EBioMedicine 40 (2019) 176–183

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Research paper

Dermoscopy diagnosis of cancerous lesions utilizing dual deep learning algorithms via visual and audio (sonification) outputs: Laboratory and prospective observational studies

Walker B.N. [a], Rehg J.M. [b], Kalra A. [c], Winters R.M. [d], Drews P. [e], Dascalu J. [f], David E.O. [g], Dascalu A. [h,*]

[a] Sonification Lab, School of Psychology, School of Interactive Computing, Georgia Institute of Technology (Walker BN), Georgia
[b] School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia
[c] Hoplabs, Atlanta, Georgia
[d] Institute of GT Sonification Lab, Georgia Technology, Atlanta, Georgia
[e] Institute for Robotics and Intelligent Machines, Georgia Institute of Technology, Atlanta, Georgia
[f] Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel
[g] Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel
[h] Department of Physiology and Pharmacology, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

ARTICLE INFO

Article history:
Received 4 December 2018
Received in revised form 23 December 2018
Accepted 11 January 2019
Available online 20 January 2019

Keywords:
Skin cancer
Deep learning
Sonification
Artificial intelligence
Dermoscopy
Melanoma
Telemedicine

ABSTRACT

Background: Early diagnosis of skin cancer lesions by dermoscopy, the gold standard in dermatological imaging, calls for a diagnostic upscale. The aim of the study was to improve the accuracy of dermoscopic skin cancer diagnosis through use of novel deep learning (DL) algorithms. An additional sonification-derived diagnostic layer was added to the visual classification to increase sensitivity.

Methods: Two parallel studies were conducted: a laboratory retrospective study (LABS, $n = 482$ biopsies) and a non-interventional prospective observational study (OBS, $n = 63$ biopsies). A training data set of biopsy-verified reports, normal and cancerous skin lesions ($n = 3954$), were used to develop a DL classifier exploring visual features (System A). The outputs of the classifier were sonified, i.e. data conversion into sound (System B). Derived sound files were analyzed by a second machine learning classifier, either as raw audio (LABS, OBS) or following conversion into spectrograms (LABS) and by image analysis and human heuristics (OBS). The OBS criteria outcomes were System A specificity and System B sensitivity as raw sounds, spectrogram areas or heuristics.

Findings: LABS employed dermoscopies, half benign half malignant, and compared the accuracy of Systems A and B. System A algorithm resulted in a ROC AUC of 0.976 (95% CI, 0.965–0.987). Secondary machine learning analysis of raw sound, FFT and Spectrogram ROC curves resulted in AUC's of 0.931 (95% CI 0.881–0.981), 0.90 (95% CI 0.838–0.963) and 0.988 (CI 95% 0.973–1.001), respectively. OBS analysis of raw sound dermoscopies by the secondary machine learning resulted in a ROC AUC of 0.819 (95% CI, 0.7956 to 0.8406). OBS image analysis of AUC for spectrograms displayed a ROC AUC of 0.808 (CI 95% 0.6945 To 0.9208). By applying a heuristic analysis of Systems A and B a sensitivity of 86% and specificity of 91% were derived in the clinical study.

Interpretation: Adding a second stage of processing, which includes a deep learning algorithm of sonification and heuristic inspection with machine learning, significantly improves diagnostic accuracy. A combined two-stage system is expected to assist clinical decisions and de-escalate the current trend of over-diagnosis of skin cancer lesions as pathological.

Fund: Bostel Technologies.

Trial Registration clinicaltrials.gov Identifier: NCT03362138

1. Introduction

Malignant melanoma (MM) is a cancer claiming about 55,000 deaths worldwide annualy [1]. The gold standard for diagnosis of skin cancer is dermoscopy [2] which results in a limited diagnostic accuracy due to the complexity of visual inputs embedded in a dermoscopy

* Corresponding author.
E-mail address: dasc@post.tau.ac.il (A. Dascalu).

https://doi.org/10.1016/j.ebiom.2019.01.028

Research in context

Evidence before this study

We searched in Pubmed and arXiv for prospective Clinical Trials using the search terms of "deep learning" or "artificial intelligence" and "melanoma" or "skin cancer". Search was conducted on Dec 15, 2017 and repeated with addition of term "prospective observational study", on Dec 18, without any finding. Laboratory studies on computer-aided diagnosis of skin cancer were published in the last years, as well as a few articles comparing retrospective laboratory data with dermatologist performance, but none was a prospective observational study or clinical trial as reiterated in an editorial on 31 Oct 2018, Dermatol Pract Concept. No study used Sonification (data conversion to sound) and deep learning in investigation of skin cancer or melanoma diagnosis.

Added value of this study

To our knowledge, this study is first to successfully test and validate in a prospective observational study the diagnostic ability of a dual Deep Learning analysis system to identify skin cancer. It is, as well, the first employment of sonification, a novel second layer of detection, in both laboratory and clinical studies, and demonstrates its additive role in improving sensitivity of detection.

Implications of all the available evidence

Although dermoscopy is the most commonly method used to inspect cancerous skin lesions, its use in the hand of clinicians calls for further improvements in sensitivity and specificity of the technique. Combining Classifier and Sonification algorithms indicate a potential clinician decision support system to be used in office or through telemedicine.

image, and its dependency on physician skills. For example, in blinded tests dermatologists achieve at the lower end of human performance a mean sensitivity for MM detection of 40% [3] and for more complex melanoma images detection is not better than chance. In clinical trials, the number of biopsies that need to be excised in order to identify one melanoma at ages <50 is 58:1 [4], and 28:1 at all ages [5].

Additional monitoring techniques are either experimental, expensive or require prolonged training periods, therefore unavailable to most dermatologists and primary care providers [6]. National skin cancer screening programs are beneficial only at a low evidence level [7], rendering accurate skin cancer diagnosis an imperative social and economical task. A Deep Learning (DL) classifier can be utilized in order to interpret complex visual data through image feature extraction and pattern analysis, such as to diagnose diabetic retinopathy of retinal fundus [8] and identifying head CT scan abnormalities [9]. DL classifiers in dermatology use can achieve a diagnostic performance equal or superior to dermatologists' accuracy [10,11].

We report on a DL classifier (System A) developed and trained to visually analyze dermoscopy images, in order to identify cancerous skin lesions, either pigmented (MM or dysplastic nevi, a clinical mimicker of MM) or skin carcinomas. Classification of a lesion is dichotomous, as malignant or benign, and enables a clinical decision support system indicating the requirement for a biopsy. This single-stage DL system is an effective diagnostic aid, on its own. Diagnostic accuracy was further boosted by a novel analysis technology (System B), in which output from the DL classifier is systematically converted into sound ("sonification" [12]), and then the sound file is classified as indicating a malignant or benign lesion.

The aim of this study was to test the diagnostic ability of a novel two-stage bedside skin cancer diagnosis system. Lesion images were captured by a dermoscope attached to a mobile phone and submitted via a purpose-built application to the classifier operating in the cloud. Instantaneous diagnosis was returned to bedside from Systems A and B. Diagnostic performance was tested by comparing Systems A and B diagnostic output to ground truth biopsies, in both a retrospective laboratory study and a prospective observational study.

2. Methods

2.1. Analysis approach

We utilized a convolutional neural network (CNN) architecture (System A) based on the Inception V2 network [13] to classify dermoscopic images into malignant vs. benign (binary classification) and obtain a feature representation for subsequent use in sonification. The network maps an input image into an output feature map that encodes the visual features which were found to be discriminative in classifying lesions.

2.2. Datasets and deep learning training approach

The System A DL classifier was developed using publicly-available datasets: the International Skin Imaging Collaboration (ISIC) 2017 dataset [14] (2361 images), and the Interactive Atlas of Dermoscopy [15] (IAD) dataset (2000 dermoscopy images and 800 context images, i.e. non-dermoscopic regular photos). Images in each of these datasets are labeled as either a melanoma or benign lesion based on pathology report. As a consequence, our DL lesion analysis method is predicting the primary finding from histopathology based solely on the lesion image. Caffe library [16] was employed to train the Inception V2 model parameters using stochastic gradient descent. Data augmentation was used to expand the available training images, i.e.transformations at random for each image were selected prior to forming each minibatch. The transformations were flips, rotations, and crops, which are meant to encourage translational and rotational invariance. Flips were either horizontal or vertical, around the midline of the image. Rotation angles were chosen at random. The centerpoint for cropping was selected at random, but was constrained so that it always contained the lesion. Training began with a pretrained Inception V2 model which was trained on the ImageNet dataset [17]. We then performed fine tuning of the model using 800 context images from the IAD dataset. Since context images can provide useful discriminative cues for dermoscopic image analysis multi-task learning was performed, which has been shown to improve the performance of deep network models [18].

2.3. Sonification of data

Sonification is the representation of data using non-speech [19]. The data here were the weighted activations of all of the 1024 nodes in the penultimate layer of the DL classifier, which were used to generate sounds in several distinct ways. In the sonification design discussed here, a k-means clustering algorithm [20] was used to cluster the 1024 node activations into groups of related observations. The K-means algorithm was initialized by randomly choosing N data points without replacement to constitute the initial cluster centers, where N is the number of clusters. In order to address the sensitivity to initialization, K-means was run 100 times, each with a different random starting point. The clustering solution with the lowest error (i.e. the one that maximizes the likelihood of the data) was chosen as the final model. Cluster centroids represented by individual pitches and malignant "alert" sounds were mapped onto loudness, timbre, and duration of a sonification, thus an audio signal for each of the centroids of data was derived, providing for an audio output that acoustically differentiated the malignant from benign lesions. The overall effect of this particular

sonification approach is to provide global information about the image, and also about how it compares diagnostically to clusters of known images that are already in the database.

2.4. Classification by sonification and secondary machine learning

The sonification algorithms are designed to allow a listener to differentiate the sound of different classes of lesions. This "diagnosis-by-ear" has been successful in our developmental stages, and we anticipated that it could become a powerful diagnostic tool, akin to the widely used stethoscope. In clinical settings, however, ambient noise can preclude the use of audio output and this motivated our development of an alternative quantification methodology. Thus, we developed a method to systematically inspect the sonification output visually for lesion diagnosis. A secondary machine learning system was developed to diagnose lesions by analyzing FFTs and spectrograms derived from the sonification output. Dermoscopy images ($n = 482$, half benign, and half malignant, all randomly selected) from the database of images that the System A classifier is built on were used to generate audio files using the k-means sonification algorithm (Supercollider v. 3.8.0). For each audio file, visual plots were produced (Sigview software, v.3.1.1.0; SignalLab,e.K., Germany) of the audio amplitude, the FFT of the audio, and the spectrogram (see Fig. 1).

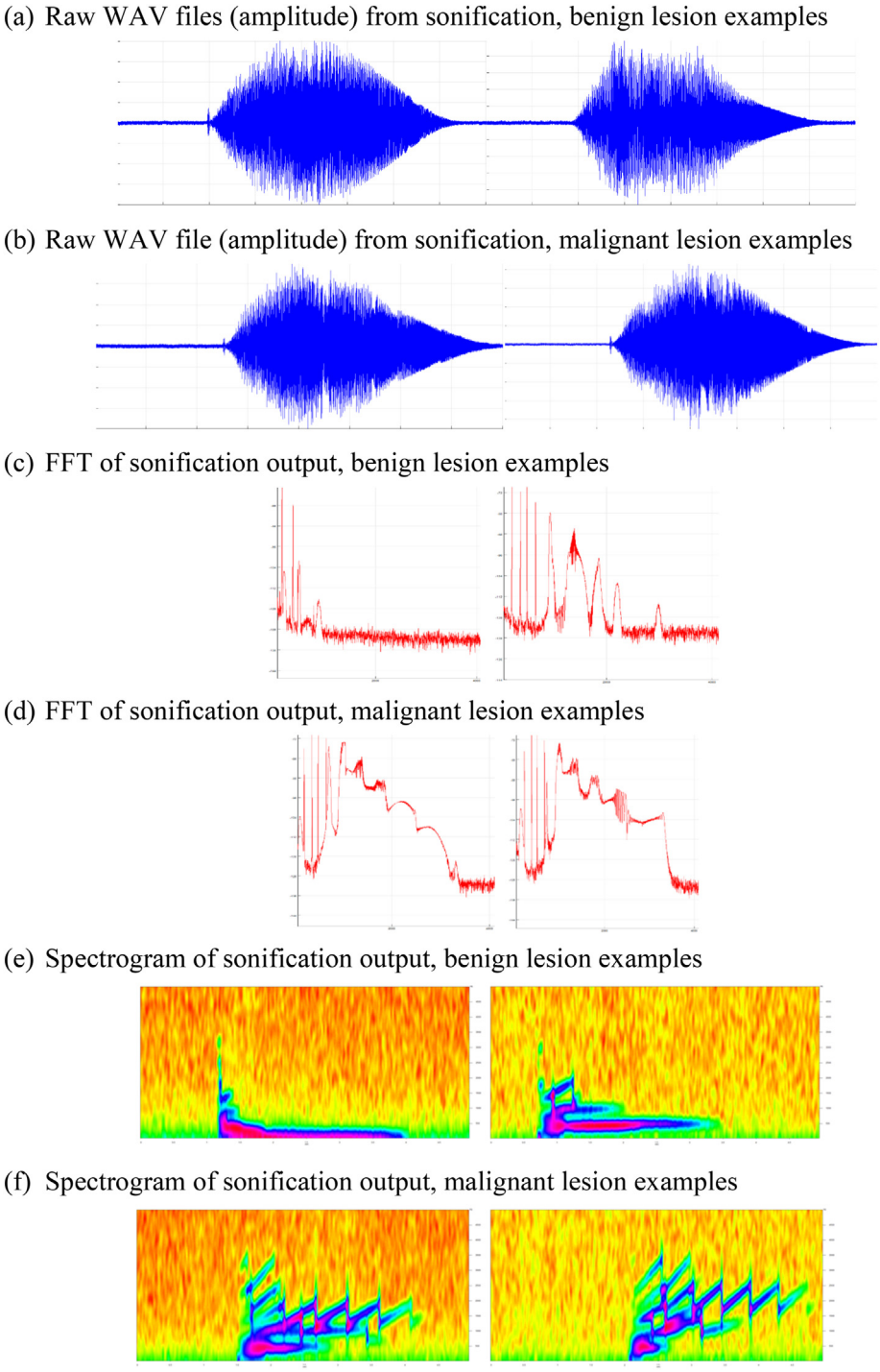

Fig. 1. Visual representations of sonification audio files.

Three separate versions of this secondary classifier, each with identical CNN architectures, were employed in order to explore the automated diagnosis of skin cancer based on the audio, FFT or spectrogram derived from the sonification. All three classifiers were trained against the ground truth diagnosis in the database, using a 80% random single split of the samples (training set).The remaining 20% of the set were held back and later used for validation (test set). All three classifiers normalize the input (zero-mean and divide by standard deviation), and dropout is used for regularization.

Raw audio classifier, LABS: each raw WAV file is single-channel (mono) audio, produced via the sonification algorithm, with sample rate of 44,100 Hz and a duration of 5 s, for a total of 220,500 data points per file. By averaging each 10 consecutive samples, we reduced the input size to 22,050 values. We used a 1-dimensional CNN, with input size 22,050, first convolutional layer with 32 filters of size 1×5; max-pooling layer with size 10; second convolutional layer with 64 filters; max-pooling layer with size 10; a fully connected layer with 128 neurons; and output softmax layer with 2 neurons. This model obtained a validation accuracy of 86.6%.

Raw audio classifier, clinical study: methodology was similar, with a sample duration of 3 s, a total of 132,300 data points per file, a reduced the input size 13,230 values. The 1-dimensional CNN was identical obtaining a validation accuracy of 80.8%.

FFT classifier, LABS: The image files were visual depictions of the FFT of the audio files. We used 2 convolutional layers, the first with 32 filters, and the second with 64 filters. Each convolutional layer was followed by a max-pooling layer of size 2×2. The two convolutional layers were followed by a fully connected layer with 128 neurons, and output softmax layer with 2 neurons. This model obtained a validation accuracy of 82.3%.

Spectrogram classifier, LABS: An identical CNN architecture to the one used for FFT was deployed, with the input files being images of the spectrograms, yielding a validation accuracy of 92.8%.

2.5. Laboratory retrospective study (LABS)

To compare and quantitatively evaluate the three secondary classifiers, we completed a laboratory study using an $= 482$ sample of images from the database. For each image, the System A model was applied and an audio file was generated from its output representation using the sonification algorithm; then, for each audio file an FFT and a spectrogram were produced. These resultant files were then submitted to the secondary machine learning classifiers described above. Performance of the classifiers was quantified by the area under the curve (AUC) of the receiver operating characteristic curve (ROC). This LABS study would serve as a retrospective assessment of the effectiveness of the sonification plus secondary classification approach and compare it to the initial DL classifier (System A).

2.6. Prospective observational study

2.6.1. Study population

An open, prospective, non-interventional prospective observational study (OBS) was conducted in a dermatologic clinic (AD, Tel Aviv, IL). The clinical trial was approved by the institutional review board of Maccabi Healthcare, Israel (protocol Aq 16,842/2017), clinicaltrials.gov Identifier: NCT03362138. Enrollment occurred between 18th Dec 2017 and 23thAug 2018. Inclusion criteria were: age 18 years and older, a suspected malignant lesion identified by a dermatologist through dermoscopy resulting in clinical management of referral to biopsy, and patients' consent to participate in the study. Exclusion criteria were a non-intact skin, >15 hairs per dermoscopic field, performance of an unsolicited biopsy by surgeon (shave), and lesion location within 1 cm of the eye or mucosae surfaces. A total of 68 consecutive biopsy reports were received, 63 being eligible by inclusion criteria.

2.6.2. Prospective observational study design

Subsequent to a clinical decision to biopsy, patient was referred to surgeon and asked to participate in the study by signing the consent form. A dermoscope (DL4, 3 Gen, TX, US) attached to a smartphone (iPhone 6) was used through a purpose-built application (HopLabs, Atlanta, GA, US) for acquiring a dermoscopic image of a suspected lesion which was transmitted securely to a server (HopLabs, Atlanta, GA, USA) via a mobile network. Participant ID was transferred as consecutive numbers, without other patient details. Images were processed on the server by the DL algorithm (System A), and the DL outputs were further processed by the sonification algorithm (System B), as previously detailed. A clinical diagnosis, benign or malignant, appears on the smartphone screen within 6–8 s from acquiring the dermoscopic image, alongside controls to play the sonification audio.

2.6.3. Validation of sonification output

Raw sound files were derived for each dermoscopic image and analyzed by a secondary learning machine for discerning malignancy. Audio files were turned into spectrograms and the AUC of each patient spectrogram were determined (ImageJ, v 1.51j8, NIH) to be further plotted versus biopsy reports.

During the LABS study (results discussed below), clear visual differences for FFT and spectrogram plots were evident for malignant versus benign lesions (Fig. 1 c-f). The obviousness of the features, visually, suggested that a set of diagnosis rules or heuristics may be determinable, so that a human could make the diagnosis without the need for a secondary machine learning algorithm.

Therefore, the sonification procedure was completed again with the new images from the OBS: for each image a sonification audio file and a spectrogram were produced. For each of these images, the frequency range, number of frequency components above 3000 Hz, and the number of saw-tooth wave components was determined. As a result of this systematic evaluation, malignancy was defined for the OBS as: [1] a spectrogram with components of >3000 Hz frequency; and/or [2] four or more saw-tooth wave spikes (typically with declining peak heights). These diagnostic heuristics are used to define the System B classifier based on "heuristic inspection" in which a human expert makes a diagnosis using the sonification-derived heuristics, following on the automated System A classifier output.

"Success" for the new system would be detection of malignancies at a Sensitivity of at least 75% for System A and 85% for System B results, as validated by biopsy (Sensitivity is the percentage of correctly diagnosed malignancies, i.e., true positive/positive diagnoses). Sensitivity metrics are based on the performance of dermatologists with "easy to recognize" class dermoscopies [3], a $72\% \pm 11$ endpoint, and our Deep Learning Sonified output was a 85% sensitivity endpoint (>1 SD of first endpoint). An additional metric of success was a Specificity of at least 33% for Classifier and Sonification, as compared to biopsy (Specificity is the percentage of correctly identified normal nevi, i.e., true negative/negative diagnoses). Specificity value are identical to a previous field test study [21].

2.7. Statistical analysis

Baseline and demographic characteristics were summarized by standard descriptive summaries. All statistical tests used in this study (SigmaPlot v10.0, Systat Software, SanJose, CA) were 2-sided and a p value <.05 was considered significant. Receiver Operating Characteristic (ROC) curves were used to compare the DL results to ground truth biopsies. In the ROCs, sensitivity, the true positive rate, was plotted on the y-axis versus [1-Specificity], the false positive rate, on the x-axis. AUC for such a plot has a maximum value of 1.0, and is a standard performance metric in the machine learning literature. A minimal clinical sample size of 36 patients for estimating sensitivity is required assuming a 0.40 proportion for clinician group (null hypothesis), a DL sensitivity of 0.75, a statistical power of 0.80 and alpha of 0.05 (Sigmaplot for

Windows, V 10.0, Systat Software, San Jose, Ca, USA). Idem, assuming a 0.10 proportion for clinician group, a DL sensitivity of 0.33, a statistical power of 0.80 and alpha of 0.05 a sample size of 58 patients is required for specificity measurement.

3. Results

3.1. Laboratory study results

A total of 482 dermoscopies were tested versus ground truth biopsies to determine the diagnostic abilities of secondary classifiers based on raw sound, FFT, spectrograms and the DL classifier. For the classifier operating on raw sound waves (Fig. 1 a, b), an AUC of 0.931 (95% CI 0.881–0.981), was achieved (Fig. 2a), yielding a remarkable automated diagnostic ability.

Unlike the raw sound waves, FFT and spectrograms exhibit visually-discernible differences between benign and malignant dermoscopies, which is the result of the intentional sonification design, for example using a saw-tooth wave to sonify images that are classified by System A as malignant. FFT of benign and malignant origins (Fig. 1 c, d) show a > 3000 Hz sound frequency, as well as a larger area under the FFT curve. When it comes to the visual spectrograms, malignant biopsies (unlike benign biopsies; Fig. 1 e, f) often also display a characteristic pattern of multiple saw-tooth peaks, declining in amplitude over time.

Applying the secondary classifiers to diagnose malignancy for FFT, spectrograms, and the original DL classifier (System A), resulting ROC curve AUCs (Fig. 2) were 0.90 (95% CI 0.838–0.963), 0.988 (CI 95% 0.973–1.00), and 0.976 (95% CI, 0.965–0.987), respectively (Fig. 2 b, c, d). From the AUC of 0.99, above, it is concluded that secondary classification of sonification spectrograms possesses the most sensitive means of diagnostic accuracy. This considerably attenuates the false negative results that are typical of current skin cancer diagnosis.

3.2. Prospective observational study results

The OBS findings provide an independent field test of the LABS results for the classifiers. As shown in Table 1, a total of 63 biopsies were analyzed. Fig. 3a depicts the smartphone application, which was used for acquiring images (via an attached dermoscope) and for displaying the System A diagnosis and sonification playback controls.

The LABS dermoscopies used melanomas as a major training indicator of pigmented nevi malignancy. The clinical testing, however, encountered mostly dysplastic nevi ($n = 14$) and only two MM due to a small sample size, which are more of a diagnostic challenge as compared to melanomas due to fine details of malignancy, which mimic but are not MM. The degree of clinical dermoscopic dysplasia of all lesions rendered a mandatory excision under suspicion of malignancy. See representative clinical examples of the dysplastic nevi excised which were identified (Fig. 3b-f) and of those not recognized by System B (Fig. 3g).

Major markers of malignancy are shared between LABS and OBS images of dermoscopies: benign lesions (Fig. 4 a,b) display a low FFT y-axis span and do not display a > 3000 Hz frequency, contrary to malignant dermoscopies (Fig. 4 c,d). Spectrograms of benign (Fig. 4e,f) and malignant skin lesions (Fig. 4 g,h) conform to the 3000 Hz threshold and show the multiple saw-tooth pattern. The differences are obvious in most, though not all, of the biopsied lesions.

Sonification diagnostic output was validated by three independent methodologies: raw sound DL, area measurement and heuristics.

Fig. 4i represents the raw sound analysis by a secondary machine learning algorithm. A ROC curve AUC of 0.819 (95% confidence interval 0.7956 to 0.8406) reconfirms the accuracy of sonification as a diagnostic test.

Fig. 4j is based on measurements of each patient's spectrogram AUC by image analysis and plotting its area versus ground truth pathology reports. A ROC curve AUC of 0.808 (CI 95% 0.6945 to 0.9208) was derived. It is concluded that Spectrograms AUCs, although a static measure which disregards dynamic shifts in frequency and time, are a promising objective criteria for identification of malignancy.

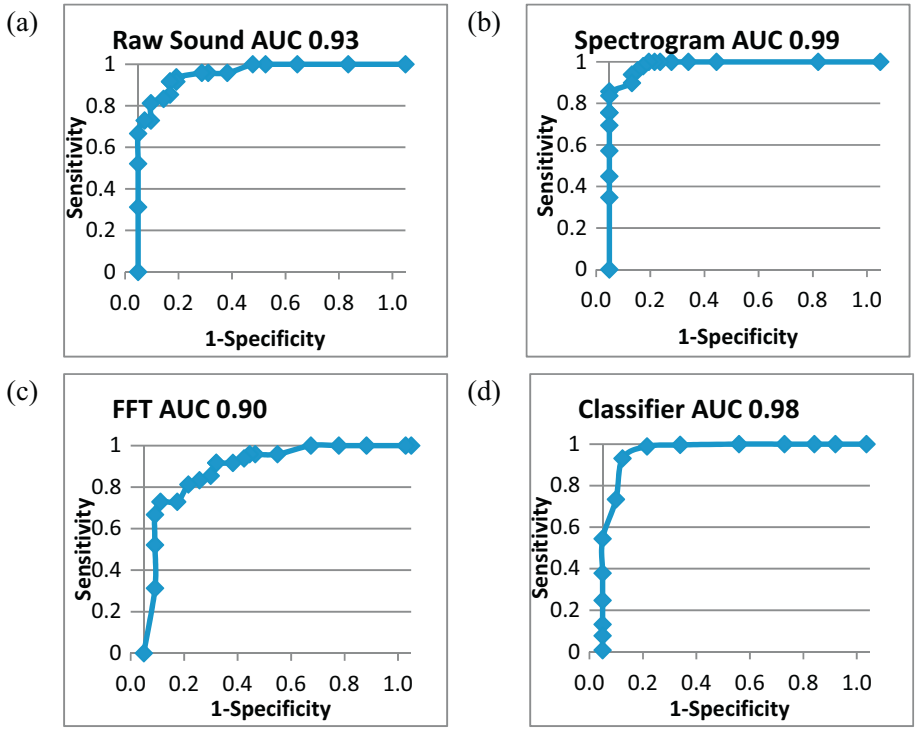

Fig. 2. Receiver Operating Characteristic Curves and Area Under the Curve for the Secondary Machine Learning System Applied to Sonification Outputs (a-c), and DL Classifier (System A) in (d).

Table 1
Epidemiologic data and characteristics of lesions.

Characteristics	No. 63
Study population	
Patients	63
Lesions	63
Age, mean (range)	50.4 ± 14.9 (18–87)
Sex	
Male	34
Female	29
Race	
Caucasian	100%
Anatomic Site	
Face	11
Trunk	31
Extremities	11
Diagnosis	
Benign Nevus	35
Skin Cancer	28
Dysplastic Nevus	14
Atypical Spitz Nevus	1
Melanoma	2
Basal Cell Carcinoma	5
Squamous Cell Carcinoma	6

The study performance of System A (DL classifier) and System B (sonification and heuristic inspection) were compared. System A achieved a 91% specificity (classifier identified 32/35 of benign lesions), accompanied by a drop in sensitivity up to 50% as compared to previous LABS. System B achieved a sensitivity of 86% (heuristic inspection correctly identified 24/28 of all skin cancers) and a specificity of 69%. System B identified 11/11 skin carcinomas as opposed to only 7/11 to be recognized by system A. The positive and negative predictive values of the combined System A specificity and System B sensitivity were 88.9% for both values. Therefore, System A seems to excel in specificity; System B excels in sensitivity and grossly replicates the LABS. The combined use of System A and B as a 2-stage clinical assistance achieves a superhuman accuracy. In conclusion, upon evaluating clinical results of System B use by different methods, OBS sonification confirmed LABS results under a field test, in spite of fewer available malignancy clues.

4. Discussion

We report on a skin cancer detection system which evaluates two different inputs derived from a dermoscopy image: visual features determined via deep learning (System A); and sonification of deep learning node activations followed by human or machine classification (System B). A laboratory study (LABS) and a prospective observational study (OBS) each confirm the accuracy level of this decision support system. In both LABS and OBS, System A is highly specific and System B is highly sensitive. Combination of the two systems potentially facilitates clinical diagnosis.

All skin carcinomas should be excised and pigmented lesions defined as atypical nevi, a clinical diagnosis, are removed out of concern of melanomas due to diagnostic uncertainty. The pathological report classifies nevi as either melanoma, dysplastic nevi (a heuristic grading into mild, moderate or severe) or normal nevi. A posteriori, only moderate or severe dysplastic nevi should be excised [22], a difficult clinical diagnosis, especially in view of the overlap between.

Our System A LABS specificity results are consistent with previously published experimental data [23,24]. System A prospective clinical testing achieved a heuristic specificity of 91%, a figure to be reiterated by additional studies, which majorly improves on the 34% specificity by a recently reported [21] medical device. A novel contribution of our article is the use of sonification, which is rarely used as a diagnostic tool [25]. System B prevails in sensitivity (86%, heuristics) and further investigation will need to parse out exactly how sonification of a DL classifier layer, followed by secondary classification of its raw sounds and spectrogram analysis, can maximally improve accuracy diversified from System A. Furthermore, sonification detected 11/11 non pigmented skin cancers, a figure which seems to outperform recent results derived in an experimental artificial setup [26]. Accordingly, System B achieved both its primary outcomes of specificity and sensitivity. Combining further both Systems might endow a clinician with an impressive assistance tool, which surpasses presently reported dermatologist performance.

Dysplastic nevi are considered to be of malignant potential, due to their risk for developing into melanoma [27], and especially in light of current publications casting doubt on pathologists' ability to discern moderate from severe dysplastic nevi [28]. Our system was assessed

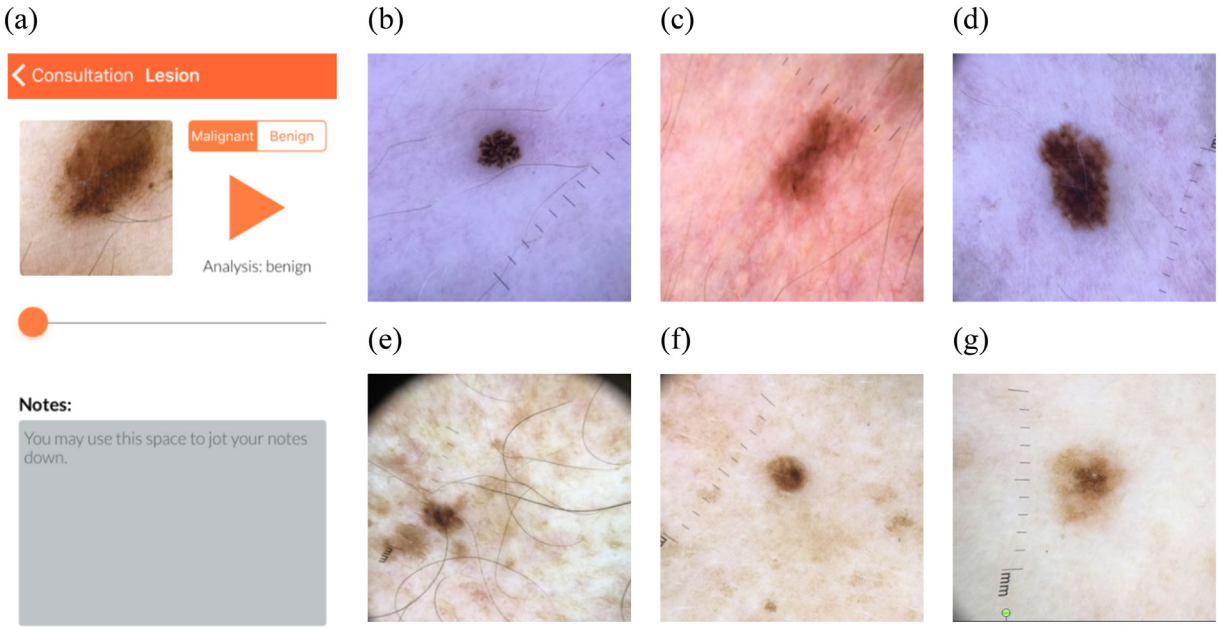

Fig. 3. Prospective observational study Example Images.

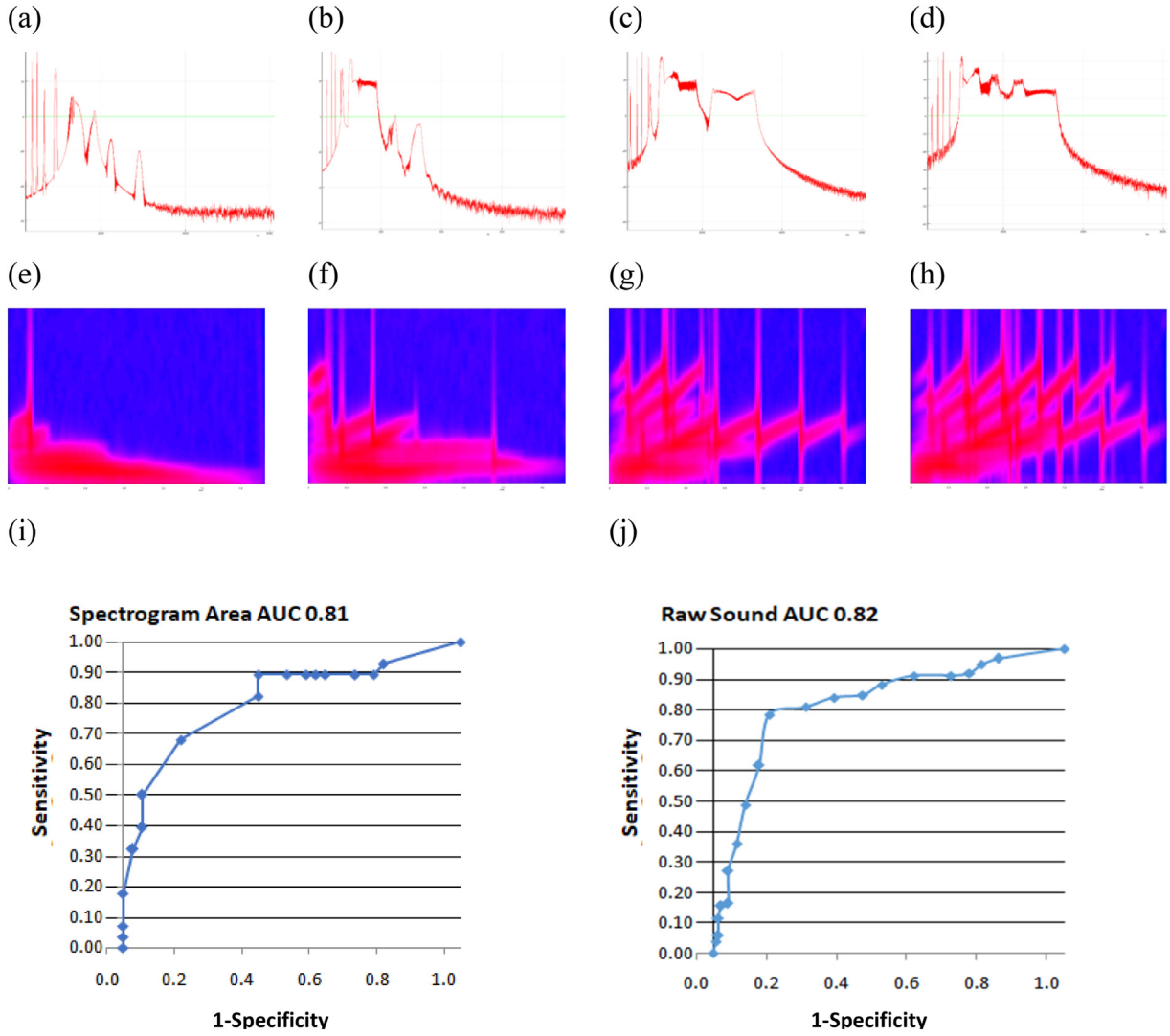

Fig. 4. Prospective observational study Output Examples.

under severe field testing, performing diagnosis of minimal dysplasia—a delicate-features dermoscopy challenge—as part of the criteria of sensitivity. We attribute System A fall off in sensitivity during OBS, as compared to LABS, to its training with a dataset composed mostly of melanomas, without fine features dysplastic nevi. A degree of caution should be exercised with estimating accuracy levels of a DL classifier, since it appears that sensitivity to malignancy are controlled by the dataset input, producing reports of a clinical sensitivity from 29% [29] to 87% [24]. These results further emphasize the high sensitivity of our System B, which identified cancerous lesions smaller than 6 mm diameter. An accurate operative telemedicine prototype as a tool for cloud-directed diagnosis is a field which might be further improved, rendering this system as a candidate for use in the detection of unimaged skin cancer [30].

As part of the initial line of thought in this project, the clinician was expected to evaluate nevi by supplementing System A by listening to the sonification output. Due to the inconvenience of sound perception at the clinic, and in order to increase accuracy, it was decided to develop visual inspection heuristics (clinical study) and a second machine learning algorithm for analyzing the sonification output (LABS, transformed into spectrograms), rendering clinician diagnosis-by-ear as optional. Two heuristic criteria seem to be critical to malignancy recognition of the spectrograms, both in the LABS and OBS: a frequency of >3000 Hz and four or more spikes of audio intensity. Turning obvious heuristics into an operative algorithm and comparison with the raw sonification sounds is a challenging task to be implemented.

The study does imply limitations. It is known that pathology reports of melanoma diagnosis are disputable in about 15% of reports [28]. Therefore, there is a potential bias of diagnosis, since all biopsies in this study were diagnosed by single pathologists. The pathologic report criteria did not disclose nevi as mild, moderate or of severe etiology, although in view of the existence of a small melanoma clinical entity, atypical features should not preclude a biopsy of irregular nevi. Mildly dysplastic nevi are not a candidate for excision, but no a priori technology can identify whether a suspicious atypical clinical lesion is a mild, moderate or severely dysplastic nevus and even pathologists are at dispute whether a nevus belongs to the spectrum of moderate to melanoma in situ,Therefore, our Systems A and B, which decide by a excise or not recommendation, include excision of all atypical nevi categories as possibly malign, This is in accordance to a 2% yield of melanoma of incompletely excised moderate dysplastic nevi at 5 years of follow up [22] which may seem significant at long range.OBS is of modest scale ($n = 63$), thus larger studies should expand on the present results. The clinical trial was performed by a single specialist in dermatology (AD), although this should not affect quality of data, especially the malignancy detection, since DL diagnosis was algorithmic and based on dermoscopy of images. It might be argued that our claimed high accuracy of melanoma detection remains to be proved. It is

assumed, but not proved, that if the System B is sensitive enough to identify fine details of pathology-diagnosed dysplastic nevi, its sensitivity will increase further with bigger melanomas which are endowed with malignant features, to a degree comparable with LABS, which was trained mostly with conspicuous melanomas.

In conclusion, a new diagnostic method for cancerous skin lesions detection, a potential method of teledermoscopy, achieved a high accuracy in a prospective study. Sonification output is a highly sensitive malignant detector of both pigmented and non pigmented skin cancer lesions as evaluated by deep learning, area measurement and heuristics identifiers. Combining sonification sensitivity with classifier might evolve into a useful decision support system for use of all physicians.

Contributors

ED and AD conceived and designed the study. All authors take responsibility for the integrity of the data and the accuracy of the data analysis. BW, JR, AK, MW, PD and ED developed the algorithms. BW, JR, ED and AD were responsible for study supervision. JD, AD and ED obtained and contributed to study interpretation and statistical analysis. All authors subsequently critically edited and revised the report. All authors read and approved the final report. The corresponding author had full access to all the data and final responsibility to submit for publication.

Declaration of interests

AD is an inventor of a patent for the system used in this study; ED reported holding patents on deep Learning, unrelated to the deep learning system in this paper. B·W, JR and AD are shareholders at Bostel LLC. BW, JR, AK, MW and PD were paid consultants. No other disclosures were reported.

Acknowledgments

The study was performed within the framework of Maccabi Healthcare Services, Il.

Role of the funding source

The funding sources had no involvement in the study design; collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

The prospective observational study was approved by the institutional review board of Maccabi Healthcare, Israel (protocol Aq 16,842/2017), clinicaltrials.gov Identifier: NCT03362138

References

[1] Schadendorf D, van Akkooi ACJ, Berking C, et al. Melanoma. Lancet 2018 Sep 15;392(10151):971–84.
[2] Carrera C, Marchetti MA, Dusza SW, et al. Validity and reliability of dermoscopic criteria used to differentiate nevi from melanoma: a web-based international dermoscopy society study. JAMA Dermatol 2016 Jul 1;152(7):798–806.
[3] Tschandl P, Hofmann L, Fink C, Kittler H, Haenssle HA. Melanomas vs. nevi in high-risk patients under long-term monitoring with digital dermatoscopy: do melanomas and nevi already differ at baseline? J Eur Acad Dermatol Venereol 2017 Jun;31(6):972–7.
[4] Matsumoto M, Secrest A, Anderson A, et al. Estimating the cost of skin cancer detection by dermatology providers in a large health care system. J Am Acad Dermatol 2018 Apr;78(4):701–9.
[5] Waldmann A, Nolte S, Geller AC, et al. Frequency of excisions and yields of malignant skin tumors in a population-based screening intervention of 360,288 whole-body examinations. Arch Dermatol 2012;148(8):903–10.
[6] Winkelmann RR, Farberg AS, Glazer AM, et al. Integrating Skin Cancer-Related Technologies into Clinical Practice. Dermatol Clin 2017 Oct;35(4):565–76.
[7] Brunssen A, Waldmann A, Eisemann N, Katalinic A. Impact of skin cancer screening and secondary prevention campaigns on skin cancer incidence and mortality: A systematic review. J Am Acad Dermatol 2017 Jan;76(1):129–39.
[8] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402–10.
[9] Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018 Oct 11(18):31643–5 pii: S0140-6736.
[10] Codella N, Cai J, Abedini M, Garnavi R, Halpern A, Smith JR. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In: Zhou L, Wang L, Wang Q, Shi Y, editors. Machine Learning in Medical Imaging. MLMI 2015. Lecture Notes in Computer Science. Cham: Springer; 2015.
[11] Takuya Yoshida M Emre Celebi, Schaefer Gerald, Iyatomi H. Simple and effective preprocessing for automated melanoma discrimination based on cytological findings. BigData 2016:3439–42.
[12] Dubus G, Bresin.. A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities. PLos One 2013 Dec;17(8):e82491.
[13] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning (ICML). France: Lille; 2015.
[14] Codella NCF, Gutman D, Celebi E, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). arXiv 2017arXiv:1710.05006.
[15] Argenziano G, Soyer HP, De Giorgi V, Piccolo D, Carli P, Delfino M, et al. Dermoscopy: A Tutorial. EDRA Medical Publishing & New Media; 2002.
[16] Jia Y, Shelhamer E, Donahue J, et al. Caffe: convolutional architecture for fast feature embedding. Proceedings of ACM International Conference Multimed; 2014. p. 675–8.
[17] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–52 (Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks.IEEE Trans Med Imaging. 2017 Apr;36(4):994–1004.
[18] Li X, Zhao L, Wei L, Yang MH, et al. Deep saliency: multi-task deep neural network model for salient object detection. IEEE Trans Image Process 2016 Aug;25(8):3919–30.
[19] Walker BN, Nees MA. Theory of sonification. In: Hermann T, Hunt A, Neuhoff J, editors. The Sonification Handbook. Berlin, Germany: Logos Publishing House; 2011. p. 9–39 [ISBN 978-3-8325-2819-5].
[20] Celebi ME, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst Appl 2013;40(1):200–10.
[21] Malvehy J, Hauschild A, Curiel-Lewandrowski C, et al. Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety. Br J Dermatol 2014 Nov;171(5):1099–107.
[22] Fleming NH, Egbert BM, Kim J, Swetter SM. Reexamining the threshold for reexcision of histologically transected dysplastic nevi. JAMA Dermatol 2016;152(12):1327–34.
[23] Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018 Jul;138(7):1529–38.
[24] Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018 Aug 1;29(8):1836–42.
[25] Poveda J, O'Sullivan M, Popovici E, Temko A. Portable neonatal EEG monitoring and sonification on an Android device. Conf Proc IEEE Eng Med Biol Soc 2017 Jul;2017:2018–21.
[26] Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol 2019;155(1):58–65.
[27] Melamed RD, Aydin IT, Rajan GS, et al. Genomic characterization of dysplastic nevi unveils implications for diagnosis of melanoma. J Invest Dermatol 2017;137(4):905.
[28] Elmore JG, Barnhill RL, Elder DE, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ 2017 Jun 28;j2813:357.
[29] Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality? J Invest Dermatol 2018 Oct;138(10):2277–9.
[30] Gendreau JL, Gemelas J, Wang M, Capu. Unimaged melanomas in store-and-forward teledermatology. Telemed J E Health 2017 Jun;23(6):517–20.