Efficient and Explainable Deep Neural Networks for Airway Symptom Detection in Support of Wearable Health Technology

René Groh, Zhengdong Lei, Lisa Martignetti, Nicole Y. K. Li-Jessen, and Andreas M. Kist**

Mobile health wearables are often embedded with small processors for signal acquisition and analysis. These embedded wearable systems are, however, limited with low available memory and computational power. Advances in machine learning, especially deep neural networks (DNNs), have been adopted for efficient and intelligent applications to overcome constrained computational environments. Herein, evolutionary algorithms are used to find novel DNNs that are accurate in classifying airway symptoms while allowing wearable deployment. As opposed to typical microphone-acoustic signals, mechano-acoustic data signals, which did not contain identifiable speech information for better privacy protection, are acquired from laboratory-generated and publicly available datasets. The optimized DNNs had a low model file size of less than 150 kB and predicted airway symptoms of interest with 81.49% accuracy on unseen data. By performing explainable AI techniques, namely occlusion experiments and class activation maps, mel-frequency bands up to 8,000 Hz are found as the most important feature for the classification. It is further found that DNN decisions are consistently relying on these specific features, fostering trust and transparency of the proposed DNNs. The proposed efficient and explainable DNN is expected to support edge computing on mechano-acoustic sensing wearables for remote, long-term monitoring of airway symptoms.

1. Introduction

The wearable device technology is widely adopted in the healthcare community. For complex disease diagnosis and monitoring, multiple physiological signals are continuously streamed to a wearable device and multiple decisions need to be intelligently made within a short time window. The integration of artificial intelligence (AI) into smart wearable devices is particularly needed for effective and accurate processing of health data at the point of care. Most wearable devices are embedded with a sensor, a microprocessor, and a limited memory flash to keep the system small and lightweight. However, such constrained computational environments make the deployment of advanced AI techniques very challenging.[1]

Cough and other audible sounds (e.g., wheezing, deviated voice quality, etc.) have been used as digital audio biomarkers for early disease detection or predicting acute exacerbations in airway diseases such as asthma, chronic obstructive pulmonary diseases (COPD), and COVID-19.[2–4] Most wearable health devices for airway diseases are built on audio sensing technology to detect aforesaid airway symptoms with an embedded microphone.[5] These acoustic microphones are often omnidirectional and capture both, a speaker's voice and surrounding sounds. Wearing a constantly recording microphone creates inevitable personal privacy concerns.

R. Groh, A. M. Kist
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Werner-von-Siemens-Straße 61, 91052 Erlangen, Germany
E-mail: andreas.kist@fau.de

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202100284.

DOI: 10.1002/aisy.202100284

Z. Lei, L. Martignetti, N. Y. K. Li-Jessen
School of Communication Sciences & Disorders
McGill University
Canada 2001 McGill College, 8th floor, Montreal, Quebec H3A 1G1, Canada
E-mail: nicole.li@mcgill.ca

N. Y. K. Li-Jessen
Department of Otolaryngology-Head & Neck Surgery
McGill University
Montreal H3A 1G1, Canada

N. Y. K. Li-Jessen
Department of Biomedical Engineering
McGill University
Montreal, Canada 2001 McGill College, 8th floor, Montreal, Quebec H3A 1G1, Canada

As a promising alternative, mechano-acoustic sensing devices, such as those of neck surface accelerometers (NSAs), are noise resistant and equally capable of generating airway health-related information.[6–9] An NSA device detects and transfers mechanical vibrations from the neck skin to electrical voltage signals. The sensor captures negligible vocal tract resonance information during phonation, which preserves a person's speech privacy.[10,11] The sensor is also insensitive to air-borne acoustic waves,[12] which ensures high-quality data acquisition due to its intrinsic anti-interference against background noise. In contrast, the attenuation of frequency information in NSA signals may make the AI classification tasks more challenging compared to that of microphone-acoustic signals.

AI and related deep learning technologies have been shown to accelerate the time course and improve the quality of disease diagnosis and treatment monitoring.[13,14] Recently, advanced AI methods have been adopted for classifying airway-related symptoms such as cough[15,16] and deviated voice quality[17] in various clinical populations. Lean models have been proposed for the detection of cough in patients suffering from chronic cough, COPD, asthma, and lung cancer.[5] Cough detection is also helpful in predicting COVID-19 infection.[4] However, these deep neural networks (DNNs) have barely been optimized for wearable devices. Further, not many algorithms are capable of multiclass classification in detecting more than one airway symptom[18] Also, given the black-box-character of AI algorithms, explainable AI has been advocated to increase trust among users and decision-makers,[19–22] especially in the development of health wearable devices.[23,24]

In this work, we aimed to optimize multiple neural network topologies using evolutionary algorithms to allow explainable, personalized airway symptom detection as well as deployable on a wearable device (**Figure 1** for study overview). In this work, research questions were: 1) Would NSA signals be on par with audio signals in terms of classification accuracy? 2) Which AI technologies would suit for classifying airway symptoms? 3) Would the proposed evolutionary algorithms be capable of optimizing DNN topologies to gain neural networks for the deployment on wearable devices operating with low memory and computing resources? 4) Would the optimized DNNs be able to cope with new, unseen datasets? 5) Would the optimized DNNs rely on specific features, i.e., frequencies of NSA signals in airway symptom classification?

2. Experimental Section

2.1. Datasets

Three individual datasets, which contained airway symptoms of interest, were curated from laboratory-generated or public sources. These datasets were from: 1) a study of reading a standard passage scripted with airway symptom productions (Rainbow Passage dataset), 2) a published study of vocal loading tests (Vocal Stress dataset)[9] and 3) a crowdsourcing COVID-19 cough sound project (COUGHVID dataset).[25]

2.1.1. Rainbow Passage Dataset

This human study was approved by McGill University Research Ethics Office (A11-B62-19A). All participants of this study gave their informed, written consent. Six female adult participants, who were vocally healthy with ages ranging between 20 and 35, were recruited for this experiment. Both audio (ICD-UX565F, Sony Inc., Japan) and NSA data were recorded simultaneously (**Figure 2**A) in a sound-proof booth. Participants were first prompted to produce isolated cough, throat clear, and dry swallow sounds. Participants were then asked to read the Rainbow Passage, which was scripted with the three airway symptoms interspersed throughout, using their conversational pitch and loudness. They read this script three times in a row (Figure 2B, Supplementary Material).

The main unit of the NSA was a printed circuit board embedded with a one-axis accelerometer (BU-27 135, Knowles Inc., IL, USA) and a custom amplification module to pre-process and transmit the signal to a recording device. A total of 294 coughs, 287 dry swallows, and 382 throat clears were obtained in this dataset. Figure 2C shows representative examples (paired audio/NSA signals and the corresponding log-mel-spectrograms) for the three symptom classes. Data were annotated by two experts who had more than five years of clinical voice evaluation experience, using a custom graphical user interface written in Python.

2.1.2. Vocal Stress Task Dataset

In addition to the Rainbow Passage dataset, we sought to obtain data samples of airway symptoms that were elicited in a relatively natural setting. Our published dataset, in which the airway symptoms were produced spontaneously by speakers during a vocal

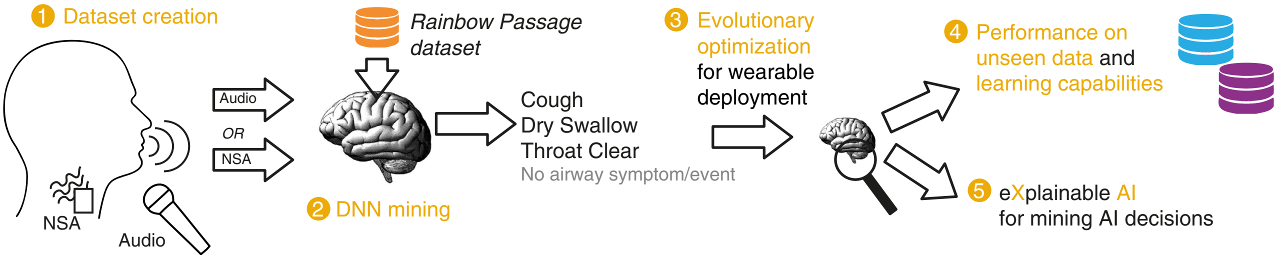
1 Dataset creation
NSA
Audio
Audio
OR
NSA
Rainbow Passage dataset
2 DNN mining
Cough
Dry Swallow
Throat Clear
No airway symptom/event
3 Evolutionary optimization for wearable deployment
4 Performance on unseen data and learning capabilities
5 eXplainable AI for mining AI decisions

Figure 1. Overview of this study as a flow diagram. Orange circles indicate milestones in the project, where the milestones 1-5 are reflected in Figures 2–6 in this study.

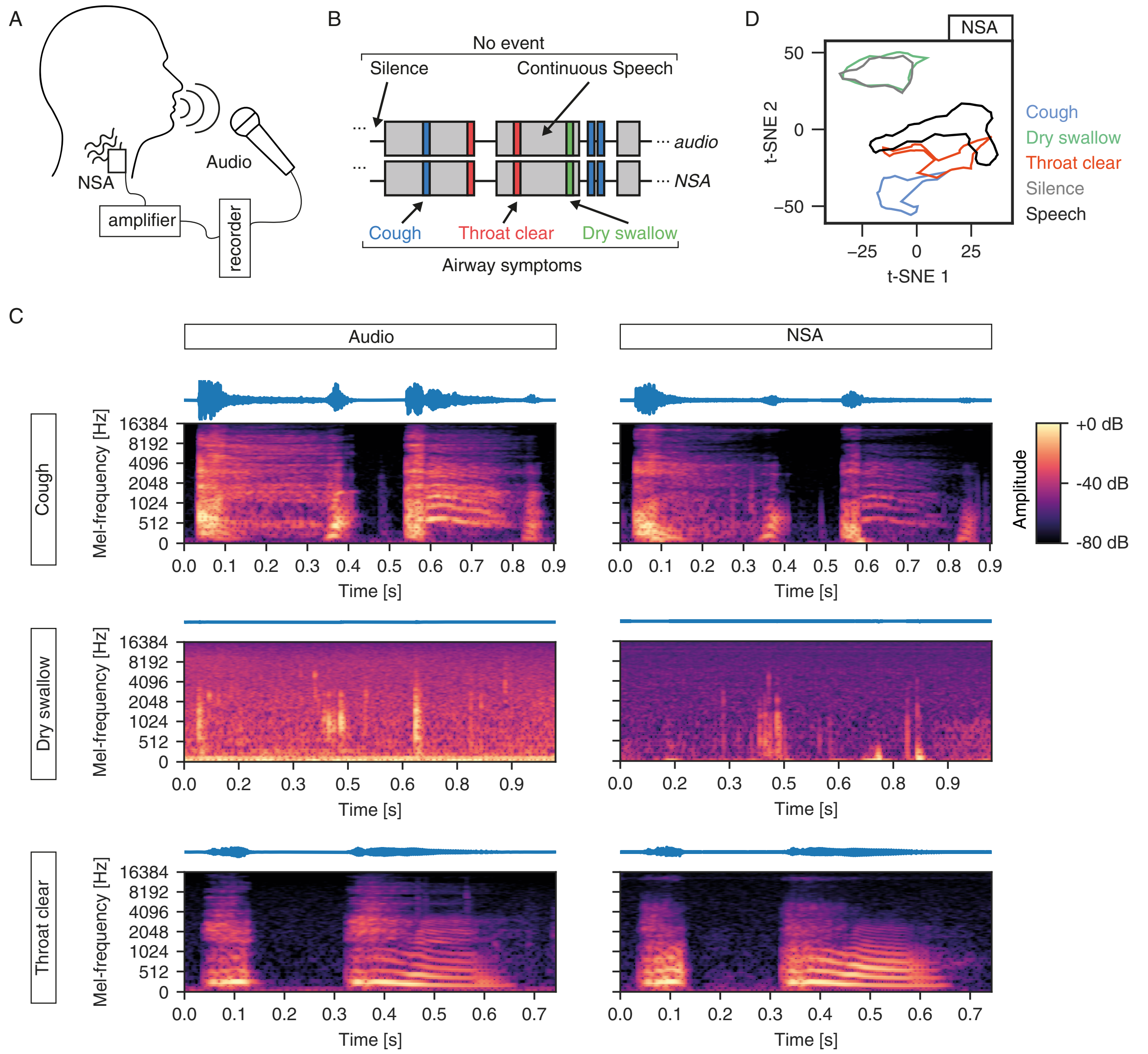

Figure 2. Airway symptoms in neck surface accelerometer (NSA) signals. A) Recording condition of the Rainbow Passage dataset. B) Schematic of the annotated paired audio and NSA data. Silence and speech are labeled both as "no event", whereas cough, throat clear, and dry swallow are distinct categories. C) Representative log-mel-spectrograms of paired audio and NSA signals for each airway symptom. D) t-SNE representation of all categories described in panel B.

stress task, was selected for this study.[9] In brief, nine female adult participants were asked to read parts of the novel "Harry Potter and the Sorcerer's Stone"[26] for up to 3 h. Both audio and NSA signals were obtained in a sound-booth environment using the same devices as those of the aforesaid Rainbow Passage dataset experiment. A total of 19 coughs, 258 dry swallows, and 11 throat clears were annotated in this dataset.

2.1.3. COUGHVID Dataset

Cough is one most common symptoms in airway disease diagnosis and monitoring. To further evaluate our AI algorithm, a highly heterogeneous dataset of coughs containing more than 20 000 recordings from all gender groups was collected from the COUGHVID crowdsourcing dataset.[25] The predictions of the classifiers were already stored in the original COUGHVID data files by the original authors.[25] We thus pre-selected the cough admissions with more than 98% classifier probability. Given that non-cough parts were also contained in the recordings, we computed the rolling standard deviation with a window size of 5,000 sampling points and an energy threshold of 8,000 to determine the onset of the cough event. As a result, a total of 3,388 cough events were obtained for the evaluation of our AI algorithms in this study. Of note, as these cough sounds were microphone audio signals, an auto-encoder DNN architecture was applied to convert the audio samples to NSA space (Figure S5, Supporting Information).

2.2. Data Preprocessing

During preprocessing, both audio and NSA data were divided into 500 ms frames. For each frame, the mel-spectrogram was calculated using 64 mel-frequency-bands, an FFT window length of 1024, a hop length of 64, an upper frequency bound of 16 384 Hz, and the HTK-formula[27] for conversion from Hertz to mel. The advantage of mel-spectrograms is that the center frequency and bandwidth of the chosen triangular filters roughly match the auditory critical band filters.[28] Using the Python package librosa,[29] each 500 ms frame resulted in a mel-spectrogram with 64 frequency points and 345 time frames.

Other preprocessing steps included calculating the log-mel-spectrogram, scaling the values in the range of −1 to 1 (min-max normalization), flipping the spectrogram such that lower frequencies were at the bottom of the spectrogram, and resizing the spectrogram to 64 × 64 data points, which we further used as an image-like object in pixels. Finally, a class label was assigned to each of the log-mel-spectrograms. If more than 70% of the 500 ms window belonged to an annotated event, the log-mel-spectrogram was labeled accordingly, i.e., Cough, Dry swallow, Throat clear, or No event. These log-mel-spectrograms and their associated labels were treated as inputs and outputs, respectively, to various DNN architectures for the classification of airway symptoms.

2.3. Data Visualization

The t-distributed stochastic neighbour embedding (t-SNE) dimensionality reduction technique[30] was used to visualize the relationship between the three airway symptoms. Input to the t-SNE algorithm were log-mel-spectrogram-derived features, including the mean, min, max, median, mode, and standard deviation of each coefficient of one mel-frequency band. These extracted features were then projected onto two t-SNE dimensions. The t-SNE algorithm was implemented using the scikit-learn library[31] with the perplexity set as 40, the learning rate as 30, and the number of iterations as 1500. We computed the alpha shape of each class and reported individual clusters to illustrate the possible overlapping of multiple classes.

2.4. Network Architectures and Training

We evaluated five state-of-the-art network architectures in the classification of cough, dry swallow, throat clearing, and no event (**Figure** 3A). We focused on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), specifically, the ResNet architecture (ResNet18 and ResNet34[32]), EfficientNetB0,[33] the MobileNetV2,[34] an Encoder–Decoder–RNN,[35] and a vanilla RNN that we developed specifically for this study. The vanilla RNN consisted of three Long Short-Term Memory (LSTM) layers[36] of 128, 64, and 32 cells, respectively, followed by a fully connected layer with softmax activation function. We included this straightforward architecture in our experiments to have a second RNN architecture as a reference.

All experiments were implemented using Google TensorFlow (version 2.5.0 with keras API) on an NVIDIA GeForce RTX 3090 GPU and an Intel Core i9-10900X CPU. During network training, the Adam optimizer[37] was used to minimize the categorical cross-entropy loss. The learning rate was 10^{-4} with an exponential decay over time. Since the entire dataset had an imbalance of events and non-events, scikit-learn was used to calculate class-dependent weights for model training.[31] Models were trained on the whole Rainbow Passage dataset and optimized with the Vocal Stress and COUGHVID datasets. As the Rainbow Passage dataset was generated from six speakers only, a six-fold

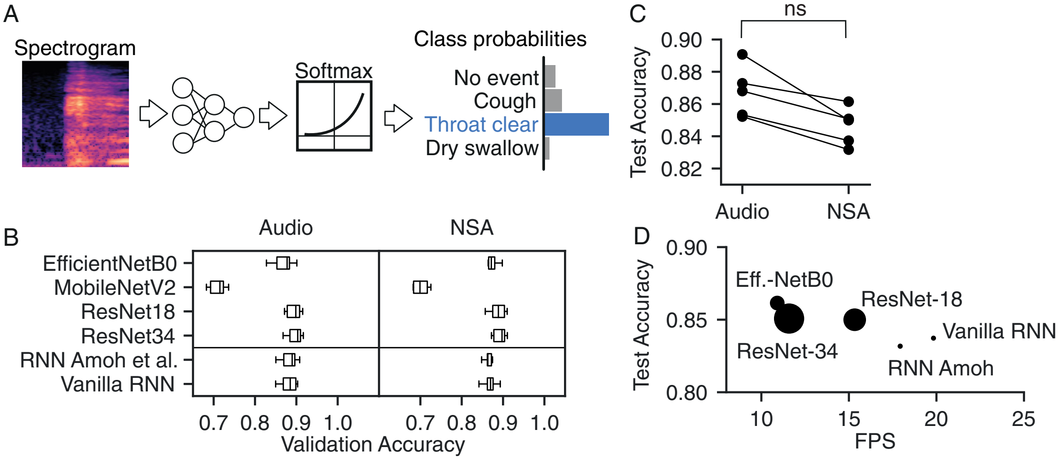

Figure 3. Training of six convolutional neural networks (CNN) and recurrent neural networks (RNN) architectures for classifying airway symptoms. A) Overview of the classification pipeline. Log-mel-spectrograms obtained as preprocessing step are the input of neural network architectures for predicting airway symptoms after softmax activation. B) Validation accuracy for different training/validation splits of the Rainbow Passage dataset for audio and NSA signals. C) Median test accuracy from all test accuracies determined by cross-validation. D) Fps during model inference and median test accuracy from cross-validation for NSA data. The size of the points correlates with the number of trainable parameters of each architecture.

cross-validation approach was used in the scheme of four speaker dataset for model training, one speaker dataset for model validation, and one speaker dataset for model testing.

2.5. Domain Adaptation

COUGHVID crowdsourcing dataset was used to further evaluate the performance of our network architectures in handling complex and heterogeneous data. Since the recordings are microphone audio samples, each COUGHVID sound sample was converted to NSA space for our application. We utilized a U-Net[38] autoencoder that was already trained with the paired audio and NSA signals from the Rainbow Passage dataset. The architecture consisted of four encoder and four decoder layers with 8, 16, 32, and 64 convolutional filters at each depth layer that was connected via skip connections. During the training, mean squared errors were minimized using the Adam optimizer with an exponentially decaying learning rate. We used the tanh-activation function in the output layer to ensure that the output log-mel-spectrograms would contain values in the range from -1 to 1 as noted in the Data Preprocessing section.

2.6. Evolutionary Optimization

To find a small and efficient CNN architecture, we utilized an evolutionary algorithm (see[39] for an overview) to select the best possible combination of neural network elements as noted in **Table 1**. In other words, we used the gene pool in Table 1 to determine a novel neural network topology that would be ideally as accurate as state-of-the-art models, but usable in wearable computing. The algorithm was allowed to evolve for 20 generations with a population size of 50 in each generation. The individuals of the first generation were created randomly. After each generation, 15 models with the highest fitness scores were selected and used for breeding the next generation's population. We further employed a mutation rate of 10% during breeding. The fitness for each architecture was calculated using the validation accuracy and the inference time and was defined as follows

$$F_i = a_i + \beta \cdot \frac{1}{t_i} \tag{1}$$

with F as fitness, a the validation accuracy, β the inference time weight, and t_i the inference time in seconds of a single frame for each architecture i. For objective 1 (O1), we set $\beta = 0$ to evolve only based on accuracy. For objective 2 (O2), we set $\beta = 0.05$ to balance accuracy and time dependence. Each architecture was trained for 12 epochs.

Table 1. Overview of the used gene pool in evolutionary optimization.

Parameter	Set of possible values
Number of convolutional layers	[1-5]
Number of convolutional filters	[8, 16, 32, 64, 128]
Convolutional filter size	[3]
Max pooling layer	[True, False]
Residual Connections	[True, False]
Batch Normalization layer	[True, False]
Activation function	[ReLU, ReLU6, LeakyReLU]

2.7. Microcontroller Deployment

To evaluate the scalability of our DNNs, we converted evolutionary optimized models to TensorFlow Lite according to standard procedures. We deployed the converted model to a development board (EdgeBadge, Adafruit Industries) as a C array and measured the inference time per single forward pass. An average of 100 single forward passes were reported herein.

2.8. Class Activation Maps and Occlusion Experiments for Explainable AI

Class activation maps (CAMs)[40,41] and occlusion experiments[42] were employed to explain neural network decisions. We created the CAMs for each log-mel-spectrogram of the test split of the Rainbow Passage dataset. We calculated the weighted sum of each output of the last convolutional layer as described in.[40] The class weights of the last network layer (fully connected layer) were used for weighting purposes. Further, we averaged all resulting CAMs of each input log-mel-spectrogram to determine which mel-frequency bands would be of high importance for classification.

The occlusion experiments were performed using a sliding window of size 16 × 16 pixels and a stride of four pixels. The values in the windowed regions were set to -1 to hide the corresponding information. We then used our trained neural networks for inference to obtain and store the corresponding prediction probabilities for each occluded log-mel spectrogram. Due to overlapping windows, we averaged the pixel values gained from the multiple predictions for reporting purposes.

2.9. Statistical Testing

Wilcoxon matched-pair tests and paired t-tests were used for sample populations with non-normal (Figure 3) and normal distribution, respectively (**Figure 4**). Bonferroni correction was used to adjust the significance level for multiple testings to reduce Type 1 error. For instance, with three groups and three comparisons (Figure 4), the corrected significance level became $\alpha_{\text{new}} = 0.05/3 = 0.017$.

3. Results

3.1. Detectable Airway Symptoms from NSA Signals

Labels were created for "no event" (continuous speech and silence) and "event" (cough, throat clear, and dry swallow) of the Rainbow Passage dataset during the expert annotation task (Figure 2B). Representative audio and NSA data pairs for the three airway symptoms are shown in Figure 2C. We found that audio and NSA data shared qualitative similarities in the low-frequency bands. Given the low-pass filter quality of the

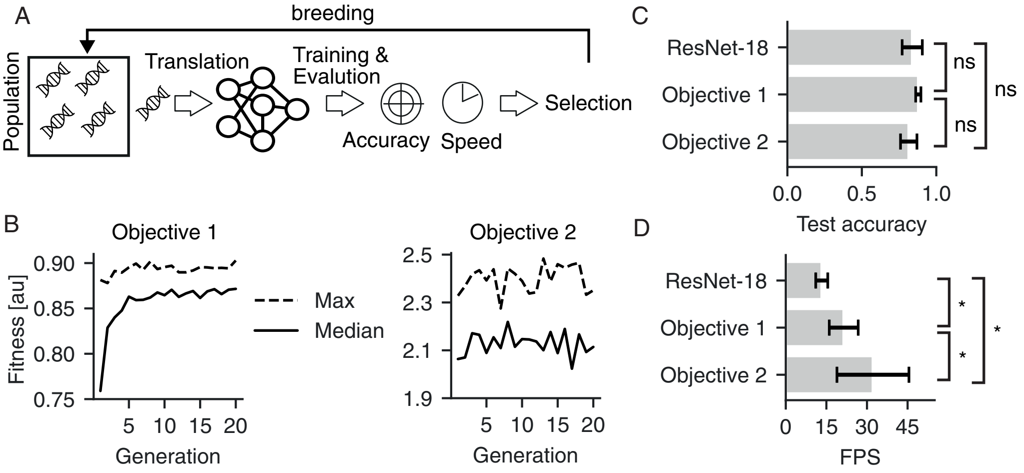

Figure 4. Evolutionary optimized deep neural networks are competitive with state-of-the-art networks. A) We used an evolutionary approach to evolve neural network topology. We selected the network parameters from a genetic pool (Table 1), translated the parameters in a trainable deep neural network, trained and evaluated each individual, and selected for the fittest. These were used to breed the next generation, employing cross-overs and random mutations. B) Convergence of fitness across generations for objective 1, i.e., optimizing for accuracy only (left panel) and objective 2, i.e., optimizing for accuracy and inference speed (right panel). C) Comparison of test accuracies from objective 1 and 2 models compared to baseline (ResNet-18) as mean ± sd. D) Comparison of inference speeds of objective 1 and 2 models compared to baseline (ResNet-18). Statistical significance is indicated with an asterisk ($p < 0.017$).

NSA, higher frequencies were less present as expected. Distinctive clusters were noted in cough, throat clear, and speech (Figure 2D) under t-SNE representation. Whereas the clusters of dry swallow and silence were highly overlapping, which might lead to difficulty in detecting dry swallow events. Based on these results, airway symptom information was reliably preserved and detected in NSA signals.

3.2. Multi-Class Classification of Airway Symptoms with DNNs

Two major DNN technologies, namely CNNs and RNNs, were evaluated for their suitability of airway symptom detection. A library of standardized log-mel-spectrograms was generated from the annotated Rainbow Passage dataset to train, validate and test each DNN by forcing the network to choose one of the four classes (Figure 3A). The examined CNNs and RNNs were found to operate within a similar validation accuracy range and were largely independent of the recording modality (Figure 3B, Figure S1 and S2, Supporting Information). MobileNetV2 was the only architecture that notably underperformed compared to other architectures (Figure 3B).

We next determined the test accuracy for each architecture using cross-validation. The median test accuracy for each architecture in airway symptom prediction was slightly worse with NSA signals but not statistically significant compared to audio signals (Wilcoxon test of paired samples, $p = 0.06$) (Figure 3C). CNN-based models were also found to be more accurate than RNN-based models. However, CNN-based models were in general slightly slower in terms of frames per second (fps) (Figure 3D). The distribution of all test accuracies across all cross-validations can be found in detail in Figure S3, Supporting Information.

Subsequently, we evaluated if the accuracy of a CNN-based model could be traded for inference speed. As a baseline, we chose the ResNet-18 architecture, as it provided a high median test accuracy of 85.0% as well as 15.3 fps in classifying airway symptoms, and was already a smaller variant of the ResNet-34 architecture. Both RNNs showed higher fps (17.9 for RNN Amoh and 19.8 for Vanilla RNN) with comparable, but lower median test accuracy (83.17% and 83.73%) to those of CNNs (Figure 3D).

In summary, we were able to show that NSA signals contained sufficient data features for airway symptom detection in combination with DNN techniques. All investigated DNNs were, however, too large for wearable deployment. We thus proceeded to optimize the network topology with the focus on CNNs next, given their superior accuracy in airway symptom classification.

3.3. CNN Topology Optimization Using an Evolutionary Algorithm

An efficient classifier is integral for mobile health wearable deployment. Here, we investigated how to optimize CNNs in a directed fashion to allow both fast and accurate classification by being wearable and deployable. Evolutionary algorithms (Figure 4A) were used to optimize CNN topology using either of the following two objective functions. Objective 1 (O1) was to maximize the validation accuracy of a CNN topology. Objective 2 (O2) was to maximize both validation accuracy and the model's processing fps. Both objective functions were found to increase their median and maximum fitness across generations (Figure 4B). Due to the evolutionary algorithm and its mutation and cross-over features, there was a gap between median and maximum possible fitness. The distribution of individual topology genetics across generations is shown in Figure S4, Supporting Information.

Paired *t*-tests were performed to compare the accuracy and inference speed across the three architectures (Figure 4C,D).

Regarding the prediction accuracy, the baseline ResNet-18, O1 and O2 achieved comparable results as 83.84%, 87.91%, and 81.49%, respectively (*t*-tests: ResNet-18 vs O1 $p = 0.145$; ResNet-18 vs O2 $p = 0.124$)(Figure 4C). With respect to the inference speed, the evolutionary algorithm was noted to boost the processing speed significantly from 13.3 to 32.2 fps (Figure 4D, **Table 2**). The O2 model was found significantly faster than the ResNet-18 and the O1 architectures (both *t*-tests: $p < 0.017$).

To test the network performance in a real-world microcontroller environment, we deployed the ResNet-18 and the optimized O1 and O2 architectures to a developmental Deep Learning-enhanced board (EdgeBadge Board, see also Methods). Unfortunately, due to the large model sizes (44 and 3.5 MB for ResNet-18 and O1, respectively), we were not able to deploy these models to the microcontroller, which was restricted to 512 kB of memory. However, once we converted and deployed the evolutionary optimized model O2 with TensorFlow Lite, we were able to gain 3.5 fps, which is considered a reasonable result for a non-optimized hardware board. In summary, the final O2 architecture consisted of 7,692 trainable parameters, which was 0.069% of ResNet-18's parameter space (11 186 692 trainable parameters), with a model file size of less than 150 kB (**Table 3**). This new O2 fitness design was shown capable to trade accuracy with inference speed in support of wearable computing (Table 2).

3.4. Adaptability of Evolutionary Optimized CNNs to New Data

We used the pre-trained O2 model on the Rainbow Passage dataset of the previous section for further analysis. To evaluate how well this evolutionary optimized CNN would be capable to adapt to new data, we performed experiments involving transfer learning and fine-tuning techniques (**Figure** 5A).

Table 2. Evolutionary algorithm results compared to the baseline model.

Architecture	Test Accuracy	FPS	Number of parameters
ResNet-18 (Baseline)	0.838 ± 0.068	13.29 ± 2.22	11 186 692
Objective 1	0.879 ± 0.017	21.4 ± 5.34	289 988
Objective 2	0.815 ± 0.055	32.2 ± 13.3	7 692

Table 3. Architecture determined by evolutionary algorithm and optimized with objective 2. Each row i is one building block with operators $\hat{\mathscr{F}}_i$, input resolution $\langle \hat{H}_i, \hat{W}_i \rangle$ and output channels $\hat{C}_i$ determined from the set of possible values shown in Table 1.

Stage i	Operators $\hat{\mathscr{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$
1	Conv3 × 3 & Relu6 & Max Pooling	64 × 64	8
2	Conv3 × 3 & Relu6 & Max Pooling	32 × 32	8
3	Conv3 × 3 & Relu6 & Max Pooling	16 × 16	32
4	Conv3 × 3 & Relu6 & Global Average Pooling	8 × 8	16
5	Dense & Softmax	16	4

The test accuracy of the Rainbow Passage dataset was preserved when fine-tuning the pre-trained O2 model with a combination of the Rainbow Passage dataset and a subset of the Vocal Stress dataset,[9] which also contains airway symptom data in NSA space. However, the test accuracy of the new dataset converged at about 0.7, suggesting that the model was only able to adapt to the additional data from the Vocal Stress dataset to a certain extent (Figure 5B, right panel). When we relied on a pure transfer learning task and used only the Vocal Stress dataset in the training process, the pre-trained model was able to learn the Vocal Stress dataset representation quickly (Figure 5B, left panel). While performing better on the test set, the test accuracy of the Rainbow Passage dataset dropped from 80% to 22.2%. In summary, we were able to show that the proposed evolutionary O2 model was capable to retain the Rainbow Passage dataset representation when used in a fine-tuning task, and to adapt to new data quickly, despite its small parameter space (Figure 5B, Table 2).

Next, we sought to test the adaptability of our proposed architecture to a separate data source. We utilized the COUGHVID dataset that contains a variety of cough audio samples gained from a crowdsourcing effort. To convert the audio samples to NSA space, a crucial step for testing the data on the O2 CNN, we trained a U-Net-like architecture with the paired audio/NSA data extracted from the Rainbow Passage dataset by minimizing the mean squared error across spectrograms (Supplementary Figure 5 for workflow). We analyzed the conversion quality using the structural similarity index measure (SSIM,[43]). Our conversion procedure was able to produce valid spectrograms in NSA space: SSIM results showed that converted NSA spectograms were closer to real NSA spectrograms (mean SSIM = 58.1%) compared to audio-derived spectrograms (mean SSIM = 37.6%).

We next analyzed the O2 model to classify the obtained NSA samples of the COUGHVID dataset. The prediction probabilities of cough samples were relatively low before transfer learning (Figure 5C, gray bars). However, by fine-tuning the model on a few samples of previously unseen data (50 converted NSA cough samples from the COUGHVID database), the test accuracy on the remaining 3,138 cough samples increased dramatically from 22.8% to 70% on average (see Figure 5C, pink bars). After 48 epochs, the model reached a similar performance for both datasets before overfitting was observed (Figure 5C, left panel).

Although the O2 model might not be robust to different sources, the model was shown to quickly adapt to new datasets. This feature is particularly useful to support personalized wearable health technology. For cases like chronic airway diseases, an individual's health data are dynamically evolved as functions of time history and personal profiles. The adaptability of the O2 model will allow continuous integration of novel data, focusing on fine-tuning, to improve its prediction accuracy when further data are supplied from individual patients.

3.5. Airway Symptoms Cluster on Specific Frequency Bands

Our next interest was to investigate if the optimized CNN O2 relied on specific frequency bands for airway symptom detection. That way, in case of confined frequencies, we would be able to

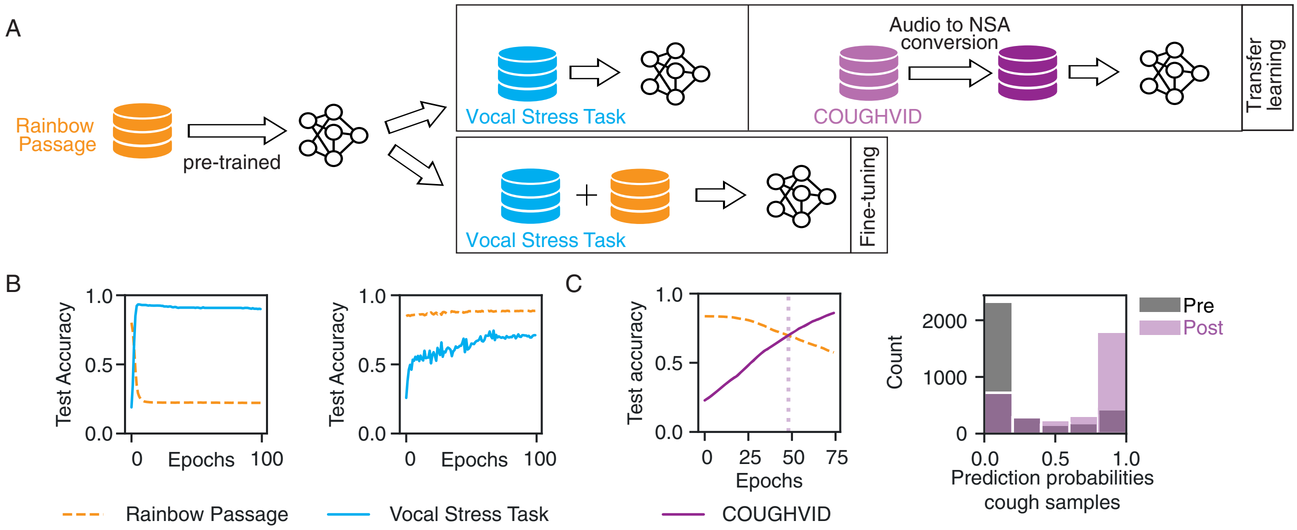

Figure 5. Transfer learning and fine-tuning determined genetic algorithm model using two additional datasets. A) Improving the pre-trained evolutionary optimized architecture O2 performance and robustness through transfer learning and fine-tuning using related datasets. B) Evolution of test accuracy over epochs when using the Vocal Stress Task dataset without (left) and with (right) the Rainbow Passage dataset for transfer learning and fine-tuning the determined genetic algorithm model. C) Transfer learning of the determined genetic algorithm architecture using the COUGHVID dataset.

further optimize the preprocessing steps and the DNNs to gain potentially even smaller or more robust models for mobile and wearable applications. We focused on two complementary approaches: occlusion experiments[42] and CAMs.[40] The occlusion experiments (**Figure 6**A, Figure S6, Supporting Information) showed that mel-frequency bands up to 8000 Hz were most important for classifying coughs, dry swallows and throat clears (Figure 6D–F). Mel-frequency bands of up to 2000 Hz were important for the correct classification of no event (Figure 6C). When analyzing the respective CAMs, we found confirming results for each event, with the exception of dry swallow (Figure 6C–F, Figure S7 and S8, Supporting Information). Class activations were higher in the same mel-frequency bands where predictions dropped off when the bands were occluded. Taken together, airway symptom features were restricted to specific frequencies, allowing not only trustworthy applications but also leaner future models.

4. Discussion

In this work, we showed that a scripted, tiny dataset was sufficient to train DNNs in the classification of airway symptoms on unseen data with satisfactory testing and validating

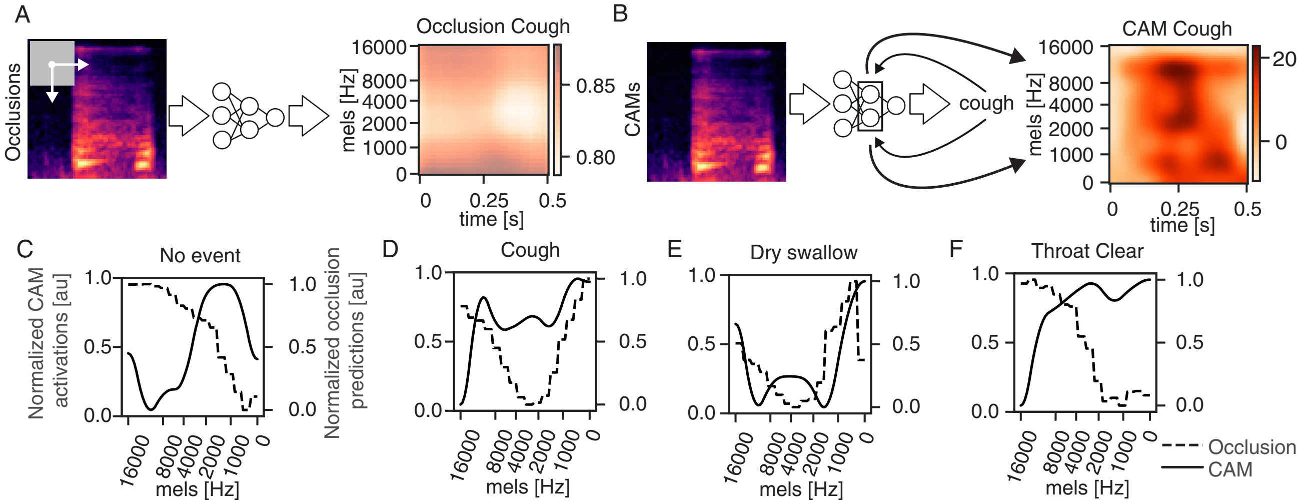

Figure 6. Specific frequency bands are important for cough detection. A) Parts of the mel-spectrogram are occluded to determine which frequencies are important for successful event prediction, here cough. B) Class activation maps are used to determine important spatiotemporal frequencies for event predictions, here cough. C–F) Normalized class activation map (CAM, solid line) and normalized occlusion prediction (dashed line) profiles for no event, cough, dry swallow, and throat clear, respectively.

accuracies. A new neural topology based on evolutionary algorithms was optimized for accuracy, inference speed, and size. In particular, this new DNN was less than 150 kB in size but achieved prediction accuracies on par with those of large state-of-the-art architectures. Such low-size model file sizes are important for downstream applications on mobile and wearable devices. Further, we found that specific frequency bands were important for airway symptom identification, which is essential for us to tailor the proposed DNNs in the future and to solidify trust in wearable health devices.

4.1. Cough Prediction

Coughing is a common symptom across multiple airway-related diseases such as asthma, COPD, and COVID-19. Cough sounds have become a useful digital audio biomarker in mobile health technologies.[15,16,35] For example, coughs were used to predict COVID-19 positivity.[44,45] In this work, we used the COUGHVID database,[25] a large crowdsourced database containing audio recordings with a wide range of perceptual audio quality of mainly cough sounds. A data cleaning strategy was tailored to extract cough sounds and convert the microphone audio signals to NSA space for model evaluation. Our proposed models were able to adapt to this diverse database. In contrast to other works, we specifically aimed for deployment of our DNNs classification algorithms on low-power, computational restrictive wearable platforms.[46,47]

4.2. Unboxing Deep Neural Networks

Explainable AI is integral to advancing and translating the technologies to clinical applications.[48,49] The occlusion experiments and CAMs (Figure 6) identified frequencies that were specific to airway symptom prediction. As a sanity check, we confirmed that log-mel-spectrograms classified as no event, which typically contained human speech signals, showed important fundamental and harmonic frequencies of up to 2000 Hz as expected in human speech. We also confirmed that CAMs were similar across architectures (such as ResNet-18 and the O2 optimized model), suggesting that the same concepts were learned, despite the fact that the latter network features less than 1% of the trainable parameters of the former. More recent model interpretation methods such as Grad-CAM[50] and DeepLift[51] can also be included in future studies as suggested by a recent review article.[52]

4.3. Limitations and Shortcomings

In this work, we achieved the first step of developing effective and explainable AI algorithms for long-term remote monitoring of airway symptoms by mechano-acoustic wearables. Despite the fact that coughs were classified with satisfactory accuracies with our CNNs, the heterogeneity of the data resulted in a large fraction of false positives. The expansion of our current training set, i.e., the Rainbow Passage dataset, may help to further improve the classification accuracy and especially robustness to various sources. Nevertheless, we were able to gain competitive results with our limited training dataset. We also noted that the genetic pool in our evolutionary algorithm was relatively limited and can be further expanded, for example, by using depth-wise convolutions[53] or compound scaling.[33] In the future, we will investigate additional topology optimizations and test the resulting topologies in a real-world application.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

N.Y.K.L.-J. and A.M.K. contributed equally to this work. The authors thank Tobias Kist for help with electronics. The authors acknowledge research grants from the Canadian Institutes of Health Research [PJT-156412], The Centre for Research on Brain, Language and Music Research (CRBLM) Incubator Awards (NLJ), and the Canada Research Chair research stipend (NLJ). The CRBLM is funded by the Government of Quebec via the Fonds de Recherche Nature et Technologies and Société et Culture. The presented content is solely the responsibility of the authors and does not necessarily represent the official views of the aforementioned funding agencies.

Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

N.L.J. and A.M.K. conceived and supervised the project. Z.L. designed the NSA board. L.M. and R.G. annotated data. R.G. analyzed the data, trained and evaluated deep neural networks, and conceived evolutionary algorithms. R.G., A.M.K., and N.L.J. wrote the article. N.L.J. and A.M.K. revised the article critically for important intellectual content.

Data Availability Statement

The data that support the findings of this study are available from corresponding authors upon reasonable request.

Keywords

airway symptoms, deep neural networks, evolutionary algorithms, wearable devices

Received: December 29, 2021
Revised: April 8, 2022
Published online: May 17, 2022

[1] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, U. R. Alo, *Expert Syst. Appl.* **2018**, *105*, 233.

[2] E. Nemati, M. M. Rahman, V. Nathan, K. Vatanparvar, J. Kuang, in *2019 IEEE/ACM Int. Conf. on Connected Health: Applications, Systems and Engineering Technologies (CHASE*), IEEE, Piscataway, NJ **2019**, pp. 15–16.

[3] C. Infante, D. Chamberlain, R. Fletcher, Y. Thorat, R. Kodgule, in *2017 IEEE Global Humanitarian Technology Conf. (GHTC)*, IEEE, Piscataway, NJ **2017**, pp. 1–10.

[4] B. W. Schuller, H. Coppock, A. Gaskell, arXiv preprint arXiv:2012.14553 **2020**.

[5] P. Kadambi, A. Mohanty, H. Ren, J. Smith, K. McGuinnes, K. Holt, A. Furtwaengler, R. Slepetys, Z. Yang, J.-S. Seo, J. Chae, Y. Cao, V. Berisha, in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Piscataway, NJ **2018**, pp. 2161–2165.

[6] K. L. Marks, J. Z. Lin, J. A. Burns, T. A. Hron, R. E. Hillman, D. D. Mehta, *J. Speech Lang. Hear. Res.* **2020**, *63*, 2202.

[7] V. M. Espinoza, D. D. Mehta, J. H. Van Stan, R. E. Hillman, M. Zañartu, *J. Speech Lang. Hear. Res.* **2020**, *63*, 2861.

[8] Z. Lei, E. Kennedy, L. Fasanella, N. Y.-K. Li-Jessen, L. Mongeau, *Appl. Sci.* **2019**, *9*, 1505.

[9] Z. Lei, L. Fasanella, L. Martignetti, N. Y.-K. Li-Jessen, L. Mongeau, *Appl. Sci.* **2020**, *10*, 3.

[10] D. D. Mehta, J. H. Van Stan, M. Zañartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortés, H. A. Cheyne, R. E. Hillman, *Front. Bioeng. Biotechnol.* **2015**, *3* 155.

[11] D. D. Mehta, J. H. Van Stan, R. E. Hillman, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **2016**, *24*, 659.

[12] M. Zanartu, J. C. Ho, S. S. Kraman, H. Pasterkamp, J. E. Huber, G. R. Wodicka, *IEEE Trans. Biomed. Eng.* **2008**, *56*, 443.

[13] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, G. Fortino, *Inf. Fusion* **2021**, *66*, 111.

[14] F. Wang, L. P. Casalino, D. Khullar, *JAMA Internal Med.* **2019**, *179*, 293.

[15] A. Kumar, K. Abhishek, M. R. Ghalib, P. Nerurkar, K. Shah, M. Chandane, S. Bhirud, D. Patel, Y. Busnel, *Trans. Emerging Telecommun. Technol.* **2020**, e4184.

[16] F. Barata, K. Kipfer, M. Weber, P. Tinschert, E. Fleisch, T. Kowatsch, in *2019 IEEE Int. Conf. on Healthcare Informatics (ICHI)*, IEEE, Piscataway, NJ **2019**, pp. 1–11.

[17] J.-Y. Lee, *Appl. Sci.* **2021**, *11*, 7149.

[18] S. Jayalakshmy, B. L. Priya, N. Kavya, in *2020 Int. Conf. on Communication, Computing and Industry 4.0 (C2I4)*, IEEE, Piscataway, NJ **2020**, pp. 1–5.

[19] D. Shin, *Int. J. Hum. Comput. Stud.* **2021**, *146* 102551.

[20] A. Adadi, M. Berrada, in *Embedded Systems and Artificial Intelligence*, Springer, Berlin **2020**, pp. 327–337.

[21] F. K. Došilović, M. Brčić, N. Hlupić, in *2018 41st Int. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE, Piscataway, NY **2018**, pp. 0210–0215.

[22] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, arXiv preprint arXiv:1712.09923 **2017**.

[23] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, *BMC Med. Inf. Decis. Making* **2020**, *20*, 1.

[24] T. Grote, P. Berens, *J. Med. Ethics* **2020**, *46*, 205.

[25] L. Orlandic, T. Teijeiro, D. Atienza, *Sci. Data* **2021**, *8*, 1.

[26] J. K. Rowling, *Harry Potter and the Sorcerer's Stone*, Bloomsbury Publishing, London **2001**.

[27] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, Cambridge **2006**.

[28] R. Islam, M. Tarique, E. Abdel-Raheem, *IEEE Access* **2020**, *8*, 66749.

[29] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, K. L. Dana, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, Viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, T. Kim, Thassilo, librosa/librosa: 0.8.1rc2, **2021**, https://doi.org/10.5281/zenodo.4792298.

[30] L. van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579.

[31] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. G. J. Grobler, R. Layton, J. V. A. Joly, B. Holt, G. Varoquaux, in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, **2013**, pp. 108–122.

[32] K. He, X. Zhang, S. Ren, J. Sun, CoRR **2015**, abs/1512.03385.

[33] (Ed: M. Tan, Q. Le, In K. Chaudhuri, R. Salakhutdinov), *Proceedings of the 36th Inter. Conf. on Machine Learning, volume 97 of Proceedings of Machine Learning Research*, PMLR, **2019**, pp. 6105–6114, http://proceedings.mlr.press/v97/tan19a.html.

[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, **2018**, pp. 4510–4520.

[35] J. Amoh, K. Odame, *IEEE Trans. Biomed. Circuits Syst.* **2016**, *10*, 1003.

[36] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9* 1735.

[37] D. P. Kingma, J. Ba, *Proc. of the 3rd Int. Conf. on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 7–9 **2015**, Conference Track Proceedings.

[38] O. Ronneberger, P. Fischer, T. Brox, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS, Vol. *9351*, Springer **2015**, 234–241.

[39] O. Kramer, *Genetic Algorithms*, Springer International Publishing, Cham, **2017**, pp. 11–19.

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, *Learning Deep Features for Discriminative Localization*, **2015**.

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, *Object Detectors Emerge In Deep Scene Cnns*, **2015**.

[42] M. D. Zeiler, R. Fergus, in *European Conference on Computer Vision*, Springer, **2014**, pp. 818–833.

[43] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *IEEE Trans. Image Process.* **2004**, *13*, 600.

[44] M. Melek, *Neural Computing and Applications* **2021**, Vol. *33*, pp. 17621–17632.

[45] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, M. Nabeel, *Inf. Med. Unlocked* **2020**, *20*, 100378.

[46] N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, F. Kawsar, in *Proc. of the 2015 Int. Workshop on Internet of Things towards Applications, IoT-App '15*, Association for Computing Machinery, New York, NY, USA, **2015**, pp. 7–12, https://doi.org/10.1145/2820975.2820980.

[47] A. Ignatov, R. Timofte, W. Chou, K. Wang, M. Wu, T. Hartley, L. V. Gool, CoRR **2018**, abs/1810.01109.

[48] E. Tjoa, C. Guan, *IEEE Trans. Neural Networks Learn. Syst.* **2021**, *32*, 4793.

[49] R. Cadario, C. Longoni, C. K. Morewedge, *Nat. Hum. Behav.* **2021**, *5*, 1636.

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, IEEE, Piscataway, NJ **2017**, 618–626.

[51] A. Shrikumar, P. Greenside, A. Kundaje, CoRR **2017**, abs/1704.02685.

[52] T. Pianpanit, S. Lolak, P. Sawangjai, T. Sudhawiyangkul, T. Wilaiprasitporn, *IEEE Sens. J.* **2021**, *21*, 22304.

[53] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, arXiv preprint arXiv:1704.04861 **2017**.