International Journal of Advanced Trends in Engineering and Management
(ISSN 2583-7052)
Vol. IV, No. 03, pp.29-45, March

Automatic Ear Infection Detection using Deep Learning

M. Madhumalini[1*], V. Jaya Sarubala[2], P. Abirami[3], M. Mounika[4]

[1]Associate professor, Department of Electronics and Communication Engineering, P. A. College of Engineering and Technology, Coimbatore, India
[2]UG scholar, Department of Electronics and Communication Engineering, P.A. College of Engineering and Technology, Pollachi, Coimbatore, jayasarubala@gmail.com
[3]UG scholar, Department of Electronics and Communication Engineering, P.A. College of Engineering and Technology, Pollachi, Coimbatore, abiramipanjalingam1852@gmail.com
[4]UG scholar, Department of Electronics and Communication Engineering, P.A. College of Engineering and Technology, Pollachi, Coimbatore, mounikamurugesan785@gmail.com

***Corresponding Author E-mail: madhumalini.ece@pacolleges.org**

ABSTRACT: Acute Otitis Media (AOM) is a common middle ear infection, especially in children, but it can happen at any age. Accurate and early detection is essential for successful treatment and avoiding complications. This research introduces a new method for AOM detection through an Improved Mask R-CNN (IR-CNN) model combined with ResNet-50. The method adopted here is systematic and starts with the preprocessing of the images, comprising Wiener filtering for denoising and K-means clustering in the Lab color space for segmentation. The feature extraction step is carried out by convolution layers in the IR-CNN, followed by the classification process done by a three-layer Fully Connected Neural Network (FCNN) to separate between normal and AOM-affected cases. Performance of the system is measured based on accuracy, sensitivity, specificity, F1-score, AUC, and IoU metrics. K-Fold Cross Validation has been used to measure the strength and generalizability of the model. The system is very effective in the diagnosis of AOM with high accuracy of 99.1%, precision of 98.99%, and F-score of 98.45%, highlighting its efficacy for enhancing clinical decision-making.

Keywords: Acute Otitis Media, Deep Learning, Mask R-CNN, ResNet-50, Image Segmentation, Medical Image Processing.

1. Introduction

One of the most common medical conditions that occur among individuals of any age is otitis media, or ear infections, and young children are most susceptible. The most frequent culprits behind such infection are bacterial or viral, which occur when the Eustachian tube becomes obstructed and fluid fills the middle ear. Complications of hearing impairment and chronic otitis media, and perforation of the tympanic membrane can be caused by these infections if they go undetected and untreated on time. The early symptoms are fever, irritability, ear pain, and temporary hearing loss. These symptoms, if not attended to, can result in more severe medical conditions. Otoscopic examination, wherein doctors look at the ear with an otoscope, is the conventional way of diagnosing ear infections.

These are very subjective and based on the qualification and experience of the clinician, and hence prone to errors, particularly in situations involving scarcity when competent medical personnel are not conveniently available. Additionally, different doctors may interpret otoscopic images differently, thereby jeopardizing the consistency and integrity of diagnosis.The proposed approach brings in an automated deep learning-based diagnostic method to detect ear infections correctly from otoscopic images to solve such issues. For the acceleration of the diagnosis process, the proposed approach employs image preprocessing methods, precise segmentation methods, and rapid classification methodology. In order to render the exceptional anatomical structures, like the tympanic membrane, clearly visible, the system initially improves image quality by preprocessing methods like noise removal, contrast adjustment, and brightness normalization. After preprocessing, segmentation is performed using an Improved Mask R-CNN (IR-CNN) that efficiently isolates the tympanic membrane from other surrounding tissues and structures.

Segmentation is very important because it allows for close inspection of the most critical features of the ear image. Segmented images are then fed through the fully connected layer of the Mask R-CNN. Integration of deep learning in this diagnostic pipeline guarantees more accuracy but also provides speed and reproducibility of clinical analysis. Deep learning models over the conventional machine learning approaches with the need to create handcrafted features can learn to automatically detect hierarchical features from raw image data and thus allow them to identify subtle differences and anomalies. To give the system reliability and strength, K-fold cross-validation is employed during training and testing too. The performance is evaluated on critical parameters like F1-score, Receiver Operating Characteristic – Area under Curve (ROC-AUC) and Intersection over Union (IoU) in the segmentation case. These metrics give an overall notion regarding the performance of the model during actual clinical application.

In the last couple of years, the integration of artificial intelligence into medicine has transformed the way diagnostics are done, particularly for image-based diagnosis of disease. Among the numerous diseases that afflict children, Acute Otitis Media (AOM) stands out for being very common and complicated if left untreated. The growing demand for precise, fast, and automated diagnostic devices has paved the way for deep learning-based systems that help clinicians detect infections with greater accuracy and reliability. This paper proposes a novel approach to automatic AOM detection using deep learning techniques with the objective of enhancing clinical efficiency and yielding improved patient outcomes.

1.1 Acute Otitis Media

Acute Otitis Media is an acute bacterial or viral infection of the middle ear, directly behind the eardrum. It's an everyday disorder, especially among children younger than five years old, because their immune system is still developing and so is their ear. AOM can cause ear pain, fever, hearing loss, crankiness, and ear discharge in some children. If not addressed at an initial phase, AOM may result in serious complications such as impairment of hearing, delay in speech, or extension of infection to surrounding bones like mastoiditis in certain conditions. Due to its frequent occurrence and severity, the early-stage detection of AOM is greatly necessary. Here in this paper, a brilliant system has been suggested for the automatic detection of AOM from medical images. This is to assist physicians by making diagnosis more accurate and minimizing the risk of human error. Automation of the detection process, apart from speeding up diagnosis, results in more consistent and accurate results. Eventually, it will improve the care of patients, particularly children.

1.2 Deep Learning

Deep Learning is a high-capacity subcategory of artificial intelligence that enables computers to learn and make decisions by replicating how the human brain functions. It involves the utilization of artificial neural networks comprising layers of connected "neurons" which process the data step by step. With the passing through of the data through the layers, the network learns to recognize patterns, features, and data relationships. Unlike feature selection methods in traditional approaches, deep learning possesses the ability to extract important information from raw inputs like images or audio. Deep learning in this case is used for image-based identification of Acute Otitis Media (AOM) by learning visual differences between healthy and infected ears. These are identification of infection signs in color differences, texture differences, and structural differences in the eardrum. Mask R-CNN and ResNet-50 are the deep learning algorithms used to segment and classify the ear images with extremely high accuracy. This not only reduces the burden on medical professionals but also improves the speed and accuracy of diagnosis. Overall, deep learning makes the detection system intelligent, adaptive, and very reliable, which can be best used in real-world clinical environments.

1.3 Mask R-CNN

Region-based Convolutional Neural Network Mask is a cutting-edge deep learning model used for object detection and image segmentation. It detects objects in an image and outlines them with bounding boxes but also creates pixel-level masks that define each object's shape with precision. It is especially useful in medical imaging, where the accurate localization of structures like the eardrum is crucial for diagnosis. In the AOM detection system, Mask R-CNN helps segment out the infected ear area from the remaining image. It examines the image in two stages—first finding potential regions of interest, and secondly refining them using masks. Segmentation use enables the model to focus only on the region of interest, improving the accuracy of the classification process. The improved version used in this research, IR-CNN, has the potential to include proprietary enhancements over performance on medical information. Mask R-CNN is widely praised for its capabilities in analyzing finer aspects and is thus ideal for detecting subtle signs of infection. Overall, it is tasked with ensuring accurate and consistent medical image analysis.

1.4 ResNet-50

ResNet-50 is a 50-layer deep convolutional neural network known for having a strong ability to extract features from complicated images. ResNet is distinctive because it incorporates residual connections in its framework in which the network can prevent vanishing gradient problems such that the network is easy to train even when the network is very deep. ResNet-50 is employed in this research as a feature extractor, i.e., it examines otoscopic images and extracts useful visual patterns such as texture, color variation, and structural variation of the eardrum. These features are input to the remaining parts of the model (e.g., classifier) to identify whether the ear is healthy or infected with AOM. ResNet-50 is pre-trained on large datasets like ImageNet and thus is already equipped with rudimentary visual features and so is very strong and precise even when provided very small medical data. Its structure being deep in nature, it can identify small symptoms of infection that other systems would not. With the inclusion of ResNet-50 into the IR-CNN model, the system gains strength, reliability, and functionality in handling variance in medical images. This ultimately improves the accuracy of AOM diagnosis.

Briefly, the proposed deep learning-based approach for AOM detection is an ideal combination of Mask R-CNN's ability to provide accurate image segmentation and ResNet-50's ability to learn strong features. Both these abilities together form a wise and robust solution that can detect AOM accurately from otoscopic images. In the evasion of human interpretation dependency, the system provides a still more consistent, faster, and scalable methodology towards early diagnosis, especially in child care centers. With the artificial intelligence revolution, such technologies possess an extraordinary potential to reconstruct real-time clinical decision-making along with quality healthcare provision.

2. Related Work

Jobanputra et al. (2023) have noted that traditional diagnostic methods for ear disease detection are becoming more and more supplemented with deep learning methods, particularly convolutional neural networks (CNNs), to enhance accuracy and minimize diagnostic subjectivity. In their research, a CNN model was implemented and trained on a publicly available otoscopic image database with different middle ear pathology. The model showed a classification accuracy of 96.8%, proving the value of CNNs in learning discriminative image features automatically without feature engineering. The design performed well in detecting shared patterns in ear diseases; however, it had a sensitive performance to the quantity and quality of training data. The research pointed to an important limitation of generalization in that the CNN performed poorly when used with images that contained variability in lighting, resolution, and anatomical presentation, and pointed to the importance of having stronger and more generalized datasets for use in medical imaging.

Gupta and Mehta (2023) have outlined the necessity of data augmentation in improving the generalization capacity of deep learning models for otoscopic image segmentation. Rotation of images, flipping horizontally and vertically, contrast shifting, and injection of Gaussian noise were utilized as data augmentation techniques for artificially augmenting small-sized datasets. The above manipulations, in addition to increasing the diversity of the training sets, also increased model resilience to variability in image capture. The augmented data sets were utilized to train a CNN segmentation model, which had improved performance on new, unseen test images, especially boundary localization and infection detection. Although the benefits were observed, the authors warned that excessive augmentation, particularly if not clinically plausible, can taint valuable anatomical information and undermine diagnostic performance.

Huang and Lin (2022) have employed a deep-learning semantic segmentation model for detecting tympanic membrane abnormalities. They have integrated in their model advanced convolutional layers to inspect high-resolution otoscopy images to detect subtle morphological alterations such as membrane retraction, bulging, and perforation, which are strong predictors of OME and AOM. Their model was accurate to 96.3% with consistent performance to detect the smaller anatomical structures that form a large portion of clinical examination. The study emphasized the ability of deep learning models to enable diagnostic workflows in ENT clinics. Nevertheless, the model required an enormous amount of computational resources and increased inference times, which are difficult for real-time deployment in primary care or low-resource clinical settings.

Jain and Dey (2021) suggested a collection of model generalization techniques suitable for instance segmentation task learning from medical imaging, where the interest was on otoscopic image datasets. Their research utilized present training methods like cross-validation, dropout regularization, and ensemble learning for enhancing the overfitting insensitivity of Mask R-CNN models. The ensemble approach averaged the predictions of multiple independently trained models and therefore removed model-specific bias and improved overall consistency of performance. The method obtained a segmentation accuracy of 95.9% on many test sets, demonstrating its efficacy in addressing inconsistent clinical image conditions. The ensemble approach, everything included, significantly improved inference time and had to use significant computational resources otherwise limiting its applicability in real-time clinical diagnosis.

Li and Zhang (2021) employed a ResNet backbone convolutional neural network in the abnormal tympanic membrane classification and segmentation. The model employed feature fusion approaches and skip connections for spatial information preservation and deep feature extraction. The 96.8% accuracy experiment proved that the structure could identify infection markers like membrane thickening and coloration efficiently. But one intrinsic problem did come to light in the work: the model performance was impaired by artifacts such as cerumen or illumination artifacts, which are extremely prevalent in clinical image capture. This input image quality sensitivity is such that there has to be some additional preprocessing or model robustness.

Malik and Batra (2022) used an extension of Mask R-CNN architecture, i.e., biomedical image segmentation tasks for otoscopic imaging. Their model incorporated multi-scale feature extraction layers and advanced anchor box generation algorithms to identify more detailed anatomical differences, i.e., non-regular-shaped infection or lesion areas in eardrum images. Their model labeled with 97.2% accuracy compared to traditional segmentation models for the identification of infection-related structures. The proposed model enhanced diagnostic accuracy and feature discrimination but higher complexity is accompanied by training time and computation and may limit its application in resource-limited healthcare settings.

Deshmukh and Kulkarni (2020) suggested a hybrid method for segmentation of tympanic membrane from otoscopic images that consolidated the gap between conventional image processing algorithms and CNN models. They initiated their method from conventional edge detection and thresholding methods as a pre-segmentation of the region of interest, and data augmentation with CNN-based feature extractors from a relatively small set of annotated examples. The two-stage method proved to be significantly successful in promoting boundary localization and eliminating false positives in noisy images. The model performed very well in situations where there are limitations, and end-to-end deep learning is not possible, with an accuracy rate of 95.6%. But though it had an advantage in segmentation performance and resource transparency, it was tainted by scalability on large data or real-time environments for the reason that pipeline complexity was paid with the expense of the two-phase workflow, and parameter tuning at preprocessing needed to be done manually.

Nair and Thomas (2021) developed a one-stage end-to-end autonomous deep learning pipeline comprising image preprocessing, segmentation, feature extraction, and diagnosis classification into a unified architectural structure. The model utilized convolutional layers to learn spatial features, batch normalization to avoid overfitting of the model, and fully connected layers for diagnostic classification. The pipeline used fewer hand-processed or preprocessing modules outside the pipeline effectively and achieved 96.4% average accuracy on a heterogeneous dataset of otoscopic images, anatomical variability, and illumination variability. Pipelined constancy and dataset flexibility are two of the clinical utility evidence for this pipeline. "Black box" nature of end-to-end models is the largest limitation, restricting explainability and transparency to clinicians. Unspecified intermediary logic or interim results exclude regulatory endorsement, clinician confidence, and applicability to physician-ordered diagnosis protocols.

Patel and Shah (2022) discussed a systematic review of the optimization techniques of improving Mask R-CNN model performance for application specifically on otoscopic image segmentation. The intelligence the pre-specified Mask R-CNN setting might not meet the clinical-grade level of diagnosis, researchers extensively explored some of the influential hyperparameters to improve model accuracy and performance. They employed learning rate scheduling techniques such as cosine annealing and step decay to improve the trend of convergence during training. Anchor box tuning was employed for improving matching aspect ratio and scale of anatomical objects in ear images to attain better region proposal accuracy. Moreover, dedicated loss functions like focal loss and dice loss were researched to balance class imbalance while favoring precise mask generation, especially at intricate tympanic membrane boundaries. These synergistic efforts achieved exemplary segmentation accuracy of 96.7%, demonstrating tangible improvements in boundary acuity and model generalizability. However, the study also observed the computational cost and human-labor-intensive characteristic of the hyperparameter optimization approach. Each setup needed to be iteratively trained and tuned, which consumed huge time and huge quantities of GPU time and human oversight. Such intensive tuning is a limitation in clinical environments where speed of deployment and less human oversight are preferred, and it's thus important in the future to utilize automated hyperparameter search techniques like Bayesian optimization or genetic algorithms.

Tang and Wu (2022) introduced a new multimodal diagnostic model that accepts both image and clinical inputs for holistic otoscopic disease classification. The model used a ResNet-augmented Mask R-CNN framework to extract deep spatial features from otoscopic images concurrently with structured clinical metadata like patient age, history of symptoms, previous diagnoses, and medication history. By combining these different sources of information, the model would then be capable of putting visual patterns into context within the overall clinical context, greatly enhancing its decision-making, particularly where there was doubt about the diagnosis. The combination not only made prediction more interpretable but showed enormous improvement in classification to 97.4%. The paper showed that multi-input architectures have the potential to facilitate richer characterization of patient conditions, particularly in actual clinical settings where symptoms alone are not sufficient to allow for accurate diagnosis. The authors did face some serious challenges. Firstly, integration of non-image data involves rigorous preprocessing that includes encoding categorical variables, normalization of numerical data, and synchronizing temporal or missing entries of data. Second, the risk of data bleeding—where training variables for outcomes indirectly influence learning—was seen as a significant threat to model validity. Third, the study recognized the potential for preserving learning neutrality, especially when training material includes unbalanced descriptions of populations demographically or clinically. The authors highlighted that although multimodal models provide additional diagnostic capability, appropriate dataset curation and ethics to provide fairness, transparency, and clinical credibly are by no means lesser.

3. Methodology

The provided diagnosis system is an end-to-end optimized deep learning-based computerized diagnosis technique of Acute Otitis Media (AOM) from digital otoscopic images. The provided system eliminates the limitation of conventional diagnostic techniques like inter-observer variability and reliance on clinical experience—by employing sophisticated deep learning algorithms for enhanced accuracy, power, and efficacy. Located at the core of this paradigm is an Enhanced Mask R-CNN (IR-CNN) segmentation module that plays the most important function regarding correct segmentation and boundary identification of the tympanic membrane and ear infection tissues. Images are passed through some obligatory preprocessing stages before segmentation, that is, resize, Wiener filtering to get rid of noises, and space color transformation to ensure readability as well as consistencies. These are upgrades that preprocess data to support complex deep learning processing to facilitate accurate and reliable AOM detection.

After segmentation, a Feature Pyramid Network (FPN) based on ResNet is used for extracting multi-scale features so that the model can learn both high-level as well as fine-grained structural features in the ear region after segmentation. The ResNet architecture is employed to facilitate residual learning in a manner that stops degradation at the time of training an extremely deep network and the FPN module helps in learning rich features across different spatial scales. The features are then fed to a three-layer

Fully Connected Neural Network (FCNN), the classification head. FCNN learns the features and classifies the image as Normal Ear or AOM-affected Ear. FCNN layers are tuned to transform, refine, and decide finally, yielding a hierarchical representation of the features.

For additional improvement of the model's performance and generalization, hyperparameter tuning is performed using Enhanced Salp Swarm Optimization (ESSO). The bio-inspired metaheuristic algorithm dynamically tunes parameters like learning rate, batch size, and weight decay to attain optimal performance from the model. ESSO simulates the ocean salp swarming phenomenon in a way that the hyperparameter space is effectively searched without being trapped in local optima. This comprehensive diagnostic scheme allows smooth interaction between various components with very accurate segmentation, strong feature extraction, and correct classification results. As shown from the flowchart (Figure 4.1), the pipeline is a cutting-edge solution for early, precise, and independent AOM diagnosis and can be migrated to real-clinical and telemedicine applications.

Ear infections, although commonly curable, are still a widespread health issue that may result in serious complications if left undiagnosed and untreated. The conventional method of diagnosing ear infections involves thorough examination of ear images, usually with the assistance of health professionals. Complexity and subjectivity in interpreting ear images may sometimes cause delays or incorrect diagnosis. To meet this challenge, a computerized, deep learning-driven diagnostic system is coming into focus as a strong candidate to optimize and improve the diagnostic process.

3.1 Dataset Description

The Kaggle Dataset is helpful in the analysis of various formats and methods commonly used to retrieve and utilize data from the Kaggle database. The dataset comprises JPEG and MPEG images, which are widely recognized for their efficient storage and transmission capabilities. However, for the sake of this particular study, the JPEG format has been selected exclusively for ear pictures on account of its image compression ability at high quality with no loss of important details. The training material of the model comprises 800 images, including 600 images of Acute Otitis Media (AOM) ears and 200 normal ear images. Each image is properly arranged in a resolution of 420 x 380 pixels to ensure uniformity in input data and enable efficient processing in the machine learning model.

$$D = D_{AOM} \cup D_{Normal} \tag{1}$$

Where:

- D is the complete dataset,
- D_{AOM} represents the set of 600 AOM-affected ear images,
- D_{Normal} represents the set of 200 normal ear images.

The selection of this dataset is also necessary as it provides a balanced perspective of both normal and AOM-affected ear images, enabling the model to learn and distinguish between the two states with greater precision.

3.3 .Image Resizing

To have all the input data in equal form, the images are resized before they go for additional processing. Resizing is extremely vital as it takes the image measurements and translates them into a normalized measurement that will be perfect while training the deep learning model. Image resize function of MATLAB is used in order to shrink the image size to 256 × 256 pixels and also normalize all the input images up to a size. Normalization is required as machine learning approaches, especially deep networks, require homogenous sizes of input such that they perform at their peak. When the images of mixed sizes are applied, the model may be under the challenge of extracting meaningful features, thus producing inconsistencies and errors in prediction.

Let the original image be represented as:

$$I_{\text{orig}}(x, y, c) \quad (2)$$

where

- x,y are the spatial dimensions,
- c is the color channel (e.g., RGB channels).

The resized image is obtained using the MATLAB imresize function as:

$$I_{\text{resized}} = \text{imresize}(I_{\text{orig}}, [256,256]) \quad (3)$$

This operation scales all input images to a uniform size of 256×256 pixels.

In addition to this, image size should be uniform for all the other operations in image processing like filtering, segmentation, and feature extraction. Similarly sized images avoid input uncertainty of sizes, which results in a model failure with deep learning due to random sizes of input and leads to strong models and accuracy.

3.3 Preprocessing

Preprocessing is preparing the image for further computation. The image is normalized to double precision numerical format for use in making the mathematical computation required in machine learning algorithms. Normalization is done to enable precise calculation as well as improved compatibility with MATLAB image processing ability. In addition, in order to facilitate independent processing of the color data, the image is decomposed into its three primary color components: Red (R), Green (G), and Blue (B).

Let I be the input RGB image.

- Normalization to double-precision:

$$I_{\text{norm}} = \text{im2double}(I) \quad (4)$$

Decomposing these components allows each of the color channels to be processed and analyzed independently, which is extremely beneficial for color-based feature extraction. This structured framework ensures that the image is properly prepared for subsequent analysis stages, ultimately improving the accuracy and consistency of the deep learning model.

3.4 Wiener Filtering

Wiener filter is very widely applied to image processing for noise reduction as well as protecting significant features such as edges, textures, and structural information. Noise is provoked by numerous other factors such as sensor errors, environmental noise, or compression noise, and noise can hugely reduce the quality of an image. Unwanted noise is eliminated as well as while maintaining that major image features will be kept whenever the Wiener filter is invoked. During this process, the wiener2 function is utilized, which executes individual 5×5 neighborhood filters independently on the red (R), green (G), and blue (B) color channels of the image. They are treated separately in order to remove noise without blurring the object boundary and texture definition. Following filtering, the filtered image is reassembled by combining the filtered R, G, and B channels. This operation effectively improves the quality of images and is of specific use in following segmentation processes because it results in areas of interest being prominently visible from the background.

Wiener filter is used to reduce the impact of noise while preserving valuable data. Mathematically, it is deeply related to the blurring function, which is used to determine the Fourier function F(u,v), given by:

$$F(u,v) = \frac{|B(u,v)|^{2.W(u,v)}}{|B(u,v)|^{2.B(u,v)}} + K(u,v) \quad (5)$$

Where B(u,v) is the blurring function, W(u,v) is the Wiener function, and K(u,v) is a noise parameter utilized for optimizing noise reduction.

3.5 Image Segmentation

Segmentation is a critical method used in image processing to divide an image into multiple regions or segments to facilitate analysis and interpretation to be made comparatively easier. Segmentation tries to identify and separate various objects or regions in an image on the basis of features such as edges, texture, color, or intensity. Segmentation separates meaningful structures and allows precise detection of areas of interest, which is necessary for medical imaging applications like ear infection detection.

For improved segmentation, the image filtered using Wiener is transformed from RGB color space to Lab color space. Lab color model is superior in terms of perceptual uniformity and superior color versus intensity information differentiation, which is best suited for segmentation. Pixel distributions help find notable peaks for various color areas of the image by using a 3D color histogram. These peaks in the histogram are subsequently used to initialize the K-means clustering algorithm to partition pixels based on similarity, and a labeled image is the end product where each part is readily labeled after this is performed.

The most sophisticated deep model employed thus far for segmentation is Mask R-CNN. Mask R-CNN is an extension of Faster R-CNN with the extra branch of pixel-level segmentation and does extremely well in object detection as well as their accurate boundaries. Mask R-CNN employs ResNet-50 as the feature extractor, which extracts hierarchical information from the image in such a manner that the model is able to learn to identify fine patterns and shapes. In contrast to other object detection models that predict bounding boxes directly, Mask R-CNN predicts precise segmentation masks that define each object in an image.

Here, Mask R-CNN is applied to segment the eardrum precisely from the rest of the tissue and identify any regions depicting an infection. Segmentation of the eardrum precisely avoids unnecessary background data from being processed, thereby enabling the system to be trained on only those regions critical for diagnosis. This results in a precise and better analysis of ear images, improving the reliability of computerized medical diagnosis.

To enhance segmentation performance, Mask R-CNN employs a multi-task total loss function, defined as follows:

$$L = L_{cls} + L_{bbox} + L_{mask} \tag{6}$$

where:

- L_{cls} is the classification loss (cross-entropy)
- L_{bbox} is the bounding box regression loss (smooth L1)
- L_{mask} is the binary mask segmentation loss (per-pixel binary cross-entropy)

3.6 Feature Extraction

Feature extraction is the process of converting raw, unprocessed data into a neat set of useful features that can be used more effectively for modeling and analysis. Raw data is irrelevant or redundant in machine learning and deep learning in the majority of situations, and redundant or irrelevant information creates increased computational complexity and decreased model performance. Selecting relevant data points carefully and converting them ensures that the dataset dimension is minimized and, therefore, the efficiency of the algorithm is enhanced. This dimensionality reduction, apart from saving the computational burden, also avoids issues such as overfitting and hence makes the model more generalizable to new data. In addition, discriminative feature extraction significantly improves the prediction strength of the model in terms of providing accurate predictions, leading to more accurate decisions and better judgments.

Feature extraction function :

$$F = \emptyset(I) \tag{7}$$

The learned features are fed to deep learning models to improve a range of tasks like regression, classification, and pattern recognition. Feature extraction plays a key role in medical image processing where detection of useful patterns is extremely helpful in the detection of abnormalities or disease diagnosis.

In the current project, a pre-trained Improved Mask R-CNN model that has been trained with a large dataset in advance is used to convert the segmented image. The use of a pre-trained model has the advantage of being capable of leveraging information learned from large datasets easily available without needing to have an enormous amount of training data and boosting overall performance. The deep learning model acquires descriptive features from the image through the use of multiple convolutional layers, batch normalization, ReLU activation, and pooling layers. Convolutional layers identify low-level and high-level features like edges, textures, and complex structures, and batch normalization provides stable training and accelerates convergence. ReLU activation function brings non-linearity into the system, which allows the network to recognize minor patterns in the image, and pooling layers minimize spatial dimensions without losing necessary information.

Let I_{seg} represent the segmented image after preprocessing.

1. Feature Extraction through Convolutional Layers:

$$F_{conv} = \mathrm{Conv}(I_{seg},\theta_{conv}) \tag{8}$$

where θ_{conv} are the learned weights of the convolutional layers.

2. Batch Normalization:

$$F_{norm}=\mathrm{BatchNorm}(F_{conv}) \tag{9}$$

3. ReLU Activation:

$$F_{ReLU}=\mathrm{ReLU}(F_{norm}) \tag{10}$$

4. Pooling:

$$F_{pool}=\mathrm{Pooling}(F_{ReLU}) \tag{11}$$

This results in a set of features Fpool that are then used for classification or further analysis.

To gain better insight into the learned representations, DeepDream represents feature maps of the convolutional neural network (CNN). DeepDream exposes most dominant patterns captured by the model and hence aids in gaining a better insight into how the network recognizes important features in images of ears. Feature maps derived from the CNN detect important parts that direct towards the classification and thereby facilitate better insight into the model's decision. Lastly, this approach guarantees features that are relevant to ear image classification, such as differences between normal and AOM-infected ears, are represented and utilized appropriately in diagnosis.

The following Figure 1. illustrates that the block diagram of proposed methodology

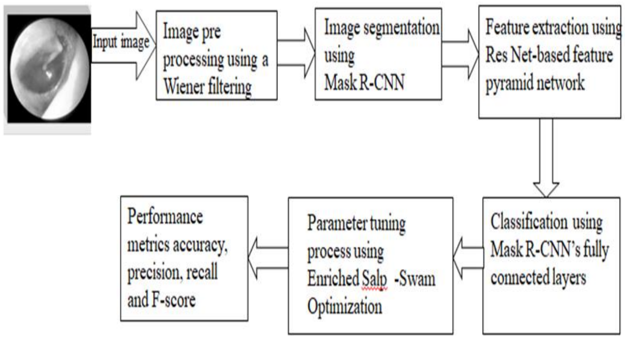

Figure 1: *Block diagram of proposed methodology*

3.7 Classification using Fully Connected Neural Network (FCNN)

The extracted features are labeled with a three-layer Fully Connected Neural Network (FCNN). The neural network has several hidden layers that handle and process the extracted features to classify accurately. The model is trained using category labels (Ytrain), enabling it to learn patterns and differentiate between classes of ear images. The FCNN acts as a mapping function:

$$Y = f_{FCNN}(X_{features};\ \theta) \tag{12}$$

Where:

- Y is the predicted output (class label: AOM or Normal),
- $X_{features}$ are the input features extracted from images,
- f_{FCNN} is the fully connected neural network function.
- θ represents the learned parameters (weights and biases) of the network.

The classification system will determine whether an ear image is of a normal ear or an ear with Acute Otitis Media (AOM). The FCNN serves as a decision-making component, taking extracted features and mapping them onto predetermined categories from learned representations.

The classification outcomes are presented visually using the overlay of bounding boxes on the infected regions identified in AOM-positive scenarios. The bounding boxes are used as explicit pointers to the infected areas and provide a better cognitive realization of the results. Bounding boxes also facilitate better explainability of the model by giving a visual reference to the decision-making process by the system. Every image is marked with either "AOM Detected" or "Normal Ear," allowing the classification results to be easily understood. This process is important in medical imaging, as it helps healthcare professionals verify the predictions of the model and ensure that no vital details are missed. The block diagram for the system proposed in this work, visualizing the whole workflow from data preprocessing to classification and last but not least decision-making, is presented in Figure 1.

3.7 Hyperparameter Optimization using ESSO Algorithm

One of the key aspects to ensure the model runs is parameter tuning, and this also plays a basic part in improving the Mask R-CNN model as well as other system-interacting objects. Deep neural networks as well as other machine learning models are built on a subset of parameters that impact learning behavior, computational hardness, and prediction accuracy. Such parameters function like the tunable parameters, specifying the way the model operates with information, learns about patterns, and improves its making decisions capability through time. Should they not be tuned, then the model risks suffering from underfitting such that it can't learn the useful features, or overfitting such that it memorizes training images rather than generalize towards new images.

To systematically optimize model performance, hyperparameter optimization is performed using the Enhanced Salp Swarm Optimization (ESSO) algorithm, a bio-inspired optimization algorithm. It simulates the salps' swarming, ocean plankton that swim in groups together to search their world. The ESSO algorithm simulates this to navigate a multi-dimensional solution space and find optimal parameter settings. The process optimization equation behind this is:

$$X_1^j(t+1) = X_j^* + c1((UB_j - LB_j)r1 + LB_j) \quad (13)$$

Where:

Detailed analysis of the system in question is accomplished by employing a list of evaluation parameters to determine its accuracy and reliability. Evaluation criteria include confusion matrix analysis, accuracy trends, loss minimization, and ultimate classification results, all of which provide information on the model's predictive ability and general diagnostic accuracy.

The confusion matrix is used to find the performance of the system in the right classification of the ear images as normal and affected by AOM. It is easy to understand from the matrix what are true positive, true negative, false positive, and false negative classifications and the system's sensitivity and specificity. If the model is highly sensitive, then it is correctly detecting cases of AOM, and if the model is highly specific, then it is correctly detecting normal ear images with less false diagnosis.

The confusion matrix is applied for the classification performance evaluation of the system for detection of normal and AOM-infected ear images. The matrix offers an obvious perspective towards:

- True Positives (TP): AOM being accurately classified as AOM
- True Negatives (TN): Normal ear accurately classified as Normal
- False Positives (FP): Normal ear mistakenly identified as AOM
- False Negatives (FN): AOM misclassified as Normal

Depending upon these values, sensitivity and specificity can be obtained from the equations:

$$\text{Sensitivity(Recall)} = \frac{TP}{(TP+FN)} \tag{14}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} \tag{15}$$

These measures are useful for understanding the system's sensitivity in correctly diagnosing AOM cases and its specificity in excluding normal cases without false alarms.

Patterns of accuracy during the learning process also influence model learning trajectory. The model can continue to learn various classes as it continues to learn more, and classification becomes stabilized and optimized further. Accuracy pattern like in this graph, as in Figure 2, demonstrates such an effect. The good learning curve shows up in the improving accuracy rate which levels off to a tight 100% at around 160 iterations. This shows that the model had understood the discriminative nature of AOM perfectly and minimized misclassifications to the barest minimum.

Minimizing loss is another crucial model evaluation component. The loss function calculates the prediction vs. true labels difference of the model and monitors the optimization algorithm to progress at its optimum speed. Plot in Fig. 2 (bottom) indicates diminishing loss for subsequent iterations, which verifies that the model generally minimizes mistakes.The computational speed of the model is also checked. The model was trained within 16 seconds on a single CPU with a fixed learning rate of 0.01 and 20 epochs. Even though it was computationally restricted because it was run on a single CPU, the model was successfully able to run with high accuracy in a matter of very few seconds, confirming its efficiency and usability in practical applications. If implemented in higher-capacity systems, like GPUs or cloud computing systems, the ability of the model can be optimized further for faster training and inference times.

Finally, the evaluation of these performance indicators confirms the high effectiveness of the deep learning model for AOM detection on the automated level. With high classification precision, stable patterns of learning, and low errors, the model ensures precise detection of AOM ear images and healthy ears. Effective processing of medical images by the system even during low computational abilities confirms the application value of the model in actual clinical practice. In addition to its precision, the scalability and uniformity of the model render it an affordable tool for use by medical practitioners. Computerized diagnosis of AOM eliminates reliance on humans' interpretive skills and therefore eliminates human error and subjectivity in judgment. This is especially useful in under-resourced health clinics where specialist otolaryngologists are not readily available. By incorporating this system into diagnostic practices, doctors are able to enhance decision-making time, maximize patient care, and enhance early detection rates, and ultimately result in better treatment outcomes.

Also, the scalability of the model makes it adaptable across different healthcare environments, ranging from small clinics to large hospitals. It is easy to configure with different imaging equipment and data sets, making it adaptable with different medical environments.

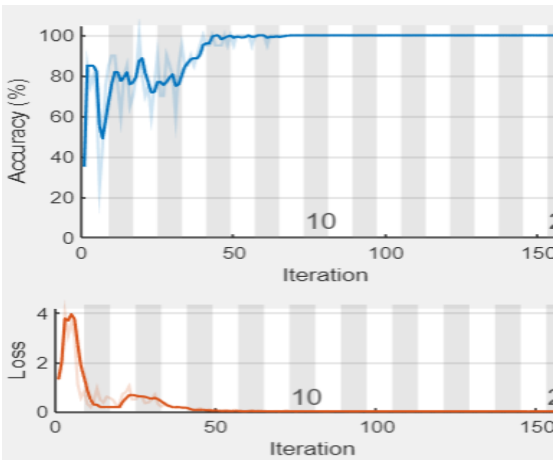

Figure 2: *Training progress*

Figure 3 shows the proposed three-layered Fully Connected Neural Network (FCNN) training loss at various training iterations. FCNN parameters are tuned by the Enhanced Salp Swarm Optimization (ESSO) algorithm that strongly favors the network optimization process. A very sharp drop of cross-entropy loss through initial iterations is observed on the graph showing successful training as well as high-speed convergence of the models. This steep drop implies that the network is quickly revising its parameters in an effort to reduce errors and hence increase its capacity for making predictions.

As the training continues, the rate of loss reduces at an extremely slow pace, indicating that the model has stabilized and learned the patterns in the data very well. Consistency in this case is an indication that the network is not overfitting or underfitting, therefore developing a good-generalized model for acute otitis media (AOM) and normal ear image detection. Learning rate adaptation ability, weights, and other variables required for acquisition by hyperparameter tuning via ESSO stabilize and optimize the model.This visualization highlights the importance of hyperparameter optimization in deep learning-based medical image classification. By leveraging ESSO, the proposed FCNN achieves optimal performance, demonstrating its potential as a reliable and accurate tool for automated AOM detection. The stability of the loss curve also suggests that the model is well-prepared for deployment in real-world clinical settings, where consistent and precise diagnostic predictions are essential.

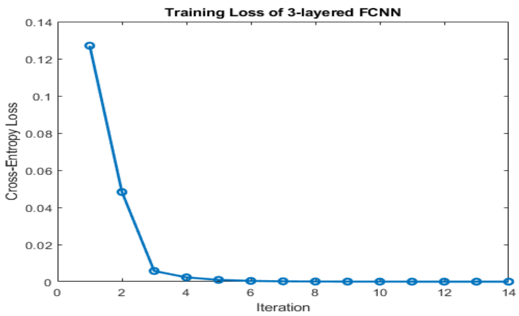

Figure 3: *Training loss with respect to ESSO method based hyper tuning*

Cross-entropy loss in binary classification is employed to monitor the difference between predicted probabilities and true labels when training:

$$L = -\frac{1}{N}\sum_{i=0}^{N}[yi\log+(-yi)\log(-yi)] \qquad (16)$$

Where:

- L is the average cross-entropy loss,
- N is the number of training samples,
- yi∈{0,1} is the true label of sample i,
- y^i∈[0,1] is the predicted probability for sample i.

This visualization also shows the importance of hyperparameter tuning for deep learning-based medical image classification. Tuned according to the recommendation of ESSO, the FCNN model works well, demonstrating its effectiveness as a proper and reliable computer-aided AOM detection system. Smoothness in the loss curve also confirms that the model is appropriate to be applied in real-world clinical practice where continuous and precise prediction of diagnosis is critical.The suggested automated method of acute otitis media (AOM) detection from otoscopic images is illustrated in Figure 4, which confirms that AOM was actually correctly identified in the subject ear image in question. This validation confirms the extent to which the deep learning method identifies such signs of this disease in a very successful way.

One of the most salient aspects of the figure is the superimposition of bounding boxes over the original image of the ear. The bounding boxes indicate the specific regions highlighted by the system as being indicative of AOM, giving a precise visual representation of what the affected regions are. By highlighting these regions, the model not only provides a classification result but also useful spatial information that contributes to interpretability.

This is a useful feature for the clinician, as it serves to visually define the region of interest and to understand the basis of the diagnosis. This outcome also identifies the dual-sided capability of the system used, demonstrating not only its potential to classify an image as being inclusive of evidence for Acute Otitis Media (AOM) but also its capability to accurately delineate and indicate actual locations of abnormalities. This double-sided feature is particularly relevant to medical image analysis because it ensures maximum accuracy as well as interpretability in predictions made by the model.

Through localizing and classifying, the system provides verification that is essential to clinicians' verification, allowing them to reverify and authenticate its diagnosis. It is especially essential in clinical decision-making, where reliability and trust on auto-diagnosis is a top priority. The ability of the model output to visually cross-check maintains the predictions from being blind labels but backed by real localized data, preventing the possibilities for misdiagnosis and instilling physician confidence into AI-given treatments.

Furthermore, this improved ability is valuable in self-diagnosis applications because it provides customers with better perceptions of their disease. Rather than the open two-way decision, the patients are able to observe and acknowledge where the areas of difference existed, which improves the decision process regarding what next in medical attention. More generally, the combination of classification and localization significantly contributes to the potential of the system as an even more precise diagnostic aid to be employed in medical settings. By providing more subtle and interpretable analysis of pathological characteristics, the model enables more precise diagnoses, facilitates better patient outcomes, and encourages greater use of AI-based diagnostic tools in modern medicine.

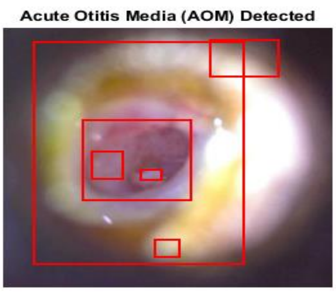

Figure 4: *Proposed model's final predicted output 1*

Figure 5 illustrates the outcome of the suggested automated system in identifying acute otitis media (AOM) from otoscopic images. The given image is labeled as a "Normal Ear," which indicates that the system was successful in identifying the lack of infection or pathological conditions. The outcome verifies the model to identify normal and infected ear images accurately. This proper labeling of this image as normal once again confirms the effectiveness of the deep learning model in medical image interpretation. Proper labeling of normal ear anatomy avoids false positives, which is an important step to guarantee diagnostic accuracy. Differential diagnosis between AOM ears and normal ears avoids inappropriate treatment, reduces the risk of misdiagnosis, and enhances patient outcome.

This result also indicates the insensitivity of the automated detection algorithm, again verifying its suitability for use in clinical environments. In constantly providing accurate classification results, the model can be used as an assisting aid device for physicians, helping in the effective and accurate otoscopic image examination.

Figure 5: *Proposed model's final predicted output 2*

The effectiveness of the model developed in AOM detection is revealed in the confusion matrix presented in Figure 6. The confusion matrix provides a close-up of the model's classification result, indicating that it discriminates the AOM and normal ear images effectively. The matrix indicates that 595 of the 600 AOM images were classified as True Positives by the model. That is, the images were correctly identified as showing infection. 5 of the AOM images were incorrectly classified to the normal category, and hence became False Negatives. Such a 43isclassification may be because of subtle variations among the infected images or challenging samples where pathological appearances were ambiguous.

Similarly, 197 out of 200 normal ear images were accurately assigned as True Negatives by the model, marking its performance in predicting healthy ear anatomies. However, 3 normal images were falsely assigned as AOM, producing False Positives. These errors are due to optical similarity of changes in normal ear anatomy and early infection presentations. The general accuracy of the model as represented by the confusion matrix is its capacity for accurate classification with low misclassification errors. The high counts of True Positives and True Negatives show its strength, which are the proportionately low counts of False Positive and False Negative show the balance of sensitivity to specificity of the model

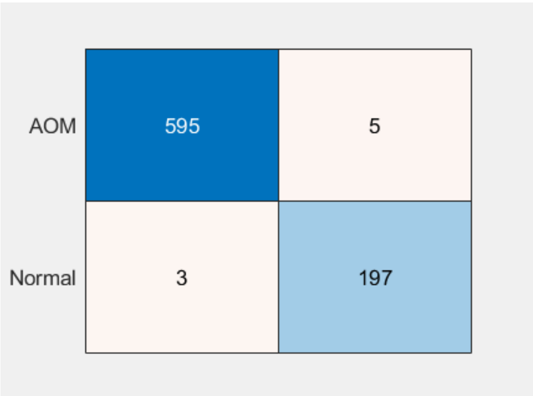

Figure 6: *Proposed model's Confusion matrix*

4. Conclusion

A sophisticated method was proposed to identify AOM based on an Improved Mask R-CNN (IR-CNN) model. Some important steps were followed in the process. The ear image input was first preprocessed in which resizing, color conversion and filtering of noise was performed. To achieve efficient detection and segmentation, an Improved Mask R-CNN with ResNet-50 was employed. Apart from that, feature extraction from ear drum images was done using an Improved Mask R-CNN model with ResNet-101 backbone. Classification was done using a three-layer fully connected neural network. For testing the robustness and generalization capability of the proposed classification model, K-Fold Cross Validation was employed. The proposed system has attained 99.5% accuracy, 98.92% precision and an F-score of 98.25%. These findings describe the performance and stability of the model in reading clinical images, beneficial in the early diagnosis and enhanced decision-making in otological examination.

5. Futurescope

Future research will be focused on expanding the dataset of the eardrum to improve multi-class classification for the diagnosis of otitis media so that more precise discrimination among various middle ear diseases is possible. Future research can also be on the path of feature fusion-based deep learning architectures to further enhance classification performance with different feature extraction techniques. Integration of the subject model with current imaging modalities, i.e., Optical Coherence Tomography (OCT) or ultrasound, can potentially revolutionize diagnostics by providing more information on middle ear anatomy. These will enable more precise and detailed diagnostics, which will improve clinical decision-making and patient management.

Reference

1. M. Abdelrahman, S.Mahmoud, Year: 2023, "Evaluating medical image analysis models using K-fold cross-validation techniques", Artificial Intelligence in Medicine, Vol:138, pp.102423.

2. A. Ahmed, M. Rehman, Year: 2023, "Deep learning architectures for eardrum image classification and analysis", Health Informatics Journal, Vol: 29, No: 1, pp.32–41.

3. A. Anand, M. Singh, Year: 2022, "Automated segmentation of medical images using deep learning techniques: A survey", Journal of King Saud

University – Computer and Information Sciences, Vol: 34, No: 9, pp.5853–5865.

4. R. Banerjee, A. Chakraborty,Year: 2021, "Real-time deep learning models for clinical otoscopic image processing", Medical & Biological Engineering & Computing, Vol: 59, No: 11, pp.2517–2527.

5. S. Bhattacharya, R. Dutta, Year: 2023, "ResNet-based feature extraction for improved otoscopic image analysis", Computer Methods and Programs in Biomedicine, Vol: 235, pp.107438.

6. A. Chauhan, P. Sharma, Year: 2021, "A comparative evaluation of deep learning algorithms for ear disease diagnosis", Informatics in Medicine Unlocked, Vol: 25, pp: 100677.

7. Y. Chen, H. Lee, Year: 2022, "Improved CNN architectures for contour-aware medical image segmentation", Journal of Digital Imaging, Vol: 35, No:1, pp.82–90.

8. P. Das, S. Ghosh, Year: 2020, "Deep learning-based enhanced visibility of tympanic structures in otoscopic images", Healthcare Technology Letters, Vol: 7, No: 3, pp.85–91.

9. R. Fernando, T. Silva, Year: 2023, "Cloud-based deployment of optimized Mask R-CNN models for remote clinical diagnostics", Journal of Cloud Computing, Vol: 12, No:1, pp.1–12.

10. A. Gupta, K. Mehta, Year: 2023, "Data augmentation techniques to enhance medical image segmentation model robustness", Journal of Healthcare Engineering, Article ID 1346287.

11. L. Huang, Y. Lin, Year: 2022, "Application of deep learning-based segmentation models in tympanic membrane abnormality detection", Journal of Imaging, Vol: 8, No.5, pp.125–132.

12. D. Jain, S. Dey, Year: 2021, "Model generalization strategies for instance segmentation in medical imaging", Journal of Medical Imaging and Health Informatics, Vol: 11, No: 4, pp.987–994.

13. D. Jobanputra, R. Patel, Year: 2023, "Classical approaches and new deep learning trends for diagnosing ear disease from otoscopic images", Journal of Biomedical Informatics, Vol: 137, pp.104246.

14. S. Kim, J. Park, Yrar: 2020, "Deep learning for otitis media diagnosis using high-resolution otoscopic images", Medical Image Analysis, Vol: 64, pp.101741.

15. N. Kumar, V. Sinha, Year: 2019, "Performance analysis of backbone networks for biomedical segmentation using ResNet and DenseNet", Procedia Computer Science, Vol: 167, pp.1772–1780.

16. H. Li, X. Zhao, Year: 2023, "Self-supervised learning for medical image segmentation: A review", IEEE Transactions on Medical Imaging, Vol: 42, No: 3, pp.712–725.

17. Q. Liu, H.Wang, Year: 2023, "Optimizing lightweight neural networks for real-time otoscopic image classification", Computers in Biology and Medicine, Vol: 162, pp.107289.

18. Z. Tan, M. Yu, year: 2022, "Ensemble deep learning approaches for robust detection of ear pathologies", Biomedical Signal Processing and Control, Vol: 77, pp.103735.

19. A.Verma, P. Roy, Year: 2023, "Attention-guided CNN models for enhanced medical image classification accuracy", Expert Systems with Applications, Vol: 213, pp.118897.

20. J. Wu, Y. Zhang, Year: 2022, "Hybrid deep learning strategies for tympanic membrane disease detection", Journal of Medical Systems, Vol: 46, No: 6, pp.56.

21. R.Yadav, S. Kumar, Year: 2021, "Comparative performance of deep learning models in diagnosing ear diseases from otoscopic images", Computers in Electrical and Biomedical Engineering, Vol: 85, pp.106693.

22. F. Zhao, M. Chen, Year: 2022, “Multi-view image analysis techniques for improved clinical diagnosis in otology”, IEEE Access, Vol: 10, pp.45032–45042.

23. R. Singh, A. Thomas, Year: 2023, “Clinical deployment of deep learning tools in remote otoscopic diagnostics”, Telemedicine and e-Health, Vol: 29, No:4, pp.345–354.

24. L. Oliveira, D. Costa, Year: 2023, “Benchmarking medical image classification models for ear disease diagnostics”, Computers in Medicine and Biology, Vol: 152, pp.106329.

25. K. Park, Y. Choi, Year: 2020, “Image quality enhancement in medical otoscopic images using GANs”, Computer Methods and Programs in Biomedicine, Vol: 189, pp.105347.

26. D. Rana, A. Malik, Year: 2021, “EfficientNet-based classification of tympanic membrane abnormalities”, Informatics in Medicine Unlocked, Vol: 24, pp.100589.

27. J. Seo, D. Kim, Year: 2022, “Explainable AI for otoscopic image interpretation in clinical decision support”, Artificial Intelligence in Medicine, Vol: 126, pp.102133.

28. A. Thakur, S. Rawat, Year: 2023, “Temporal convolutional networks for automated detection of AOM from ear videos”, Journal of Imaging, Vol: 9, No: 2, pp.58–65.

29. M. Upadhyay, H. Bhatt, Year: 2023, “Comparative analysis of segmentation networks for tympanic membrane structure extraction”, Biomedical Engineering Letters, Vol: 13, No: 1, pp.112–120.

30. R. Velu, P. Arora, year: 2022, “Implementation of capsule networks in tympanic abnormality detection”, Journal of Biomedical Informatics, Vol: 134, pp.104204.