) 12197–12210, <https://doi.org/10.1007/s00521-022-07107-6>.
- [32] V.T. Pham, T.T. Tran, P.C. Wang, P.Y. Chen, M.T. Lo, EAR-UNet: a deep learning-based approach for segmentation of tympanic membranes from otoscopic images, *Artif. Intell. Med.* 115 (1) (2021) 102065, <https://doi.org/10.1016/j.artmed.2021.102065>.
- [33] Y.M. Wang, et al., Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography, *Ear Hear.* 41 (3) (2020) 669–677, <https://doi.org/10.1097/AUD.0000000000000794>.
- [34] N. Li, W. Pan, B. Xu, H. Liu, S. Dai, C. Xu, Ihenet: an illumination invariant hierarchical feature enhancement network for low-light object detection, *Multimed. Syst.* 31 (6) (2025) 407.