SpiroConfidence: Determining the Validity of Smartphone Based Spirometry Using Machine Learning

Varun Viswanath[1], Jake Garrison[2], Shwetak Patel[1,2]

***Abstract*— Prior work has shown that smartphone spirometry can effectively measure lung function using the phone's built-in microphone and could one day play a critical role in making spirometry more usable, accessible, and cost-effective. Although traditional spirometry is performed with the guidance of a medical expert, smartphone spirometry lacks the ability to provide the patient feedback or guarantee the quality of a patient's spirometry efforts. Smartphone spirometry is particularly susceptible to poorly performed efforts because any sounds in the environment (e.g., a person's voice) or mistakes in the effort (e.g., coughs or short breaths) can invalidate the results. We introduce two approaches to analyze and estimate the quality of smartphone spirometry efforts. A gradient boosting model achieves 98.2% precision and 86.6% recall identifying invalid efforts when given expert tuned audio features, while a Gated-Convolutional Recurrent Neural Network achieves 98.3% precision and 88.0% recall and automatically develops patterns from a Mel-spectrogram, a more general audio feature.**

I. INTRODUCTION

Spirometry is the most widely employed objective measure of lung function. It is central to the diagnosis and management of chronic lung diseases, such as asthma, chronic obstructive pulmonary disease (COPD), and cystic fibrosis. However, a standard spirometer is too expensive and cumbersome to be accessible at home, particularly in low resource regions. To address this issue, work from Larson et al, SpiroSmart [1], [2], has shown it is possible to perform a spirometry test using only the audio data from the microphone of a standard smartphone. Typically, spirometry is done under the supervision of a medical professional to provide the patient with feedback on their technique to ensure there are no errors in their spirometry effort. However, this requires a nontrivial investment of time and money from both the patient and the doctor. For smartphone spirometry to be truly inexpensive and ubiquitous, it must be independent of a medical professional and must therefore automatically determine the validity of efforts using purely the audio data from the smartphone.

Prior work has shown that errors in standard spirometry efforts can be detected automatically using an expert engineered algorithm. Melia et al. [3] achieved a specificity of 95% and a sensitivity of 96% in this regard using the flow-versus-volume curve. However, they performed their analysis on 1022 spirometry curves from a traditional spirometer while our work is focused on data from a smartphone spirometer. In our analysis, precision corresponds to the positive predictive value, and recall to sensitivity.

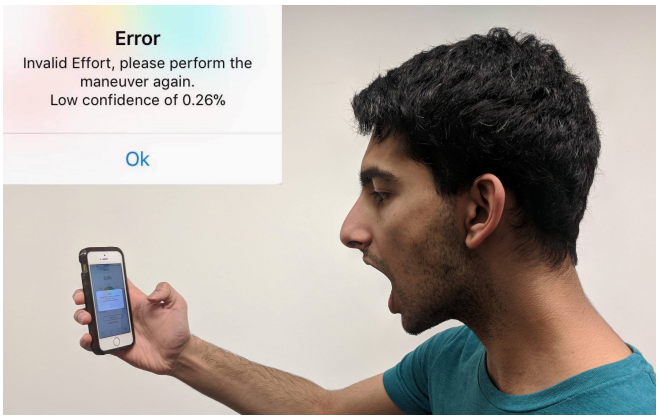

Fig. 1. A patient performs a smartphone spirometry effort. The app using our proposed method recognizes that there was too much background noise, so it rejected the effort from analysis.

SpiroSmart is a smartphone app developed by the Ubicomp Lab from the University of Washington to measure lung function. To use SpiroSmart, a patient holds a smartphone an arm's-length in front of their mouth and then forcefully exhale a full breath. The smartphone records the exhale audio. The microphone acts as an uncalibrated flow sensor because sound is a pressure wave, and pressure is related to flow.

Our work, like that of Melia et al., aims to provide automatic feedback to a patient as they perform a smartphone spirometry effort. An effort should be excluded from analysis if it is incorrectly performed or contains confounding noise. We introduce and compare several approaches to analyze the quality and validity of smartphone spirometry efforts. The first approach uses classical machine learning on an assortment of expert defined sound processing features that have been extracted from the audio recording of the patient's effort. The second approach uses neural networks on a time frequency representation of the audio to automatically find more complex patterns from the general audio features. Our secondary aim is to evaluate how suitable each model is to be run on a smartphone. An ideal model takes as little memory as possible to store while also evaluating validity in as little time as possible. Given these constraints, the Gated-CRNN model performs the best; it achieves 98.3% precision and 88.0% recall on the evaluation dataset of only valid and

*Varun Viswanath, Jake Garrison, and Shwetak Patel are with the University of Washington

[1]S. Patel is Faculty and J. Garrison is a Masters Student in the Department of Electrical Engineering, University of Washington, Seatle, WA 98105, USA

[2]S. Patel is Faculty and V. Viswanath is an Undergraduate student in the Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98105, USA

invalid smartphone spirometry efforts.

II. METHODS

A. Dataset

Our original dataset contained 26,304 audio samples that were collected from patients from clinics primarily in India. The Institutions Ethical Review Board approved all experimental procedures involving human subjects. The audio recordings were recorded with an iPhone 5 at a sampling rate of 44.1 kHz with no compression. The data was collected by having the patient hold the smartphone vertically an arm's-length in front of their mouth, take a deep breath, and then forcefully exhale while holding the phone in place.

A team labeled these recordings as either correctly performed, incorrectly performed, or inaudible. We removed the 1,166 audio recordings that contained inaudible spirometry efforts, and were left with 20,505 valid efforts and 4,633 invalid efforts. In order to balance the number of valid and invalid efforts, we collected 11,023 audio samples from a subset of a cough dataset generated at our university[4], the Urban sound dataset[5], and speech from VoxForge dataset. Coughs, urban sounds, and human speech all invalidate spirometry efforts. These audio samples also help our model generalize better to unseen invalid cases. In summary, our final dataset comprises 36,161 recordings: 20,505 correctly performed the spirometry efforts, 4,633 incorrectly performed the efforts, and 11,023 contained noises that would invalidate a spirometry effort resulting in a more balanced dataset of 20,505 valid attempts and 15656 invalid recordings.

B. Approach #1: Classical Machine Learning Models

Our first approach uses machine learning models trained on an assortment of time-series features selected by an expert with knowledge of spirometry and sound processing techniques. These features include the peak loudness, loudness at different times, the spectral envelope, the duration, the number of peaks, and the coefficients of fitting a polynomial equation to the audio envelope. We also generate a downsampled Mel-spectrogram, a time frequency representation of audio data that is commonly used for speech analysis. The Mel-spectrogram features were computed by downsampling the audio to 16 kHz and then computing 64 Mel's over a window size of 1024 samples with an overlap of 400 samples. The resulting features were downsampled to 8 time slices $\times$ 4 Mels. The final input vector has 142 values.

We designed six machine learning models to train on these features: a Naïve Bayes classifier [6], a K-Nearest Neighbors classifier, a logistic regression model with L1 regularizer, a logistic regression model with L2 regularizer, a Random Forest classifier[7], and a gradient boosting model[8].

In order to solidify our selection of expert features and provide insight as to which features most impacted the outcome, we performed a rigorous feature selection process by analyzing the weights and decision trees learned by our models. We found the following features to be most effective at classifying an effort; Exhale duration, initial loudness, room noise and envelope polynomial coefficients. This outcome makes sense given that a valid spirometry effort has a finite duration range, anything too long or too short can be rejected with confidence. Valid efforts also have a loud burst at the beginning and a small amount of background room noise. Finally the polynomial coefficients of the envelope are valuable because they effectively summarize the overall shape of the amplitude envelope.

TABLE I
ARCHITECTURES OF NEURAL NETWORK MODELS

VGG style	value	G-CRNN	value
2D Conv Layers	4	1D Conv Layers	3
Number of Filters	32 / layer	Number of Filters	64/32/16
Filter Size	(7, 7)	Kernel Size	2
Pool Size	(2, 2)	Stride Size	2
Fully Conn. Layers	4	Recurrent GRU layers	1
layer size	$\frac{1}{2}$ / layer	number of units	16

C. Approach #2: Convolutional Neural Networks

Our second approach uses convolutional neural networks trained on the Mel-spectrogram described earlier, but without the downsampling at the final step. In other words, the network was trained on a Mel-spectrogram with 128 time slices $\times$ 64 Mels of an audio signal sampled at 16 kHz. We use the full-resolution Mel-spectrogram for the neural network approach because such networks automatically downsample the spectrogram and build complex relations between their values.

We built two Neural Network models that only use the Mel-spectrogram features. One is a nine-layer VGG style convolutional neural network commonly used in image classification[9], and the other is a three-layer Gated-Convolutional Recurrent neural network[10]. The nine-layer VGG style CNN has three repetitions of two convolutional layers followed by a dense layer and a dropout layer. The Gated-CRNN has three convolutional layers followed by a Gated recurrent unit (GRU). Both models use stochastic gradient descent and have dropout of 0.3.

The VGG CNN model trains faster than the Gated-CRNN because it does not contain recurrent cells and can thus be parallelized, despite the larger number of parameters. The Gated-CRNN, having only 3 1D convolution layers and thus fewer parameters, requires much less memory and fewer computations for inference resulting in a much more suitable model for mobile phone usage.

D. Model Training

We first trained the two approaches on their respective features as described earlier. Then, to provide a more fair comparison between the classical models and the neural nets in regards to their feature spaces, we trained each approach on it's respective spectrogram. We had the machine learning models use a downsampled version of the spectrogram the neural networks use because we wanted to keep the input feature vector small, and because Mel-spectrograms have redundant information that can be downsampled without losing insight. The neural networks perform this downsampling

automatically because of the max pooling layers in their architectures'.

We trained on 85% of the dataset, the other 15% was set aside for evaluation using four fold cross-validation as well as parameter grid search to finalize hyperparameters for the classical models.

III. RESULTS

A. Evaluation Datasets

We use two different evaluations datasets. The first evaluation dataset contains a balanced number of valid and invalid smartphone spirometry efforts and includes 5% of the total data. The second evaluation dataset contains a balanced number of valid and invalid audio samples. The invalid audio samples include both poor smartphone spirometry efforts and non-smartphone spirometry effort sounds from the other datasets mentioned. This dataset includes 15% of the total data and is a superset of the first evaluation dataset.

B. Analysis

For our analysis, we frame our accuracy metrics according to the identification of an invalid test; in other words, recordings that involve a poor smartphone spirometry effort or too much noise are assigned into the positive class. As shown in Table II, both approaches tend to have higher precision than recall. This means that if an audio sample is predicted to be invalid, it is more likely to contain invalidating sound. However, if a particular audio sample is invalid, it's less likely to be predicted to contain an error. For our application we want to ensure that we minimize the number of invalid examples that are not identified, even if it means rejecting some well performed efforts because it is better to have a patient perform an effort again than allow an erroneous effort to be used for medical diagnosis.

TABLE II
PRECISION, RECALL OF MODELS ON LARGE EVALUATION DATASET CONTAINING A MIX OF NOISE SAMPLES AND SMARTPHONE SPIROMETRY EFFORTS

	All Features		Mel-Spectrogram	
	Precision	Recall	Precision	Recall
Naïve Bayes	0.943	0.725	0.821	0.717
K-Nearest Neighbors	0.969	0.812	0.970	0.829
Log Reg (L1)	0.963	0.902	0.928	0.824
Log Reg (L2)	0.936	0.875	0.927	0.837
Random Forest	**0.978**	0.905	0.966	0.886
Gradient Boosting	**0.978**	**0.916**	0.965	0.889
VGG CNN	NA	NA	**0.978**	0.927
Gated-CRNN	NA	NA	0.967	**0.928**

The best-performing model is the Gated-CRNN; it has multiple advantages over the other models. It has a recurrent layer that can take better advantage of temporal locality than any of the other models. It uses 1-dimensional convolutions that not only find patterns across the time domain more effectively than the 2-Dimensional convolutions in the VGG CNN, but also pool the Mel-spectrograms down to a reasonable feature map. Additionally, it applies attention to each time slice via GRU cells, allowing the model to isolate the specific patterns over time that characterize good versus bad smartphone spirometry efforts.

TABLE III
PRECISION, RECALL OF MODELS ON SMALL EVALUATION DATASET CONTAINING ONLY SMARTPHONE SPIROMETRY EFFORTS

	All Features		Mel-Spectrogram	
	Precision	Recall	Precision	Recall
Naïve Bayes	0.872	0.347	0.805	0.524
K-Nearest Neighbors	0.965	0.478	0.944	0.523
Log Reg (L1)	0.975	0.690	0.944	0.581
Log Reg (L2)	0.947	0.720	0.938	0.586
Random Forest	**0.989**	0.732	0.961	0.661
Gradient Boosting	0.978	**0.758**	0.963	0.614
VGG CNN	NA	NA	0.970	0.805
Gated-CRNN	NA	NA	**0.983**	**0.88**

The Gradient Boosting model is the strongest of the classical machine learning models, but Random Forest model also performs well in comparison to the other classical machine learning models. This is likely because they both build decision trees that inherently form complex chains of relationships between features. However, neither decision tree model performs as well as the neural networks because they don't intrinsically tend to take advantage of temporal or frequency locality whereas a CNN finds such patterns by sliding a convolution over the matrix of features ordered in time and frequency. The decision tree models were just as likely to associate features from the first and last time slices as the first and second time slices, resulting in complex feature relationships that are less likely to reveal pertinent information about the signal. Nevertheless, we selected the Gradient Boosting model as the best classical machine learning model because of its precision and recall as well as it's shallow decision tree takes less memory and performs faster inference than the Random Forest model's heavy, fully grown tree.

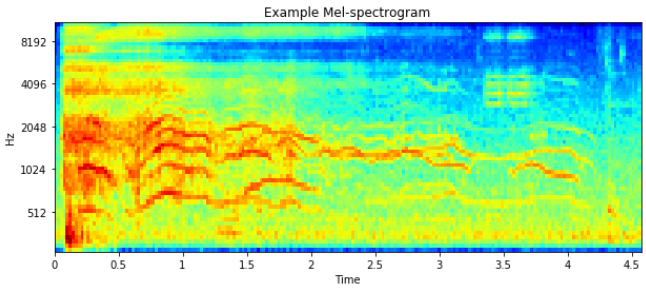

Fig. 2. An Example of the Mel-Spectrogram for a well performed smartphone spirometry effort

The small evaluation dataset is more challenging for all of the models. This is to be expected since spirometry efforts have a very specific shape; a sharp burst followed by a decay to almost nothing at the end. In the large evaluation dataset, most models relied on the very early and very late time slices in the spectrogram to verify the burst or decay to quickly

identify the invalid non-smartphone spirometry sounds. On the other hand, in the small evaluation dataset, models need more nuanced features to analyze the patterns in the middle time slices to identify poorly performed smartphone spirometry efforts. Poorly performed efforts are defined by abnormalities in the decay that must be represented by a relationship over multiple time slices [11]. The variety and subjectivity of the rules that define invalid spirometry efforts is what makes this dataset and this problem challenging. Both the neural networks and the decision tree based models stand out on the small dataset, but because decision trees do not use temporal or frequency locality to develop feature relationships, the neural networks perform the best.

When only given the Mel-spectrogram features rather than the full suite of expert crafted features, the classical machine learning models performed worse on average. The most simplistic models actually performed better in the small dataset, but this is likely because the smaller input vector allowed it to find more consistent patterns. Otherwise, the classical machine learning models perform worse without the features designed with sounds processing and biophysical spirometry expertise. The Gradient Boosting model performed best because its decision trees can extract similar information to the expert designed features. Overall, the classical machine learning models struggled to extract adequate information to consistently identify poorly performed efforts when only using the Mel-spectrogram features.

IV. CONCLUSION

The classical machine learning models were using features that sound processing experts with understanding of the biophysics of spirometry designed and the neural networks still extracted more critical information in the features they built than these experts with domain knowledge. The neural network's performance is likely attributed to their ability to develop higher complexity features such as relationships between different frequency values over multiple slices of time. The classical machine learning models form shallow relationships between *much* smaller sets of time and frequency regions and fail to capture the changes in frequency over time that distinguish poorly performed spirometry efforts.

The two neural networks were fundamentally different in a couple ways. The Gated-CRNN performs a 1-dimensional convolution where the window moves across several time slices while the VGG CNN performs 2-Dimensional convolutions across both time and frequency. This causes the Gated-CRNN to build features that are time dependent while the VGG CNN does not necessarily have this time dependency. The features that truly capture what it means for an effort to be invalid must build complex relationships over different time slices because spirometry efforts contain a sharp peak in the early time slices and then decay through the middle and later time slices. Thus this gives the Gated-CRNN one distinct advantage over the VGG CNN.

The other important difference is that the Gated-CRNN applies attention gates to each time slice in the recurrent layer. This allows the model to isolate specific patterns over time, ignoring the more chaotic or less interesting time slices of the signal that the VGG CNN likely clings to. These key differences are strong reasons to select the Gated-CRNN over the VGG CNN, in addition to its faster evaluation and its more efficient memory consumption which make it more suitable for mobile spirometry.

Although there is still room for improvement, our work has shown that neural networks can extract more information from potentially muddled signals than traditional methods using domain-specific, expert-designed features, that a Gated-CRNN taking less memory and using fewer parameters can perform validity checking more effectively than a very deep convolutional neural network, and that it is possible to provide the necessary expert level validity feedback for smartphone-based spirometry efforts.

We are actively integrating the results of this work in our larger effort in developing a smartphone spirometer. We plan to classify the type of error in an invalid Spirometry effort similarly to how a clinician may coach the patient. With this improvement, the patient can have an idea on how to improve their results, for example the model may tell the patient to head to a quieter room, or exhale for longer. In addition, we aim to optimize our model for offline mobile use so it can be used in environments lacking fast Internet. This work is essential for ensuring the quality of mobile spirometry to match those of clinical spirometers.

REFERENCES

[1] E. C. Larson, M. Goel, G. Boriello, S. Heltshe, M. Rosenfeld, and S. N. Patel, "SpiroSmart: using a microphone to measure lung function on a mobile phone," *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, p. 280, 2012.

[2] M. Goel, E. Saba, M. Stiber, E. Whitmire, J. Fromm, E. C. Larson, G. Borriello, and S. N. Patel, "SpiroCall: Measuring Lung Function over a Phone Call," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5675–5685, 2016.

[3] U. Melia, F. Burgos, M. Vallverdú, F. Velickovski, M. Lluch-Ariet, J. Roca, and P. Caminal, "Algorithm for Automatic Forced Spirometry Quality Assessment: Technological Developments," *PLOS One*, 2014.

[4] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*, p. 375, 2011.

[5] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22st ACM International Conference on Multimedia (ACM-MM'14)*, (Orlando, FL, USA), Nov. 2014.

[6] H. Zhang, "The Optimality of Naive Bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference FLAIRS 2004*, vol. 1, no. 2, pp. 1 – 6, 2004.

[7] T. K. A. T. B. L. Ho, "Random Decision Forests," *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 278–282, 1995.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, pp. 3149–3157, 2017.

[9] K. Simonyan and A. Sizzerman, "Very Deep Convolutional Networks For Large-Scale Image Recognition," *ICLR*, 2015.

[10] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *Computing Research Repository*, 2017.

[11] A. Z. Luo, E. Whitmire, J. W. Stout, D. Martenson, and S. Patel, "Automatic characterization of user errors in spirometry," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 4239–4242, 2017.