Enhanced Feature Representation for Otoscopic Image Classification Using Supervised Contrastive Learning

Mesut Şeker

Department of Electrical and Electronics Engineering, Dicle University, Diyarbakir 21280, Turkey
mesut.seker@dicle.edu.tr

Abstract. Accurate interpretation of otoscopic images is essential for the early detection of middle-ear disorders, yet diagnostic variability among clinicians and the scarcity of large, annotated datasets make automated screening a challenging task. Traditional deep learning methods, typically based on fine-tuning convolutional networks with cross-entropy loss, may not fully capture inter-class relationships in small datasets. This study investigates whether supervised contrastive learning can produce more discriminative feature representations for otoscopic image classification compared to conventional fine-tuning. Using the publicly available Eardrum Dataset, images were reorganized into two categories, normal (534 images) and abnormal (391 images), and split into training (70%), validation (15%), and test (15%) sets using stratified sampling. Two pipelines were compared: (1) a baseline model fine-tuning a pretrained ResNet-50 using cross-entropy loss, and (2) a supervised contrastive learning approach, where a ResNet-50 encoder was trained to minimize intra-class distance and maximize inter-class separation. Contrastive pairs were generated via strong augmentations including random cropping, color jitter, grayscale conversion, flipping, and erasing. After training, the encoder was frozen, and a linear classifier was trained on the learned embeddings. A temperature ablation (τ = 0.03, 0.07, 0.10, 0.20) was performed to assess sensitivity. The contrastive learning model achieved superior performance, reaching 85.82% accuracy and 85.36% F1-score compared to the baseline's 82.27% accuracy and 81.43% F1-score. The improvement was most evident in detecting abnormal cases. These findings demonstrate that a supervised contrastive learning stage can enhance diagnostic reliability in otoscopic image classification, particularly under limited and imbalanced data conditions.

Keywords: otoscopy, deep learning, contrastive learning, fine-tune, supervised learning.

1 Introduction

Otitis media (OM) is a common inflammatory disease of the middle ear and an important global health problem, especially in children, where a large proportion

R. Rzayev et al. (eds.), *Proceedings of the International Conference on Current Problems in Engineering and Applied Sciences (ICCPEAS 2025)*, Advances in Engineering Research 299,
https://doi.org/10.2991/978-94-6239-668-5_31

experience at least one episode in early childhood and recurrent infections may lead to hearing loss, delayed speech, and reduced quality of life [1, 2]. The clinical spectrum includes acute otitis media (AOM) with tympanic membrane bulging and acute inflammation, otitis media with effusion (OME) characterized by fluid behind the membrane without prominent symptoms, and chronic suppurative otitis media (CSOM), which is associated with perforation, persistent discharge, and long-term structural damage; in addition, earwax, although not an OM subtype, can obscure the eardrum and complicate evaluation [3]. Diagnosis in routine practice relies mainly on visual inspection of the tympanic membrane via otoscopy, but this process is subjective and affected by clinician experience, image quality, lighting, and the subtle nature of early findings. Despite the availability of diagnostic guidelines describing key membrane features such as bulging, inflammation, opacity, and perforation, adherence varies across practitioners, and OM remains prone to misdiagnosis or delayed diagnosis, contributing to inappropriate antibiotic use and missed opportunities for timely treatment [4].

In recent years, computer-aided diagnostic (CAD) systems and machine learning approaches have been explored to support decision-making in otolaryngology, with several studies applying classical machine learning, handcrafted features, or deep convolutional neural networks (CNNs) to classify otoscopic images of OM [5]. Although many of these methods report promising accuracy, they often depend on labor-intensive feature engineering, non-public or single-center datasets, and show limited generalization across different populations and imaging conditions; moreover, standard deep learning models trained solely with cross-entropy loss can struggle to learn robust, discriminative representations when data are scarce or class distributions are imbalanced, as is common in otoscopy. These limitations highlight the need for more reliable and data-efficient computational strategies for tympanic membrane assessment. Deep representation learning, and in particular supervised contrastive learning, offers a way to explicitly pull together embeddings from the same class and push apart different classes, thereby improving class separability under constrained data regimes. Building on this idea, the present study evaluates supervised contrastive learning against conventional fine-tuning for automated classification of otoscopic images.

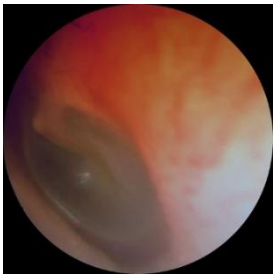
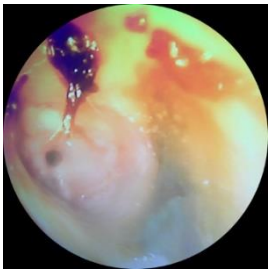
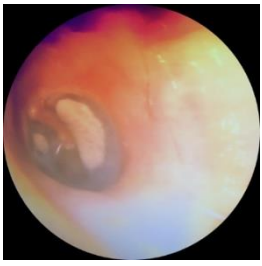
Recent work on automated otitis media (OM) diagnosis with deep learning broadly falls into two lines: image-based and video-based approaches. Most studies use single otoscopic images and train convolutional networks (e.g., Xception, MobileNet, EfficientNet, ensembles) to distinguish between normal ears and different OM subtypes, often reporting very high accuracies on their own institutional datasets[6, 7]. However, these image-based models are usually built on large, curated datasets and tend to lose performance when tested on external data due to changes in device type, lighting, and acquisition conditions. A smaller but growing group of studies instead uses full otoscopic video recordings and exploits temporal information, for example with recurrent networks or video transformers, to capture membrane motion and fluid levels that are important for diagnosing effusion [8]. These video-based methods help overcome some limitations of single-image models, but they introduce additional complexity in data collection, annotation, and computation.

2 Material and Methods

2.1 Dataset

We reorganized a publicly available tympanic membrane (TM) image collection [3] into a binary classification task: normal vs abnormal. We retained all images labeled Normal TM as the normal class (n = 535). The abnormal class aggregates clinically relevant TM conditions with pathology indicators: Acute Otitis Media (AOM) (n = 119), Chronic Suppurative Otitis Media (CSOM) (n = 63), Earwax/cerumen impaction (n = 140), Otitis externa (n = 41), and Tympanosclerosis (n = 28), yielding 391 abnormal images in total. Rest of the categories were excluded to avoid extreme imbalance and because they do not consistently reflect disease state in the same sense. All labels used in this study come from the dataset's expert annotations; we did not alter per-image labels beyond the binary consolidation. The final working dataset thus contains 926 images (535 normal, 391 abnormal). Table 1 shows the class distribution used in our experiments.

Table 1. Image samples from normal (n=534) and abnormal classes (n=391)

		
Normal TM (n=534)	**Acute OM (n=119)**	**CSOM (n=63)**
		
Otitis externa (n=41)	**Tympanoskleros (n=28)**	**Earwax (n=140)**

2.2 Proposed Method

All experiments use the same training (70%), validation (15%), and test (15%) partitions generated with stratified sampling. Two pipelines are compared. First, a

classical baseline model fine-tunes a ResNet-50 pretrained on ImageNet by replacing the classification head and training the full network with cross-entropy loss [9]. The entire network is fine-tuned using cross-entropy with label smoothing 0.1, LR = 3×10^{-4}, weight decay = 0.05, for 40 epochs. Second, we first learn class-aware representations and then train a simple classifier [10]. A ResNet-50 encoder (same initialization) outputs features via an identity head, followed by a projection head (512 → 128). For the contrastive stage, each training image is turned into two strongly augmented views (Random Resized Crop with larger scale range, flip, heavy Color Jitter with p=0.8, Random Grayscale p=0.2, small rotation, Normalize, Random Erasing). We optimize a supervised contrastive loss that pulls same-class pairs together and pushes different-class pairs apart, for 30 epochs with *AdamW* (LR = 3×10^{-4}, WD = 0.05). We run a temperature ablation with $\tau \in \{0.03, 0.07, 0.10, 0.20\}$, logging the SupCon loss curve for each τ. Next, we freeze the encoder and train a linear classifier (single fully connected layer) for 20 epochs using cross-entropy with label smoothing 0.1 (LR = 1×10^{-3}, WD = 0.05). For fairness, the linear-probe stage uses the same normalization and standard classification augmentations as Experiment 1. For each τ, we save the best validation model, evaluate on the same test split, and report the full metric set and confusion matrix. The overall steps are illustrated in Fig. 1.

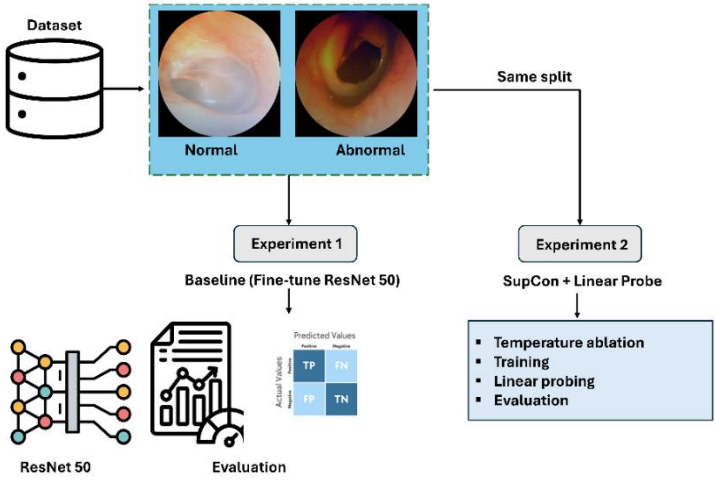


Fig. 1. Methodological pipeline of the proposed framework

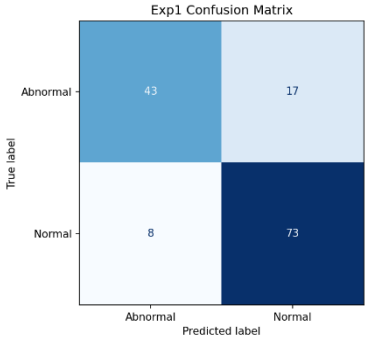
3 Results and Discussion

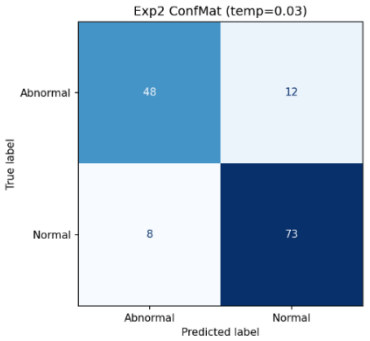
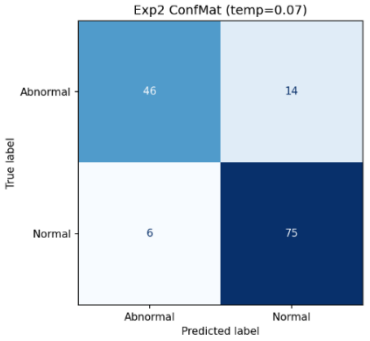
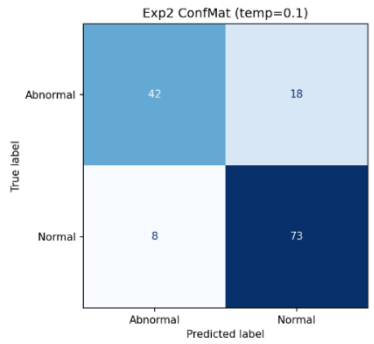
The experimental results in Table 2. show that supervised contrastive learning provides a clear performance improvement over the classical fine-tuning baseline. The ResNet-50 baseline model achieved 82.26% accuracy and an 81.42% F1-score, with the confusion matrix indicating that many misclassifications occurred in the abnormal class. This limitation is clinically important, as insufficient sensitivity to

pathological cases may lead to delayed diagnosis or improper treatment. In comparison, the supervised contrastive learning pipeline consistently outperformed the baseline across most of the temperature values. The best results were obtained at $\tau = 0.03–0.07$, reaching 85.81% accuracy and 85.18–85.35% F1-score, representing an improvement of approximately 3–4%. The gains were most evident in abnormal case detection, confirming that contrastive objectives help the encoder learn more discriminative and well-separated feature representations. Higher temperatures (e.g., $\tau = 0.1$ or $\tau = 0.2$) resulted in decreased performance, demonstrating the importance of appropriate temperature tuning.

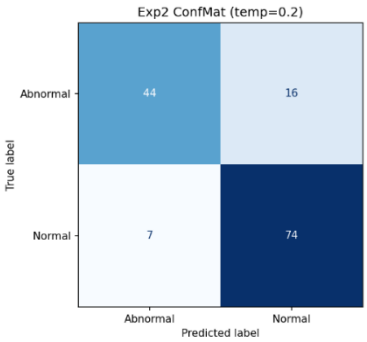
When compared with the study [3] by Comert et al. (2020), which uses the same dataset but performs multiclass classification, differences in performance are expected. Their multiclass setting is more challenging due to substantial overlap between OM subtypes and limited samples for each category. Our binary formulation simplifies the task and provides more stable class boundaries, explaining the higher accuracy obtained in this study. Nonetheless, their work offers more detailed clinical categorization, while ours focuses on improving robustness under limited data conditions. Although contrastive learning strengthens representation quality, the approach still depends on augmentation strength, dataset size, and hyperparameter selection. Expanding the dataset, validating across different imaging devices, and adopting contrastive methods for multiclass OM classification represent meaningful directions for future work. Overall, the findings demonstrate that incorporating a supervised contrastive stage prior to classification is an effective strategy to enhance otoscopic image analysis, particularly when datasets are small and class imbalance is present.

Table 2. Confusion matrices and performance metrics of baseline (Exp1) and SupCon (Exp2) model within ablation results

Models	Confusion Matrix	Performance metrics
Baseline		**Precision:** 82.71% **Recall:** 80.89% **F1-score:** 81.42% **Accuracy:** 82.26%

SupCon (τ=0.03)		**Precision:** 85.79% **Recall:** 85.06% **F1-score:** 85.35% **Accuracy:** 85.81%
SupCon (τ=0.07)		**Precision:** 86.36% **Recall:** 84.62% **F1-score:** 85.18% **Accuracy:** 85.81%
SupCon (τ=0.1)		**Precision:** 82.10% **Recall:** 80.06% **F1-score:** 80.62% **Accuracy:** 81.56%

SupCon (τ=0.2)



Precision: 84.24%
Recall: 82.34%
F1-score: 82.91%
Accuracy: 83.68%

4 Conclusion

This study shows that adding a supervised contrastive pretraining stage to a standard ResNet-50 pipeline yields more discriminative representations for otoscopic image classification than direct fine-tuning alone. Current results suggest that contrastive objectives promote better inter-class separation and more balanced decision boundaries under limited, imbalanced data. Practically, the method is simple to adopt - requiring only strong augmentations, a projection head, and linear probing - yet delivers consistent benefits without large datasets or domain-specific pretraining. Future work will expand to multi-class labeling and evaluate cross-domain generalization across devices and clinical sites, aiming to translate these gains into robust, real-world screening tools.

Acknowledgments. The author would like to thank the creators of the publicly available tympanic membrane dataset on (https://www.kaggle.com/datasets/erdalbasaran/eardrum-dataset-otitis-media) for making their data accessible to the research community.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Shah-Becker, S., Carr, M.M.: Current management and referral patterns of pediatricians for acute otitis media. Int. J. Pediatr. Otorhinolaryngol. 113, 19–21 (2018). https://doi.org/10.1016/j.ijporl.2018.06.036.
2. van Uum, R.T., Venekamp, R.P., Sjoukes, A., van de Pol, A.C., de Wit, G.A., Schilder, A.G.M., Damoiseaux, R.A.M.J.: Optimising pain management in children with acute otitis media through a primary care-based multifaceted educational intervention: study protocol

for a cluster randomised controlled trial. Trials. 19, 501 (2018). https://doi.org/10.1186/s13063-018-2880-4.
3. Zafer, C.: Fusing fine-tuned deep features for recognizing different tympanic membranes. Biocybern. Biomed. Eng. 40, 40–51 (2020). https://doi.org/10.1016/j.bbe.2019.11.001.
4. Camalan, S., Moberly, A.C., Teknos, T., Essig, G., Elmaraghy, C., Taj-Schaal, N., Gurcan, M.N.: OtoPair: Combining Right and Left Eardrum Otoscopy Images to Improve the Accuracy of Automated Image Analysis. Appl. Sci. 11, 1831 (2021). https://doi.org/10.3390/app11041831.
5. Başaran, E., Cömert, Z., Çelik, Y.: Convolutional neural network approach for automatic tympanic membrane detection and classification. Biomed. Signal Process. Control. 56, 101734 (2020). https://doi.org/10.1016/j.bspc.2019.101734.
6. Akyol, K., Uçar, E., Atila, Ü., Uçar, M.: An ensemble approach for classification of tympanic membrane conditions using soft voting classifier. Multimed. Tools Appl. 83, 77809–77830 (2024). https://doi.org/10.1007/s11042-024-18631-z.
7. Wu, Z., Lin, Z., Li, L., Pan, H., Chen, G., Fu, Y., Qiu, Q.: Deep Learning for Classification of Pediatric Otitis Media. Laryngoscope. 131, (2021). https://doi.org/10.1002/lary.29302.
8. Lu, H., Camalan, S., Elmaraghy, C., Moberly, A.C., Gurcan, M.N.: A video classification method for diagnosing ear diseases using otoscope imaging. In: Astley, S.M. and Wismüller, A. (eds.) Medical Imaging 2025: Computer-Aided Diagnosis. p. 117. SPIE (2025). https://doi.org/10.1117/12.3046822.
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90.
10. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised Contrastive Learning. (2021).