Pediatric Otoscopy Video Screening With Shift Contrastive Anomaly Detection

Weiyao Wang[1]*, Aniruddha Tamhane[1], Christine Santos[2], John R. Rzasa[3], James H. Clark[4], Therese L. Canares[2] and Mathias Unberath[1]

[1] Department of Computer Science, Johns Hopkins University School of Engineering, Baltimore, MA, United States, [2] Department of Pediatric, Johns Hopkins University School of Medicine, Baltimore, MA, United States, [3] Robert E. Fischell Institute for Biomedical Devices, University of Maryland, College Park, MA, United States, [4] Department of Otolaryngology, Johns Hopkins University School of Medicine, Baltimore, MA, United States

Ear related concerns and symptoms represent the leading indication for seeking pediatric healthcare attention. Despite the high incidence of such encounters, the diagnostic process of commonly encountered diseases of the middle and external presents a significant challenge. Much of this challenge stems from the lack of cost effective diagnostic testing, which necessitates the presence or absence of ear pathology to be determined clinically. Research has, however, demonstrated considerable variation among clinicians in their ability to accurately diagnose and consequently manage ear pathology. With recent advances in computer vision and machine learning, there is an increasing interest in helping clinicians to accurately diagnose middle and external ear pathology with computer-aided systems. It has been shown that AI has the capacity to analyze a single clinical image captured during the examination of the ear canal and eardrum from which it can determine the likelihood of a pathognomonic pattern for a specific diagnosis being present. The capture of such an image can, however, be challenging especially to inexperienced clinicians. To help mitigate this technical challenge, we have developed and tested a method using video sequences. The videos were collected using a commercially available otoscope smartphone attachment in an urban, tertiary-care pediatric emergency department. We present a two stage method that first, identifies valid frames by detecting and extracting ear drum patches from the video sequence, and second, performs the proposed shift contrastive anomaly detection (SCAD) to flag the otoscopy video sequences as normal or abnormal. Our method achieves an AUROC of 88.0% on the patient level and also outperforms the average of a group of 25 clinicians in a comparative study, which is the largest of such published to date. We conclude that the presented method achieves a promising first step toward the automated analysis of otoscopy video.

Keywords: anomaly detection, pediatric healthcare, otoscope, pediatric healthcare, self-supervised learning, deep learning

1. INTRODUCTION

Ear related concerns constitute the leading cause for seeking pediatric healthcare attention in the USA. Given the current lack of cost-effective confirmatory testing, accurate diagnosis and subsequent management depend on visual detection of characteristic findings during otoscope examination. Despite the frequency of such encounters, the medical literature suggests that the diagnostic accuracy for such pathology is only around 46–56%. In alignment with the bias toward over-diagnosis of ear disease, it is currently estimated that between 25–50% of all antibiotics prescribed for ear disease are not indicated (1–3). Beyond risking unnecessary medical complications and the downstream unintended consequence of potential antibiotic resistance, over-diagnosis of ear disease adds an estimated $59 million in unnecessary healthcare spending in the US per annum (4). Computer-aided diagnosis on otoscopy images (5) has been suggested as a potential tool to improve the care of ear disease. Previous studies have mainly focused on applying machine learning to classify images of the eardrum. (6) compare support vector machine (SVM), k-nearest neighbor (k-NN), and decision trees on predicting ear conditions with feature extracted by filter bank, discrete cosine transform (DCT), and color coherence vector (CCV) (6). More recently, deep learning has been applied to classify otoscopy images. Zafer studies the combination of the fused fine-tuned deep features and SVM model applied to 3 diagnostic class image classifications (7).

Furthermore, (8) compare the performance of a variety of CNN architectures and find that DenseNet (9) produces the best result in the classification of 3 diagnostic classes. The authors also use Grad-CAM (10) to visualize the important regions in the image for the model.

In the study by (11), the authors compare the performance of applying 9 kinds of ImageNet (12) pre-trained CNN networks to 6 class otoscopy image classification and then further improve the performance by ensembling the output of 2 networks. The pre-mentioned deep learning based methods overall show great performance in the respective settings, achieving high accuracy ranging from 94 to 99%.

The accuracy reported by previous study, even in multi-class settings, is remarkable, yet, all previous approaches process high quality still images of the eardrum. Unfortunately, in a clinical setting the capture of high quality still images can be challenging in practice, especially given that pediatric patients are often moving and uncooperative with the examination. The use of a single image also increased the chance of failing to capture a clinically accurate representation of the anatomy due to an incomplete view. A method that analyzes otoscopy video sequences rather than still images could help overcome the aforementioned shortcomings.

The straightforward approach to have the whole video sequence as input is to train a 3D (2D images plus temporal dimension) CNN mapping from videos to labels. However, that would require (1) heuristics to harmonize videos to a preset length, and (2) large amounts of labeled data to avoid over-fitting, especially considering that the class frequency in any otoscopy video dataset is likely to be heavily skewed toward normal cases. The present study addresses these challenges by framing otoscopy interpretation as a video anomaly detection problem. Under our setting, the model is trained with normal videos only and will flag videos that diverge from the normal as an anomaly during testing.

The contribution of this study is 2-fold. First, we devise a two stage model to apply on full video sequences collected in clinics, which differentiates our approach from all previous methods that only considered high quality still images. A region detector first extracts eardrum patches from video frames that are then used to perform anomaly detection in the following stage. This architecture restricts anomaly detection to semantically useful regions. Second, we develop the shift contrastive anomaly detection (SCAD) method that leverages color-jitter-based distributional shift-transformation to improve the separability between normal and abnormal data. Tailored to our application, our self-supervision task enforces the model to leverage subtle color features to identify abnormalities during testing. Even when the amount of training videos is relatively small, our approach generates superior binary (normal/abnormal) otoscopy video screening results compared to both a baseline method and the average of a group of clinicians. We believe that this study constitutes a promising first step toward achieving algorithmic decision support for the diagnosis of pathology of the middle and external ear and the development of a smart tool that might aid clinicians in the accurate diagnosis and management of common ear disease.

2. RELATED STUDY

The task of anomaly detection (13, 14), also referred to as out-of-distribution or novelty detection depending on the context, is to identify unfitted data samples. This is a crucial task in many real world applications such as detecting system malfunction, financial fraud, and health issues. Anomaly detection is a useful but challenging task, because anomalies are hard to define. In a supervised anomaly detection setting, abnormal samples are provided to the model in the training stage to give guidance on what qualifies as an "anomaly." However, there could be many reasons a data sample is abnormal such that collecting a representative amount of abnormal samples is a hard problem itself. This is especially challenging in medical applications, where the dataset is very likely to the heavily skewed toward the normal class. Unsupervised anomaly detection (UAD), on the other hand, does not need normal/abnormal supervision signals during training. Current approaches usually utilize only the normal data during training, which makes it more appealing to applications with a skewed dataset. Therefore, in the scope of this study, we focus only on the UAD setting where the detector can only access samples from the normal data distribution during training. Moreover, in many real world applications, the task is to perform UAD on high-dimensional imagery data (e.g., in our case, detecting anomalies from endoscopic videos), which makes the problem even more challenging.

A rich set of methods (13) has been proposed to study the UAD problem on imagery data, approaching the problem with four main paradigms: (i) Density-based methods first estimate the normal data density using methods, including Gaussian Mixture and Energy Based Models, and then detect data with

a low estimated density as an anomaly. DAGMM (15) and DSEBM (16) are methods that belong to this category. (ii) One-class classifier-based methods fit a classifier, e.g., Deep SVDD (17) and DROC (18), to separate normal and all other data and then use it to detect anomalies. (iii) Reconstruction-based techniques learn a reconstruction model, e.g., AnoGAN (19), of normal images and detect anomalies as samples with high reconstruction error. (iv) Self-supervised-based methods learn a feature extractor using self-supervised tasks, such as distinguishing whether or not a certain transformation has been applied to the image (20). Then, during learning features of augmented versions of the same image should be closer than features of different images (21). Upon convergence, anomaly detection is performed on the extracted features. Notable methods along this line include SVD-RND (22), CutPaste (23), CSI (24), SSD (25), PANDA (26), and MSC (27). UAD has also been applied to medical imaging (28) across many domains, including X-ray (29, 30), CT (31, 32), MRI (33–35), and endoscopy (36) datasets.

Anomalies in medical imaging tend to be more subtle and only reside in a small region of the whole image, which makes it particularly challenging to apply UAD in this field. To address this problem, (37) propose to connect Context-encoding (CE) and Variational Autoencoders (VAE) to combine both reconstruction and density based anomaly scores. Doing so shows superior performances in three brain MRI image datasets. To directly target subtle anomalies in small regions, there are efforts to develop self-supervised tasks that are tailored to a certain anomaly group. It is desirable to apply the artificial transformation of data that is similar to real anomaly so that the self-supervised task can force the model to learn features that are discriminative enough to detect anomalies in testing. As an example, CutPaste (23) creates discontinuous defects by cutting a region of an image and pasting it to another location. While this method works well for industrial visual inspection, the introduced sharp discontinuity is uncommon in medical imagery making this method less useful for anomaly detection in medical data. Alternatively, foreign patch interpolation (FPI) (38) creates artificial defects by interpolating a small local patch with a foreign patch of the image. A network is trained to estimate the pixel-wise interpolating factor from the synthesized image. The estimated interpolating factor is then used as an anomaly score during testing. Success is demonstrated using Brain MRI and abdominal CT image datasets. Poisson image interpolation (PII) (39) extends this idea by using Poison image editing to blend in the foreign patch instead of direct linear interpolation, targeting more subtle and continuous irregularities. Improvement in performance is shown using chest X-ray and fetal ultrasound datasets. Meanwhile, both FPI and PII utilize self-supervised tasks targeted at regional shape anomaly which is less applicable to our problem where the subtle color anomaly is more crucial.

3. METHODOLOGY

The otoscope video database is skewed toward the non-disease status given the practice of capturing bilateral ear examinations

and occurrence of non-ear disease presenting with symptoms such as ear pain. Dataset imbalance such as this can present a significant challenge for the development of robust design boundaries, particularly when relying on small data sets. To overcome such limitations, we propose to use UAD approach to provide computer-aided screening for otoscopy video. The formal definition of this problem is as follows.

We are given a large training dataset $\mathcal{D}_{train}$ comprising only normal video sequences and a smaller testing dataset $\mathcal{D}_{test}$ comprising both normal and abnormal video sequences. Given the training dataset, our objective is to learn an anomaly score function $A(v)$ with video v as input and detect the abnormal videos as an anomaly during testing. Ideally, the function learns to map normal videos to small anomaly scores and abnormal videos to large anomaly scores. During testing, we threshold the score, where $A(s) > \psi$ indicates an anomaly.
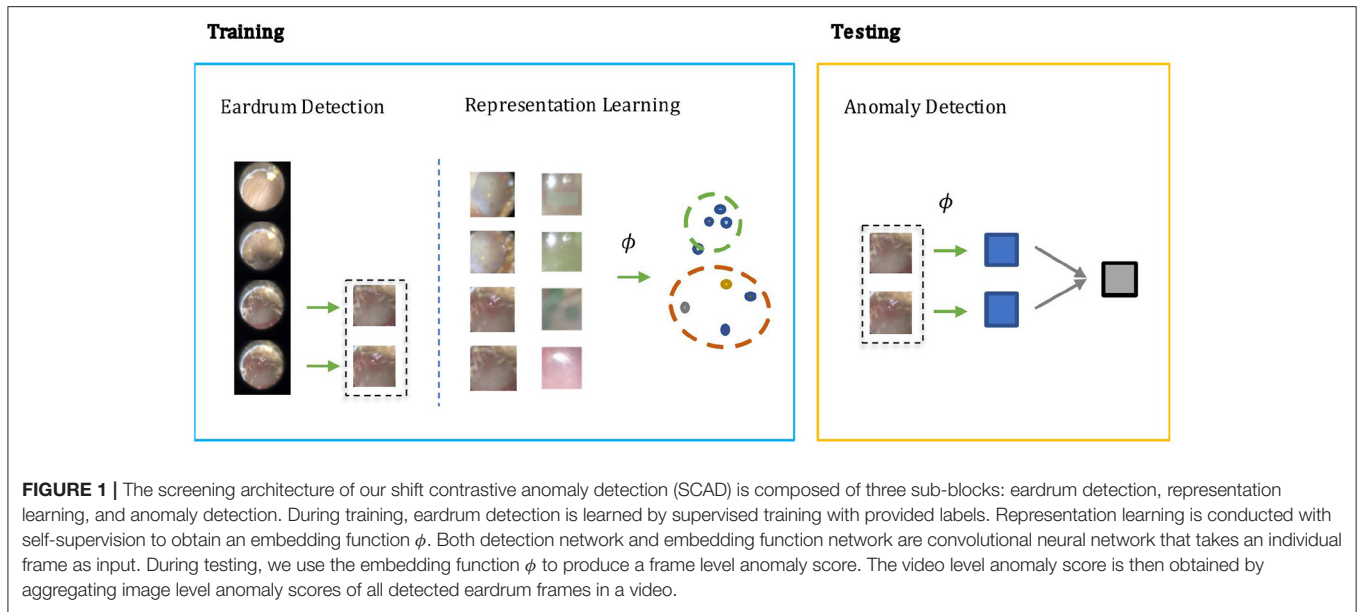
In otoscopy video diagnosis, the temporal relation between frames is not as influential as in action recognition tasks. Whether or not a video is abnormal is directly determined by the existence of abnormal frames therein. With this observation, we turn the video anomaly detection problem into a frame anomaly detection problem with an additional video level aggregation function. Moreover, only some of the frames in the video are vital to diagnosis, and in those frames, only the region that visualizes the eardrum. Therefore, it would be desirable to only conduct anomaly detection on such frames, and the eardrum region in particular. As shown in **Figure 1**, our method has two steps during training: Step 1) Supervised learning of a detection module that extracts eardrum patches x from given video v using provided labels. Step 2) Self-supervised representation learning on the labeled region of interest to obtain an embedding function that ideally projects normal and abnormal samples to distinct regions of the embedding space. During testing, we compute anomaly scores using the detected patches and the learned embedding function.

3.1. Eardrum Detection

To detect the frames and regions showing the eardrum, we use a convolutional neural network and train it in a supervised manner using image-label pairs. Different from the standard detection problem where each image may have multiple instances, we can only have at most one eardrum in each frame. Therefore, we do not use a standard object detection architecture and rather choose to have a single convolutional neural network to predict the binary label and corresponding bounding box. We define $\mathcal{L}_{cls}$ as the cross entropy loss of the groundtruth label and the classification score and define $\mathcal{L}_{local}$ as the $L1$ loss of the groundtruth bounding box and the predicted bounding box. Then, the overall loss is simply combining the classification loss and localization loss.

$$\mathcal{L}_{detect} = \mathcal{L}_{cls} + \mathcal{L}_{local} \tag{1}$$

Given our problem setup, we train the detection module on the normal training videos only. Even though the model does not use abnormal videos during training, we find it generalizes well to abnormal frames during testing as shown in Section 4.2.

FIGURE 1 | The screening architecture of our shift contrastive anomaly detection (SCAD) is composed of three sub-blocks: eardrum detection, representation learning, and anomaly detection. During training, eardrum detection is learned by supervised training with provided labels. Representation learning is conducted with self-supervision to obtain an embedding function ϕ. Both detection network and embedding function network are convolutional neural network that takes an individual frame as input. During testing, we use the embedding function ϕ to produce a frame level anomaly score. The video level anomaly score is then obtained by aggregating image level anomaly scores of all detected eardrum frames in a video.

3.2. Shift Contrastive Anomaly Detection

Contrastive Learning: Contrastive learning (21, 40) has become the top performing self-supervised learning method in recent years. In this paradigm, the first step of the training procedure is to sample a minibatch of size N and perform augmentation twice to each sample x_i to obtain $(x_{i'}, x_{i''})$ termed as positive pair, producing $2N$ samples in total. All images are passed through a feature extractor and then features are typically scaled to the unit sphere by l_2 normalization to get a representation ϕ. The contrastive loss to pair $(x_{i'}, x_{i''})$ is then defined as

$$\mathcal{L}_{con}(x_{i'}, x_{i''}) = -\log \frac{\exp(\phi(x_{i'}) \cdot \phi(x_{i''})/\tau)}{\sum_{m=1}^{2N} \mathbb{1}[i' \neq m] \cdot \exp(\phi(x_{i'}) \cdot \phi(x_m)/\tau)}, \tag{2}$$

where τ is a temperature hyper-parameter.

The contrastive learning objective pulls $x_{i'}$ close to $x_{i''}$ and pushes all other samples away from $x_{i'}$. Intuitively, this will make the learned embedding capture a latent structure that is meaningful enough to separate samples from one another and thus be beneficial for downstream tasks. This paradigm shows great success in self-supervised training for image recognition but has inherent problems for anomaly detection. To minimize the loss function, the angles between positive and negative samples need to be maximized even though both samples are from the normal class. This results in a scenario where representations of normal samples span across the whole unit sphere. During testing, anomaly samples could, therefore, be projected onto a location that is close to normal embeddings, which challenges the paradigm of anomaly detection.

Mean-Shifted Contrastive Loss: To adapt contrastive learning to anomaly detection, MSC (27) proposes the alternative mean-shifted contrastive loss. Rather than directly minimizing the contrastive loss in the representation space, MSC constructs a mean-shifted counterpart, by subtracting the center c of the whole training set and then normalizing to it the unit sphere. For

a given sample x, the mean-shifted embedding is defined as

$$\theta(x) = \frac{\phi(x) - c}{\left\| \phi(x) - c \right\|}. \tag{3}$$

The mean-shifted loss is then constructed by applying typical contrastive loss on this mean-shifted embedding space:

$$\mathcal{L}_{msc}(x', x'') = \mathcal{L}_{con}(\theta(x'), \theta(x'')). \tag{4}$$

Because the mean-shifted representation is normalized around c, the training no longer spans the normal samples across the whole unit sphere of $\phi(x)$. This makes anomaly samples more separable from normal samples in the $\phi(x)$ space.
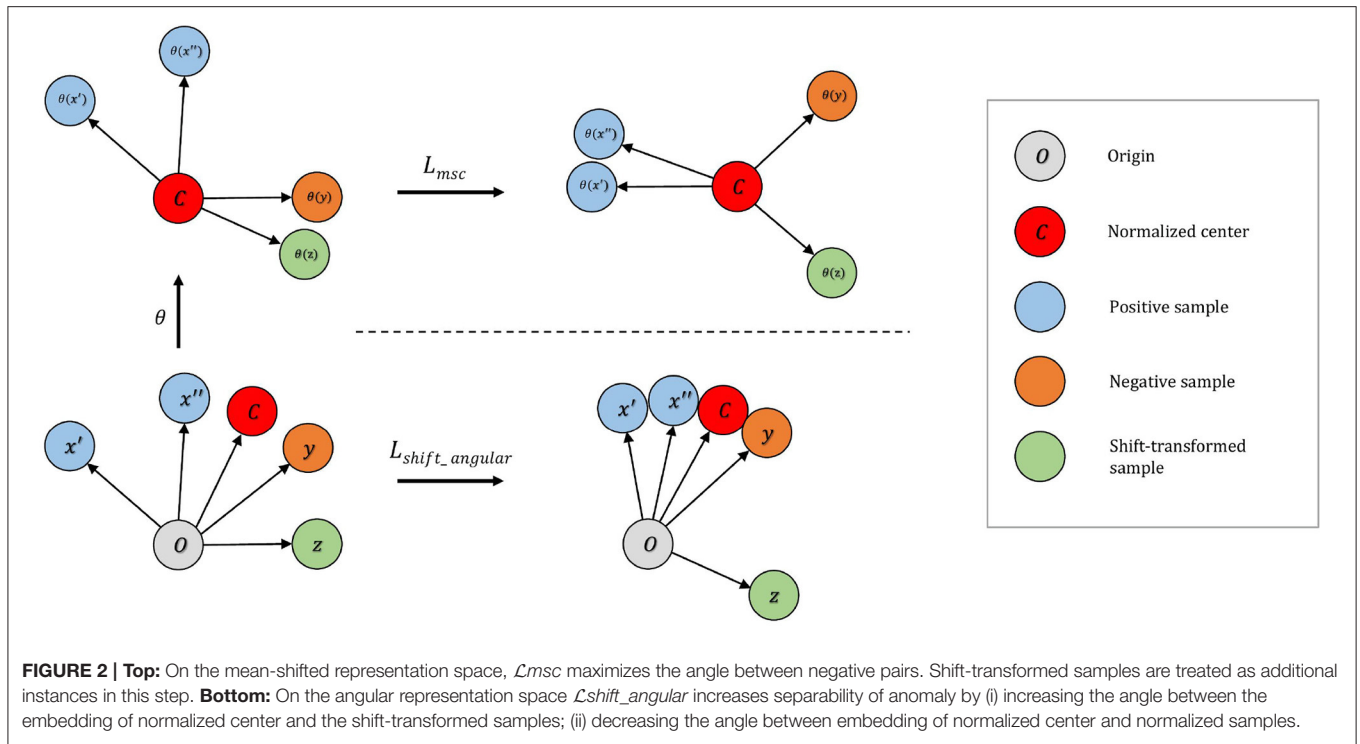
Distributional Shift Angular Loss: To further improve separability, MSC (27) uses additional angular center loss

$$\mathcal{L}_{angular} = -\phi(x) \cdot c, \tag{5}$$

to shrink normal samples around the normalized center. The assumption is that normal data lying in a small region around the center will be more discriminative. The total loss that MSC uses then becomes $\mathcal{L}_{angular} + \mathcal{L}_{angular}$.

This approach works well in datasets composed of regular images (e.g., CIFAR dataset (41)), where images from different classes are treated as an anomaly. Empirically, we find that this strategy does not perform as well in our medical imaging application. The main reason is that we are having much smaller semantic variations between normal and abnormal images so that their respective embeddings are still close to each other even after training using mean-shifted contrastive and angular center loss.

To address this issue, we develop the distributional shift angular center loss to further enhance separability between normal and abnormal samples. As shown in **Figure 2**, our method constructs additional shift-transformed samples z from

FIGURE 2 | Top: On the mean-shifted representation space, $\mathcal{L}msc$ maximizes the angle between negative pairs. Shift-transformed samples are treated as additional instances in this step. **Bottom:** On the angular representation space $\mathcal{L}shift\_angular$ increases separability of anomaly by (i) increasing the angle between the embedding of normalized center and the shift-transformed samples; (ii) decreasing the angle between embedding of normalized center and normalized samples.

x and then pushes $\phi(z)$ away from the normalized training center c. Shift-transformed samples z are created by applying distributional shift-transformations on x so that z is no longer from the original data distribution. The intuition is straightforward: not only should the normal samples occupy a small region around the center but also the samples not drawn from the original distribution cannot be too close to the center either. Hinge loss is applied to the angle between center c and shift-augmented samples z. The loss function takes the form:

$$\mathcal{L}_{shift\_angular} = -\phi(x) \cdot c + max(0, 1 - \phi(z) \cdot c). \quad (6)$$

Ideally, these shift-transformed samples z are semantically similar to real anomalies. In the context of otoscopy images, color is an important feature in that infection may lead to the change of color of the whole or part of the eardrum. Thus, we employ three variants of color-jitter as the distributional shift-transformation. They are color-jitter-random-cut (CJ-RC), color-jitter-random-region (CJ-RR), and color-jitter-whole-frame (CJ-WF). CJ-RC uses a random rectangle as a mask of transformation. CJ-RR is constructed by interpolating the color-jittered image and the original image using pixel-wise weights for each image. We first initialize the pixel-wise weights to be all zero, and then pick random points to be filled with one, and apply a Gaussian filter to obtain smoothed weights. CJ-WF simply applies color-jitter to the whole frame. **Figure 3** presents illustrations of the three augmentation strategies.

Final Loss: We combine (i) the mean-shifted contrastive loss $\mathcal{L}_{msc}$ and (ii) the angular loss with distributional shift augmentations $\mathcal{L}_{shift\_angular}$:

$$\mathcal{L}_{final} = \mathcal{L}_{msc} + \mathcal{L}_{shift\_angular}. \quad (7)$$

This loss function is tailored to this specific medical imaging application and enjoys the best of both objectives. The mean-shifted loss makes the features to be representative of the images and the angular loss encourages normal and abnormal instances to be distant from each other in the feature space. Therefore, the combination of these two losses achieves superior performances, which is demonstrated through ablation in Section 4.3.
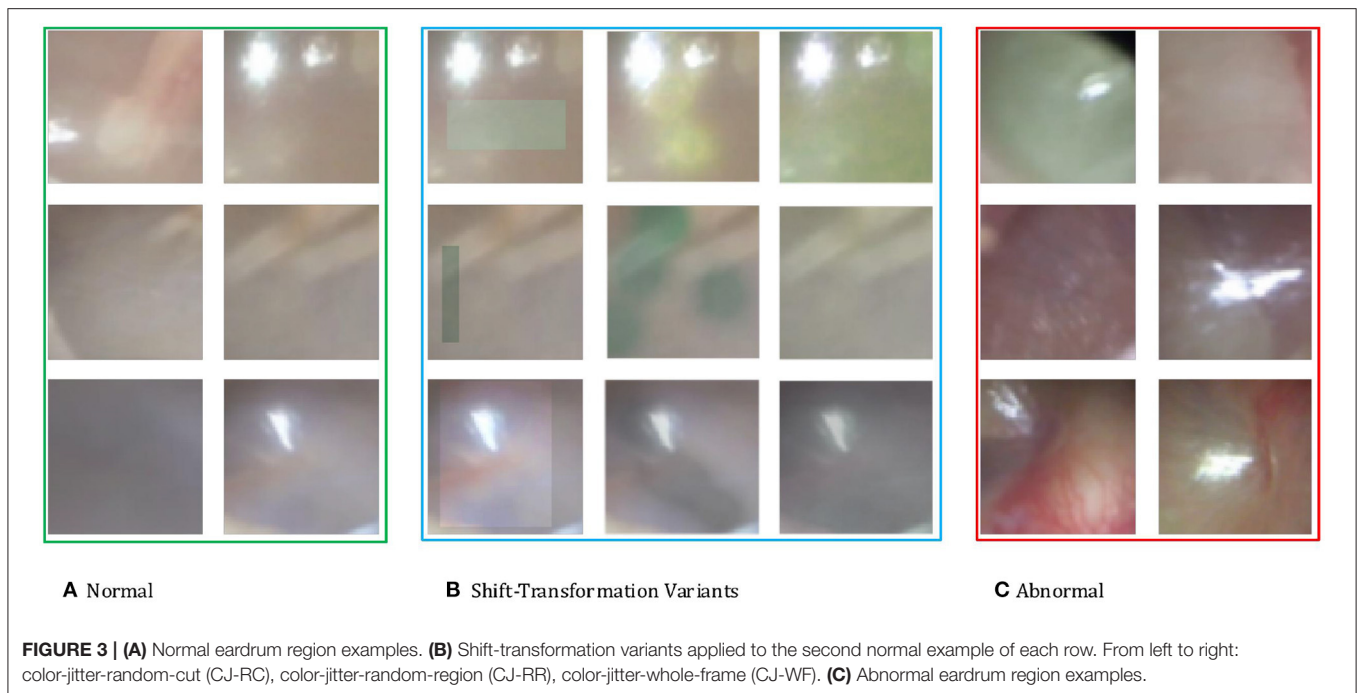
Frame Level Anomaly Score: In order to classify a sample as normal or abnormal, we use a simple criterion based on kNN. kNN predicts the label of a data point by ensembling the k closest labeled samples according to some distance measure. In this study, we use the cosine distance between the features of the target image x and those of all training images. The anomaly score is then given by:

$$a(x) = \Sigma_{\phi(y) \in N_k(x)}(1 - \phi(x) \cdot \phi(y)), \quad (8)$$

where $N_k(x)$ denotes the k nearest features to $\phi(x)$ in the training feature set $\{\phi(z)\}_{z \in \mathcal{X}_{train}}$.

Video Level Anomaly Score: We consider a relatively simple method to aggregate the frame level anomaly score to construct the video level anomaly score. The aggregation function is simply taking the average of the frame level anomaly score across all detected eardrum patches. The video level anomaly score is defined as

$$A(s) = \overline{\{a(x)\}}, \quad (9)$$

FIGURE 3 | (A) Normal eardrum region examples. **(B)** Shift-transformation variants applied to the second normal example of each row. From left to right: color-jitter-random-cut (CJ-RC), color-jitter-random-region (CJ-RR), color-jitter-whole-frame (CJ-WF). **(C)** Abnormal eardrum region examples.

where $\{a(x)\}$ represents the set of frame level anomaly scores for all frames in the given video s.

In summary, during training, we learn a frame level embedding function ϕ with all individual normal frames in the training set. During testing, we compute the kNN-based anomaly score using the embedding feature from ϕ for all detected frames. We then aggregate the predicted anomaly score for each detected frame in the testing video to produce a video level anomaly score A through averaging. Such a score is expected to be small for normal videos and large for abnormal videos. Any frame that is substantially different from normal frames in the training set (e.g., a red, bulging eardrum) would lead to a large distance to training examples in embedding space and then produce a large image and video level anomaly score. The threshold can be selected to meet the desired clinical requirement (e.g., a certain specificity value) and in practice is selected using the validation dataset consisting of both normal and abnormal videos.

4. EXPERIMENTAL RESULTS

4.1. Data Preparation

We collected a total of 100 otoscopy videos from pediatric patients that were seen for various conditions in an urban, tertiary-care pediatric emergency department that treats over 34,000 patients, aged between 0 and 21 years of age per year. Patients were recruited as a convenience sample based on their willingness to participate in the study. At this site, patient ages range from 0 to 22 years. Study protocols were approved by the local institutional review board. The length for each video ranges from 5 to 40 s. Twenty seven to thirty frames per second are extracted from the video. Two otolaryngologists (ears, nose, and throat specialists) annotated every video, and

TABLE 1 | Detection performance evaluated by accuracy% under different IoU thresholds.

	Accu (IoU > .50)	Accu (IoU > .75)	Accu (IoU > .90)
Normal	77.3	69.7	37.8
Abnormal	82.1	79.3	34.4
Overall	80.8	73.1	35.7

consensus agreement was used as ground truth. While the expert annotations are more granular, due to the small size of the dataset, we decided to restrict the analysis to two classes and each video is assigned into either the normal or abnormal category. Additional to the binary categorical label on video level, we also annotate frame level binary labels and bounding boxes for each appearance of the eardrum, where the eardrum occupies at least 20% of the image size in width and height. Of the 100 videos, 80 are labeled normal and 20 are labeled abnormal. We assign 60 normal videos to the training set, 10 to the validation set, and 10 to the testing set. For anomaly videos, we assign 10 to the validation set and 10 to the testing set. The testing set is used for the evaluation of the algorithms and comparison with human clinicians. We do not have overly apparent abnormal examples (e.g., the whole eardrum is filled with blood) during testing which makes it a challenging comparative study with human clinicians.

Video capture of ear exam was performed using Cellscope Oto attached to a first-generation iPhone SE. Maintaining relative uniformity of video settings such as color is critical to the accuracy of the AI algorithm. Clinical deployment of such technology, therefore, presents a significant challenge given the *ad hoc* selection of otoscopes used in the clinical setting and

perhaps more importantly the overwhelming reliance on non-digital otoscopes in clinical practice. As a result, we opted to use Cellscope Oto as an affordable solution that can potentially be deployed at scale. The Cellscope Oto device is designed to attach to an ear speculum and provides an additional light source but relies on the iPhone camera to capture the exam. We do note that the image and video quality captured on this device is not as high as reported in previous studies (6, 42), especially with respect to sharpness and clarity. Unlike Cellscope Oto, however, more advanced high quality otoscopes with digital video capacity are often cost prohibitive to wider adoption.

4.2. Eardrum Detection

We first train a model for eardrum detection to extract eardrum patches. We use Resnet-101 as the backbone network and choose the stochastic gradient descent (SGD) optimizer with a $1e - 3$ learning rate, 0.9 momentum, and $1e - 4$ weight decay. The model is trained using all frames from the normal training set videos for 5,000 epochs with a batch size of 128. The input images are resized to 256 by 256 and randomly cropped to 224 by 224, followed by a random selection of data augmentation of color-jitter, cutout, rotation, and shearing.

Results: Since under our setup there is at most one instance of a single class, we evaluate the performance of our model by the accuracy of eardrum detection under different intersection over union (IoU) (43) thresholds. A correct detection is defined as a frame that produces a correct frame-level classification and, if the given frame shows the eardrum, also produces a bounding box with IoU score larger than a certain threshold in that given frame. As shown in **Table 1**, our model produces high accuracy across varied IoU thresholds. A visualization of eardrum detection in video sequences can be found in **Figure 4**. Note that even though the model is supervised only on instances from training normal videos, it generalizes well to both normal and abnormal videos in the testing set. This indicates that the network extracts features that are more consistent with the image semantic and agnostics to the details on the eardrum.

4.3. Anomaly Detection

Following the architecture suggested by (27), we construct the feature extractor ϕ to be the ResNet-101 convolutional neural network followed by an additional l2 normalization layer. We augment the normal samples by sequentially applying 224 by 224-pixel crop from a randomly resized image and random horizontal flips. The model is initialized with the ImageNet pre-trained model and then the two last blocks of the ResNet-101 are fine-tuned for 5,000 epochs with the loss function in Equation (7). The temperature τ in Equation (2) is set to be 0.25. We use an SGD optimizer with a learning right of $1e - 5$ weight decay of $5e - 5$ and no momentum. For each minibatch size of 60, we sample one frame from each video to prevent similar frames treated as negative pairs. During training, we use the groundtruth bounding box to extract patches from videos. We use KNN (k=2) in computing the frame-level anomaly score. We conduct experiments with three random seeds for evaluations.

Results: We evaluate our methods on the anomaly detection task in the previously introduced otoscopy video dataset. We adopt the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision Recall Curve (AUPRC) as anomaly detection performance score. We compare our approaches against the recently proposed top performing method MSC (27). The training and testing procedures are the same as ours except that it does not utilize shift-transformations and uses the angular loss (3) instead of shift angular loss (4) in computing the final loss function. In **Table 2**, we present a comparison among three variants of SCAD and three state of the art methods SSD (25), PANDA (26), MSC (27) evaluated by AUROC and AUPRC. We see that the SCAD-CJ-WF variation performs the best, reaching an AUROC of 88.0% and AUPRC

TABLE 2 | Results evaluated by AUROC% and AUPRC%.

Method	AUROC	AUPRC
SSD (25)	54.3 ± 9.6	55.2 ± 6.4
PANDA (26)	60.7 ± 0.6	58.2 ± 0.3
MSC (27)	64.3 ± 4.0	60.7 ± 2.4
SCAD-CJ-RC (ours)	78.0 ± 1.0	69.1 ± 1.1
SCAD-CJ-RR (ours)	82.3 ± 1.2	74.4 ± 1.1
SCAD-CJ-WF (ours)	$\mathbf{88.0 \pm 1.0}$	$\mathbf{87.9 \pm 1.8}$

We report the means and SD over three trials. The bold values indicate best results.
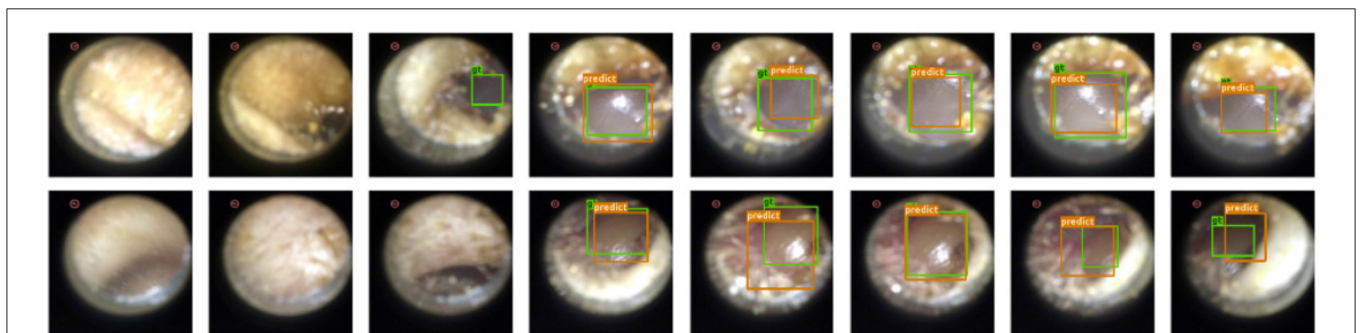

FIGURE 4 | Our detection module successfully extracts eardrum regions in video sequences. Green bounding box represents groundtruth eardrum region; orange bounding box represents detected eardrum region. **Top Row:** Normal video example. **Bottom Row:** Abnormal video example.

of 87.9% on average. This suggests that (i) color is indeed a very important feature in separating anomaly from normal in otoscopy images, and that (ii) color jitter on the whole image can produce semantically more similar samples to the real anomaly. This could be due to that both random cut and random regions introduce artificial patterns on the image that are utilized in the self-supervised training but are not seen in the abnormal samples during testing. To intuitively visualize the feature embedding produced by different methods, we plot the latent embedding for all testing frames and the corresponding kNN based frame-level anomaly score in **Figure 5**. As we can see, SCAD-CJ-WF generates the embedding that is most separable both qualitatively and quantitatively as measured by the anomaly score.

Training Objective: The individual effect of each loss component is presented in **Table 3**. We note that neither $\mathcal{L}_{msc}$ or $\mathcal{L}_{angular}$ individually performs well in our dataset. $\mathcal{L}_{shift\_angular}$ outperforms all other individual objectives and combining it with $\mathcal{L}_{msc}$ results in further improvement.

4.4. Comparison With Human Clinicians

To further evaluate our performance in a real-world setting, we performed a cross-sectional study comparing the performance of clinicians vs. our model. The study was approved by the local institutional review board.

Study Setting: Clinicians who routinely evaluate eardrums by otoscopy in a primary care, urgent care, or emergency medicine setting were invited to participate in this study. Qualifying clinicians included physicians (MD/DO), nurse practitioners (NP), and physician assistants (PAs). The is clinician group reflects frontline healthcare providers who typically diagnose and
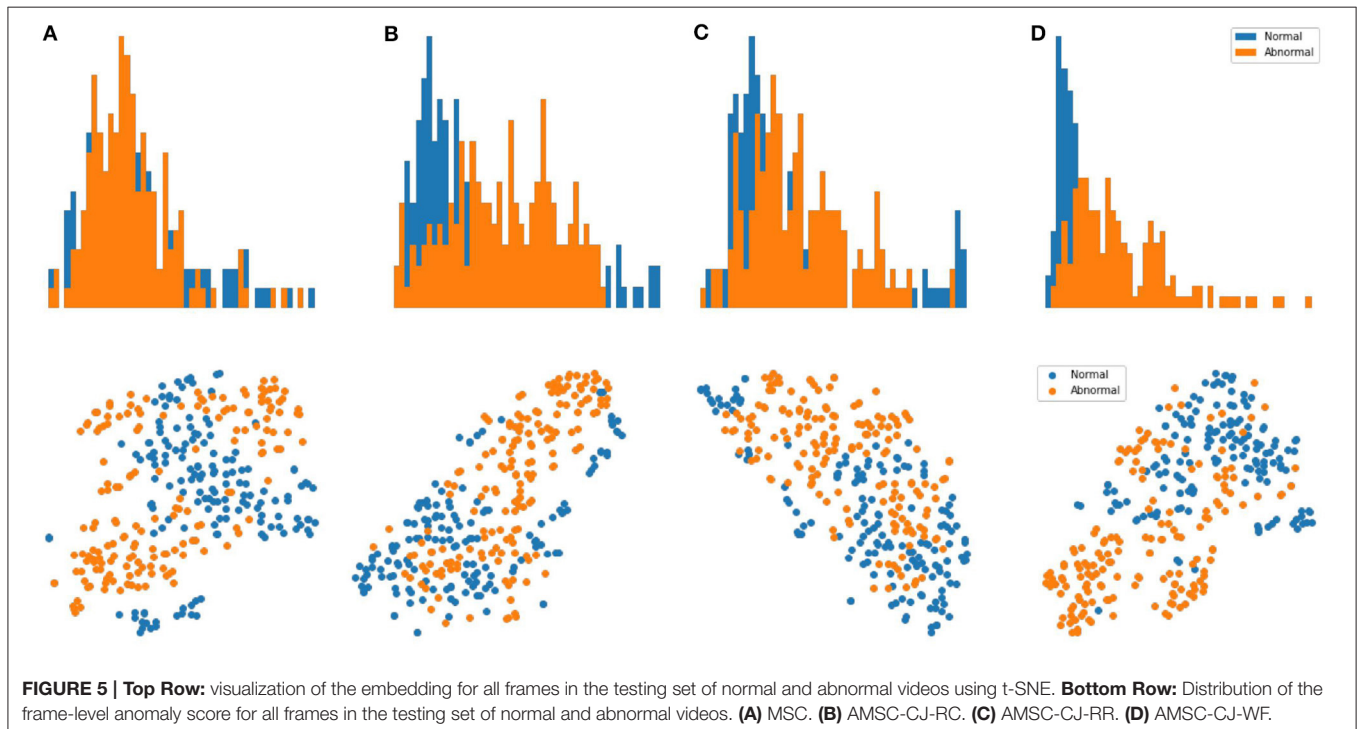
manage ear disease. Exclusion criteria were practicing outside of the United states, employment by the Johns Hopkins University or Johns Hopkins Hospital, visual impairment limiting video analysis, or non-English speakers. Recruitment took place through digital messaging *via* professional email mailing lists and professional social media groups during November 2020.

Study Procedure: Ten normal and ten videos designated as acute otitis media (AOM) by ground truth were selected as the testing set of images, as described above. Clinicians completed an online survey in which they were instructed to review the 20 eardrum videos without any further clinical information provided. Following each video, participants were asked to indicate whether the eardrums appeared normal or abnormal. They were also asked to rate their level of confidence in the accuracy of their diagnosis using a Likert scale (1= not at all confident to 4 = very confident). Participants were also asked to report the number of times they viewed each video. Participant responses were collected and managed using REDCap electronic data capture tools (44) hosted at Johns Hopkins University.

TABLE 3 | Training objective ablation study (AUROC% and AUPRC%).

Metric	$\mathcal{L}_{msc}$	$\mathcal{L}_{angular}$	$\mathcal{L}_{shift\_angular}$	$\mathcal{L}_{msc} + \mathcal{L}_{shift\_angular}$
AUROC	59.3 ± 0.6	57.7 ± 4.0	78.0 ± 1.0	$\mathbf{88.0 \pm 1.0}$
AUPRC	55.1 ± 1.6	54.5 ± 3.0	69.8 ± 0.8	$\mathbf{87.9 \pm 1.8}$

$\mathcal{L}_{shift\_angular}$ is of the SCAD-CJ-WF variant. We report the means and SD over three trials. Note that the MSC method uses $\mathcal{L}_{msc} + \mathcal{L}_{angular}$ as loss function and the result is presented in **Table 2**. The bold values indicate best results.

FIGURE 5 | **Top Row:** visualization of the embedding for all frames in the testing set of normal and abnormal videos using t-SNE. **Bottom Row:** Distribution of the frame-level anomaly score for all frames in the testing set of normal and abnormal videos. **(A)** MSC. **(B)** AMSC-CJ-RC. **(C)** AMSC-CJ-RR. **(D)** AMSC-CJ-WF.

Study Population: A total of 59 clinicians were recruited and assessed for eligibility. After removing ineligible, declined, or incomplete surveys, 25 surveys were included for analysis. The process is shown in the consort flow diagram in **Figure 6**. Participant demographics are summarized in **Table 4**. All respondents are early-to-mid career physicians with the majority being female and practicing pediatrics.

Results: Clinician confidence has a mean of 3 and a SD of 0.5 and the view count per video has a mean of 1.5 and an SD of 0.5. We use linear regression to analyze the correlation of both clinicians' confidence and view count with their score. No statistically significant correlation between clinician confidence and the accuracy score ($P=.07$) or with view count and score ($P=.32$) were found (a P-value of $\leq.05$ was considered statistically significant).

In **Table 5**, we compare anomaly detection methods against clinicians in accuracy, sensitivity, specificity, and precision. We compute the metrics for all individual clinicians and then compute the mean and SD of the group. We choose the decision threshold of machine learning approaches of which the sensitivity (TPR) on the validation set to be 90%, and then report the metric values on the testing set. Compared to clinicians, our methods perform better and achieve lower variances across all metrics. We further notice that SCAD-CJ-WF outperforms both clinicians and all other methods in most metrics.

In **Figure 7**, we provide further visualizations comparing our best performing model SCAD-CJ-WF to clinicians. All seeds of our model produced a ROC curve that exceeds the average clinician response. Controlling the same False Positive Rate, our model outperforms the clinician performance in True Positive Rate (Sensitivity) by a large margin from 61.5 to 90%. Besides the improvement in average performance, our model also generates

more consistent predictions. We can see from the figure that different clinicians produce results that are not consistent with each other. While the best clinician included in our study did achieve a perfect True and False Positive Rate, the majority of clinician evaluations are scattered with a large variation. The wide variability and lower accuracy of clinician's scores from this sample are comparable to prior reports of pediatricians and general practitioners at identifying AOM from otoscopy (45, 46). This variability between clinician scores reflects the clinical challenge of accurately diagnosing AOM. Individual factors such as a clinicians' training and experience, combined with patient factors such as cooperativeness and a non-obstructed view contribute to a successful diagnosis. In contrast, our model across different seeds generated more consistent and more accurate results. Given the clinician variability to diagnose AOM

TABLE 4 | Clinician demographics and corresponding accuracy (mean% and SD%) by groups.

Characteristic	Sample size	Accuracy
Sex		
Female	15 (60%)	73.0 ± 15.1
Male	10 (40%)	60.0 ± 17.5
Current Practice		
General Pediatrics	12 (48%)	73.0 ± 15.1
Pediatric Specialty	5 (20%)	70.0 ± 22.4
General Internal Medicine	3 (12%)	55.0 ± 8.7
Others	5 (20%)	67.0 ± 8.4

Other practices include emergency medicine, family medicine, and urgent care.
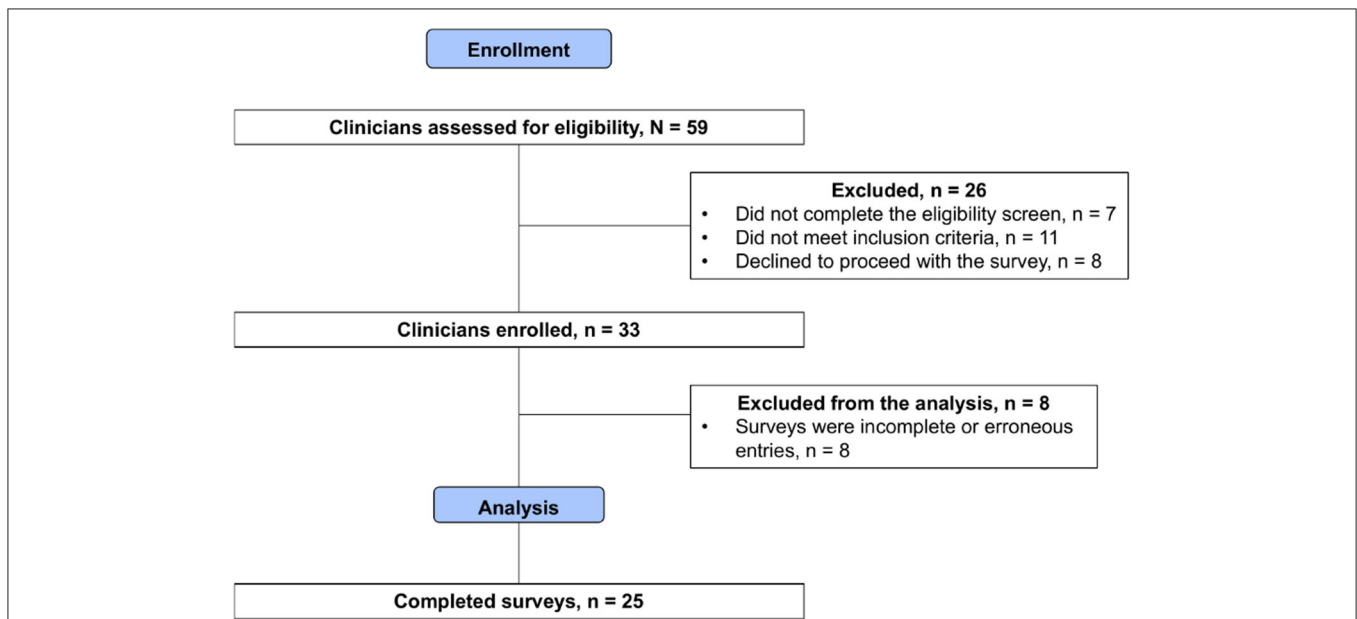

FIGURE 6 | Consort flow diagram of clinician enrollment. After removing ineligible samples, 25 surveys were included for comparative analysis.

TABLE 5 | Comparison of machine learning models against a group of clinicians (n = 25) in performances (mean and standard deviation of accuracy%, sensitivity%, specificity%, and precision%) on the testing set.

Method	Accuracy	Sensitivity	Specificity	Precision
Clinicians	67.8 ± 17.0	63.2 ± 31.8	72.4 ± 17.4	64.5 ± 24.3
SCAD Variants				
SCAD-CJ-RC (ours)	73.3 ± 2.9	70.0 ± 0.0	76.7 ± 5.8	75.2 ± 4.5
SCAD-CJ-RR (ours)	78.3 ± 7.6	$\mathbf{90.0 \pm 0.0}$	66.7 ± 15.3	73.7 ± 8.8
SCAD-CJ-WF (ours)	$\mathbf{81.7 \pm 2.9}$	83.3 ± 5.8	$\mathbf{80.0 \pm 0.0}$	$\mathbf{80.6 \pm 1.0}$

Machine learning models' decision thresholds are chosen by setting the sensitivity (TPR) on the validation set to be 90%. We report the means and SD among the group of clinicians as well as each model over three trials. The bold values indicate best results.
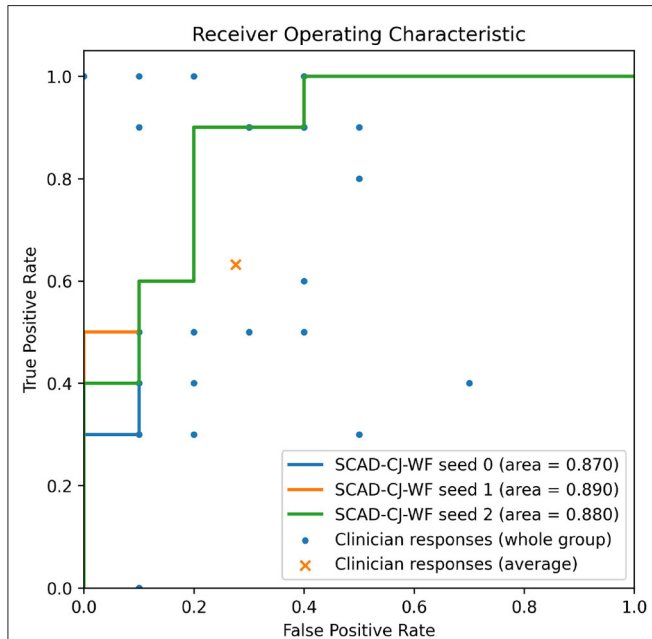
FIGURE 7 | Overlaying clinicians' TPR/FPR samples (n = 25) and group average on SCAD-CF-WF's Receiver Operative Characteristic (ROC) curve. Multiple clinicians' responses may have the same TPR/FPR values.

illustrated in our comparative study and previous studies, there is significant value to improve diagnosis in a clinical setting by adopting deep learning based anomaly detection screening like that of the present study.

Discussion: Our presented method shows promising performance in anomaly detection on otoscopy video sequences. The whole workflow runs at 15 ms per frame on a single NVIDIA Quadro RTX 6000 GPU. This suggests the feasibility of adopting computer-aided diagnostics to real time detect and report anomalies may assist front-line clinicians in primary care and eventually even patients at home.

The eventual successful deployment of such technology necessitates more than the successful development of a diagnostic AI algorithm with high accuracy. For such technology to integrate within current healthcare models and work flow, there is a need to create a whole new technological infrastructure.

The implementation of such technology beyond its obvious need to be adopted by healthcare providers, perhaps the more pressing questions where the required financial investment for the creation of this infrastructure will come from.

Meanwhile, our study exhibits several limitations at the current stage. (i) The dataset is relatively small, skewed toward normal, and collected through convenience sampling. As such our data serve to provide some additional support for the proof of concept but does not negate the need for a larger and more rigorously collected data-set before such technology becomes clinically viable. The development of a large video database would likely further improve the algorithm's diagnostic performance and enable training of specific diagnoses as opposed to being limited to detecting normal vs. abnormal ear exams. (ii) Cellscope Oto does not permit pneumotoscopy (blow air at the eardrum to evaluate for movement) which can also serve as a critical step for accurate diagnosis of ear disease. (iii) In our comparative study, clinicians were not provided with patient history which might increase their diagnostic accuracy in a clinical setting.

5. CONCLUSION

In this study, we present a two-stage anomaly detection method that is designed to perform normal/abnormal classification for otoscopy videos and that can be developed based on a small dataset and highly skewed class distribution. We demonstrate that our method outperforms baseline algorithms and the average of 25 clinicians, constituting a promising step toward a computer-aided diagnosis of the middle and external ear pathology. The further development of computer-aided diagnosis methods, like ours, could contribute to timely and high-quality management of ear disease.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the institutional review board restrict data to be accessed only by approved personals. Requests to access the datasets should be directed to Weiyao Wang, wwang121@jhu.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Johns Hopkins Medicine Institutional Review Boards. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

WW, AT, and MU contribute to this study on the algorithm design, implementation, and evaluation. CS, JC, and TC contribute to this study on the clinical comparative

study's organizing, data collection, and processing. JC contributes to this study on the hardware integration. All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

1. Gurnaney H, Spor D, Johnson DG, Propp R. Diagnostic accuracy and the observation option in acute otitis media: the Capital Region Otitis Project. *Int J Pediatr Otorhinolaryngol.* (2004) 68:1315–25. doi: 10.1016/j.ijporl.2004.05.005

2. Brinker Jr DL, MacGeorge EL, Hackman N. Diagnostic accuracy, prescription behavior, and watchful waiting efficacy for pediatric acute otitis media. *Clin Pediatr.* (2019) 58:60–5. doi: 10.1177/0009922818806312

3. Poole NM, Shapiro DJ, Fleming-Dutra KE, Hicks LA, Hersh AL, Kronman MP. Antibiotic prescribing for children in United States emergency departments: 2009–2014. *Pediatrics.* (2019) 143:e20181056. doi: 10.1542/peds.2018-1056

4. Rosenfeld RM, Culpepper L, Doyle KJ, Grundfast KM, Hoberman A, Kenna MA, et al. Clinical practice guideline: otitis media with effusion. *Otolaryngol Head Neck Surgery.* (2004) 130:S95–S118. doi: 10.1177/0194599815623467

5. Canares T, Wang W, Unberath M, Clark J. Artificial intelligence to diagnose ear disease using otoscopic image analysis: a review. *J Investig Med.* (2021) 14:jim-2021-001870. doi: 10.1136/jim-2021-001870

6. Viscaino M, Maass JC, Delano PH, Torrente M, Stott C, Cheein FA. Computer-aided diagnosis of external and middle ear conditions: a machine learning approach. *PLoS ONE.* (2020) 15: e0229226. doi: 10.1371/journal.pone.0229226

7. Zafer C. Fusing fine-tuned deep features for recognizing different tympanic membranes. *Biocybern Biomed Eng.* (2020) 40:40–51. doi: 10.1016/j.bbe.2019.11.001

8. Khan MA, Kwon S, Choo J, Hong SM, Kang SH, Park IH, et al. Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Netw.* (2020) 126:384–94. doi: 10.1016/j.neunet.2020.03.023

9. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Honolulu, HI (2017). p. 4700–08.

10. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision.* Venice (2017). p. 618–26.

11. Cha D, Pae C, Seong SB, Choi JY, Park HJ. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine.* (2019) 45:606–14. doi: 10.1016/j.ebiom.2019.06.050

12. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition.* Miami, FL: IEEE (2009). p. 248–55.

13. Chalapathy R, Chawla S. Deep learning for anomaly detection: a survey. *arXiv [preprint] arXiv:190103407.* (2019).

14. Ruff L, Kauffmann JR, Vandermeulen RA, Montavon G, Samek W, Kloft M, et al. A unifying review of deep and shallow anomaly detection. *Proc IEEE.* (2021) 109: 756–95. doi: 10.1109/JPROC.2021.3052449

15. Zong B, Song Q, Min MR, Cheng W, Lumezanu C, Cho D, et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: *International Conference on Learning Representations.* Vancouver, BC (2018).

16. Zhai S, Cheng Y, Lu W, Zhang Z. Deep structured energy based models for anomaly detection. In: *International Conference on Machine Learning.* New York, NY: PMLR (2016). p. 1100–9.

17. Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui SA, Binder A, et al. Deep one-class classification. In: Dy J, and Krause A, editors. *Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research.* Vol. 80. Stockholm: PMLR (2018). p. 4393–402. Available online at: http://proceedings.mlr.press/v80/ruff18a.html

18. Sohn K, Li CL, Yoon J, Jin M, Pfister T. Learning and evaluating representations for deep one-class classification. *arXiv [preprint] arXiv:201102578.* (2020).

19. Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: *International Conference on Information Processing in Medical Imaging.* Boone, NC: Springer (2017). p. 146–57.

20. Kolesnikov A, Zhai X, Beyer L. Revisiting self-supervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Long Beach, CA (2019). p. 1920–9.

21. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning.* PMLR (2020). p. 1597–607.

22. Choi S, Chung SY. Novelty detection via blurring. *arXiv [preprint] arXiv:191111943.* (2019).

23. Li CL, Sohn K, Yoon J, Pfister T. CutPaste: self-supervised learning for anomaly detection and localization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Nashville, TN (2021).

24. Tack J, Mo S, Jeong J, Shin J. CSI: novelty detection via contrastive learning on distributionally shifted instances. In: *Advances in Neural Information Processing Systems.* Curran Associates (2020).

25. Sehwag V, Chiang M, Mittali P. SSD: a unified framework for self-supervised outlier detection. In: *International Conference on Learning Representations.* (2021).

26. Reiss T, Cohen N, Bergman L, Hoshen Y. PANDA: adapting pretrained features for anomaly detection and segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Nashville, TN (2021). p. 2806–14.

27. Reiss T, Hoshen Y. Mean-shifted contrastive loss for anomaly detection. *arXiv [preprint] arXiv:210603844.* (2021).

28. Fernando T, Gammulle H, Denman S, Sridharan S, Fookes C. Deep learning for medical anomaly detection—a survey. *arXiv [preprint] arXiv:201202364.* (2020).

29. Davletshina D, Melnychuk V, Tran V, Singla H, Berrendorf M, Faerman E, et al. Unsupervised anomaly detection for X-ray images. *arXiv [preprint] arXiv:200110883.* (2020).

30. Bozorgtabar B, Mahapatra D, Vray G, Thiran JP. Anomaly Detection on X-Rays Using Self-Supervised Aggregation Learning. *arXiv [preprint] arXiv:201009856.* (2020).

31. Sato D, Hanaoka S, Nomura Y, Takenaga T, Miki S, Yoshikawa T, et al. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes. In: *Medical Imaging 2018: Computer-Aided Diagnosis.* Vol. 10575. Houston, TX: International Society for Optics and Photonics (2018). p. 105751P.

32. Pawlowski N, Lee MC, Rajchl M, McDonagh S, Ferrante E, Kamnitsas K, et al. *Unsupervised lesion Detection in Brain CT Using Bayesian Convolutional Autoencoders.* Amsterdam: Medical Imaging with Deep Learning Abstract Track (2018). Available online at: https://openreview.net/references/pdf?id=S1hpzoisz

33. Baur C, Denner S, Wiestler B, Navab N, Albarqouni S. Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Med Image Anal.* (2021) 14:101952. doi: 10.1016/j.media.2020.101952

34. Han C, Rundo L, Murao K, Noguchi T, Shimahara Y, Milacski ZÁ, et al. MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. *BMC Bioinf.* (2021) 22:1–20. doi: 10.1186/s12859-020-03936-1

35. Baur C, Graf R, Wiestler B, Albarqouni S, Navab N. SteGANomaly: Inhibiting CycleGAN steganography for unsupervised anomaly detection in brain MRI. In: *International Conference on Medical Image*

Computing and Computer-Assisted Intervention. Lima: Springer (2020). p. 718–27.

36. Liu Y, Tian Y, Maicas G, Pu LZ, Singh R, Verjans JW, et al. Unsupervised dual adversarial learning for anomaly detection in colonoscopy video frames. *arXiv [preprint] arXiv*:191010345. (2019).

37. Zimmerer D, Kohl SA, Petersen J, Isensee F, Maier-Hein KH. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint.* arXiv:181205941. (2018).

38. Tan J, Hou B, Batten J, Qiu H, Kainz B. Detecting outliers with foreign patch interpolation. *arXiv [preprint] arXiv*:201104197. (2020).

39. Tan J, Hou B, Day T, Simpson J, Rueckert D, Kainz B. Detecting outliers with poisson image interpolation. *arXiv [preprint] arXiv*:210702622. (2021).

40. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* Seattle, WA (2020).

41. Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images. (2009). Available online at: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

42. Livingstone D, Talai AS, Chau J. Building an Otoscopic screening prototype tool using deep learning. *J Otolaryngol Head Neck Surg.* (2019) 48:66. doi: 10.1186/s40463-019-0389-9

43. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bounding box regression. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Long Beach, CA (2019). p. 658–66.

44. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inf.* (2009) 42:2377–381. doi: 10.1016/j.jbi.2008.08.010

45. Pichichero ME, Poole MD. Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination. *Int J Pediatr Otorhinolaryngol.* (2005) 69:361–6. doi: 10.1016/j.ijporl.2004.10.013

46. Pichichero ME, Poole MD. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. *Arch Pediatr Adolesc Med.* (2001) 155:1137–42. doi: 10.1001/archpedi.155.10.1137