HEARING ARTIFICIAL INTELLIGENCE: SONIFICATION GUIDELINES & RESULTS FROM A CASE-STUDY IN MELANOMA DIAGNOSIS

R. Michael Winters

GT Center for Music Technology
Georgia Institute of Technology
Atlanta GA
mikewinters@gatech.edu

Ankur Kalra

Hop Labs
Atlanta, GA
ankur@hoplabs.com

Bruce N. Walker

GT Sonification Lab
Georgia Institute of Technology
Atlanta, GA
bruce.walker@gatech.edu

ABSTRACT

The applications of artificial intelligence are becoming more and more prevalent in everyday life. Although many AI systems can operate autonomously, their goal is often assisting humans. Knowledge from the AI system must somehow be *perceptualized*. Towards this goal, we present a case-study in the application of data-driven non-speech audio for melanoma diagnosis. A physician photographs a suspicious skin lesion, triggering a sonification of the system's penultimate classification layer. We iterated on sonification strategies and coalesced around designs representing three general approaches. We tested each in a group of novice listeners (n=7) for mean sensitivity, specificity, and learning effects. The mean accuracy was greatest for a simple model, but a trained dermatologist preferred a perceptually compressed model of the full classification layer. We discovered that training the AI on sonifications from this model improved accuracy further. We argue for perceptual compression as a general technique and for a comprehensible number of simultaneous streams.

1. INTRODUCTION

Artificial Intelligence (AI) algorithms are becoming an increasingly important part of interacting with computers [1]. Today, almost every major content provider uses machine learning, deep learning, or artificial intelligence more generally to produce their final product.

In spite of the complexity and sophistication that is required to produce a well-functioning AI system, often the information needs to be displayed to a human recipient. In these contexts, an important layer of the AI system is the perceptualization of the machine knowledge. This perceptualization can take many sensory, linguistic, or cognitive forms, and the best way to communicate will depend upon human-factors such as the context, expertise, and task goals.

In this paper, we describe a context where an AI system assists a human in the diagnosis of skin cancer from photographs of suspicious skin areas (lesions). A doctor takes a photograph of a suspicious area on their patient's skin, triggering an analysis phase by the AI system. Once the image has been processed, it generates a sonification that represents what has been sensed/classified in the image—good and bad. The doctor then uses this sound, in addition to other factors such as the patient's medical history, to determine if further tests (biopsy) or treatment is indicated.

We describe our design process for creating sounds for this AI system, which included three sonification designs and a user study with novice listeners. After describing the context around the work, we present the three designs in the order that they were created. We describe the study that we administered and our results, then finish with general design guidelines for working with AI systems that may prove useful in similar contexts.

2. BACKGROUND CONTEXT

Listening has formed a vital component of medical practice. Indeed, auscultation has been considered the first "imaging" technology [2], and the stethoscope is still routinely used by general practitioners. Doctors are trained listeners.

We worked with an algorithm that has been developed to identify melanomas from photos of skin lesions [3]. The algorithm was a deep learning convolutional neural network, and was trained on thousands of images. The algorithm was designed to produce a binary classification output: benign or malignant.

A simple auditory display strategy would be to read out a "benign" or "malignant" diagnosis for a given input image. However, we sought to use a more sophisticated sonification to provide additional information and context. We reasoned that if the sonification targeted the more subtle information behind the course benign/malignant classification, a listener might be able to understand more of the nuance behind the given classification. For example, each image might produce a unique aural signature that helps convey why the algorithm decided on its final classification.

For the purposes of design, we targeted the penultimate layer in the AI system. While the final layer of the network had a binary classification, the layer before that had 1024 nodes, each with an associated weight and image-dependent activation. Although the full system contained hundreds of layers and loops, our choice to use the penultimate layer came from the desire to have the most direct and information rich layer available. This layer also made it easy to use the final classification output.

3. DESIGN PROCESS

In the process of designing the sonification algorithm, we went through several design iterations, which manifested in three distinct design strategies. The three designs all used the penultimate

https://doi.org/10.21785/icad2019.021

layer, but differed in their underlying goal, sound design and mapping strategy.

We made a graphical user interface (GUI) to assist our exploration of the dataset, sampling of the sonification strategies, and our evaluation (See Fig. 1). In the "Training Mode," the GUI displays the image of the suspicious skin region in the upper left, the result of the classification in the upper right, and a graph representing the activations of "benign" and "malignant" nodes on the bottom. For any image, the user could listen to any of the three sonification designs by pressing a button, and control the playback speed using a knob. In the evaluation mode, the image was hidden, and the user diagnosed based only on what they heard.

4. DESIGN #1

The first sonification system used a rapid parameter mapping approach (i.e. granular synthesis) to directly sonify the 1024 nodes in the penultimate layer. Because the data were rendered in an unprocessed format, we nicknamed this design "Raw."

In this approach, the nodes were first sorted by descriptive power. Using all of the images in the dataset, we quantified each node based upon their descriptive power for either benign or malignant diagnosis. Once quantified they were sorted such that the nodes that were most positively associated with the benign images were at the beginning, and the nodes that were most positively associated with the malignant images were at the end. With this ordering in place, each node was mapped to a note whose loudness and duration was determined by the strength of that node for a given image. For example, if a given node had a strength of 1 for a given image, the note assigned to that node would play at full volume for 100ms. If the same node were to have a strength of 0 for a different image, it would have no volume and would have no duration.

In order for each node to be played with most clarity, each node was associated with a unique frequency. These frequencies were evenly distributed across the frequency spectrum according to a logarithmic mapping (i.e., mostly linear in musical note space). This choice allowed each octave to have an equal number of pitches within that octave.

In order to play all of the notes for a given image so that they might be heard, the notes were spaced out in time such that they were triggered in short succession. The amount of time between note onsets was increased in order to produce more clarity of notes, but the amount of time was decreased to limit the total amount of time that a note would ring for.

The notes were played in ascending order from low pitches (generally mapped to benign lesions) to high pitches (malignant lesions), creating a total upward glissando sound as each image played through all of its nodes.

4.1. Analysis

All together, the approach was successful in producing sounds that were different for each image. However, the overall sound was quite chromatic and dissonant due to the closely spaced notes in both frequency space and time.

Furthermore, in order to determine whether a given sound corresponded to a benign or malignant image, it was necessary to do a type of aural weight analysis, where the total amount of low frequency volume was weighed against the total amount of high frequency volume. This process was necessary because each image had a combination of low and high notes that reflected the 1024 nodes of the penultimate layer. Rapidly playing through all nodes did not make the perceptual identification easier. If anything, the resulting effect was to render the choice more difficult. For example, by hearing a mixture of high and low notes (a true reflection of the layer), a listener might be less confident when making a decision regarding whether the image was benign or malignant. By comparison to a simple mathematical number that could be calculated from every image, this approach ultimately seemed to be less useful.

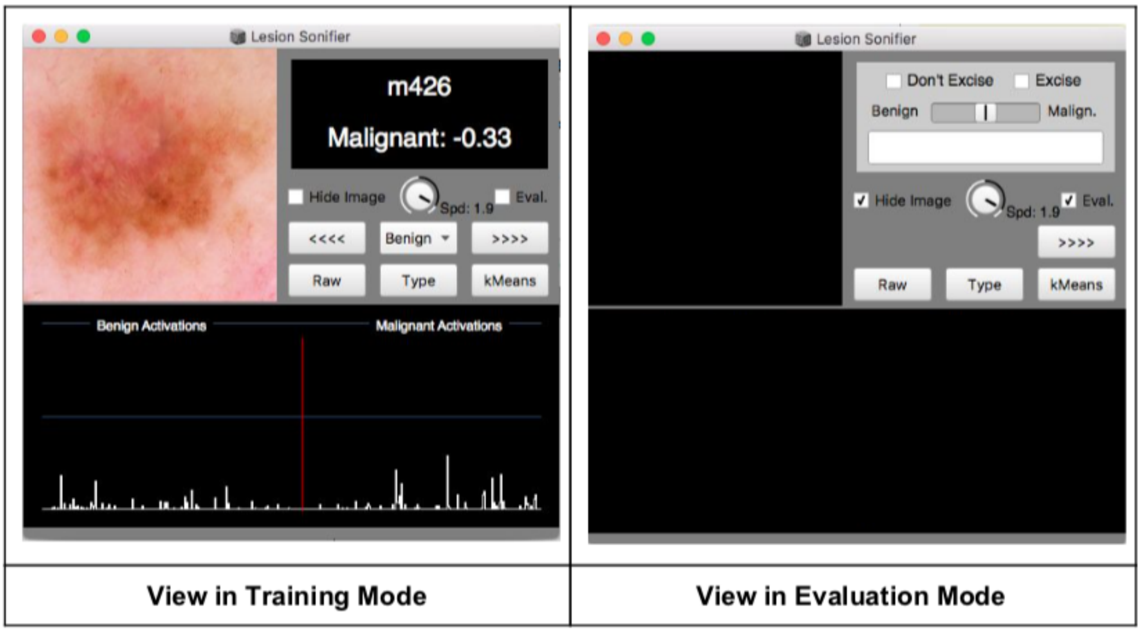
5. DESIGN #2

The second design reflected a desire to make the ultimate decision of benign or malignant more clearly audible, decreasing the amount of learning and time required to make accurate aural diagnosis. To accomplish this goal, we reasoned that the sound of a malignant melanoma should be very clearly different from a benign lesion, and should have sonic qualities of loudness, roughness, dissonance, fear, or in general, "badness." This would contrast to a benign lesion, which would sound more easy-going, clear, consonant and quiet. Furthermore, the goodness or badness of the sound should correspond to the actual certainty that a given lesion would be benign or malignant. Because this design was designed to make the benign or malignant classification clear, we nicknamed it "Type."

With these design goals in place, we drew upon auditory cues from the music emotion literature [4], specifically emotion sonification [5]. In our strategy, a sonic space was modeled that would be controlled by a single "goodness"-to-"badness" dimension that was calculated by summing all of the activations and weights of the first design strategy and applying a scaling based upon the probability of correct diagnosis. This number would be either positive (benign), or negative (malignant), and magnitude would increase linearly with confidence. Because this one dimension controlled many sound parameters, this design was a one-to-many mapping strategy [6].

The timbre used as the basis of the sonification was created using modal synthesis with fixed resonant modes and decay times. In this design, the sound of a benign lesion was a simple timbre that would strike the first note, wait a few moments, and then play a note a perfect fifth above it. The decay time was controlled by the "goodness" of the classification such that a long decay time meant that a lesion was good/benign, and a short decay time indicated bad/malignant. Additionally, the amount of time in between notes was also controlled by the same dimension. The amount of time in between the two strikes was a direct indication of the confidence of being benign. For example, a lesion that was classified with confidence as being benign may have 1.5 seconds of gap between the two sounds, whereas a lesion classified as benign, but with less confidence may have 0.3 seconds of gap between the two notes. Perceptually, benign lesions sounded more "relaxed."

The sound of a lesion that was classified as malignant would sound "bad" using additional auditory cues. Continuing from the benign sound model, the decay time of each of the two strikes would be short, and the time in between the two strikes would be short as well. However, the two strikes were allowed to echo through the sound model, while simultaneously being frequency and amplitude modulated. These modulations, combined with the echo, created a sound that was aggressive, having many attacks in short succession, general roughness, and frequency instability. As



View in Training Mode

View in Evaluation Mode

Figure 1: The GUI used for interactively exploring the dataset and sampling the three sonification designs. In the training mode, the sonification strategies are paired with an image of the skin region and AI-output. In the evaluation mode, the user makes decisions based upon the sonification alone.

with the "good" sounds, the cues used for "badness" increased in magnitude when lesions were classified as being more malignant, and decreased in magnitude when lesions were classified as being more benign.

5.1. Analysis

Having created the sonification model that represented a continuous "goodness"-to-"badness" scale, we felt that the sounds themselves were able to communicate these high-level constructs. We reasoned that the difference would be easy to explain to an untrained listener. Furthermore, the emotion-laden auditory cues would contribute to a more "embodied" [7] or tangible sonic character compared to a spoken classification.

However, the design also had weaknesses. By relying upon a single continuous number for sonification, the sound was not able to provide as much detail as the first design. The nuances in the soundscape in this design were not due to subtle differences in the data, but were instead completely determined by the magnitude of a single number. For example, if the user did not like the sound, or didn't want to use it, a simple real number could replace the sound without any information loss. Thus, the sonification in this case might be able to communicate clearly the goodness or badness, but perhaps not provide much (if any) additional unseen information.

6. DESIGN #3

After producing the first two designs, we sought a third design that could capture what we already learned from the first two and produce something in between. The third design would represent some of the subtleties of the underlying 1024 weights, but include clear acoustic cues that would differentiate benign and malignant lesions. Such a design would offer a mid-way point between the first two designs.

The initial idea for the third design came through a brainstorming session with the team members that made the deep learning classifier. We decided to look into ways to intelligently reduce dimensionality down from 1024, without dropping all the way back down to the final binary classification layer. In the end, we used a clustering approach, where a given lesion would be described by its distance to N different cluster centers. By analyzing all of the ground truth data using this method, we determined how descriptive each cluster was for being either malignant or benign in its diagnosis. For example, if a cluster center reliably predicted a malignant diagnosis with 95% accuracy, we reasoned that being close to this cluster center should have a very bad sound. Similarly, if a different cluster center had reliably predicted benign diagnosis with a 95% accuracy, it should contribute a sound that was peaceful and relaxing. Because of the clustering algorithm we used as part of this design, we nicknamed it "kMeans."

Using this approach, we decided to use fewer than 20 cluster centers, and ordered them according to their ability to predict a benign diagnosis. Each of these were then assigned to a pitch, with each pitch being a fourth above the previous pitch in ascending order. By separating the notes in ascending perfect fourths, we were assured that each cluster center would have a unique pitch, and that the overall tone produced would not be associated with any familiar chord (which would include combinations of thirds). Furthermore, by not stacking the notes on 5ths, the overall range from lowest to highest note was smaller and more compact in pitch-space.

For any particular lesion, the underlying data would be the distance to all cluster centers. However, because a large distance meant that the lesion was not well described by that cluster center, we inverted that value to produce a new parameter: "closeness." Closeness became the variable being sonified, and was mapped to the duration of the note. If a lesion's image was close to any of the cluster centers, the sound of those cluster centers would play for a relatively long time (i.e. up to 2s), compared to clusters that were far.

Because the clusters were ordered according to their ability to predict benign images and stacked in ascending order from a base note, this meant that low notes were (again) associated with benign images and high notes were associated with malignant images. However, after listening to these, we felt that there should be additional cues for cluster centers that were malignant, that would make them not only higher in pitch but also clearly differentiated in timbre. Furthermore, we thought that it would be useful for those notes to also sound more urgent or salient. Therefore, we used a different waveform to frequency modulate the malignant cluster center sounds. The depth of the modulation was fixed relative to the center frequency, but the speed of modulation was greater for cluster centers that were more malignant.

6.1. Analysis

After producing the third design, we felt that we had produced the strongest design yet. By wrapping the design in a GUI, we were able to distribute it to a physician with experience in diagnosis. His feedback was that the sound was able to highlight very minute features in the lesion image. What he was hearing was probably the cluster centers that were not as strongly predictive, and therefore included a slower frequency modulation that might appear in a context of many sounding benign cluster centroids.

7. DEMONSTRATION VIDEOS

We made three demo videos (one for each design) and posted them online.[1][2][3] In each video, a listener uses the GUI (Fig. 1) in Training Mode to hear examples of images from different classification zones. For example, in the video "Design 2 - Type", the user begins by sampling images that have been classified "Malignant 3" (very likely malignant). At 0:42, the user switches to sampling images classified as "Benign 3" (very benign). At 1:04, the user switches to samples classified as "Borderline 0" (equally probable benign or malignant). At 1:28, the user switches to images classified as "Malignant 1" (possibly malignant). Finally at 1:45, the user switches to images classified as "Benign 1" (possibly benign). The corresponding sonification accompanies each new image.

8. STUDY

8.1. Study Purpose & Overview

Given that one physician can become very effective in utilizing the sonification tools, the question arises as to how much practice or training is required for a listener to become proficient in utilizing the sonification output for diagnostic purposes, and whether there is a difference in learnability for the three different sonification approaches. In order to study the learnability of the sonifications, we conducted a training study.

We performed a small controlled ("lab-like") study, to assess the effectiveness and learnability of the sonifications developed in this project. This was a small, initial study focused on the sonification specifically, and not on the entire diagnostic apparatus.

Listeners (not medically trained, but otherwise representative of future medical listeners) were trained to associate sonification sounds to labels (e.g., very bad, neutral, very good), then tested to assess the effectiveness of the training. They also provided subjective feedback about the sonifications.

We were looking at how intuitively the sounds represent the concepts, and how easily the listeners could learn to associate the sounds with the concepts.

Since we developed three novel sonification strategies, all different from each other, the listeners interacted with each of the sonification approaches, in random order. This within-subjects study design allowed us to compare the sonification designs for intuitiveness and training ease. We expected the three sonification approaches to differ not only in how quickly they could be learned, but the way performance evolved with practice.

8.2. Participants

Participants included seven adults (3 male, 4 female), aged 25-35; all had completed a 4-year college degree. These participants were not physicians, but were meant to be somewhat representative of medical students or (young) physicians in many respects (educated, other than medical training).

These participants were recruited in a major US city using a "friends and family" approach. All completed confidentiality agreements, but in any case they did not see any dermatology images, nor were they told anything about the ultimate purpose of the project. Participants were paid $50 for their participation.

8.3. Apparatus

The study was conducted in a commercial office space, generally on weekends and after normal business hours, to maintain confidentiality and ensure a quiet testing environment. Participants interacted with a laptop computer connected to an external monitor and external mouse and keyboard. Participants used high fidelity headphones. A bespoke software program written in SuperCollider provided a GUI through which the sounds could be played and responses recorded. Along with the sound controls, the software presented a word very bad, bad, neutral, good, very good, or a number -3, -2, -1, 0, +1, +2, +3.

8.4. Proceedure

Participants completed a brief demographics form, and executed a confidentiality / non-disclosure agreement. Within one encounter, they completed three sessions, with each session consisting of three blocks of trials. Each block of trials consisted of 21 training trials, followed by 21 testing trials. During the training phase of a block, participants saw the number (or word) and listened to the sound. During the testing phase of a block, participants heard the sound and then selected a number (or word) that they felt represented the sounds "goodness. Data about the responses were recorded for each trial, in every block and session; along with how many times a sound was listened to and the time spent at each stage of each trial.

[1] [Design #1 Example Video:] https://youtu.be/McBoGHIy7qg

[2] [Design #2 Example Video:] https://youtu.be/ay3UpoemiZs

[3] [Design #3 Example Video:] https://youtu.be/4ZZKx9FhYBk

8.5. Results

8.5.1. Sensitivity & Specificity

Sensitivity, here, relates to the number of correctly identified malignant lesions. It is also known as the hit rate, and loosely corresponds to the notion of accuracy in the classification task. In the case of diagnosing melanomas, it is important to catch all the malignant lesions. On the other side of the same coin, specificity refers to how often (or how rarely) a benign lesion is correctly classified as benign. Thus, it is also known as the correct rejection rate, and is important since a mis-classification of a benign lesion as malignant can lead to unnecessary tests and stress to the patient. It is clearly desirable, whenever possible, to have both a high sensitivity, and a high specificity. However, in practical applications, it is often necessary to prioritize one or the other of these performance metrics.

In this study, the sensitivity was calculated for each participant, for each sound design, and for each subsection (i.e., each block of trials). Then, the mean sensitivity was calculated across participants (i.e., collapsing on participant), for each subsection and each sonification design. Similarly, specificity was calculated for each participant for each subsection and for each sonification design. Then, mean specificity was calculated for each subsection for each sound design.

The results for mean sensitivity are presented in Figure 2, and the results for the mean specificity is displayed in Figure 3.

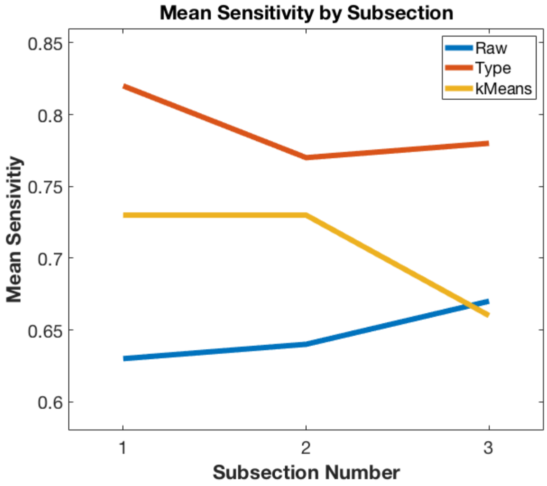


Figure 2: The mean sensitivity of each design accross the three subsections of the study.

8.5.2. Analysis

Within-subjects analyses of the comparisons between sessions was one of the primary measures. For our purposes we looked mostly for evidence in support of the sonification intuitiveness, but did not need statistically reliable measures.

Based upon the mean values across participants, we found that the Type sonification had the overall highest sensitivity, and the Raw design had the lowest. The kMeans and Type designs both had comparable specificity, but the Raw design was much lower.

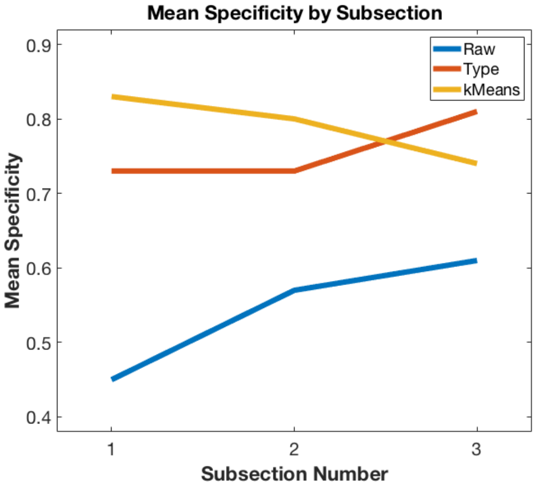


Figure 3: The mean specificity of each design accross the three subsections of the study.

Based upon the mean values across subsections, we found some effects of learning in the Raw design, but not in the Type and kMeans designs. For both Sensitivity and Specificity, the kMeans approach became more difficult with time. For the Type approach, sensitivity decreased somewhat after each subsection, but the specificity increased.

8.5.3. Additional Data

In addition to the performance data described above, we collected data about preferences via a post-task questionnaire. Responses for numerical questions were generally provided via a Lickert-style scale, on which 1=Completely DISAGREE; and 7=Completely AGREE.

In general, the small sample size (n=7) means that we are not really able to make statistically reliable conclusions about the preferences data. We can, however, see that in this sample of participants, there was preference for (and dislike for) each of the sound designs. There was no unanimous favorite. There was, however, less support for sound design #1, especially when it came to asking how easy it was to interpret and understand.

9. GENERAL DISCUSSION

9.1. Study

Our study was largely confirmatory of our design intuitions. Namely, the Raw approach was less useful than the Type or kMeans designs, and the Type approach provided the greatest overall accuracy for novice listeners. Although novices performed well with the kMeans approach initially, their performance decreased with time. This change may be due to them hearing more detail and nuance as they learned. A future study with expert dermatologists and images of the lesions, might produce different results.

9.2. Expert Feedback

In addition to the results from the study, our design process included almost daily interactions with a trained dermatologist. This dermatologist was invested in sonification for the domain, and would explore the dataset using the GUI after each design iteration, offering his insights and support as a specialist. From the dermatologist's experience, we learned that the third design was very "precise," often revealing nuances that were also quite subtle in the photograph.

9.3. Sonification as Layer

One of the unexpected outcomes of our work was finding that when we trained the classifier on audio from the third design, accuracy was increased relative to the AI algorithm alone [3]. The process of perceptualizing the information had in a sense formed another compression layer, which removed noise from the data and increased signal. In the future, we think that designing the outputs of an AI algorithm to be interpretable by a perceptual system (such as the auditory system), might be an effective strategy for boosting performance in an AI system.

9.4. Comprehension Guidelines

In our work, we explored different ways of perceptualizing information in the penultimate layer of a AI algorithm. Based upon our experience, we recommend the approach used in our third design. In this design, we applied compression in the form of a clustering algorithm prior to sonification. Although a mathematical algorithm might not be limited by the number of nodes or dimensions it can utilize, the same is not true for the human perceptual system. In our view, a successful compression algorithm will reduce the number of simultaneous streams to a number that will maximize listening comprehension [8]. For example, when sonifying knowledge in a complex AI system, first reduce the information space to a subset of 10-15 dimensions, of which only 2-5 will be prominent for any given input.

10. ACKNOWLEDGMENT

We'd like to thank Dr. Avi Dascalu for his engaged listening and critical feedback during the iterative design phase. We are also grateful for the 7 participants who contributed their data to the study. BosTel Technologies LLC provided funding for this project.

11. REFERENCES

[1] S. Russell and P. Norvig, *Aritificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, N.J.: Prentice Hall, 2009.

[2] J. Sterne, *The Audible Past: Cultural Origins of Sound Reproduction*. Durham, NC: Duke University Press, 2003.

[3] B. N. Walker, J. M. Rehg, A. Kalra, R. M. Winters, P. Drews, J. Dascalu, E. O. David, and A. Dascalu, "Dermoscopy diagnosis of cancerous lesions utilizing dual deep learning algorithms via visual and audio (sonification) outputs : Laboratory and prospective observational studies," *EBioMedicine*, vol. 40, pp. 176–183, 2019.

[4] P. N. Juslin and J. A. Sloboda, Eds., *Handbook of Music And Emotion: Theory, Research, Applications*. New York, NY: Oxford University Press, 2010.

[5] R. M. Winters and M. M. Wanderley, "Sonification of Emotion: Strategies for Continuous Auditory Display of Arousal and Valence," in *Proceedings of the 3rd International Conference on Music and Emotion*, Jyväskylä, 6 2013.

[6] A. Hunt, M. M. Wanderley, and M. Paradis, "The Importance of Parameter Mapping in Electronic Instrument Design," *Journal of New Music Research*, vol. 32, no. 4, pp. 429–440, 2003.

[7] S. Roddy and D. Furlong, "Embodied Aesthetics in Auditory Display," *Organised Sound*, vol. 19, no. 1, pp. 70–77, 2014.

[8] J. H. Schuett and B. N. Walker, "Measuring comprehension in sonification tasks that have multiple data streams," in *Proceedings of the 8th Audio Mostly Conference on - AM '13*, Piteå, 9 2013.