A Novel Framework for Medical Learning: Using AI Based Grad-CAM for Improving Otitis Media Diagnosis

A Novel Framework for Medical Learning: Using AI Based Grad-CAM for Improving Otitis Media Diagnosis

Andres Godoy

A Thesis in the Field of Biotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2023

Abstract

Otitis Media (OM) and its sub-categories of pathology are the number one pathology in children. Diagnosis is very difficult as it requires visual inspection the tympanic membrane of a child, which is in view for only a few seconds during a clinical exam. Improving diagnosis requires the transfer of visual insights which is a complex learning and training task.

A validated method to understand visual task insights has been to use eye-tracking as a surrogate for neural attention. Eye tracking data can be represented in the form of a heat-map or a visual saliency map. Considering the power and benefits of using state-of-the-art Machine Learning techniques in diagnosing visual pathology, our purpose is to derive a heat-map from a Machine Learning algorithm that acts as an "expert", and to provide these heat-maps for medical students with the final aim of understanding if this improves medical learning, specifically for OM.

Our results indicate a significant improvement in diagnostic performance when showing medical students heat-maps derived from machine learning models, in conjunction to traditional teaching tutorials when compared to a control group not exposed to the heat-maps. This research provides a simple, cost-effective proof-of-concept framework to enhance the diagnostic accuracy and training speed for medical student as well as contribute in bridging the disparity gap in diagnostic accuracy of otitis media amongst practitioners.

Dedication

To my daughter Olivia. Stay hungry. Stay foolish.

Acknowledgments

I would like to thank Dr. Marcos Goycoolea as well as Dr. Maria José Herrera at Universidad de los Andes, for their expert opinion, and help. I would also like to thank Dr. Jose Godoy for his expert revision as well as Dr. Masha Hareli for her time and contribution for this study.

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

Otitis media (OM) encompasses various inflammatory processes, including acute otitis media (AOM) and otitis media with effusion (OME) in the middle ear, all with different clinical manifestations. It is a relevant pediatric disease as it represents the number-one infection in children (Teele et al., 1989) (Schappert, 1992), with an estimated 80% of children experiencing one episode or more before the age of three (Auinger et al., 2003). OM contributes to over 6 to 10 million clinic visits and is the primary indication for the use of antimicrobials in children (Suaya et al., 2018) (Nelson et al., 1987) (Owings et al., 1998).

When undertreated, the morbidity of OM represents the leading cause of hearing loss and surgery in children, associated with billions of dollars annually in costs. Its overtreatment is equally problematic as misuse of antibiotics is a leading cause of increased resistance and treatment failure.

Accurate and consistent diagnosis has been problematic. Its error and over-diagnosis, based on symptoms and signs, is estimated at least to be one-third.

With the advent of artificial intelligence, new machine-learning algorithms have been proposed as a diagnostic alternative to a physician in an effort to increase diagnostic consistency and accuracy.

Several research teams have successfully applied various machine learning techniques to diagnose otitis media (Kuruvilla et al., 2013) (Livingstone et al., 2020)

(Monroy et al., 2019) (Shie et al., 2014) (Tran et al., 2018) (Wang et al., 2020) (Viscaino et al., 2020).

Recently, a new clinical framework for machine learning algorithms has emerged, that of computer aided diagnosis (CAD), in which both physician and the algorithm participate in the clinical diagnosis. When evaluating a patient, the physician is presented with the algorithms' prediction and its respective probability. The physician then makes a diagnostic decision with these additional data points.

Several important limitations remain to the widespread utilization of machine learning algorithms regardless of their application method. One limitation is that machine-learning-based predictions are considered a "black-box" solution. The "black-box" term refers to the inability to determine the motives to which an algorithm arrived at a particular prediction. This is often referred to as model explainability or interpretability.

Several strategies have been proposed to solve this issue. A widely accepted solution is to utilize visual saliency maps or *heat-maps* which reflect the specific visual weights in terms of what the algorithm focuses on each time it makes a diagnostic decision. One of the most utilized techniques in this regard is known as Grad-CAM.

Obtaining a Grad-CAM after a machine learning model has been created is now a standard procedure, which resolves in part the model interpretability concerns commented on earlier. We now understand the value of deriving *heat-maps* from a model once it converges into a successful algorithm. Evaluating these *heat-maps* provides trust in the specific areas of interests of the algorithm, which helps in avoiding algorithmic bias, especially when the data used to generate the model does not represent a homogenous population. The algorithm can then be deployed as a binary outcome in conjunction with a physician, creating a successful computer aided diagnostic framework (CAD).

The essence of the CAD method is for the algorithm to be deployed while the physician is providing the diagnosis. The physician ultimately decides the outcome but is provided with an algorithmic *hint* in terms of what the algorithm *thinks* is the correct diagnosis. In this context, a limitation of the CAD method is the dependance on the algorithm during a clinical visit. This dependance translates into the need for sophisticated computer hardware to be at the clinical site, or in the case of a *cloud* deployment method, for the clinic to be tethered to a high-speed internet access which would stream the visual findings to the *cloud* and obtain a predictive outcome in return.

These limitations are considerable, more so when we consider the role of AI which is to provide enhanced medical diagnostics at scale; these limitations could increase healthcare inequalities rather than improve them.

We could solve the deployment issues related to the CAD method by simply using the algorithm to *train* a physician, and then a physician could be deployed at scale without any of the above limitations. We know that when we examine the CAD method in the context of medical training, the data reveals an increased performance for the physician. Therefore, a critical question arises. Can this performance be maintained in the absence of the algorithm? If we could use an algorithm to train a physician, we could not only increase a physicians' diagnostic ability, but perhaps more importantly, we could solve the deployment issues related to the CAD method and thus improve all healthcare outcomes without the tethered technological dependence.

We have seen therefore the change in the clinical deployment strategies for algorithms. Initially, algorithms were deployed to replace physicians, while now they are being deployed to work side-by-side with physicians. The final ideal step would be to use algorithms to train physicians.

Our interest therefore converges into how to effectively transfer the insights of the algorithm to a physician, in the hopes of improving their own clinical outcomes. Current CAD methods utilize binary predictions and considering current experimental setups, it is difficult to attribute the increase in diagnostic performance to an actual improvement in clinical understanding or learning. The increase performance could be intuitively explained by a clinician revising or double-checking their initial diagnosis when a mismatch with the model occurs.

Considering the use of an algorithms predictions is a black-box model, we can instead use the model interpretability framework, that of deriving a heat-map, to not only evaluate the model when it is being created, but potentially as the key method to transfer an algorithms insight to a physician.

The reason behind why a heat-map could be used as the key resource to transfer information relates to how we process visual information and hence neural attention.

Although our brain *sees* an image in full, our sensory retina can obtain high quality input only via a particular region of the retina called the *fovea*. Therefore, our eyes *foveate* to capture a scene in high definition. The critical component in this process is the order in which our brains determine how to capture a visual scene; this order is not arbitrary. In the context of uncertainty and limited time, our brains instruct our eyes to first capture the most relevant aspects of a visual scene.

The movements of these focal points result in a scan path or attention map. As such, eye movements are a proxy of neural attention processing. This neural attention process can vary and be optimized based on the visual task at hand.

For a visual diagnostic task, our brains must be able to guide our eyes towards the areas of interest that represent features and targets based on general medical knowledge and our prior visual experience. Thus, how we visually scan an image, meaning which aspects we fix our gaze on, reveal our understanding of where we think we should seek the relevant information for the task at hand.

This visual scan path can be visually represented in multiple ways, such as by blurring out unimported regions of the image, or as in this study, by providing a gradient-color based *heat-map.* Importantly, the scan-path we will provide in this experiment is derived not from actual human experts but from the machine learning model.

We hypothesized that by viewing these *heat-map*, a non-expert can effectively understand, transfer, and ultimately acquire new diagnostic experience in contrast to traditional learning strategies based on transfer of insights via text.

To our knowledge, these *heat-maps* have never been used to understand their impact and potential value in the context of medical training, specifically in the context of OM.

Finally, this research is significant because not only can it provide a potential proof-of-concept solution to aid in the diagnostic disparity for otitis media amongst practitioners, but it can generalize for other diagnostic entities which are now being evaluated through the lens of machine learning and CAD strategies.

Furthermore, any scenario that enhances algorithmic model explainability, ultimately aids in decreasing bias. In an era of increasing reliance on algorithmic diagnostics, it is a moral and ethical duty to avoid compounding existing societal inequities through algorithms, but rather to utilize these fairly so that they can benefit a homogeneous population.

Chapter II.

Definition of Terms

Otitis media (OM): Any inflammatory process of the middle ear.

Machine learning: The use and or development of computer systems that are able to learn and adapt without following human derived instructions, thus obtaining insights from patterns and data.

Computer aided diagnosis (CAD): Computational systems that assist a physician in the interpretation of a medical image. Typically, they are deployed in an existing clinical framework in which a machine-learning model provides a predictive output which the clinician can incorporate into their decision matrix.

Model interpretability: The process through which a researcher can understand and interpret predictions made by a machine learning model.

Grad-CAM: It is a *heat-map* derived from a machine-learning model. More specifically, it is a gradient-weighted activation map (Grad-CAM) of a machine-learning model. It can be the equivalent of a saliency map which reflect the weighted importance of each area of interest within a particular image for the given diagnostic model.

Chapter III.

Research Methods

Participants

We obtained Harvard IRB approval on September 8th to conduct our questionnaire-based study. Data collection occurred at Universidad de los Andes and at Universidad de Chile during September 2022. A group of 93 randomly selected medical students in their last year of school were chosen and 5 were excluded due to incomplete questionnaire data entry.

Students were randomly divided into two groups of 44 participants each, with an equal female-to-male ratio. Individuals were excluded if they reported clinical experience in otolaryngology, such as a clinical clerkship or rotation. Clinical experience was an excluding factor in order for the experiment to capture the performance change in a novice when exposed to the heat-maps derived from a state-of-the-art diagnostic algorithm.

Instrument

In preparation of the experiment, we utilized existing public datasets of ear pathology and normal otoscopic images to derive a state-of-the-art machine learning algorithm for diagnosing OM. We then generated a Grad-CAM visual saliency map for this algorithm. We inspected the visual Grad-CAM map so that it would reflect well-known anatomical points that are significant for detecting otitis media. These Grad-CAM heat maps were also validated by experts.

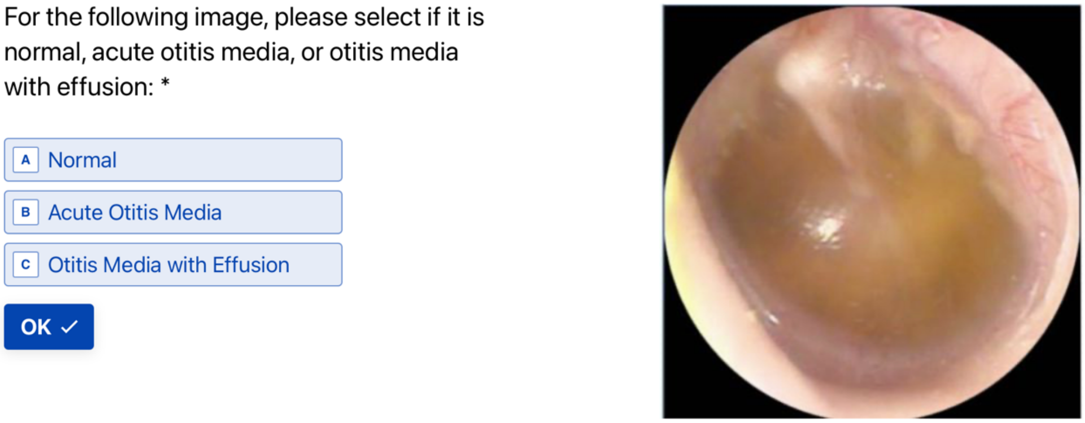
In terms of the experimental setup, thirty photographs representing otoscopies were randomly selected from an existing public clinical archive of ear pathology using search terms TM, OME, and AOM. The images were then randomly selected for inclusion using a random sampling without replacement approach. These thirty photographs were subdivided into two sets of fifteen photographs.

Each photographic set was used to construct a visual diagnostic questionnaire. The two visual diagnostic questionnaires contained different photographs in order to avoid re-sampling bias but were symmetrical in terms of the diagnostic categories. As such, the questionnaires were composed of five photographs of a normal otoscopy, five corresponding to acute otitis media, and five corresponding to otitis media with effusion. For each photograph presented, subjects were asked to mark the photograph either as normal, acute otitis media or otitis media with effusion. Each correct answer was counted as one point, and all 15 questions were summed to provide a final score which we then turned into a percentage for further analysis. Correct answers were provided by experts.

Procedure

After recruitment, we obtained informed consent as approved by the IRB. Prior to beginning the study, subjects were placed into two groups. Placement was conducted by a random computer algorithm. Group 1 represents the control group and Group 2 the experimental group. Both groups first answered a brief self-reporting knowledge questionnaire to ensure they met our inclusion criteria.

Both groups began the first set of the visual questionnaire (Figure 1 – Page 10). The results of this questionnaire were tabulated into our pre-intervention results. This allowed us to have an understanding of diagnostic accuracy prior to our intervention.



For the following image, please select if it is normal, acute otitis media, or otitis media with effusion: *

A Normal

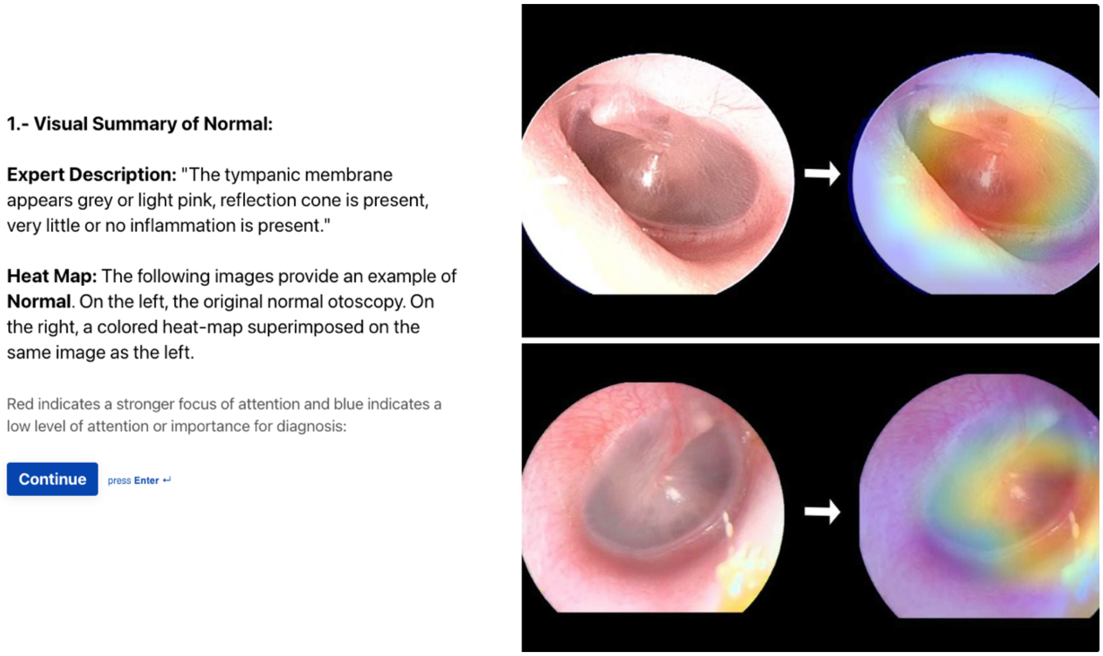
B Acute Otitis Media

C Otitis Media with Effusion

Figure 1. Example of a test question.

After the first questionnaire, both groups continued into a self-paced instructional tutorial on middle ear pathology. This instructional tutorial was vetted by an expert panel and was constructed based on a traditional teaching framework.

After the instructional tutorial, Group 2 continued into the experimental intervention. We showed Group 2 participants photographs of normal, acute otitis media and otitis media with effusion alongside corresponding “attention” maps in the form of heat maps (Figures 2-4 – Pages 11-13).



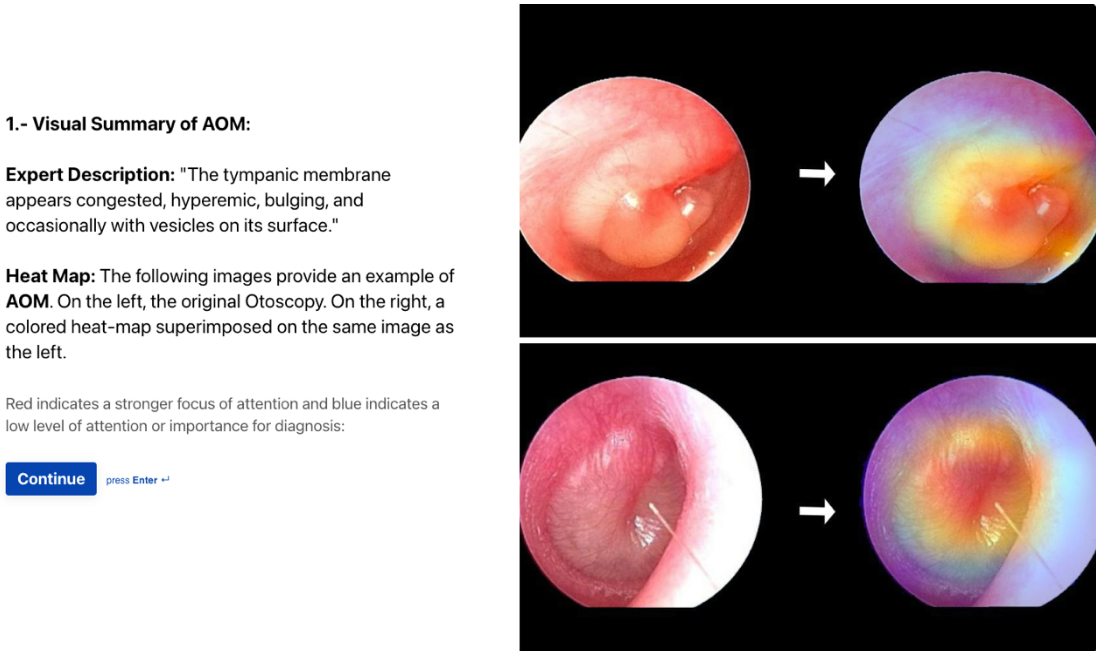
1.- Visual Summary of Normal:

Expert Description: "The tympanic membrane appears grey or light pink, reflection cone is present, very little or no inflammation is present."

Heat Map: The following images provide an example of **Normal**. On the left, the original normal otoscopy. On the right, a colored heat-map superimposed on the same image as the left.

Red indicates a stronger focus of attention and blue indicates a low level of attention or importance for diagnosis:

Figure 2. Example of a normal otoscopy and derived Grad-Cam Heat-Map.



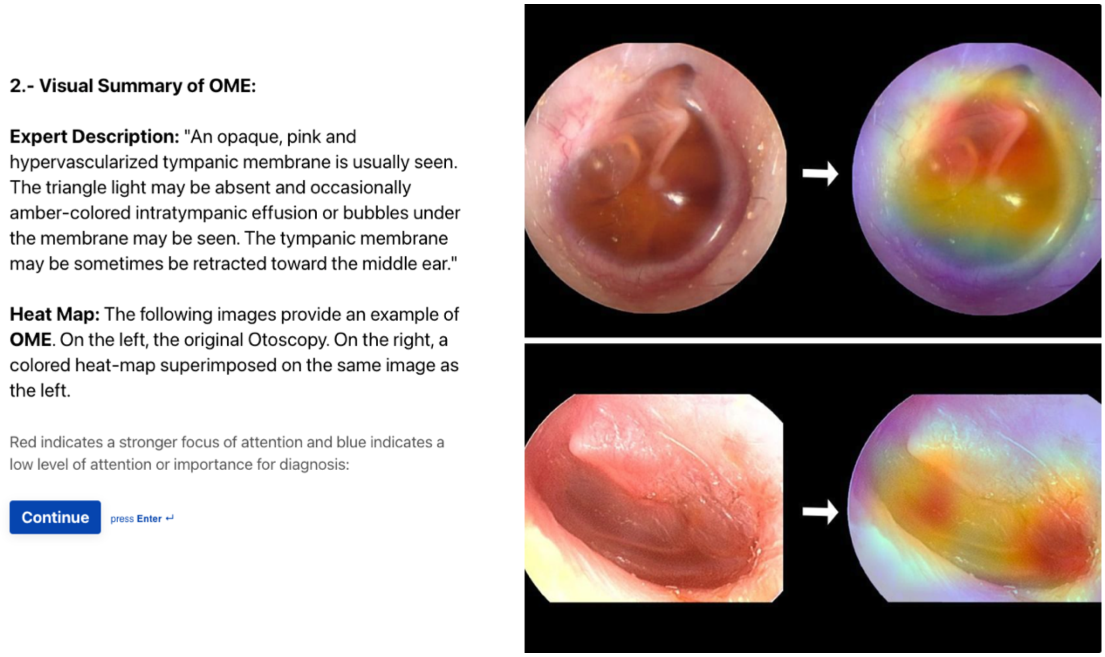
1.- Visual Summary of AOM:

Expert Description: "The tympanic membrane appears congested, hyperemic, bulging, and occasionally with vesicles on its surface."

Heat Map: The following images provide an example of **AOM**. On the left, the original Otoscopy. On the right, a colored heat-map superimposed on the same image as the left.

Red indicates a stronger focus of attention and blue indicates a low level of attention or importance for diagnosis:

Figure 3. Example of a AOM otoscopy and derived Grad-Cam Heat-Map.



2.- Visual Summary of OME:

Expert Description: "An opaque, pink and hypervascularized tympanic membrane is usually seen. The triangle light may be absent and occasionally amber-colored intratympanic effusion or bubbles under the membrane may be seen. The tympanic membrane may be sometimes be retracted toward the middle ear."

Heat Map: The following images provide an example of **OME**. On the left, the original Otoscopy. On the right, a colored heat-map superimposed on the same image as the left.

Red indicates a stronger focus of attention and blue indicates a low level of attention or importance for diagnosis:

Figure 4. Example of a OME otoscopy and derived Grad-Cam Heat-Map.

Both groups finalized with the second set of the photographic survey, and we tabulated these as our post interventional results, maintaining each groups' data separate.

Data analysis

The data was maintained in an Excel spreadsheet. Statistical analysis was performed using Stata. Our survey results were correspondingly tabulated and listed based on pre (questionnaire 1) and post score (questionnaire) results as well as on group status. We evaluated our data for normal distribution and then followed up with unpaired t-tests to compare the means between both groups, pre-exposure and post-exposure in terms of diagnostic accuracy post intervention. Then, we continued with paired t-test to compare the means within each group, pre-exposure and post-exposure.

Chapter IV.

Results

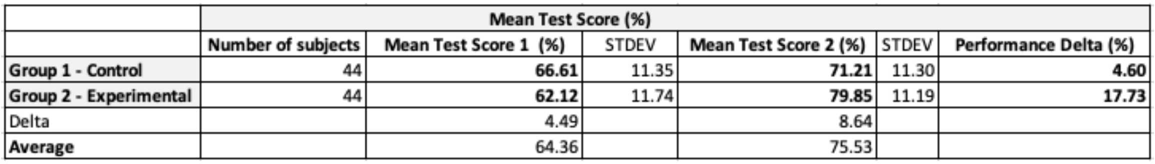
In total, 88 questionnaires were successfully completed across both groups. Each group was comprised of 44 medical students, with an equal proportion of male-to-female ratio. All medical students were in their last year of medical school. Average survey response time was of 6 minutes and 57 seconds. Mean test scores were tabulated and presented as percentage ranging from 0 to 100.

All participants prior to the experiment graded their experience and knowledge on OM. When asked if they knew and understood OM, 100 percent of participants responded yes.

Participants were then asked to rate their own level of knowledge from 1-to-10, 10 being very confident and knowledgeable about the pathology, and 1 having no information about the pathology. Average self-assessment rating was of 7-out-of 10. If we analyze these assessments by group, the control group average was 7.2 and the experimental group average was of 6.8.

In terms of test scores, both groups averaged 64.36 percent pre-exposure. The average post exposure (questionnaire two) for both groups was of 75.53 percent.

Table 1. Test Scores by Group.



	Mean Test Score (%)					
	Number of subjects	**Mean Test Score 1 (%)**	STDEV	**Mean Test Score 2 (%)**	STDEV	**Performance Delta (%)**
Group 1 - Control	44	**66.61**	11.35	**71.21**	11.30	**4.60**
Group 2 - Experimental	44	**62.12**	11.74	**79.85**	11.19	**17.73**
Delta		4.49		8.64		
Average		64.36		75.53		

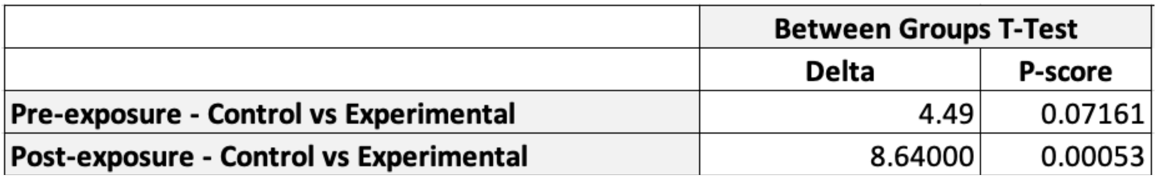
Specifically, for the control group, questionnaire one was 66.61 percent followed by 71.21 for questionnaire two. For the experimental group, questionnaire one was 62.12 percent followed by 79.85 percent for questionnaire two (Table 1 – Page 14).

In terms of assessing the individual pathological entities, across both groups correct diagnosis for "Normal" was 73.3 percent, for "AOM" correct diagnosis was 83.3 percent while for "OME", correct diagnosis was 46.6 percent.

We assessed for normal distribution and then proceeded to evaluate the differences in performance. To test our hypothesis our interest was to determine if the described increase in performance between each group was statistically significant. For this, we conducted an un-paired t-test of two tails, with equal variance, to compare the difference between each group.

First, when we compare the control group with the experimental group for questionnaire one (pre-exposure), the difference of 4.49 percent was found too not be statistically significant. (p = 0.07) at an alpha of 0.05. This is important because it reflects equal knowledge level for both groups.

Table 2. T-Tests Between Groups.



	Between Groups T-Test	
	Delta	**P-score**
Pre-exposure - Control vs Experimental	4.49	0.07161
Post-exposure - Control vs Experimental	8.64000	0.00053

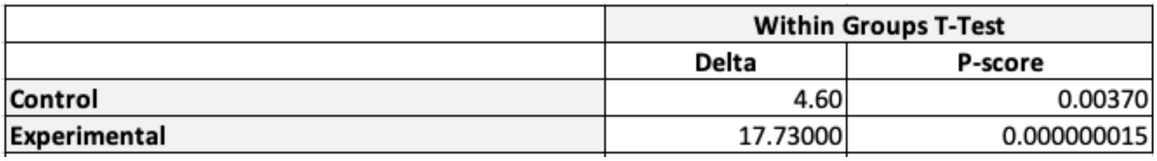
Second, when we compare the mean test scores for questionnaire two (post-exposure) between both groups, the difference of 8.64 was found to be statistically significant (p = 0.0005) at an alpha of 0.05.

This was the critical analysis of the experiment because we expected both groups to increase their performance for questionnaire two, but importantly, we expected a higher increase for those exposed to the experimental method being proposed. The difference is highly statistically significant, hence the group exposed to the novel learning technique increased their performance significantly more than the control group.

We then analyzed the differences within each group. For this, we conducted a paired t-test of two tails, with equal variance, to compare the difference within each group.

For the control group, the performance increase from 66.61 percent to 71.21 percent of 4.60 percentage points was statistically significant (p = 0.0037). For the experimental group, the performance increase was from 62.12 percent to 79.85 percent, resulting in a delta increase of 17.73 percent. This difference was found to be statistically significant (p = 0.000000014).

Table 3. T-Tests Within Groups.



	Within Groups T-Test	
	Delta	P-score
Control	4.60	0.00370
Experimental	17.73000	0.000000015

The statistically significant improvement within each group was expected. Both groups were presented with a standard online tutorial explaining the basics of otoscopy diagnosis and pathology. As such, our interest was not the direction of the improvement per se, but rather the magnitude of the improvement. This magnitude, when compared between the groups was also found to be statistically significant.

Chapter V.

Discussion

Otitis Media

The ear as a sensory organ can be subdivided into outer, middle, and interior portions. The external ear forms the external ear canal, and the TM divides the external ear from the middle ear portion. The TM can be described as a thin membrane of connective tissue covered by the skin on the exterior surface and mucosa on the interior surface. The TMs' primary function is to transfer sound waves to the middle ear, which can be amplified and modulated before passing to the inner ear.

Any inflammatory process of the middle ear can be described as OM. The nature of the inflammation and its acuteness can give way to further sub-types of OM, such as otitis media with effusion (OME), adhesive OM, chronic OM, or acute OM (AOM) to name a few.

In a seven-year follow-up study conducted by Teele et al. (1989), they estimated that by one year of age, 62% of the children had greater or equal to one episode of acute otitis media, and 17% had greater than or equal to 3 episodes. By three years of age, 83% had greater than or equal to one episode of AOM. The incidence of OM depends in part on climate and other factors such as pollution status and population vaccination level.

Today, the morbidity associated with OM has decreased due to increased medical awareness, vaccination, and widespread adoption of antibiotics. However, given the magnitude and risks associated with both under and overdiagnosis of the disease, it remains a significant public issue.

Clinical Diagnosis of Otitis Media and Challenges

Considering that the middle ear is not feasible to directly observe, most diagnostic criteria derive from visually inspecting a child's TM, as a clinical proxy of the middle ear, and inferring OMs' status and diagnostic sub-type.

More specifically, a clinical examination involves four distinct aspects regarding the tympanic membrane: color, position, mobility, and perforation status. Color should be typically described as pale gray, while an opaque or blue TM can represent middle ear effusion.

In OME, the position should be either retracted or neutral, while in AOM, the TM is frequently bulging. Mobility requires a pneumatic otoscopy assessment of the TM. If the mobility of the TM is impaired, this is more consistent with OME. Finally, perforations indicate chronic middle ear pathology regardless of infection and inflammation status.

The most basic examination is an otoscopy to visually inspect the TM, in which color, position, and perforation can be evaluated. An otoscopy can also be evaluated with tympanometry which measures the changes in the acoustic impedance of the TM to changes in air pressure. Finally, acoustic reflectometry can be done, which measures the reflected sound from the TM. The more sound is reflected, the greater the chance of a middle ear effusion. All these specific visual and tympano-metric data points can be accompanied by general signs and symptoms of an infection such as headache, fever, irritability, and loss of appetite, as well as otalgia and otorrhea. These visual signs and exams are then integrated with the patients’ symptoms and medical history to provide a diagnosis.

The difficulty in assessing the tympanic membrane status is due to a physical limitation in observing it, specifically in the context of a sick child in which observation can lasts a few seconds at best. Also, several different medical experts typically establish this diagnostic entity. Each one of these has varying degrees of expertise, specifically in terms of ear pathology.

Consistent diagnostic accuracy does not surpass 70% across the wide variety of medical professionals who must interphase with this diagnostic entity, including but not limited to primary care, pediatric, emergency, and otolaryngology specialists (Pichichero et al., 2001) (Pichichero et al., 2002) (Jones et al., 2003) (Young et al., 2009) (Rosenfeld et al., 2002) (Wu et al., 2021).

Bluestone et al. (1979) and colleagues assessed the sensitivity and specificity of OM diagnosis between an otolaryngologist and pediatricians. Otolaryngologists have a crucial learning advantage in that they routinely practice a surgical procedure called "myringotomy," in which the tympanic membrane is surgically incised, and the content of the middle ear is observed. This procedure is considered the gold standard in diagnosing otitis media.

The researchers speculated that the visual link between viewing the tympanic membrane and the excising of it to view the contents of the middle ear played an important visual learning feedback loop. They found that the diagnostic specificity of two pediatricians were 25% and 64%, in contrast to 90% of an otolaryngologist. Naturally, this was a small observational study.

There are also, as expected, differences between different physician specialties as well as within each group in terms of the respective training experiences. Aronzon et al.

(2004) compared differing experience levels between different medical specialties in diagnosing OM pathology against the gold standard of surgically obtaining a fluid sample of the middle ear. The groups were medical students, internal medicine residents, junior resident otolaryngologists, senior resident otolaryngologists, attending otolaryngologists and finally a group that provided diagnosis under microscopic evaluation prior to the surgical procedure.

When considering either tympanogram, images, or a combination of both, sensitivity was similar between otolaryngologist residents and attendings. The specificity between the groups was statistically different. Notably, specificity improved significantly with seniority in all cases. Also, less trained physicians were more likely to over-diagnose.

We must also consider that an otolaryngologist does not evaluate the majority of OM diagnostic cases. The consequence of non-expert consultations results in a significant health care burden, either due to excess referrals and patient assessments, or due to overdiagnosis and the unwarranted use of antibiotics.

Machine Learning Algorithms for Diagnosis of Acute Otitis Media

Machine learning algorithms aimed at improving diagnosis have exploded in the past five years. These technological breakthroughs can be explained due to the intersection of several advancements. On one end, increased computational power and reduced cost have allowed anyone to run state-of-the-art machine-learning techniques.

Machine learning is heavily reliant on the data set provided in order to provide good results. The increasing use of imaging capturing hardware and sensors has resulted in an explosion of the available data. Finally, the machine-learning ecosystem is open, and new advancements quickly permeate researchers, enabling quick iterative progress of the methods and techniques.

These machine learning algorithms are typically designed as predictive models without an expert participating in their feature design, in contrast to traditional and prior generation algorithms based on expert feature design and correlation frameworks. As such, current machine-learning algorithms are relatively simple to generate, and their power relies on the quality and the quantity of the imaging data to which they are exposed. After a model is generated, it can be tested and further perfected on untrained data.

Deep learning is a subset of machine learning in which the learning and training process occurs in a sequence of consecutive multi-layer neural networks or nodes (modeled after our brain).

In the case of Otitis Media, hundreds of algorithms have been proposed to date. The most advanced state-of-the-art algorithms can accurately diagnose over 90% of new unseen cases. Livingstone et al. (2020) published an algorithm for OM that displayed accuracy of 88.7%. In contrast, the mean physician accuracy in this study was significantly lower (58.9%). Work conducted by Viscaino et al. (2020) using a different machine learning technique achieved a mean accuracy of 94% using four distinct diagnostic sub-entities for middle ear pathology.

Tsutsumi et al. (2021) published a novel neural network in which several pathological entities for middle ear were included, that of, AOM, Normal, Chronic Suppurative Otitis Media, Cerumen as well as Otitis Externa. They applied the technique of transfer learning yielding a model with robust results. The overall AUC-ROC published was of 0.91 and the AUC-ROC for each category was of 0.85 for Normal, 0.89 for AOM, 0.79 for Chronic Suppurative Otitis Media, 0.97 for Cerumen and 0.98 for Otitis Externa.

Transfer learning is a specific technique in which a machine learning model can be generated using a fraction of the data normally required to build a robust ML algorithm. Typically, thousands or millions of images per category are required to train a machine learning model, so that it can correctly adapt its weights, learn, and ultimately discriminate between the categories effectively.

In the context of creating a novel algorithm, creating a visual otoscopic database of millions of images each representing unique categories is very difficult to achieve if not impossible. Fortunately, the transfer learning technique was created to solve this problem.

In transfer learning, a new algorithm can be trained using the specified weights of an existing algorithm. It is common to use algorithms that have already been trained in visual categorization tasks and which are highly effective. Some of the most popular algorithms for this are ResNet, MobileNet, VGG and Inception. By starting out with the existing weights of these models, the system is already highly skilled at visual categorization. When presented with the new data, in this case otoscopic images and new categories, it can be further fine-tuned for this task.

In a variation of traditional studies, the authors Tsutsumi et al. (2021) created a live website with a *drag-and-drop* interphase, in which any clinician could freely use. In this use-case, physicians could benefit from an additional consulting reference, and the researches could benefit with additional data, which can then be used further fine-tune and perfect the model, creating a virtuous cycle.

In a recent meta-analysis review (Cao et al., 2022), the authors objective was to "systematically evaluate the development of Machine Learning Models (ML) and compare their diagnostic accuracy for the classification of Middle Ear Disorders (MED) using Tympanic Membrane (TM) images". The authors found 16 studies, and therefore unique ML models, that met their inclusion criteria. These models were generated with an aggregate of 20254 TM images between the studies. The stated accuracy of the meta-analysis for the studies ranged from 76 percent to 98.26 percent. Authors stated that aggregate sensitivity and specificity across the studies were 93 percent and 85 percent, respectively. As this meta-analysis revealed, the performance metrics of the included studies are strong and state-of-the-art.

The above authors also noted that despite all 16 studies showing robust performance metrics, none of these have been deployed in a live clinical healthcare scenario. As such, the authors suggested prospective clinical trials in order to provide additional data points in terms of the true behavior in a decision-making context.

A common difficulty across all proposed algorithm has to do with choosing specific labeling categories to be either included or excluded in the model. In the case of OM, many sub-variations such as OME in different stages or other trivial clinical findings such as cerumen in the ear canal or tympanostomy tubes can deeply affect the performance while building the algorithm as well as the potential live deployment of these algorithms. For example, if an author does not include tympanostomy tubes as part of the diagnostic categories and the algorithm is deployed live, it will be unable to interpret the clinical context of a pediatric patient with tubes, which is a common treatment (tube placements) for patients that have experienced OME. Therefore, we must caution the interpretation of these results, as they must only be viewed in the context of the labels and data set used to construct them, and not extrapolate their potential effectiveness to a real clinical setting.

It is important to consider, that in most machine-learning algorithms for otitis media, the labeling of the images is done by an expert panel of physicians. This labeling process, as we have discussed, is error prone. The diagnostic accuracy even within expert otolaryngologist is not perfect nor necessarily consistent, as it depends on years of expertise amongst many other factors.

The gold standard in OM classification is to surgically enter the middle ear, obtain a fluid sample, and assess the fluid's inflammatory, bacteriological, and virological laboratory status. Matthew et al. (2021) took this approach in a recently published study. The authors built a visual database of otoscopy matched to the corresponding surgical evaluation of the middle ear fluid. With this dataset, the researchers achieved a mean image classification accuracy of 83.8% using a deep network model. These results are auspicious and continue to showcase the enormous potential of machine learning in overcoming diagnostic hurdles.

Overall, these new state-of-the-art OM focused algorithms could provide a fast and potentially cost-effective solution to the diagnostic issues involved in otitis media as new research and further improvements prepare these models for live deployment.

Machine learning algorithms are now being studied not simply as clinician replacements but also as an enhancement to existing clinical workflows.

In a recent study by the Stanford Machine Learning Group, Bien et al. (2018) developed an algorithm to predict knee pathology based on MRI exam data. They probed the utility of providing the algorithm's predictions to radiologists during the diagnostic clinical interpretation stage. Their results indicate that the performance of radiologists, when coupled with the algorithms assistance, significantly reduces the rate at which a normal MRI scan can be misclassified as abnormal. This approach of assisted diagnosis can be of particular value for OM given that ruling out a normal ear exam can reduce antibiotic use, medical consultation time, and referrals, with all its social implications such as parental leave from work, to mention a few.

Existing studies in the clinical diagnostics for OM, in particular, indicate that non-experts tend to overdiagnose; hence the utility of this proposed machine learning assisted method or computer aided diagnosis (CAD).

Regarding CAD in otitis media, Byun et al. (2021) showed an increase in the diagnostic accuracy of medical residents, consistent with research on other medical domains. There is increasing data that supports the utilization of machine learning algorithms in an assistive diagnostic setting.

A natural extension of CAD would be to explore its potential in enhancing medical training within non-expert clinicians. In a recent multi-center study by Byun et al. (2021), the authors conducted a controlled experiment in which they evaluated the diagnostic accuracy with and without the use of a computer-aided diagnosis (CAD) system within a group of young physicians without prior expert knowledge (2 years or less of graduation). The diagnostic accuracy of the residents improved significantly when they used the computer-aided diagnosis system.

As we continue to expand the scope of how we integrate these algorithms both as a clinical and learning tool, it becomes relevant to fully acknowledge both their limitations as well as to understand the mechanisms by which they are improving medical performance.

Machine Learning Limitations and Model Interpretability

The significant limitation in understanding how these algorithms make their predictions cannot be overstated. The issue of model interpretability is even more critical in the medical domain, not only because it allows a physician to understand the model decision framework and build trust, but also because it allows all users to understand its implicit bias based on the nature of the original data it was fed.

These biases can have profound ethical considerations in medical management when minorities or under-represented patients are evaluated using these algorithms.

The real-world inequalities of healthcare are encrypted in a silent signature, that of data. Data is a mirror of the healthcare realities and inequalities. For example, if an algorithm is created based on a particular emergency room data, this data will represent the population related to that specific emergency room. This demographic cluster has its own unique set of healthcare attributes based on the dynamic interplay of local environmental factors and genetics. It is highly unlikely that this demographic is representative of the population at large, which is heterogenous and complex.

With the data, regardless of the origin, predictive models can be generated. At first inspection, these models can seem to be very effective. The effectiveness is often evaluated by separating a percentage of the data which is not exposed to the model. Once the model is built, it is then tested against this *unseen* data and graded based on performance. The issue in this approach is that the performance is correlated to the data used which as stated above, is unlikely to reflect other demographics.

Furthermore, currently there are no agreed upon methods for assessing data quality. Although we have seen very promising results in terms of diagnostic performance metrics, these algorithms are created very differently (despite all applying varying methods of ML); we lack standard protocols in terms of initial data collection and data quality evaluation.

Once a machine learning algorithm is created and tested, it can be readily deployed at a global scale. The problem is now global, in that automated diagnostic choices are being made without the correct interpretation of the local healthcare realities.

These biases which are encoded in the data, are a threat to equal treatment and healthcare access. Instead of provided convenient cost-effective healthcare, they can in fact, if deployed incorrectly, increase marginalization of at-risk cohorts of patients.

Another context of model explainability relates to inform consent. As is the case with any traditional medical diagnosis, a physician is required to explain the risks and benefits of a particular diagnostic procedure.

Considering the opaque nature of the AI based model, this can become real challenge, not only due to the underlying nature of the data used to create the model, but more importantly in understanding how this predictive model will behave in the specific context of a patient.

The need for new algorithms has increased the need and value of healthcare data. High income patients can have access to healthcare that has more robust privacy frameworks and protections in place, while lower income patients are more vulnerable to their privacy rights being abused. The increasing role of algorithms in healthcare is substantially changing the data pipeline.

This new emerging data pipeline is risking universal human rights of patients such as privacy, confidentiality, and informed consent. New players in the healthcare space such as tech startups and other actors are playing and increasingly important role in this new era of medicine, yet without the checks and regulation that traditional medical systems have had.

Finally, there are also significant legal considerations that must be considered. When using a CAD framework, it is unclear to what extent in the diagnosis was influenced by the machine learning model. As such, it remains elusive if these types of diagnosis should be disclosed to patients in particular regarding liability. When there is malpractice, to what extent is the algorithm to blame versus the physician.

Machine Learning Weights and Model Interpretability

Several strategies have been proposed to solve these issues. A widely accepted solution is to analyze the models' weights or *activation layers*. The weights of a visual machine learning model determine which areas of an image should be more important in determining a particular prediction or diagnosis.

It is important to take a step back and understand the overall procedure of how a visual machine learning model works. This is important if we aim to trust machines with diagnostics. Considering machines are numerical processing entities and can't process visual information as humans, a machine learning model turns images into numbers. The visual data is encoded by representing each visual pixel as a number alongside the height and the length of the image. All these elements as well as the color scale of the image determine the final numerical matrix which represents an individual image.

In the context of a predictive diagnostic model, the algorithm then multiplies this matrix (which represent the image or input data) with other matrixes, which can be represented as the weights. This multiplication will either help the model make better predictions or it will make the model worse.

A cost function is introduced so that the model can evaluate its performance after each multiplication and understand how far or close it is from the ideal predictions. If the model is far, it will adjust the weights of the matrix and try again. After hundreds or thousands of iterations of multiplications and weight adjustments, the model will converge onto an accurate model (depending on the nature of the initial data provided).

The weights thus contain the “secret” how a model arrives at its results. These weights as we have discussed are numbers or matrixes to be more precise, hence of little utility to us as such. The solution for us to interpret these weights is to do what the computer initially did to the image, but in reverse. We can turn the weights of a model into a visual gradient. This visual gradient can then be superimposed on a particular image by mapping the weights to the pixels of an image of interest creating a *heat-map.* We can now finally view an image and visually interpret it as the machine learning model is doing so. Grad-CAM is a specific method in which a *heat-map* can be created from an algorithm and is considered an industry standard.

Heat Maps as a Framework for Enhancing Medical Learning

Considering algorithms can enhance our diagnostic abilities, as research has shown, how can we probe these in order to learn from them? The challenge in learning from an algorithm can be explained by how an algorithm *thinks*.

As commented above, an algorithm *thinks* in the form of numbers and equations, but a clinician forms a diagnostic opinion via visual inspection. Importantly, this visual inspection is not random. Each time we view an image, our eyes focus on portions of the image or regions of interests.

Although our brain *sees* an image in full, our sensory retina can obtain high quality input only via a particular region of the retina called the *fovea*. Therefore, our eyes constantly rotate and move through a visual scene in order to *foveate* or capture the scene in high definition.

The critical component in this process is the order in which our brains determine how to capture a visual scene; this order is not arbitrary. In the context of uncertainty and limited time, our brains instruct our eyes to first capture the most relevant aspects of a visual scene.

The movements of these focal points result in a scan path or attention map. As such, eye movements are a proxy of neural attention processing. This neural attention process can vary and be optimized based on the visual task at hand. For a visual diagnostic task, our brains must be able to guide our eyes towards the areas of interest that represent features and targets based on general medical knowledge and our prior visual experience. Thus, how we visually scan an image, meaning which aspects we fix our gaze on, reveal our understanding of where we think we should seek the relevant information for the task at hand.

As such, when observing a tympanic membrane to derive a diagnosis, our scan paths indicate our level of identification and comprehension of the physiology and related pathology of the system being evaluated. The mentioned study by Bluestone et al. (1979) provides the first insights into the relevance of using visual feedback loops to improve medical diagnosis.

Considering otolaryngologists routinely practice a surgical procedure called "myringotomy," in which they can correlate and extrapolate the TM status with the actual content of the middle ear, the authors comment on the importance of this visual feedback as a driving factor which could explain the diagnostic differences between an otolaryngologist and pediatrician, or any other physician that does not practice “myringotomy” procedures.

In other medical domains in which a visual diagnosis is required, studies have compared the differences between expert and novice in terms of their scan-paths and visual fixations, and noticed significant differences, mainly on task efficiency. In this study, we aim to understand the value of providing a scan-path to a novice and evaluate their learning ability.

The scan-path can be visually represented in multiple ways, such as by blurring out unimported regions of the image, or as in this study, by providing a gradient-color based *heat-map.* Importantly, the scan-path we provide is derived not from actual human experts but from the machine learning model.

In the context of current machine learning practices, Grad-CAM derived visual saliency maps have been traditionally used to *probe* the algorithm during its development stage. If it is consistent with accepted medical heuristics, and the algorithm performs well (accurate and consistent) in new un-seen data, then Grad-CAM images are typically not further used during the live deployment of the algorithm, either when the algorithm works by itself or in conjunction with a physician (CAD method).

As stated above, we can approximate gradient-weighted activation (Grad-CAM) maps to a human scan path. It reveals everything the model thinks is essential within the image and where it should preferentially fixate its attention on. We hypothesized that by viewing these *heat-maps*, a non-expert can effectively understand, transfer, and ultimately acquire new diagnostic experience in contrast to traditional learning strategies based on the transfer of insights via text.

Results of Experiment as a Framework for Enhancing Medical Learning

The overall averages we obtained as results are considered a realistic representation of diagnostic accuracy in the context of a medical student without extensive clinical practice. When comparing questionnaire 1 with questionnaire 2, both groups on average increased their diagnostic accuracy. This was expected as we included between the questionnaires, a tutorial explaining how to diagnose middle ear pathology. Notably, the differences between both groups were statistically significant (Table 2 – Page 15). The differences within the control group, although statistically significant, (an increase of 4.6 percent) was of a smaller magnitude when compared to the differences within the experimental group (an increase of 17.73 percent). As such, the net increase which can be attributed to the experimental setup is of 8.64 percent (experimental increase minus control increase). We further analyzed this net increase, and it was found to be statistically significant (Table 2 – Page 15).

When evaluating the individual pathological entities, we can see that AOM seems to be the *easiest* to correctly diagnose. This is not surprising given that AOM is an extreme clinical scenario with very vivid visual translations in terms of the tympanic membrane and the surrounding anatomical space. The difficulty in treatment regarding OM in general, is not so much in specifically recognizing AOM, which would require antibiotic treatment as standard-of-care, but rather in the false-positives which would under standard-of-care receive antibiotics for an OM pathology that does not require it (for example in OME), leading to antibiotic resistance and decreasing healthcare outcomes globally.

OME was the entity with the worst diagnostic performance, on average with 46.6 percent across both groups. This is significant because OME is a very distinct and a different clinical scenario from AOM.

Not recognizing OME can lead to chronic ear pathologies which can result in permanent deafness and neural developmental delays. In our experiment, 81.3 percent of diagnostic errors for OME were selected as AOM. This again reflects the diagnostic difficulties in OM. Intuitively we could expect this 81.3 percent of these diagnostic errors to have been given antibiotics, while the other 18.7 percent would be incorrectly selected as normal, which could also lead to long term hearing problems.

The authors Tarpada et al., (2017) published a systematic review outlining some of the main trends in digital learning for both medical students and otolaryngology residents. They specifically looked at e-learning as an approach to solve the lack of consistent traditional training in medical school.

Some of the advantages of e-learning are that it is adaptable, updatable, and perhaps more importantly, accessible. E-learning has the advantage that it can incorporate modern teaching techniques such as spaced repetition, which is optimized to improve memory retention, as well as adapt the training to the individual knowledge-set a medical student might already have.

The authors included 12 studies that met inclusion criteria as established by the PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). The authors noted that in the digital learning approach, in five out of eight studies that included medical students as a subject group, results showed an increase in performance as well as improved subjective ratings of satisfaction. These results are significant because the proposed new learning technique we advocate could be easily incorporated into all existing varieties of online teaching methods, further enhancing the advantages of e-learning.

Our results are in line with a landmark study commented earlier published by Aronzon et al. (2004). The authors compared the following groups: medical students, internal medicine residents, junior resident otolaryngologists, senior resident otolaryngologists, attending otolaryngologists and a final group that provided diagnosis under microscopic evaluation prior to the surgical procedure. They compared the diagnostic performance of OM pathology between these groups against the gold standard of evaluating the contents of the middle ear.

The authors found no statistically significant difference in the diagnostic sensitivity, with group performance ranging from 89 to 92 percent, which is considered a very good diagnostic metric. Similarly, in our experiment, both the control group as well as the experimental group were effective in *ruling in* AOM (sensitivity) with performance ranging from 80 to 90 percent. In the study by Aronzon et al. (2004), the authors comment on the statistically significant differences in diagnostic sensitivity, which ranged from a reported 34 to 79 percent. In our experiment, specificity was also low, with OME being labeled as AOM being the main sources of errors and false positives (specificity). The importance in low specificity cannot be understated, considering it is likely the main source of unwarranted antibiotic prescription, leading to increase healthcare costs, bacterial resistance, poor health outcomes, and overall increased patient burden.

Our experiment revealed an increase of 17.73 (Table 3 – Page 16) percentage points in overall diagnostic performance and of 8.64 percent net increase (Table 2 – Page 15). If this diagnostic improvement can be sustained in a real-life scenario, it would translate potentially to millions in healthcare cost saved, and into a significant healthcare improvement for the millions of children who suffer from OM pathology each year.

An important practical consideration are the deployment strategies associated to this new proposed framework in contrast to current machine learning algorithms. For an algorithm to be deployed in a live clinical scenario, it must be able to appropriately handle common situations, for example ear canal cerumen. The algorithm must be trained in all the common varieties of clinical findings beyond the strict academic categorization of OM entities, otherwise, the algorithm is at risk of being sensitive but having very low specificity.

In a normal clinical scenario, it is common for a general medical practitioner, family medicine or junior medical resident to establish an initial screening test. If abnormal findings are observed, then a second opinion is typically requested in which an otolaryngologist can evaluate and perform a second otoscopy.

This combination is ideal because we have learned that sensitivity tends to be good even within unexperienced physicians, while specificity increases considerably within medical specialties such as otolaryngology as well as with seniority or increased expertise. If we consider the replacement of this framework for an algorithm, given current performance metrics, we know they are highly accurate and sensitive, but we have no data that supports their actual performance in a live patient scenario. This is significant because if an algorithm cannot fully replace the above framework, it would only introduce increased healthcare costs and potential bias, outweighing the benefits.

Our proposed novel training framework on the other hand, is not only simple to implement, but should outperform an algorithm in a live clinical scenario. This is because it relies on a physician, which are trained to integrate and contextualize the information they are observing and assessing. As such, in places where expert assessment by an otolaryngologist is not possible, the increase in diagnostic performance could result in significant improvements in terms of healthcare outcomes.

In a study by Blomgren et al. (2003), AOM diagnosis was compared between general practitioners and otolaryngologists. The authors found only a 64 percent of diagnostic agreement. The authors also noted that general practitioners were much more likely to diagnose AOM then an expert otolaryngologist, leading to a reported increase in incidence and antibiotic prescriptions. These reported data points align with our experiment, in which the non-expert medical students, when they made a diagnostic error, were much more likely to categorize the otoscopy as AOM than another entity.

In an interesting experiment by Rosenfeld et al. (2002), the authors measured physician confidence as a self-rated survey in terms of OM diagnosis, as well as how close the physicians followed formal diagnostic guidelines for their patients. The authors reported an overall certainty of diagnosis for AOM as 90 percent (self-reported confidence). According to the guidelines, no more than 70 percent of the cases warranted AOM diagnosis. If a physician was *certain* of the diagnosis, regardless of if it was correct or not, they were 50 percent more likely to prescribe antibiotics.

The above study highlights how valuable continual clinical education is. We must educate clinical practitioners to remain cognizant of the challenges for diagnosis and to remain vigilant in terms of further education and clinical knowledge improvement, otherwise, clinical practitioners are vulnerable to personal biases which can affect healthcare outcomes.

Our research not only highlights the difficulties in diagnostics, but also the potential improvements that can be achieved with the novel heat-map technique in a short training session.

A recent paper titled "Machine Learning for Accurate Intraoperative Pediatric Middle Ear Effusion Diagnosis" explored the challenges and potential benefits of applying machine learning for OME (Crowson et al., 2021). The authors used a transfer learning technique as stated above based on the ResNet neural architecture and they presented a model with state-of-the-art performance with an Area Under the ROC stated at 0.93. We highlight the importance of OME in this study, as well as in our experimental results, because if we aim to achieve better diagnostic accuracy, and an improved learning framework, it is vital to fully grasp, both at the machine learning level, as well as at the clinical setting, the importance of distinguishing the unique states of OME.

In our experiment, the control group was exposed to a traditional text-based tutorial. A medical student in their final year already understands normal physiology and anatomy. As such, they are able to correlate a given pathology to the corresponding clinical representations in the form of signs and symptoms when presented with new information in the form of a tutorial.

Furthermore, they can *learn* new pathological entities and quickly map them onto known physiological responses to injury, such as inflammation. Perhaps most importantly, a last year medical student is able to re-create a timeline in which all the normal and altered states occur in, and the particular progression of each. For this experiment, the experimental group was, in addition to the above, able to visually recognize these items in a visual hierarchy.

Visual hierarchy refers to the particular emphasis or importance of certain visual areas within the otoscopy image. For example, although inflammation is a key component to distinguish OM from a normal state, inflammation represents different pathological entities based on the location of the inflammation. The tympanic membrane with inflammation is a very different clinical and diagnostic entity than inflammation limited to the outer ear canal. As such, these location-based insights can be effectively *transferred* and learned, based on the gradient heat-map method.

The gradient is a color-coded method of representing less to more important and location is represented by placing these colors on top of the areas of interest within the image. This color-coded method is not arbitrary, and it is built based on our understanding of visual saliency.

Visual saliency is based on specific visual features to which we are neurological programmed to emphasize, such as color, edges, contrasts, and shapes to mention a few.

It is important to mention the top-bottom versus bottom-up approach for saliency. For example, the top-down system is rendered by our frontal lobe, and it specifically directs our visual attention to elements we recognize as important based on past experiences, information, or theory we might have and is a conscience experience. For example, in the case of an otoscopy, intuitively, a medical student will be *searching* for areas of the tympanic membrane which are distorted, either bulging out or retracted, as they know from theory this is indicative of pathology. This is the top-down system influencing the visual search path.

On the other hand, there might be specific colors within the otoscopy which are unexpected or do not match the expected color pallet. When this happens, visual attention will be drawn to those pixels in an effort to understand and extract information from that area. This is the bottom-up system and is automatic.

When an expert otolaryngologist views an otoscopy image, the main system at play is the top-down system. The combination of experience and knowledge results in the training of how to visually scan an otoscopy image to collect the most important visual cues which represent the correct proxies for the underlying pathological process. An expert will also be less distracted by visually salient items, which have no significance in the clinical physiology of the patient.

A medical student with little experience on the other hand, we expect their visual scan path to be dominated by the bottom-up system. This system will scan the image based on salient colors, edges, and contrasts within the otoscopy, in an effort to obtain information from the image.

Considering a medical student has little prior information on how and where to scan the image, the frontal lobe, top-down system will be in idle mode until it learns and can assert its preferential scan over the bottom-up system.

In this experiment, the heat-maps used can be analogous to the visual path of an expert in which the top-down system is guiding the scanning of the image. Importantly, instead of a human expert, it was generated by a machine learning algorithm, which has the benefit of scale, cost, and deployment.

The method of transferring these visual paths to a novice (medical student) is the following. We expect a novice to engage their bottom-up system, therefore, for their visual path to be determined by salient colors, edges, borders and. As such, we play to this fact and we transform the locations of an otoscopy, which the algorithm has determined is very important in containing relevant information for the diagnosis, into specific colors which we know will capture the attention of the novice viewer since they are using a bottom-up approach.

For example, if the location is important, red will be added. If the location is not important, blue will be added. These colors are mapped onto the otoscopic image in the form of a gradient from red (more important) to blue (least important), and a transparency factor is added, so that the underlying otoscopy can still be *seen* despite this color map superimposed.

We can therefore alter the expected visual path of the medical student, by re-directing their attention to spatially relevant areas, and through this method, transfer knowledge. This experiment validates this approach.

A benefit of this training approach is consistency. Medical training is highly dependent on the relationship of the mentor as well as on the type and volume of patients.

It is often the case that a particular medical center where a medical student or resident-in-training is rotating, receives a specific demographic of patients, which may or not be of interest for training purposes. The approach of using a *heat-map* derived form an algorithm, is in providing a consistent training which can complement any traditional live training scenario.

Overall, our results validate our hypothesis. There are statistically significant differences in diagnostic performance between the two groups, and critically, the group exposed to the heat-maps performed better than the control group. A follow-up study would be required to evaluate long-term performance retention.

Other Use Cases of Framework for Enhancing Medical Learning

In a variation of this framework, it could easily be expanded to all current pathological entities which are dependent on visual diagnosis. For example, skin cancer is the number one cancer in the world. Melanoma, a specific type of skin cancer is the culprit of the majority of deaths and disease burden associated with skin cancer.

Screening for melanoma is done through a visual inspection exam where the *a-b-c-d-e* of the lesion are noted. A is for asymmetry, b is for borders, c is for color, d is for diameter and e for evolution. A physician is trained to understand the normal and abnormal of the above 5 components.

Today, as with OM, machine learning algorithms have been developed to replace the medical evaluation for melanoma screening; these are highly sensitive and specific. Similar to the experiment proposed, a *heat-map* derived from these state-of-the-art melanoma detection algorithms could be used to train medical students or general medical practitioners. This could be particularly useful in cases where deploying algorithms for screening is not practical nor possible.

Radiology is another field where most of the training occurs through the mapping of knowledge to a visual scene. The experimental approach suggested could be of value not only to in-training radiologists, but to general medical practitioners who must interphase with basic radiological interpretations such as chest x-rays or abdominal imagining.

In an emergency room setting, it is very common to practice eco-graphic live imaging in patients. This is particularly important for acute abdomen syndrome which can be caused by a multitude of life-threatening entities such as appendicitis or a ruptured ectopic pregnancy.

The interpretation of the ecographical findings is key in making a correct surgical decision and hence potentially in saving a patients' life. As such, a live *heat-map* from an algorithm could provide consistent and effective training for all medical practitioners in this context.

The current the rate of innovation in these algorithms continues to accelerate, and an increasing percentage of medical practitioners are beginning to use these predictions in conjunction with their existing diagnostic frameworks. For these motives, it is of paramount importance to fully understand how we can harness these improvements to boost learning and use these new breakthroughs responsibly.

Overall, we propose a simple, cost-effective method which can be used to readily train and teach both medical students and junior physicians on the nuanced pathological varieties that affect the middle ear.

Chapter VI.

Research Limitations

There are several limitations to this study. As a proof of concept, we utilized a limited amount of ear pathology in terms of photographs. Therefore, the study did not include all the sub-type of OM or their specific diagnostic entities but instead focused on AOM versus OME versus normal status.

Also, despite including a washout period of 5 minutes, it is beyond the scope of this study to understand the permanence of these differences, assuming they are present, within the groups. Differences could be attributed to simple visual memorization and not necessarily an increased understanding of the clinical pathology. A follow-up study ideally could analyze the diagnostic performance on a spaced-out approach that could consider long term memory retention.

A potential follow-up experiment could repeat this setup but with eye-tracking technology. This could potentially provide an objective neurological proxy of how attention changes and hence how information is acquired, transferred, and learned by comparing the medical students scan path pre and post exposure. In this study, we are using the test grades, pre and post exposure, as a metric of information acquisition.

Furthermore, a more complex variation of the above would be to analyze live magnetic resonance imaging when conducting the diagnostic test. This would provide a live objective neural understanding of the cognitive processes at work and the differences in the cognitive processing when viewing the visual heat-maps in contrast to learning from traditional word based medical tutorials.

More extensive objective analysis would be suggested as a follow-up study to understand further the scope and magnitude of utilizing saliency maps derived from machine-learning models as a framework for medical training.

References

Aronzon, A., Ross, A.T., Kazahaya, K., Ishii, M. (2004). Diagnosis of middle ear disease using tympanograms and digital imaging. *Otolaryngol Head Neck Surg, 131(6), 917-20.*

Auinger, P., Lanphear, B.P., Kalkwarf, H.J., Mansour, M.E. (2003). Trends in otitis media among children in the United States. *Pediatrics, 112(3 pt 1), 514-520.*

- The authors acknowledge the known increase in rates of AOM and OM before 1988.
- They analyzed from 1988 to 1994 to verify if this increase continues, and in fact, they can establish this.
- They also comment on population groups based on socio-economic status, and impoverished children reflect the group with the most significant increase.

Bien, N., Rajpurkar, P., Ball, R.L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B.N., Yeom, K.W., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D.F., Beaulieu, C.F., Riley, G.M., Stewart, R.J., Blankenberg, F.G., Larson, D.B., Jones, R.H., Langlotz, C.P., Ng, A.Y., Lungren, M.P. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med, 15(11), e1002699.*

Bluestone, C.D., Cantekin, E.I. (1979). Design factors in the characterization and identification of otitis media and certain related conditions. *Ann Otol Rhinol Laryngol, 88, 13-27.*

Blomgren K, Pitkäranta A. (2003) Is it possible to diagnose acute otitis media accurately in primary health care? *Fam Pract. Oct;20(5):524-7.*

Byun, H., Yu, S., Oh, J., Bae, J., Yoon, M.S., Lee, S.H., Chung, J.H., Kim, T.H. (2021). An Assistive Role of a Machine Learning Network in Diagnosis of Middle Ear Diseases. *J. Clin.Med, 10, 3198.*

Cao, Z., Chen, F., Grais, E. M., Yue, F., Cai, Y., Swanepoel, D. W., & Zhao, F. (2022). Machine Learning in Diagnosing Middle Ear Disorders Using Tympanic Membrane Images: A Meta-Analysis. *The Laryngoscope*

Crowson, M.G., Hartnick, C.J., Diercks, G.R., Gallagher, T.Q., Fracchia, M.S., Setlur, J., Cohen, M.S. (2021). Machine Learning for Accurate Intraoperative Pediatric Middle Ear Effusion Diagnosis. *Pediatrics, 147(4), e2020034546.*

Jones, W.S., Kaleida, P.H. (2003). How helpful is pneumatic otoscopy in improving diagnostic accuracy? *Pediatrics, 112, 510–513.*

Kuruvilla, A., Shaikh, N., Hoberman, A., Kovacevic, J. (2013). Automated diagnosis of otitis media: vocabulary and grammar. *Int J Biomed Imaging, 2013, 327515.*

Livingstone, D., Chau, J. (2020). Otoscopic diagnosis using computer vision: an automated machine learning approach. *Laryngoscope, 130(6), 1408–1413.*

Monroy, G.L., Won, J., Dsouza, R., *et al.* (2019). Automated classification platform for the identification of otitis media using optical coherence tomography. *NPJ Digit Med, 2, 22.*

Nelson, W.L., Kuritsky, J.N., Kennedy, D.L. (1987). Outpatient pediatric antibiotics use in the US: trends and therapy for otitis media, 1977- 1986. *27th Interscience Conference on Antimicrobial Agents and Chemotherapy, Washington, DC: American Society for Microbiology, 4.*

Owings, M.F., Kozak, L.J. (1998). Ambulatory and inpatient procedures in the United States, 1996. *Vital Health Stat, 13(139), 44.*

Pichichero, M.E., Poole, M.D. (2001). Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. *Arch Pediatr Adolesc Med, 155(10), 1137–1142.*

Pichichero, M.E., Poole MD. (2002). Diagnostic accuracy, tympanocentesis training performance, and antibiotic selection by pediatric residents in management of otitis media. *Pediatrics, 110(6), 1064–1070.*

Rosenfeld, R.M. (2002). Diagnostic certainty for acute otitis media. *Int J Pediatr Otorhinolaryngol, 64(2), 89-95.*

Schappert, S.M. (1992). Office visits for otitis media: United States, 1975-1990. *Adv Data Vital Health Stat, 214, 1-20.*

- The authors compared the incidence of OM reported by NHIS with office visits.
- Both metrics reveal a significant increase from 1975 to 1990 in the range of 60% (incidence) to 80% office visits.

- Although the study does not provide a causal framework, it does provide interesting insights in which the increased office visits can also be reflected and probably related to the increase in incidence.

Shie, C.K., Chang, H.T., Fan, F.C., Chen, C.J., Fang, T.Y., Wang, P.C. (2014). A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media. *Annu Int Conf IEEE Eng Med Biol Soc, 2014, 4655–4658.*

Suaya, J.A., Gessner, B.D., Fung, S., Vuocolo, S., Scaife, J., Swerdlow, D.L., Isturiz, R.E., Arguedas, A.G. (2018*).* Acute otitis media, antimicrobial prescriptions, and medical expenses among children in the United States during 2011-2016. *Vaccine, 39(49), 7479-7486.*

- The authors analyze public health costs and trends regarding antimicrobial usage.
- Overall, OM decreased during the 2011-2016 period analyzed, resulting in a 5.6 billion approximated decrease in direct medical expenditures.
- This decrease, the authors mainly relate to success in vaccination and awareness.
- Importantly, even though there is a reduction in antibiotic usage for otitis media, average prescriptions per visit remained stable. What they did notice was a change in the antimicrobials used.

Tarpada, Hsueh, W. D., & Gibber, M. J. (2017). Resident and student education in otolaryngology: A 10-year update on e-learning. *The Laryngoscope*, *127*(7), *E219–E224*

Teele, D.W., Klein, J.O., Rosner, B. (1989). Epidemiology of otitis media during the first seven years
of life in children in Greater Boston: a prospective cohort study. *J Infect Dis, 160(2), 83-94.*

- Provides a general context of the epidemiology of otitis media, from the vantage point of patient age.

- Most risk is concentrated in the first year. However, when analyzed cumulatively, by three years of age, 83% had greater than or equal to one episode of AOM.

- Authors also analyzed risk factors and determined that male gender, sibling history of recurrent AOM, early occurrence of AOM, and not being breastfed are predisposing factors.

Tran, T.T., Fang, T.Y., Pham, V.T., Lin, C., Wang, P.C., Lo, M.T. (2018). Development of an automatic diagnostic algorithm for pediatric otitis media. *Otol Neurotol, 39(8*), *1060–1065.*

Tsutsumi, K.; Goshtasbi, K.; Risbud, A.; Khosravi, P.; Pang, J.; Lin, H., et al. (2021). A Web-Based Deep Learning Model for Automated Diagnosis of Otoscopic Images. *Otology & neurotology : official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology*, 42(9), e1382-e1388.

Viscaino, M., Maass, J.C., Delano, P.H., Torrente, M., Stott, C., Auat Cheein, F. (2020). Computer-aided diagnosis of external and middle ear conditions: A machine learning approach. *PLoS One, 15(3*)*, e0229226.*

- Analyzed the utility of a CAM method for improving physician medical diagnostic accuracy for OM

Wang, Y.M., Li, Y., Cheng, Y.S., *et al*. (2020). Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography. *Ear Hear, 41(3*), *669–677.*

Wu, Z., Lin, Z., Li, L., Pan, H., Chen, G., Fu, Y., Qiu, Q. (2021). Deep Learning for Classification of Pediatric Otitis Media. *Laryngoscope, 131(7), E2344-E2351.*

Young, D.E., Ten Cate, W.J.F., Ahmad, Z., Morton, R.P. (2009). The accuracy of otomicroscopy for the diagnosis of paediatric middle ear effusions. *Int J Pediatr Otorhinolaryngol, 73(6), 825–828.*