IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 24, 2021, accepted November 21, 2021, date of publication December 1, 2021,
date of current version December 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3132133

Computer-Aided Ear Diagnosis System Based on CNN-LSTM Hybrid Learning Framework for Video Otoscopy Examination

MICHELLE VISCAINO[1,2], (Graduate Student Member, IEEE), JUAN C. MAASS[3,4,5], PAUL H. DELANO[2,5,6], AND FERNANDO AUAT CHEEIN[1,2], (Senior Member, IEEE)

[1]Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso 2340000, Chile
[2]Advanced Center of Electrical and Electronic Engineering, Valparaíso 2340000, Chile
[3]Interdisciplinary Program of Physiology and Biophysics, Institute of Biomedical Sciences (ICBM), Faculty of Medicine, Universidad de Chile, Santiago 8380453, Chile
[4]Department of Surgery, Clínica Alemana de Santiago, Unit of Otolaryngology, Facultad de Medicina Clínica Alemana-Universidad del Desarrollo, Santiago 7610658, Chile
[5]Department of Otolaryngology, Hospital Clínico Universidad de Chile, Faculty of Medicine, Universidad de Chile, Santiago 3659, Chile
[6]Department of Neuroscience, Faculty of Medicine, Universidad de Chile, Santiago 3659, Chile

Corresponding author: Fernando Auat Cheein (fernando.auat@usm.cl)

This work was supported in part by the Chilean National Agency for Research and Development (ANID) (ex CONICYT) under Grant FB0008; in part by the CONICYT-PCHA/Doctorado Nacional/2018-21181420; and in part by the Federico Santa Maria Technical University (UTFSM), Chile, under Grant DGIIP-PIIC-28/2021.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Scientific Research Ethics Committee of Clinical Hospital of the University of Chile under Application No. 65 - 996/18, and performed in line with the national regulations and the declaration of Helsinki, revision 2013.

ABSTRACT Ear disorders are among the most common diseases treated in primary care, with a high percentage of non-relevant referrals. The conventional diagnostic procedure is done by a visual examination of the ear canal and tympanic membrane. Consequently, the accuracy of the diagnosis is affected by observer-observer variation, depending on the technical skill and experiences of the physician as well as on the subjective bias of the observer. This situation impacts the proper implementation of treatments, increases health costs, and can lead to serious health complications. To eliminate subjectivity and enhance diagnostic accuracy, we present a diagnostic tool for nine ear conditions in a computer-aided diagnosis scheme. We propose a hybrid learning framework based on convolutional and recurrent neural networks for video otoscopy analysis. The proposed method first extracts the deep features of each relevant frame from the video. Then, a Long Short-term Memory network is introduced to learn spatial sequential data by analyzing deep features for a certain time interval. We carried out the study in collaboration with the Clinical Hospital of the University of Chile and included 875 subjects in a period of 12 months (continuous). The experiments were conducted on a new video otoscopy dataset and showed high performance in terms of accuracy (98.15%), precision (91.94%), sensitivity (91.67%), specificity (98.96%), and F1-score (91.51%). To the best of our knowledge, the proposed system is capable of predicting more diagnoses of ear conditions known to date with high performance. Our system is designed to assist in a real otoscopy examination by analyzing a sequence of images instead of a still image as previous state-of-the-art works. This advantage allows it to provide a comprehensive diagnosis of both eardrum and ear canal diseases.

INDEX TERMS Computer-aided diagnosis, convolutional neural network, deep learning, ear diseases, LSTM, otolaryngology, transfer learning.

I. INTRODUCTION

Ear diseases are among the most frequent pathologies treated in primary care with a high percentage of non-relevant referrals, including tympanic membrane and external auditory canal abnormalities [2], [3]. Generally, the diagnosis of ear diseases is carried out by a medical interview and an otoscopic examination. Otolaryngologists and general practitioners perform otoscopic examinations daily as a part of routine care [4]. However, the diagnosis using common tools

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

such as otoscopy or even otoendoscopy is susceptible to misdiagnosis due to its dependence on the technical skills and experience of the physician as well as the observer subjective bias [5]. According to the studies presented in [6] and [7], the correct diagnosis rate of non-otolaryngologists such as pediatricians is 50%, and although the rate increases when the diagnosis is performed by an otolaryngologist (73%), these values are still low. A misdiagnosis impairs the appropriate implementation of treatments and health outcomes in severe physical, psychological, financial, and legal complications. The latter means that there is an opportunity to introduce new tools to support the physician in making correct decisions and improving diagnostic accuracy.

Computer-aided diagnosis system (CADx) with clinical purposes has been under development for many years [8]. CADx systems allow the characterization of a region and serve as a second opinion to physicians who take the final decision [9]. Conventionally, CADx is based on computer algorithms designed by domain experts for particular feature extraction from medical digitalized data [10]. However, models based on deep neural networks (DNNs) have had a significant impact on enhancing the performance of CADx systems [11].

One of the most popular DNNs architectures is the convolutional neural networks (CNNs). They have been widely used for pattern recognition and image classification [12]. Their design allows for processing data using spatial information by taking 2D or 3D images as input. In classification and detection tasks, CNNs have achieved high performance in terms of inference times and detection rates outperform even traditional computer vision methods [13]. However, the major limitation to apply CNNs is the large amount of labeled data needed to obtain accurate and generalizable models [12] (in medicine, such data should be carefully reviewed and labeled by experts in the field). To overcome the latter issue, the transfer learning technique has been successfully applied to use pre-trained models on huge datasets (e.g., ImageNet that contains more than millions of natural images) [14]. Such a technique consists of transferring the knowledge (weights) learned during the training using millions of samples to a different but related problem (e.g., image classification).

Recurrent neural networks (RNN) are another important DNN architecture that have been successfully applied in sequence analysis. They contain blocks of connected neurons with input units, hidden units, and output units which share the same weights across all steps [15]. Thus, there is a reduced number of trainable parameters compared to CNN models. The main advantage of RNN is the ability to process data sequentially at each time, considering two sources of input –present and recent past– to provide the output of the new data [12]. However, the RNN standard model suffers from vanishing and exploding gradient problems [16]. An improved version of RNN that overcomes these limitations is the Long Short-Term Memory (LSTM) network. The architecture of LSTM also includes a memory cell which is capable of maintaining long-term dependencies. The information on the memory cell is carefully regulated by nonlinear functions typically called gates [17]. LSTM networks have shown remarkable effectiveness in language processing tasks, and voice recognition [16].

Applications that involve sequential data use an LSTM network for their ability to learn long-term dependencies. However, the LSTM networks are designed to deal with one-dimensional data [17]. When they are used to process video or sequential images, the spatial features of the input data are lost. Therefore, in applications such as video action recognition or video classification where both spatial-temporal features are relevant, a combined architecture of CNN and LSTM has been used [15], [18]. One architecture that combines the capabilities of CNN and the LSTM network is ConvLSTM [19]. It is a variation of the LSTM network that applies convolution operation instead of only multiplications both at the input-to-state transition and state-to-state transitions. The latter solves the initial input vectorization problem and allows 2D or 3D processing. Another approach that allows learning capabilities to complement each other is the combined CNN-LSTM architecture. Such architecture stacks one block after another. The first part of the model (i.e., CNN) extracts the relevant features of the input data. Then, the results are flattened in a 1-D tensor to be used as input for the second part of the model (i.e., LSTM) [18]. Comparison results of the two approaches show that both are suitable for spatial-temporal data modeling. Nevertheless, the time processing and memory resources of a ConvLSTM increase compared to CNN-LSTM depending on the size of the input image or the length of sequences [20]. Furthermore, the CNN-LSTM approach has been used for classification tasks as action recognition [15], [18] due to the computational capacity of the feature map given by the first part of the model. Whereas ConvLSTM fits better in prediction tasks where the input is 2D or 3D sequential data [20].

II. RELATED WORK

In recent years, the interest in developing computer-aided diagnosis systems for ear pathologies has been growing. Early approaches addressed a binary classification to determine whether an image of the tympanic membrane is normal or abnormal [21] (and the references therein). Models with multi-class approaches have also been developed to predict otitis media diseases such as chronic otitis media, acute otitis media, and otitis media with effusion [1], [31]. Such schemes used features extractors as local descriptors of texture or color, and conventional machine learning models (e.g., support vector machine). Although the performance of the proposed models has been improving, the datasets used are small in order to generalize results.

Other recently developed schemes have used CNN to classify the image of the eardrum considering a binary approach [24], [25] or multi-class problem [23], [26]. However, the pathologies studied are mostly limited to otitis media and its derivations. A more complete study was presented

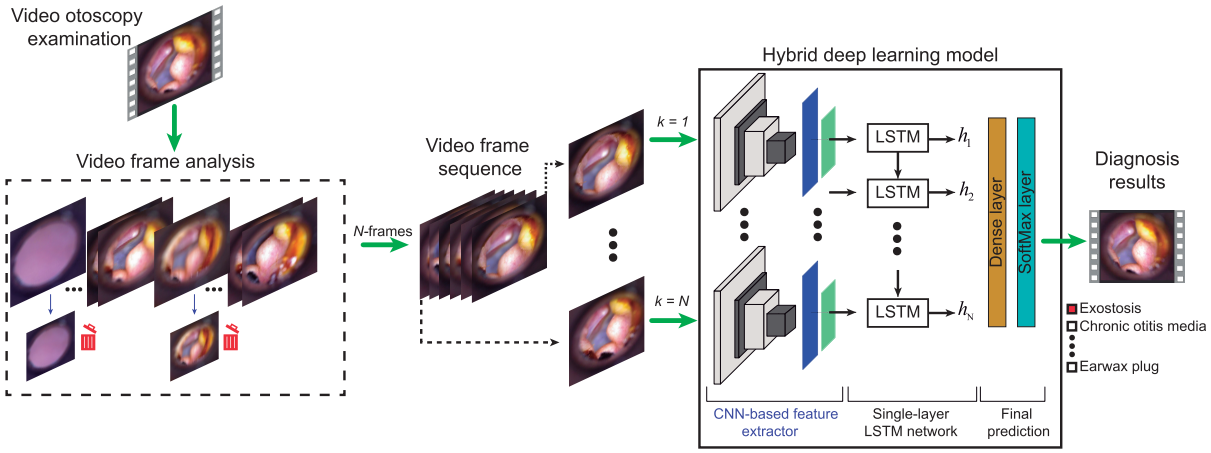

FIGURE 1. Deep learning framework to infer the diagnosis of the middle ear and ear canal diseases. The spatial sequential data obtained through CNN are passed to an LSTM model to produce the final decision (i.e., the diagnosis).

in [7] where other relevant pathologies such as tympanic perforation or attic retraction were also included. The accuracy achieved in previous studies remained in the range of 84% to 95% covering up to 6 middle ear pathologies. Not only diagnosis-oriented work has been developed, but the work presented in [44] also showed that CNNs are a useful tool to identify the tympanic membrane side and tympanic perforation with an average accuracy of 97% and 91% respectively.

The models previously proposed for the diagnosis of ear diseases based on still images of the eardrum, are a good starting point but are impractical in the clinical environment. During the ear examination, the physician observes the ear canal and tympanic membrane from different points of view to achieve a diagnosis. When using a digital otoscope, this means acquiring a certain number of images in a time interval (video) that must be analysed by the learning architecture to reach a complete diagnosis (i.e. describe the condition of both the ear canal and the tympanic membrane). In this paper, we propose a computer-aided diagnosis scheme for the classification of nine ear conditions on video otoscopy. This scheme is based on a hybrid learning model that exploits the ability of CNNs to extract deep features as well as LSTM networks to learn hidden patterns in spatial sequential data (i.e., from videos). We use a CNN model via transfer learning for feature extraction following by a customized single layer LSTM network and a final SoftMax layer to predict the class (diagnosis). To select the CNN model, we compare four state-of-the-art baselines deep learning architectures for image classification. The proposed scheme was trained, validated, and tested in a new large video otoscopy dataset, created in collaboration with physicians from the Otolaryngology the Clinical Hospital of the University of Chile (HCUCH). The contributions of this paper are summarized as follows:

1. We took into account the ear examination workflow to propose a computer-aided diagnostic system for nine ear conditions ranging from typical ear canal conditions such as earwax plug to more critical diseases of the tympanic membrane such as chronic otitis media. The system capable of predicting more diagnoses for ear conditions known to date.
2. We introduce a hybrid framework to exploit the main advantages of CNN and RNN (LSTM) networks in spatial sequential data to provide a robust learning model for video otoscopy classification.
3. We analyse the performance of the proposed learning framework by varying the amount of information it receives (frames), and establish an adequate working point for the prediction of ear pathologies in a video sequence. In this sense, we aim to eliminate redundant information while maintaining adequate performance.

The paper is organized as follows: the video otoscopy dataset, as well as the learning proposed framework, are described and analysed in detail in Section III. The results are presented in Section IV, and the discussion showing the strengths and limitations of the proposed system is discussed in Section V. Finally, the conclusions of our work are presented in Section VI.

III. METHODS

The proposed computer-aided diagnosis scheme consists of two stages: video frame evaluation and video sequence prediction, as shown in Fig. 1. In the first stage, non-relevant frames are discarded and the level of blurring of each frame is evaluated. A sequence of sharp frames is then formed and passed through a CNN model to extract deep features. Finally, a single-layer LSTM network trained on the spatial data from the video sequence predicts the diagnosis.

A. VIDEO OTOSCOPY DATASET

Our medical collaborators were responsible for the design and execution of the protocol to conduct the data collection. The protocol (number 65 - 996/18) was reviewed and approved by the Scientific Research Ethics Committee of the HCUCH.

The cohort included 875 patients who consulted the HCUCH for otolaryngological care between December 2018 until December 2019. Patients were selected if they presented clinical features related to any of the following ear conditions: normal, myringosclerosis, monomeric or dimeric (mono-dimeric) membrane, chronic otitis media (COM), otitis media with effusion (OME), external otitis (EO), exostosis of the ear canal, tympanic membrane retraction (TMR), earwax plug, acute otitis media (AOM), osteoma osteoid, and foreign body in the ear. The category "other" was also included to address fewer common diseases such as myringitis.

Three data acquisition equipment was placed in different consultation rooms of the HCUCH to record the otoscopic examination. Each kit consisted of a digital otoscope (DE500 Firefly with a 4 mm and 5 mm speculum) connected to a computer. The videos were recorded at 20 fps (frames per second) with a 640 × 480-pixel resolution.

The otolaryngologist recorded two videos per patient (one per ear) as well as the corresponding diagnosis. In cases where the patient showed more than one ear condition, the specialist registered up to three different diagnoses in order of priority.

Almost 50% of the patients presented normal ear conditions during data collection, which resulted in an unbalanced dataset. To avoid bias in the results, a similar number of videos were selected for all classes. Also, four categories were discarded because the number of videos recorded was less than 10 (insufficient for training stages).

TABLE 1. Distribution of video otoscopy dataset.

Ear condition	Total videos	Videos used
Normal	712	54
Earwax plug	223	59
Chronic otitis media	130	51
Myringosclerosis	110	49
Mono-dimeric	60	35
Tympanic membrane retraction	59	29
Otitis media with effusion	58	36
Exostosis	53	47
External otitis	27	25
Total	1423	385

Table 1 shows the number of videos recorded by pathology as well as the number of videos used for this study. To avoid the multi-label cases, we selected those videos that presented a single diagnose. The resulting video otoscopy dataset consists of 385 video otoscopic examinations from 341 patients. An image example obtained from the digital otoscope per each ear condition is shown in Fig. 2.

B. VIDEO FRAME EVALUATION

One of the typical behavior of the physician during the examination was starting the exploration recording before inserting the otoscope into the ear canal. Unfortunately, such action produced non-relevant frames for this study (i.e., frames that do not show the ear canal or tympanic membrane). Therefore, those frames were discarded using a histogram-based algorithm, which analyzed the domain of interest of the image.

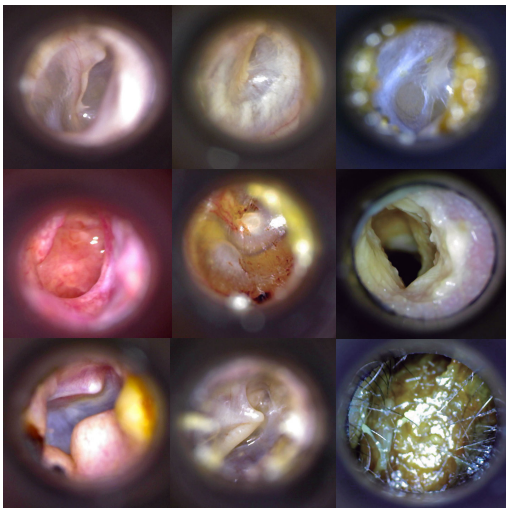

FIGURE 2. Eardrum image samples in the video otoscopy dataset. From left to right, first row: normal, myringosclerosis, mono-dimeric; second row: chronic otitis media, otitis media with effusion, external otitis; third row: exostosis of ear canal, tympanic retraction and earwax.

Another important aspect to keep in mind during the exam is that the physician handled the otoscope to obtain different eardrum views, enriching the dataset with spatial information and producing some undesirable blurred frames. Therefore, we preprocessed the data to discard non-relevant frames and evaluate the level of blurring of the frames.

1) IMAGE DOMAIN EVALUATION

To determine whether an image belongs to the domain of interest or not, we followed the steps proposed in Algorithm 1. The input is the structure B_k which contains frames extracted from the video sequence. The variable th_R is a predefined threshold (line 1) that allows us to add or discard the current frame. First, we initialize the empty B_N structure (line 2) in which the frames within the domain of

Algorithm 1 Algorithm for Image Pre-Processing: Discarding Unnecessary Images

Input: B_k {k extracted frames from video otoscopy}
1: Let th_R design parameter
2: initialize structure $B_N \leftarrow \emptyset$
3: $I_R = \text{FindImageReference}(B_k)$
4: $H_R = \text{CalcHistogram}(I_R)$ {Q with a pdf $q(x)$}
5: **for** $j = 1$ to k **do** {k: no. of video frames}
6: $I_j = B_k(j)$
7: $H_j = \text{CalcHistogram}(I_j)$
8: $D_{KL} = \text{CalcDKLscore}(H_j, H_R)$
9: **if** $(D_{KL} \leq th_R)$ **then**
10: Append I_j to B_N
11: **end if**
12: **end for**
Output: B_N {N keyframes}

interest will be added. Because the frames of no interest are at the beginning or end of each video (as a consequence of the diagnostic procedure), we assume that the central frame of the sequence B_k shows the eardrum. Therefore, the variable I_R represents the central frame from the sequence frames (line 3). Then, in line 4, we compute the histogram of I_R, which is used as a reference to determine the difference between two frames by comparing their histograms. Finally, the histogram of each frame is compared to the reference (lines 6-8) using the Kullback-Leibler divergence score –also known as relative entropy– [28].

The relative entropy D_{KL} is used to compare two probability distributions. Let P and Q two discrete probability distributions defined on the same probability space χ, the D_{KL} score is defined as follows:

$$D_{KL}(P, Q) = \sum_{x \in \chi} P(x) \, log\left(\frac{P(x)}{Q(x)}\right) \tag{1}$$

The D_{KL} score results in a positive quantity and is equal to 0 if and only if $P = Q$ almost everywhere [28].

If the D_{KL} score exceeds the th_R threshold, the image is considered not relevant and is discarded (lines 9-11). At the end of this stage, a sequence of N frames is obtained showing the tympanic membrane and the ear canal (keyframes).

2) IMAGE BLUR EVALUATION

During the acquisition stage, blurry frames occurred due to the camera's movement or to certain regions of the scene that were outside the camera's focus. Although there are frames with non-uniform blurring, we are interested in evaluating those frames with uniform blurring by analyzing the whole image.

Algorithm 2 shows the steps to evaluate frame blur. The input is the sequence of keyframes obtained in the previous stage. We define the th_B threshold (line 1) as a design parameter that has been fine-tuned on a subset of the video otoscopy dataset. The variable B_k is an empty structure (line 2) where incoming data (video frames and blur evaluation) will be stored according to the following conditions. First, we compute the DFT for each keyframe using the Fast Fourier Transform (FFT) algorithm (lines 4 and 5). Then, we shift the zero-frequency component (DC component) to match the center of the image and remove the low frequencies (lines 6 and 7). In the next step (line 8), we put the DC component back in the original position and compute the inverse FFT to reconstruct the image. The magnitude spectrum of the reconstructed image is computed (line 9), and its mean value is compared with a certain threshold value (th_B). If the mean value exceeds the threshold, the image is considered sharp, otherwise blurred. Finally, both the image and blur evaluation are saved in the B_L structure.

After blurriness evaluation, the complete video sequence was discarded if it had an excessive number of blurred frames (>60% of total frames).

Algorithm 2 Algorithm for Image Pre-Processing: Blurriness Detection

Input: B_N {N keyframes}

1: Let th_B design parameter
2: initialize structure $B_L \leftarrow \emptyset$
3: **for** $i = 1$ to N **do** {N: no. of keyframes}
4: $I_i = B_N(i)$ {i_{th} frame}
5: $I_i^f = \text{ComputeFFT}(I_i)$
6: $I_i^c = \text{CenterFFT}(I_i^f)$
7: $I_i^z = \text{RemoveLowFrequencies}(I_i^c)$
8: $I_i^r = \text{ReconstructInverseFFT}(I_i^z)$
9: $S = \text{MagnitudSpectrum}(I_i^r)$
10: **if** $(mean(S) \geq th_B)$ **then**
11: Append I_i, *sharp* to B_L
12: **else**
13: Append I_i, *blur* to B_L
14: **end if**
15: **end for**

Output: B_L {N frames, blur evaluation}

C. HYBRID DEEP LEARNING FRAMEWORK

The architecture is composed of two separate steps. In the first step, the spatial features of each keyframe are extracted using a pre-trained CNN model. Such deep features of the N frames constitute the sequential spatial data that are passed through the LSTM network for final prediction. We discuss each step in detail below.

1) FEATURES EXTRACTION VIA CNN MODEL

The CNN model takes advantage of the spatial and configuration information of the input data by not resorting to initial vectorization. Therefore, it is widely used in applications where the input data is 2D or 3D images because it exploits hidden spatial information among neighbouring pixels.

In classification tasks, CNN models can be interpreted as two sub-process: deep feature extraction and classification. The former includes a set of convolutional layers interspersed with polling layers, and the next stage uses a combination of fully connected and SoftMax layers. In this way, the discriminative features of the input data obtained from the extraction step are summarized into a one-dimensional vector which feeds a fully connected layer and then a SoftMax layer (the architecture can have more than one hidden layer in the last part).

Our proposal includes the Inception-V3 [29] and ResNet50 [30] models as baselines since they have been previously used for tympanic image classification [7], [24], [26]. In addition, we incorporate two more recent state-of-the-art models DenseNet-121 [31] and EfficienNet-B4 [32]. Table 2 summarizes the main characteristics of the four CNN models implemented. The inception model incorporates kernels of different sizes within the same layer to effectively recognize salient features in the image (i.e., inception modules) [29]. ResNet model includes shortcut connections, typically called skip connections, that allow very deep architectures to be

TABLE 2. Main characteristics of the four CNN models implemented.

CNN	Deep	Input size	# FC nodes	Parameters (Millions)
Inception-V3	159	299 × 299	2048	23
ResNet-50	50	224 × 224	2048	25
DenseNet-121	121	224 × 224	1024	8
EfficientNet-B4	-	380 × 380	1796	19

trained without the problem of vanishing gradients [30]. DenseNet model uses dense connections between layers (via dense blocks), where all layers are directly connected. The latter allows feature maps to be concatenated rather than aggregated [31]. Therefore the models require fewer parameters than other CNNs, although the training time is longer. In the EfficientNet model, all dimensions (depth, width, and resolution) of the network uniformly scales using a compound coefficient instead of arbitrary scales as in the conventional practice [32].

2) TRANSFER LEARNING

Although our dataset is large compared to previous state-of-the-art [1], [7], [23], [26], [44], training a CNN model from scratch requires more samples and a high computational cost. Furthermore, the use of features extractors from pre-trained CNN models with proper fine-tuning can overperform training from scratch according to [11]. Therefore, we use transfer learning to perform feature extraction of the individual N frames obtained after the preprocessing stage.

In a naive approach, the input frame is propagated forward until the final pooling layer. As a result, the output values from the pooling layer are extracted as a feature vector. The dimension of the feature vector depends on the fully connected nodes (see Table 2) in the final layer before the softmax layer. We also explored fine-tuning process to retrain the final layers –considering an ascending order– and force the network to learn patterns in the domain of our dataset.

In this sense, we performed an ablation study on the different steps for transfer learning, as shown in Fig. 3. First, we trained only the final set of fully connected layers to adjust the model's head to our dataset, whereas the rest remain frozen (i.e., keep the ImageNet weights). After the fully connected layers have started to learn patterns in our dataset, we paused training, unfreeze one by one layer or group of layers depending on each model (e.g., inception modules, residual blocks, or dense blocks), and continued training. For the latter, we decreased the learning rate (almost 10%) to avoid drastically changing the weights learned by previous convolutional layers.

Since we use the CNN model to extract features, we propagate forward the input frame until the final pooling layer of the fine-tuned model and extracte the feature vector. We repeat the process for all the N frames from each video sequence.

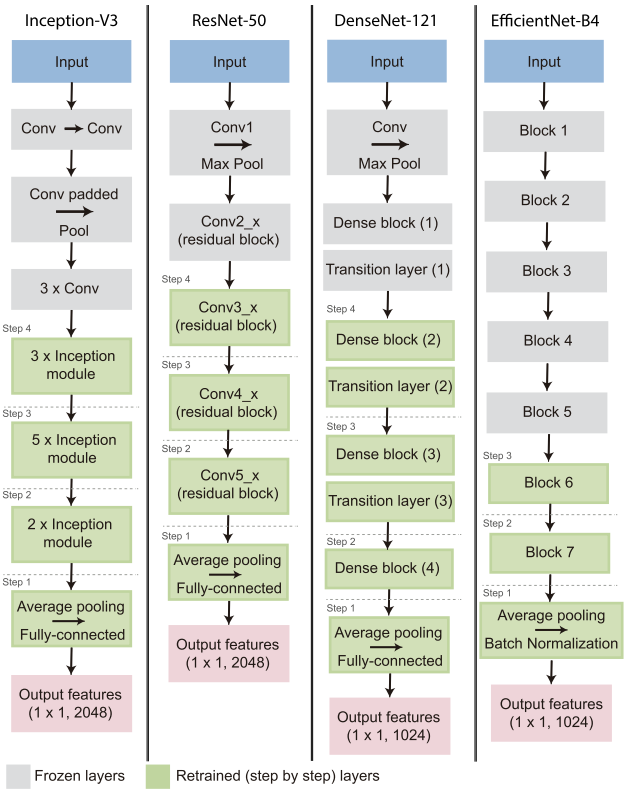

FIGURE 3. The fine-tuning process for the four CNN models. First, the fully connected layers started to learn patterns in our dataset, remaining the rest frozen with the ImageNet weights. Then, we paused training, unfreeze the next layer or group of layers, and continued training with a lower learning rate. Finally, we extracted the feature vector for each frame by propagating forward the input frame until the final pooling layer of the fine-tuned model.

3) LONG SHORT-TERM MEMORY (LSTM) NETWORK

The LSTM network is an improved version of the RNN, capable of learning long-term dependencies from sequential data [33]. It is composed of blocks with input, output, and forget gates controlling the long term sequence pattern identification. The gates are adjusted by a sigmoid unit that learns during training where it is to open and close.

Equations (2) to (5) explain the operations performed in the LSTM unit, where X_t is the input at time t (i.e., features from frames of the video otoscopy), C_t represents the content of the memory, and H_t represents the output state. Then, f_t is the output of the forget gate at time t which through a sigmoid function σ decides what information will be removed or kept in the cell (2). Then, the input gate i_t (3) generates the current memory cell C_t by computing the new candidate memory unit $\bar{C}_t$, and combining it with the old memory from the output of the forget gate (4), (5).

$$f_t = \sigma\left(W_f\left[x_t, h_{t-1}\right] + b_f\right) \tag{2}$$

$$i_t = \sigma\left(W_i\left[x_t, h_{t-1}\right] + b_i\right) \tag{3}$$

$$\bar{C}_t = \tanh\left(W_c\left[x_t, h_{t-1}\right]\right) + b_c \tag{4}$$

$$C_t = i_t \cdot \bar{C}_t + f_t \cdot C_{t-1} \tag{5}$$

Finally, the output of the LSTM unit h_t (7) is determined by filtering the current state memory using the information in the input gate (6). As the prediction of the diagnosis does not need of the intermediate output from the LSTM, we make a final decision by applying SoftMax layer on the final state of the network (8).

$$o_t = \sigma\left(W_o\left[x_t, h_{t-1}\right] + b_o\right) \tag{6}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{7}$$

$$final\_state = SoftMax\left(H_t\right) \tag{8}$$

We propose a single LSTM layer with the number of units equal to a number of fully connected nodes in the CNN (see Table 2). Then, we add a dense layer with 512 neurons. To prevent overfitting, we also include a dropout of 0.5. Finally, the model includes a SoftMax layer to assign a class (i.e., diagnosis).

D. TRAINING AND TESTING

The duration of the otoscopy videos varies according to the time required by the physician to perform ear examination. Our dataset contains videos ranging in duration from 12 to 90 s approximately. To tackle the variable duration, we divide each video into fixed-length sequences. This action causes an increase in the number of samples used for training and testing that depends on choice of sequence length (L). Considering the lower limit on the recording time, the length after preprocessing stage is 200 frames (10 s at 20 fps).

The video otoscopy dataset is randomly divided into training/validation samples and testing set, with no overlap. Following the settings as in [34], [35], the ratio between training, validation and testing sets is 70/20/10 (percentage of total data respectively). The results reported in this paper are derived after an average of 10 trials with the respective dataset partitioning.

As shown in Fig. 4, the image acquired with the otoscope has areas that are useless for analysis. Therefore, we crop the image considering the region of interest (ROI) as the circle delimited by the otoscope specula. For this purpose, we use the Hough transform as detailed in [36]. Then, the images are resized according to Table 2 to fit the input size requirements of each deep feature extractor. All images are also normalized via subtracting their mean and dividing by their standard deviation.

1) CLASS BALANCE

To minimize the risk of bias against minority classes (e.g., external otitis or tympanic membrane retraction), we balance the classes by oversampling the examples from minority classes.

A naive method is duplicated random examples from the minority class but does not provide any additional information to the model, and it can cause overfitting. Another option is to generate artificial samples using synthesis-based approaches such as SMOTE [37]. However, there is a gap between the synthetic and real samples, leading the models to learn the wrong information from the synthetic samples [38]. The latter is particularly critical in medical applications.

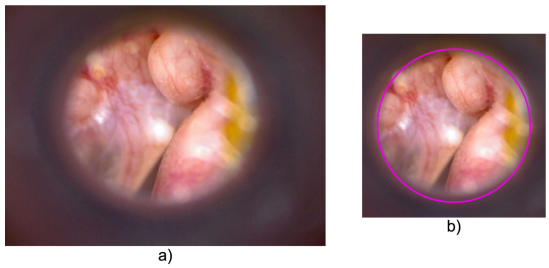

FIGURE 4. Example of images: a) raw image acquired by the digital otoscope, b) ROI-based cropped and resized image.

In this context, we use a transformation-based approach that applies a certain number of transformation operations on existing samples to generate additional samples. Such an approach is a method of data augmentation that has been widely used in deep learning models to increase the robustness and prevent the model from overfitting [9]

Table 3 shows the data distribution for the training, validation, and testing sets, as well as their respective means. To balance the data, we ensure that both the validation and testing sets have the same amount of data as their mean (33 and 16 samples, respectively) by adding (or removing) samples from the training set as appropriate. In consequence, we generated additional samples to balance the training set by considering the samples in the majority class. The transformation operations were random rotation at a certain angle range (-20 to 20 degrees) and vertical and horizontal flip.

TABLE 3. Video clips of length L frames.

Ear condition	Sequences L=200		
	Training 70%	Validation 20%	Testing 10%
Normal	141	40	20
Earwax plug	121	35	17
Chronic otitis media	95	27	14
Myringosclerosis	155	44	22
Mono-dimeric	63	18	9
Tympanic membrane retraction	151	43	22
Otitis media with effusion	86	25	12
Exostosis	104	30	15
External otitis	119	34	17
Total	1035	296	148
Mean	115	33	16

The number of samples for each class after class balance are training 173, validation 33 and testing 16.

E. PERFORMANCE EVALUATION

Although the overall performance of CNN models is quantified by accuracy, it could be affected by class distribution [40]. Therefore, we also include other evaluation metrics such as precision, specificity, and F1-score that do not depend on the class distribution. In medical applications, a relevant metric is sensitivity (also known as recall) which becomes even more important than accuracy because it quantifies the

ability of the model to reject false negatives. The expressions for computing each evaluation metric using macro-averaging method [41] are presented below:

$$Precision = \frac{\sum_{i=1}^{c} \frac{TP_i}{TP_i+FP_i}}{c} \tag{9}$$

$$Sensitivity/recall = \frac{\sum_{i=1}^{c} \frac{TP_i}{TP_i+FN_i}}{c} \tag{10}$$

$$Specificity = \frac{\sum_{i=1}^{c} \frac{TN_i}{TN_i+FP_i}}{c} \tag{11}$$

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{12}$$

where c represents the number of classes ($i = 1, \ldots, 9$), true positives (TP) means all instances correctly classified into a certain class, true negative (TN) means instances correctly classified as not belonging class, false positive (FP) means all instances incorrectly classified as belonging to a class and false-negative (FN) means all instances that were incorrectly classified as not belonging class. Although these terms are defined within a binary classification problem, their concept can be extended through the use of the one against all methods, as reported in [41].

IV. EXPERIMENTAL RESULTS

At first, we present the implementation details to run the experiments as well as the results of transfer learning during fine-tuning of the model. Then, we present the results of the performance comparison of the four implemented hybrid models. We also analyse how subsampling the data to eliminate redundancy of information can affect the performance of the model. Finally, we report the results of the final system per ear condition.

A. IMPLEMENTATION SET-UP

The proposed scheme is implemented using TensorFlow framework. The principal settings of the server include an Intel 2.9-GHz CPU and NVIDIA GeForce GTX1080 GPU.

1) FINE-TUNING CNN MODELS

The fine-tunning process of CNN models is conducted following the steps given by the ablation study proposed in subsection III-C2. We unfreeze layers or group of layers as long as the accuracy of the validation set increased. Otherwise, we stop the training process and use the model to extract features from the image sequence. As shown in Fig. 5, we plot the learning curve of each model to detect changes in the performance of the model by tracking the accuracy of the validation set. Both models, Inception-V3 and EfficientNet-B4, show similar behavior as can be seen in Fig. 5a, and Fig. 5d respectively. There are no changes in the performance after epoch 50, where the third inception module was unfreezing for the Inception-V3 model and block six was unfreezing for the EfficientNet-B4 model (step 3 in Fig. 3). On the other hand, the learning curve of ResNet-50 shows that the accuracy decrease after the third residual block

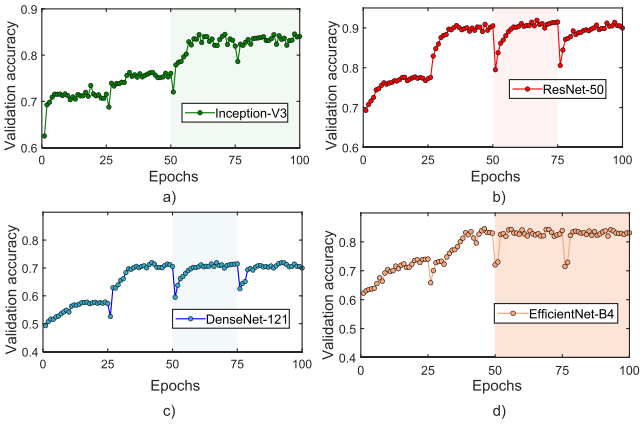

FIGURE 5. Learning curve of validation accuracy during ablation study on four steps of transfer learning.

(step 4 in Fig. 3) was unfreezing. Finally, the validation accuracy of DenseNet-121 remains constant after unfreezing the third dense block corresponding to step 3 in Fig. 3.

We set the hyperparameters in a preliminary trial using a smaller dataset. They are adjusted from values given in [29]–[32] for the Inception-V3, ResNet-50, DenseNet-121, and EfficientNet-B4 respectively. We use Adam stochastic optimization [39] with a learning rate of 10^{-5} during the first training stage and 10^{-6} for the next training stages. The batch size is 64 with 25 epochs per step.

2) CNN-LSTM MODEL

To train the complete model, we use Adam stochastic optimization with a learning rate of 10^{-5}. The batch size is 32 with 100 epochs. During training, we also use the early stopping option to avoid overfitting and improve the model's generalization.

B. HYBRID DEEP MODEL EVALUATION

As detailed in Table 4, we evaluated and compared the four hybrid models to determine the best configuration that fits our data in terms of accuracy, precision, recall, specificity, and F1-score. Although all models are suitable for predicting the diagnosis of the nine ear conditions, the ResNet-50 + LSTM hybrid model outperforms all models. Such a model achieved the highest average accuracy of 97.8%, followed by the EfficientNet-B4 + LSTM with 97.7%, Inception-V3 + LSTM with 97.2%, and finally, DenseNet-121 + LSTM with 96.6%.

In the context of computer science, accuracy is a reliable metric in studies with balanced databases. However, in medical applications, other metrics such as recall or specificity are even more relevant. The recall metric measures the ability of the algorithm to correctly identify the positive cases, whereas the specificity measures the ability of the algorithm to identify the negative cases correctly. In this sense, both models ResNet-50 + LSTM (90.3% recall, 98.8% specificity) and

TABLE 4. Performance comparison of all implemented hybrid models. The evaluation was conducted on the testing set.

Method	Accuracy %	Precision %	Recall %	Specificity %	F1-score %
Inception-V3+LSTM	97.22	88.69	87.50	98.44	86.85
ResNet-50+LSTM	**97.84**	**90.54**	**90.28**	**98.78**	**90.13**
EfficientNet-B4+LSTM	97.69	90.12	89.58	98.70	89.43
DenseNet-121+LSTM	96.60	87.19	84.72	98.09	84.63

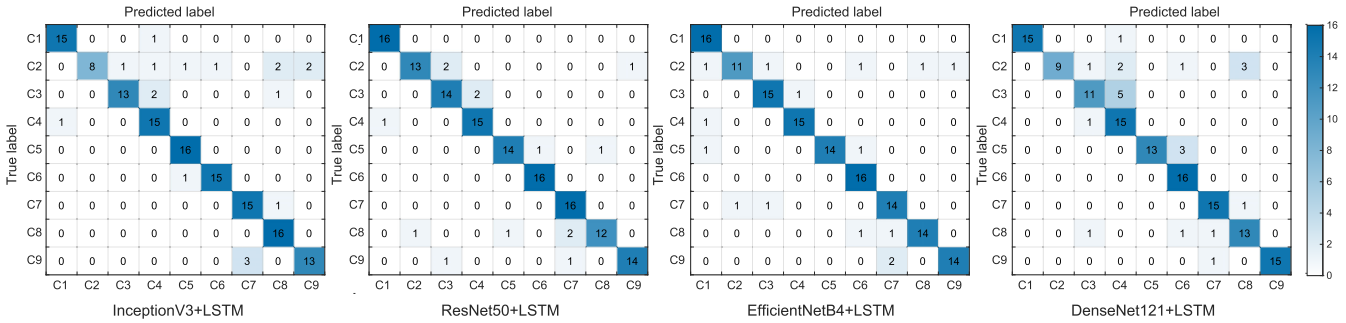

FIGURE 6. Confusion matrices for each hybrid model: Inception-V3 + LSTM, ResNet-50 + LSTM, EfficientNet-B4 + LSTM, and DenseNet-121 + LSTM. C1: exostosis; C2: myringosclerosis; C3: mono-dimeric; C4: normal; C5: external otitis; C6: chronic otitis media; C7: otitis media with effusion; C8: tympanic membrane retraction; and C9: earwax plug.

EfficientNet-B4 + LSTM (89.6% recall, 98.7% specificity) also presented high performance.

The evaluation metrics presented in Table 4 were calculated from the confusion matrices of each model, which are presented in Fig.6. As can be seen, the ResNet-50 + LSTM and EfficientNet-B4 models show the best average performance for all nine ear conditions. However, the Inception-V3 + LSTM model best identifies the ear condition *tympanic retraction* (16 out of 16 cases); whereas the DenseNet-121 + LSTM model best identifies the ear condition *earwax plug* (15 of 16 cases).

Based on the results obtained, the model selected for further analysis is ResNet-50 + LSTM.

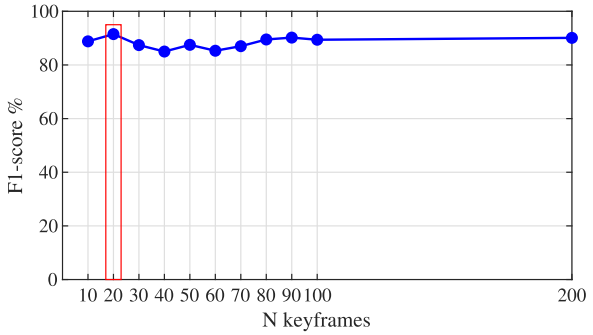

FIGURE 7. F1-score when the number of keyframes (N) is varied. F1-score is maximum when $N = 20$ frames.

C. ELIMINATING REDUNDANT DATA

We analyse the option of reducing frames to eliminate redundant information and find the appropriate point of operation of the proposed scheme. To avoid problems related to non-uniformly sampled data in LSTM networks, which are explicitly sensitive to temporal variation, we uniformly subsample the frames within the video sequence. To verify the impact of subsampling on model performance, we perform several experiments by varying the number of frames $N = 10, 20, 30, \ldots, 100$. Figure 7 shows the F1-score of the models for each variation, and we also include the score when the model uses all frames ($N = 200$). As can be seen, the highest value is obtained by using 20 of the 200 frames available in the sequence. Although the data was acquired at 20 fps, the network only needs two fps to predict the diagnosis with suitable performance. The F1-score remains approximately constant from 30 to 70 frames. It increases from $N = 80$, and no significant changes are observed when using $N = 90$, 100, and 200 frames. Although we increase the amount of information (from 10 to 200 frames) that passes through the network, it may not necessarily be helpful to the model. Therefore, we can decrease the amount of information needed without compromising the performance of the proposed network.

We also compute the remaining evaluation metrics given by (9)-(12) and they are presented in Table 5. As can be seen, the best performance in terms of accuracy (98.15%), precision (91.94%), recall (91.67%), specificity (98.96%), and F1-score (91.51%) is still when $N = 20$ frames.

To verify the performance of the proposed computer-aided system on all nine ear conditions, we have plotted the Receiver Operating Characteristic (ROC) curves from their respective confusion matrix as shown in Fig. 8. It reveals that the proposed scheme can give a promising diagnosis on all nine ear conditions. Furthermore, we compute the AUC (area under the curve) as well as precision, recall, specificity, and F1-score per ear condition shown in Table 6. As can be seen, the system achieves high performance when classifying

TABLE 5. System performance comparison when N varies using testing set after 10 repetitions with the respective data partitioning.

N	Accuracy %	Precision %	Recall %	Specificity %	F1-score %
10	97.53	89.06	88.89	98.61	88.79
20	**98.15**	**91.94**	**91.67**	**98.96**	**91.51**
30	97.22	88.71	87.50	98.44	87.40
40	96.76	86.17	85.42	98.18	84.97
50	97.22	88.26	97.50	98.44	87.48
60	96,76	85,97	85,42	98,18	85,30
70	97,15	87,92	86,81	98,44	86,97
80	97,69	89,80	89,58	98,70	89,51
90	97,84	90,58	90,28	98,78	90,20
100	97,69	90,45	88,89	98,78	89,37
200	97.84	90.54	90.28	98.78	90.13

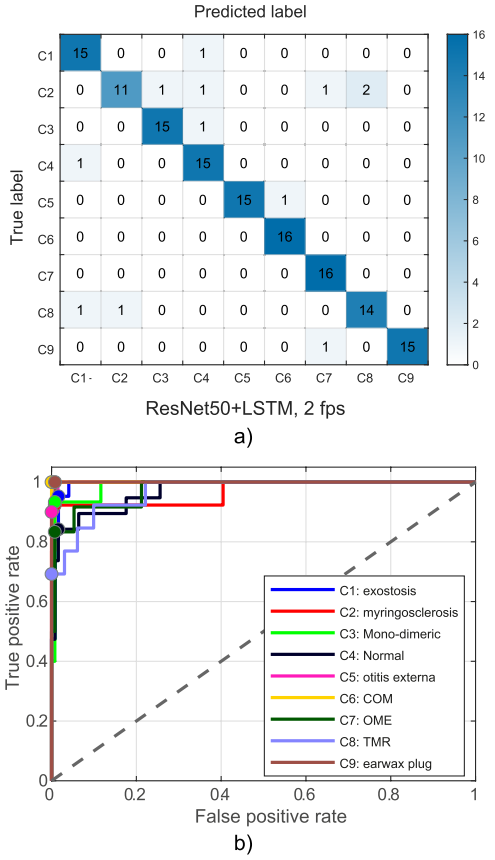

FIGURE 8. a) confusion matrix and b) ROC curves of the diagnoses generated by the proposed computer-aided system on the testing set for all nine ear conditions.

earwax plug, chronic otitis media, mono-dimeric, and otitis media with effusion. It also gives an acceptable diagnosis on myringosclerosis, exostosis, normal ear, and tympanic membrane retraction.

V. DISCUSSIONS

According to the National Academies of Sciences, Engineering, and Medicine in the USA [42], a diagnostic error will occur in the care of nearly all patients in their lifetime, sometimes with devastating consequences. An accurate diagnosis is important to receive appropriate treatments and avoid serious physical, psychological, and financial consequences.

In medicine, there are several areas with subjective diagnostic procedures that depend almost exclusively on the skills and experience of the physician. Motivated by this, different assistance systems are being developed to present objective methods that can serve the physician as a second opinion (e.g., CADx systems).

In particular, a study [43] showed that the diagnostic accuracy of a set of 12 ENT specialists is about 73% on otitis media diagnosis. Such value decreases to 50% when the diagnosis is made by a non-specialist physician. The study also showed that the diagnostic confidence is only 80% on average. These results demonstrate that the diagnosis of diseases such as otitis media is not trivial even for a specialist.

The proposed architecture based on a hybrid learning framework (CNN + LSTM) classified middle ear and ear canal pathologies into nine categories with an average precision of 91.65%, average recall 89.78%, average specificity 98.67% and average F1-score 90.30%. We have included nine of the most frequent ear diseases, carefully selected based on the criteria of ENT specialists of the Clinical Hospital of the University of Chile. However, other relevant pathologies such as acute otitis media or osteoid ostema were discarded due to the limited amount of data collected for training.

Previous studies [7], [24]–[26] have proposed models that classify still images of the eardrum into up to six categories, the accuracy achieved is in the range of 84% to 93%. However, access to the databases of these studies is restricted or the number of samples per class is limited. Therefore, we created a video otoscopy dataset of 875 patients from which 365 video otoscopies were effectively used.

It is not possible to make a one-to-one comparison between the previous studies due to differences in the experiments and the restriction on the use of data. We summarize in Table 7, the main characteristics of the related state of the art. The models proposed to date use a still image classification scheme which is impractical for the reality of medical workflow in the area of otolaryngology. The physician examines the eardrum area and the system generates a certain diagnosis, however by continuing the inspection, it is possible to find some suggestive finding of another disease that could be discarded with the analysis on the first instance. Therefore, the analysis of the entire image sequence (video) proposes a more complete diagnosis.

Since there was no standardization for data acquisition, the videos presented variations on white balance or lighting that in turn leads to inconsistent color of eardrum or ear canal. Camera exposure may not optimally focus on eardrum in case of a tortuous external auditory canal blocking the eardrum. Further, the eardrum is not always in the centre of the image, and some flipping frames or with rotation were produced.

Although we performed an analysis based on subsampling to reduce redundant information, a better approach might use

TABLE 6. System performance comparison per class using testing set after 10 repetitions with the respective data partitioning.

Ear condition	Accuracy %	Precision %	Recall %	Specificity %	F1-score %	AUC
Exostosis	97,92	88,24	93,75	98,44	90,91	0.99
Myringosclerosis	95,83	91,67	68,75	99,22	78,57	0.96
Mono-dimeric	98,61	93,75	93,75	99,22	93,75	0.98
Normal	97,22	83,33	93,75	97,66	88,24	0.97
External otitis	99,31	100,00	93,75	100,00	96,77	0.99
Chronic otitis media (COM)	99,31	94,12	100,00	99,22	96,97	1.00
Otitis media with effusion (OME)	98,61	88,89	100,00	98,44	94,12	0.97
Tympanic membrane retraction (TMR)	97,20	87,50	87,50	98,44	87,50	0.96
Earwax plug	99,31	100,00	93,75	100,00	96,77	0.99

TABLE 7. A comparison of the related studies.

Previous studies	Method	# of pathologies	Accuracy %	Sensitivity %	Specificity %	F1-score %
Myburgh et al. [1]	Feature-based description, DT, Neural Networks	5	86.80	87.00	96.40	-
Zafert [23]	CNN features, kNN, DT, SVM	4	98.74	98.23	99.38	98.34
Viscaino et al. [36]	Texture and color descriptors, kNN, DT, SVM	4	93.90	87.80	95.90	-
Cha et al. [7]	Nine public CNN models	6	93.67	-	-	-
Yeon Lee et al. [44]	Class activation map, customize CNN	2	97.90	99.30	96.30	-
Azam Khan et al. [26]	Five public CNN models	3	94.90	95.00	-	-
Bacsaran et al. [24]	Six public CNN models	2	90.48	86.84	93.50	-
Our proposed system	**hybrid model (pre-trained CNN and single layer LSTM)**	**9**	**98.15**	**91.67**	**98.96**	**91.51**

a video summarization technique based on features to keep the relevant frames. As a result, non-uniformly sampled data sequences can be obtained. In this sense, the LSTM network will need to incorporate additional gates to include more information, such as the time variable [45].

The proposed system could be used by general practitioners, otolaryngology specialist or healthcare professionals to provide them a second objective opinion about diagnosis of middle ear pathologies.

VI. CONCLUSION

We developed a computer-aided diagnosis system for nine ear conditions: normal, myringosclerosis, monomeric or dimeric membrane, chronic otitis media, otitis media with effusion, external otitis, exostosis of ear canal, tympanic membrane retraction and earwax plug. The system achieved an average precision of 91.65%, average recall of 89.78%, average specificity of 98.67%, and average F1-score of 90.30%, outperforming the current state of the art in the above pathologies. The system might be used for general practitioner to make more accurate diagnosis using a low-cost system based on a hybrid learning framework. As future work, we will start measuring the physician workload when using our interface in consultation rooms.

ACKNOWLEDGMENT

The authors would like to thank the University of Chile, Department of Otolaryngology, personnel and physicians for the help provided in patients recruitment and video acquisition.

REFERENCES

[1] H. C. Myburgh, S. Jose, D. W. Swanepoel, and C. Laurent, "Towards low cost automated smartphone- and cloud-based otitis media diagnosis," *Biomed. Signal Process. Control*, vol. 39, pp. 34–52, Jan. 2018.

[2] C. Senaras, A. C. Moberly, T. Teknos, G. F. Essig, C. A. Elmaraghy, N. F. Taj-schaal, L. Yu, and M. N. Gurcan, "System and method of otoscopy image analysis to diagnose ear pathology," U.S. Patent App. 16 329 903, Jul. 18, 2019.

[3] J. S. Earwood, T. Rogers, and N. A. Rathjen, "Ear pain: Diagnosing common and uncommon causes," *Amer. Family Physician*, vol. 97, no. 1, pp. 20–27, 2018.

[4] A. M. Bur, M. Shew, and J. New, "Artificial intelligence for the otolaryngologist: A state of the art review," *Otolaryngology–Head Neck Surg.*, vol. 160, no. 4, pp. 603–611, Apr. 2019.

[5] M. G. Crowson, J. Ranisau, A. Eskander, A. Babier, B. Xu, R. R. Kahmke, J. M. Chen, and T. C. Y. Chan, "A contemporary review of machine learning in otolaryngology–head and neck surgery," *Laryngoscope*, vol. 130, no. 1, pp. 45–51, 2019.

[6] M. E. Pichichero, "Diagnostic accuracy of otitis media and tympanocentesis skills assessment among pediatricians," *Eur. J. Clin. Microbiol. Infectious Diseases*, vol. 22, no. 9, pp. 519–524, Sep. 2003.

[7] D. Cha, C. Pae, S.-B. Seong, J. Y. Choi, and H.-J. Park, "Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database," *EBioMedicine*, vol. 45, pp. 606–614, Jul. 2019.

[8] M. L. Giger, "Machine learning in medical imaging," *J. Amer. College Radiol.*, vol. 15, no. 3, pp. 512–520, Mar. 2018.

[9] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[10] J. Gutiérrez-Martínez, C. Pineda, H. Sandoval, and A. Bernal-González, "Computer-aided diagnosis in rheumatic diseases using ultrasound: An overview," *Clin. Rheumatol.*, vol. 39, pp. 993–1005, Nov. 2019.

[11] H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken, and Y. Shin, "Improving automatic polyp detection using CNN by exploiting temporal dependency in colonoscopy video," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 180–193, Jan. 2020.

[12] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.

[13] J. P. Vasconez, J. Delpiano, S. Vougioukas, and F. Auat Cheein, "Comparison of convolutional neural networks in fruit detection and counting: A comprehensive evaluation," *Comput. Electron. Agricult.*, vol. 173, Jun. 2020, Art. no. 105348.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[15] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[16] C. Feng, A. Elazab, P. Yang, T. Wang, F. Zhou, H. Hu, X. Xiao, and B. Lei, "Deep learning framework for Alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM," *IEEE Access*, vol. 7, pp. 63605–63618, 2019.

[17] K. Greff, R. K. Srivastava, J. Koutnìk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[18] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.

[19] S. H. I. Xingjian, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[20] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4580–4584.

[21] C. Senaras, A. C. Moberly, T. Teknos, G. Essig, C. Elmaraghy, N. Taj-Schaal, L. Yu, and M. Gurcan, "Autoscope: Automated otoscopy image analysis to diagnose ear pathology and use of clinically motivated eardrum features," *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 101341X.

[22] Y.-K. Huang and C.-P. Huang, "A depth-first search algorithm based otoscope application for real-time otitis media image interpretation," in *Proc. 18th Int. Conf. Parallel Distrib. Comput., Appl. Technol. (PDCAT)*, Dec. 2017, pp. 170–175.

[23] C. Zafer, "Fusing fine-tuned deep features for recognizing different tympanic membranes," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 40–51, Jan. 2020.

[24] E. Başaran, Z. Cömert, and Y. Çelik, "Convolutional neural network approach for automatic tympanic membrane detection and classification," *Biomed. Signal Process. Control*, vol. 56, Feb. 2020, Art. no. 101734.

[25] C. Senaras, A. C. Moberly, T. Teknos, G. Essig, C. Elmaraghy, N. Taj-Schaal, L. Yua, and M. N. Gurcan, "Detection of eardrum abnormalities using ensemble deep learning approaches," *Proc. SPIE*, vol. 10575, Feb. 2018, Art. no. 105751A.

[26] M. A. Khan, S. Kwon, J. Choo, S. M. Hong, S. H. Kang, I.-H. Park, S. K. Kim, and S. J. Hong, "Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks," *Neural Netw.*, vol. 126, pp. 384–394, Jun. 2020.

[27] R. Liu, Z. Li, and J. Jia, "Image partial blur detection and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[28] J. Morio and M. Balesdent, *Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems: A Practical Approach*. Sawston, U.K.: Woodhead Publishing, 2015.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[32] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] Q. Qi, Y. Li, J. Wang, H. Zheng, Y. Huang, X. Ding, and G. K. Rohde, "Label-efficient breast cancer histopathological image classification," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2108–2116, Sep. 2019.

[35] H. Wang, H. Jia, L. Lu, and Y. Xia, "Thorax-net: An attention regularized deep neural network for classification of thoracic diseases on chest radiography," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 475–485, Feb. 2020.

[36] M. Viscaino, J. C. Maass, P. H. Delano, M. Torrente, C. Stott, and F. Auat Cheein, "Computer-aided diagnosis of external and middle ear conditions: A machine learning approach," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0229226.

[37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[38] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller, "Snore-GANs: Improving automatic snore sound classification with synthesized data," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 300–310, Jan. 2020.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[40] H. Almeida, M.-J. Meurs, L. Kosseim, G. Butler, and A. Tsang, "Machine learning for biomedical literature triage," *PLoS ONE*, vol. 9, no. 12, Dec. 2014, Art. no. e115892.

[41] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[42] E. A. McGlynn, K. M. McDonald, and C. K. Cassel, "Measurement is essential for improving diagnosis and reducing diagnostic error: A report from the institute of medicine," *Jama*, vol. 314, no. 23, pp. 2501–2502, 2015.

[43] M. E. Pichichero and M. D. Poole, "Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination," *Int. J. Pediatric Otorhinolaryngol.*, vol. 69, no. 3, pp. 361–366, Mar. 2005.

[44] J. Y. Lee, S.-H. Choi, and J. W. Chung, "Automated classification of the tympanic membrane using a convolutional neural network," *Appl. Sci.*, vol. 9, no. 9, p. 1827, May 2019.

[45] S. O. Sahin and S. S. Kozat, "Nonuniformly sampled data processing using LSTM networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1452–1461, May 2019.

MICHELLE VISCAINO (Graduate Student Member, IEEE) received the B.S. degree in electronic engineering from Escuela Politécnica Nacional (EPN), Quito, Ecuador, in 2016. She is currently pursuing the Ph.D. degree in electronic engineering with Universidad Técnica Federico Santa María. From 2016 to 2018, she was involved in research and teaching duties with EPN. Her research interests include machine learning, deep learning, computer-aided diagnosis, E-health, and robotics. She received the Best National Innovation/Technology Development Award granted by the Institute of Electrical and Electronics Engineers in Chile, in 2019 and 2020. Other relevant awards include Best Original Work in 76th Chilean Congress of Otorhinolaryngology, in 2020, and Grand Prize winner of the 5-Minute Video Clip Contest at 2021 IEEE International Conference on Image Processing.

JUAN C. MAASS received the M.D., Otolaryngologist (ENT), and Ph.D. degrees from the University of Chile, Santiago, Chile, in 2001, 2005, and 2010, respectively. Since 2013, he has been an Assistant Professor with the Institute of Biomedical Sciences, University of Chile; and an Assistant Professor with the Otolaryngology Department, Clinical Hospital Universidad de Chile. Since 2008, he has been an Otolaryngologist with the Clínica Alemana de Santiago, Chile. He authored more than 30 international and national publications and eight high-impact review articles. His research interests include inner ear regeneration, development of hearing loss genetics, genomics otolaryngology, otology and neuro-otology, and sudden deafness artificial intelligence-aided diagnosis. He has been granted more than 12 competitive research grants in his scientific career, including ANID funding (FONDECYT) and international grants.

PAUL H. DELANO received the M.D., Ph.D., and Otolaryngologist (ENT) degrees from the University of Chile, Santiago, Chile, in 2001, 2006, and 2010, respectively. Since 2010, he has been an Assistant Professor with the Physiology and Biophysics Program, Biomedical Science Institute (ICBM), University of Chile; and an Assistant Professor with the Otolaryngology Department, Clinical Hospital Universidad de Chile. In 2020, he was promoted to a Full Professor in otolaryngology and neuroscience with the Otolaryngology and Neuroscience Department, University of Chile. He authored 41 international publications, including high impact and prestigious journals, such as *PNAS*, *Journal of Neuroscience,* and *Scientific Reports*. In addition, he has published more than 30 national articles in Spanish, two book chapters, and one e-book. He has been granted more than ten competitive research grants in his scientific career, including ANID funding (FONDECYT, ANILLO, FONDEF, and PCI) and international grants. His research interests include translational research, between basic, and clinical investigation in hearing and related fields.

FERNANDO AUAT CHEEIN (Senior Member, IEEE) received the B.S. degree in electronics engineering from the Universidad Nacional de Tucumán, Tucuman, Argentina, in 2002, and the M.S. and Ph.D. degrees in control systems from the Universidad Nacional de San Juan, San Juan, Argentina, in 2005 and 2009, respectively. Since 2011, he has been an Assistant Professor with the Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaiso, Chile. He is also the Founder of the Autonomous and Industrial Robotics Research Group (GRAI), Advanced Center for Electrical and Electronic Engineering, Chile. He has authored three book chapters and published more than 150 journal articles. His research interests include autonomous navigation, agricultural robotics, and estimation theory. He received the International Joint Conference on Biomedical Engineering Systems and Technologies Best Paper Award, in 2008, the Best National Innovation/Technology Development Award granted by the Institute of Electrical and Electronics Engineers in Chile, in 2015, 2019, and 2020; and the Outstanding Teaching Award granted by Universidad Técnica Federico Santa María, from 2012 to 2014.