FluSense: A Contactless Syndromic Surveillance Platform for Influenza-Like Illness in Hospital Waiting Areas

FORSAD AL HOSSAIN, University of Massachusetts Amherst, USA
ANDREW A. LOVER, University of Massachusetts Amherst, USA
GEORGE A. COREY, University of Massachusetts Amherst, USA
NICHOLAS G. REICH, University of Massachusetts Amherst, USA
TAUHIDUR RAHMAN, University of Massachusetts Amherst, USA

We developed a contactless syndromic surveillance platform *FluSense* that aims to expand the current paradigm of influenza-like illness (ILI) surveillance by capturing crowd-level bio-clinical signals directly related to physical symptoms of ILI from hospital waiting areas in an unobtrusive and privacy-sensitive manner. FluSense consists of a novel edge-computing sensor system, models and data processing pipelines to track crowd behaviors and influenza-related indicators, such as coughs, and to predict daily ILI and laboratory-confirmed influenza caseloads. *FluSense* uses a microphone array and a thermal camera along with a neural computing engine to passively and continuously characterize speech and cough sounds along with changes in crowd density on the edge in a real-time manner. We conducted an IRB-approved 7 month-long study from December 10, 2018 to July 12, 2019 where we deployed *FluSense* in four public waiting areas within the hospital of a large university. During this period, the *FluSense* platform collected and analyzed more than 350,000 waiting room thermal images and 21 million non-speech audio samples from the hospital waiting areas. *FluSense* can accurately predict daily patient counts with a Pearson correlation coefficient of 0.95. We also compared signals from *FluSense* with the gold standard laboratory-confirmed influenza case data obtained in the same facility and found that our sensor-based features are strongly correlated with laboratory-confirmed influenza trends.

CCS Concepts: • **Applied computing → Life and medical sciences**;

Additional Key Words and Phrases: Contactless Sensing, Influenza Surveillance, Edge Computing, Crowd Behavior Mining

ACM Reference Format:
Forsad Al Hossain, Andrew A. Lover, George A. Corey, Nicholas G. Reich, and Tauhidur Rahman. 2020. FluSense: A Contactless Syndromic Surveillance Platform for Influenza-Like Illness in Hospital Waiting Areas. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 1 (March 2020), 28 pages. https://doi.org/10.1145/3381014

1 INTRODUCTION

Influenza, a contagious respiratory viral disease, causes acute illness that typically impacts the nose, throat, and lungs. In the United States, the Centers for Disease Control and Prevention (CDC) estimates that influenza infections lead to 4,000,000-23,000,000 medical visits and 12,000-79,000 deaths each season [50]. The estimated annual economic impact of these infections is $47 and $150 billion in the US alone [31]. Current programs for

Authors' addresses: Forsad Al Hossain, University of Massachusetts Amherst, Amherst, MA, 01002, USA, falhossain@cs.umass.edu; Andrew A. Lover, University of Massachusetts Amherst, Amherst, MA, 01002, USA, alover@umass.edu; George A. Corey, University of Massachusetts Amherst, Amherst, MA, 01002, USA, gcorey@uhs.umass.edu; Nicholas G. Reich, University of Massachusetts Amherst, Amherst, MA, 01002, USA, nick@schoolph.umass.edu; Tauhidur Rahman, University of Massachusetts Amherst, Amherst, MA, 01002, USA, trahman@cs.umass.edu.

2474-9567/2020/3-ART1 $15.00
https://doi.org/10.1145/3381014

influenza surveillance and forecasting by the Center for Disease Control and Prevention (CDC) rely primarily on aggregating reports of outpatient "influenza-like illness" (ILI) from state-level public health agencies for monitoring influenza burden during the course of the transmission season. These reports are based on data from sentinel reporting sites, which include hospitals, and selected outpatient clinics. For reporting purposes, ILI is defined as a fever of 100 °F (37.8 °C) or greater, plus a sore throat and/or cough. ILI levels at the state, regional, and national levels are monitored weekly by CDC throughout the transmission season, and thresholds have been established to classify seasons based on severity [8]. Additionally, virologic surveillance is collected by clinical laboratories affiliated with the CDC, and provides detailed information about the relative prevalence of influenza types and subtypes at a given timepoint [49]. The CDC also monitors hospitalization and mortality rates due to influenza. However, there are several important limitations with these existing surveillance systems. Most notably, there is a substantial lag time in the reporting of these data [35]. For example, reports of outpatient visits routinely are only available 7-14 days after the visit itself, and during holiday periods this delay can be extended. This time lag is primarily due to the complexities of how virological testing and how hospital patient reports are handled and recorded. This time lag is a major threat for public health as an epidemic event has the potential to expand and spread rapidly without detection during this period.

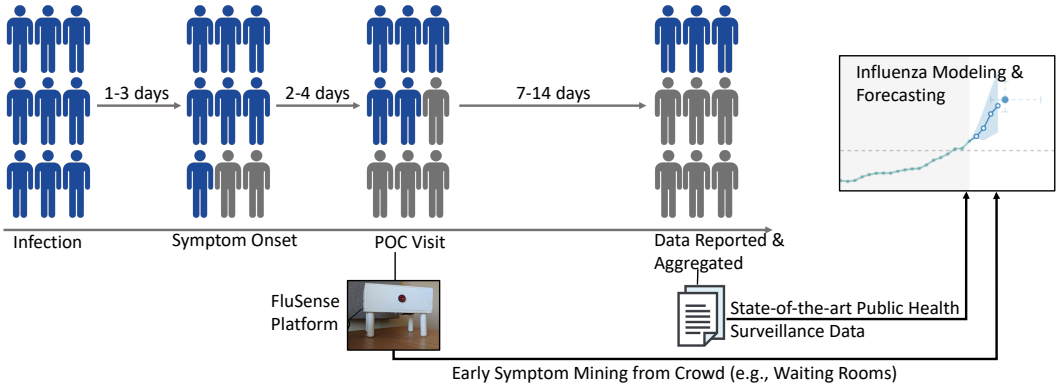

Fig. 1. Comparison of target populations and reporting delays in the proposed contactless syndromic surveillance platform FluSense versus current public health surveillance. FluSense could capture ILI symptom-related information 7-14 days earlier than the current ILI surveillance mechanisms.

Recently there has been increasing research incorporating non-traditional data sources into infectious disease forecasting ("digital epidemiology") in order to address the shortcomings of traditional epidemiological surveillance data. These sources include climate data [41]; social media [4, 9, 42]; internet search trends; [5, 13, 14, 36], satellite imagery [34]; and smartphone data [12]. Some data sources may only be available in specific contexts, e.g. drug market sales and may be unavailable or unreliable in resource-limited settings. For Internet-based data sources (i.e., search trends or social media data), it is unclear how much of the observed "signal" reflects actual changes in incidence as opposed to "noise", caused by changes in information environment (e.g., increased public awareness). For example, Google Flu Trends (GFT) faced significant criticisms for dramatically overestimated prediction and lack of reproducibility and reliability during the 2012-13 flu season [25]. Many of these non-traditional data sources have shown some promise but are inherently limited in that they do not directly measure biological signals of infection or associated physical symptoms. If we examine the trajectory of a hypothetical influenza epidemic (Figure 1), a variety of symptoms including cough, fever (often with chills), sore throat, and nasal symptoms appears within 1-3 days of infection and a substantial proportion of these patients may visit a point-of-care

(POC) facility (e.g., a local clinic or a hospital) within 2-4 days of the symptom onset. Currently, no automatic or readily-scalable methods exist for real-time measurement of these physical symptoms of ILI directly from a crowd. This research seeks to develop and validate a novel syndromic surveillance system that captures bio-clinical signals directly related to physical symptoms of influenza-like-illness (ILI) in defined geographic spaces from the hospital waiting areas.

This paper documents the deployment of our sensor platform, *FluSense*, within a university healthcare system. The common complaints and symptoms of many respiratory infections include nasal obstruction and discharge, sore and scratchy throat, hoarseness, and cough [19]. When influenza is circulating within a community, patients with influenza typically show symptoms of cough within 48 hours of initial infection. A recent study found that the best multivariable predictors of influenza infections were cough and fever with a positive predictive value of 79% ($p<0.001$) at an individual level [32]. We show that the total daily cough counts exhibited strong correlations with laboratory-confirmed influenza infections on campus. Additionally, the thermal camera images paired with a neural network model was able to accurately estimate the total number of patients seen at the clinic each day, which as then used to quantify incidence rates (e.g., total cough count per crowd size) that is highly informative of the daily ILI and confirmed influenza case counts. This study provides vital proof-of-concept data on this new technology platform, and highlights the potential of a wide-scale deployment (i.e., beyond hospital waiting rooms) for practical public health response.

The early symptom-related information captured by the FluSense sensor array data could provide valuable additional and complementary information to current influenza prediction efforts (e.g. CDC FluSight Challenge). Figure1 illustrates how such a FluSense sensor array can capture very early symptoms related to influenza at a population-level from the waiting room populations and how such a system can be potentially integrated to existing state-of-the-art influenza modeling and forecasting efforts to potentially increase system sensitivity. Moreover, the overall intent of this system is to capture data *outside* clinical contexts for estimation of infections within the general population- which is currently not captured by any routine data streams. In summary, our contributions are:

- Non-speech body sounds, like coughs, passively and unobtrusively captured from crowds along with the patient size estimates from thermal imaging in the hospital waiting areas together contain important epidemiological information about the intensity and severity of respiratory illness at a population-level. To date, the majority of clinical research on influenza linking to cough has been done at the individual-level in the context of clinical diagnosis. **To the best of our knowledge, this is the first work that establishes the link between cough sensing and influenza-like-illness trends at the scale of a college/university community.**
- **The FluSense platform processes low-cost microphone array and thermal imaging data on the edge with a Raspberry Pi and a neural computing engine** (Intel Movidius) while storing no personally identifiable information. It is capable of running the deep learning-based acoustic models and thermal imaging-based crowd density estimation algorithms in real-time.
- **We implemented a rigorous field study (Dec 10, 2018 to Jul 12, 2019) where we deployed *FluSense* in four public waiting rooms within the university health service of the University of Massachusetts Amherst, a research and land-grant university with more than 30,000 students.** During this deployment, we collected more than 350,000 waiting room thermal images and 21,230,450 non-speech audio snippets from the hospital waiting rooms. We partially annotated both of these audio snippets and thermal images libraries to provide a rich dataset for the broader community use in public health, computer and information science. We also aim to release our additional annotations/labels of existing publicly available dataset along with a partially labeled data collected from our hospital waiting room deployment.

- **After development of a laboratory cough model with state of the art techniques, it was adapted and thoroughly evaluated in real world clinical settings.** This real-world cough model shows resilience to complex noises and diverse settings and achieves an F1 score of 88%. Thermal imaging-based crowd density estimates were found to have a strong correlation with the reference daily waiting room patient counts from hospital records (with a Pearson correlation coefficient of 0.95). These data were then used to estimate a cough incidence rate [2] (e.g., total cough count per estimated crowd size), which is then used for ILI modeling.
- A critical scientific question that we addressed herein is how the captured physical symptom intensity and crowd behavior measured in a crowded public space (i.e., the hospital waiting areas) can be linked to Influenza-Like Illness (ILI) severity within the same community. **Our results suggest that multiple and complementary sets of *FluSense* signals exhibited strong correlations with both ILI counts and laboratory-confirmed influenza infections on the same campus.**

2 RELATED WORK

2.1 Influenza Forecasting with Population-Level Information

Accurate real-time forecasting of infectious disease outbreaks is of paramount importance to medical practitioners, public health professionals, and policymakers as it can inform targeted prevention and interventional strategies including increased health care staffing or vector control. Current efforts in infectious disease forecasting rely on statistical models to forecast key values of epidemics, such as incidence in a particular week or the cumulative incidence over a season. In the case of influenza these models have relied on epidemiological data from existing public health ILI surveillance which has major limitations including coarse geographic boundaries, limited reporting schedules and an undesirable lag time between the collection of clinical data and subsequent availability for flu forecasting.

As discussed above, to address the shortcomings of the traditional epidemiological surveillance techniques, researchers have introduced novel digital data streams including climate data, social media [4, 9, 42] and internet search trends [5, 13, 14, 36], satellite imagery [34], and smartphone data [12] for ILI severity modeling. Many of these non-traditional data sources have shown promise but are inherently limited in that they do not directly measure the infectious process and symptomology. Our approach circumvents these problems by capturing physical symptom of ILI directly from a crowd (i.e., hospital waiting room crowd) with the proposed contactless sensor system and then map this information to ILI severity at a community-level.

2.2 Cough Modeling

Several recent studies have explored audio-based cough recognition algorithms. For example, Mel-frequency cepstral coefficient (MFCC) along with Hidden Markov Model (HMM) has been used to train cough recognition models [30, 44, 52]. Larson *et al.*, and Amoh and Odame have used Spectrogram based feature to train the cough recognition models [6, 24]. Other acoustic features including LPC [7], Hilbert Marginal Spectrum [26] or Gammatone Cepstral Coefficient [28] have also been used in conjunction with both static (e.g., Random Forest, Support Vector Machine) and temporal models (e.g., Hidden Markov Model, Recurrent Neural Network). More recently, several studies have also explored different Convolutional Neural Network (CNN) architectures [6]. Building on top of this rich literature, in this work we have developed and validated a cough recognition model that is optimized for noisy and crowded public spaces, specifically in hospital waiting rooms.

There are several limitations of the existing approaches that limit the utility of these models for public health applications. For example, the training and testing data used in these studies were fairly limited in terms of the size and diversity of the participant pool. A majority of these studies included fewer than 20 people and cough data were collected only from the specific patient population (e.g., asthma patients). In this work, we have compiled and labeled a large audio dataset consisting of different abnormal upper respiratory sounds including

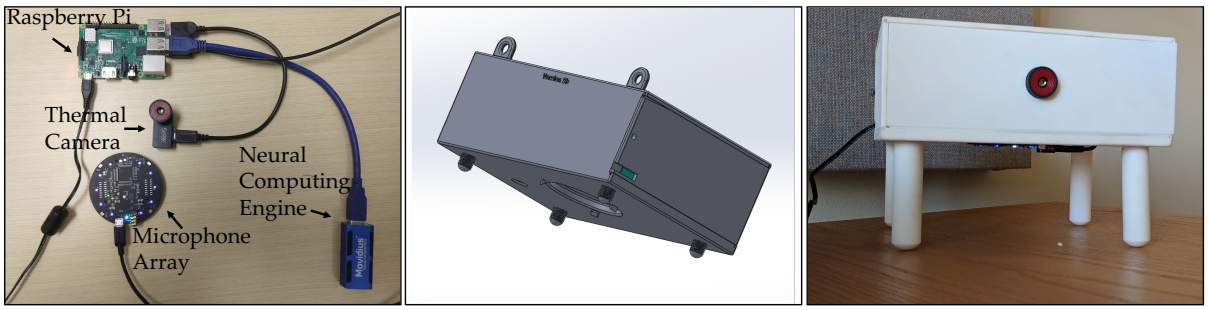

Fig. 2. Illustration of electronic components (left) the 3D mechanical design of the sensor box/enclosure (middle) and the deployed FluSense platform (right).

Table 1. Illustrative example of the metadata retained by the audio processing pipeline

Time stamp	Probability of being speech	Probability of being cough	Audio snippet (raw waveform)	Direction of arrival
15 Dec 2018 11:34:05	0.001	0.996	Saved	85 deg
15 Dec 2018 11:45:55	0.990	0.005	Not Saved	25 deg

cough, sneezes and throat clearing. In total, we have manually categorized approximately 170 hours of audio data including cough events from a wide variety of individuals in diverse acoustic environments. Moreover, we have rigorously evaluated the performance of these cough models with different augmentation techniques simulating different challenging scenarios (i.e., accounting for different background noise and room acoustics). Lastly, we have collected over 21,000,000 non-speech audio snippets during our 7-month long clinical deployment study which includes actual cough events in the four hospital waiting rooms. Our optimized cough model achieves an 88% F1 score with the real world hospital waiting area data.

3 FLUSENSE: A CONTACTLESS SENSING PLATFORM

The contactless sensing framework, FluSense, consists of a microphone array and a thermal camera to capture different waiting room crowd behaviors including cough and speech activities along with waiting room patient counts (figure 2).

The Flusense platform consists of several modules, including:

- ReSpeaker Microphone Array (version 2.0) [48]: A microphone array with 4 microphones and built-in high performance chipset.
- Seek CompactPRO [47]: A thermal camera able to capture thermal images with a $320x240$ pixel resolution and a 32 degree field of view.
- Intel Neural Compute Stick 2 [3]: A computing hardware that uses Intel Movidius Myriad X Vision Processing Unit (VPU) for efficient deployment of deep learning models on the edge.
- Raspberry Pi: This serves as the control platform to synchronize all the attached sensors and devices.

3.1 Audio Processing

To ensure privacy in the hospital waiting areas, all audio data are processed in real-time into 1-second blocks immediately as the raw audio signal is sampled. A high-fidelity binary classifier for both speech and cough (i.e., Speech vs. Else, Cough vs. Else) then classifies every 1-second audio block. A detailed description of the cough

and speech recognition model is provided in section 5. No audio data are retained if any speech-like sounds are detected within the 1-second audio snippet. The time-stamped output from these classifiers (confidence score or log-likelihood) is also stored as metadata for later evaluations. Table 1 shows an example of anonymized metadata information recorded from the cough- and speech-recognition engine. FluSense also stores all non-speech snippets onto the local hard drive using two-stage encryption. While storing the data in the local hard drive is neither necessary for the FluSense platform, we opted to include this feature in anticipation that the non-speech 1-second audio snippets would allow us to validate the cough model (trained on different pre-collected datasets) in real waiting room setting which is presented in Section 5.4. Lastly, the direction of arrival information from the microphone array is also stored.

3.2 Thermal Imaging

A low-cost Seek CompactPRO thermal camera was used to collect thermal images once a minute during routine patient hours within the waiting areas of the hospital which were then stored on the local hard drive with two-stage encryption. A detailed description of the thermal imaging-based crowd estimation algorithm is provided in section 6. FluSense processes all data in real-time while maintaining a small memory and energy footprint. A detailed power, memory and throughput benchmarking analysis can be found in Appendix A.

4 PILOT STUDIES IN A CLINICAL SETTING

The IRB-approved study involves non-invasive, anonymized, passive data collection using the contactless mobile sensing and edge computing platform (as shown in figure 2) from all the individuals (e.g., patient, patient chaperone, waiting room attendant) present in four public waiting areas within the university health service. The University Health Service (UHS) provides comprehensive medical care to the students, faculty, and staff, along with their spouses, domestic partners, and dependents. Data collection occurred from December 10, 2018 to July 12, 2019. During this period our system ran continuously every day from 8 a.m. to 8 p.m. on weekdays and from 11 a.m. to 5 p.m. on weekends (the general hours of the health offices).

Figure 3 shows an overview of the three waiting areas within the hospital and the sensor boxes deployed in these spaces. The contactless mobile sensor boxes were securely attached to the walls within the waiting areas, with placements to maximize the coverage of the thermal camera over the sitting arrangement and to minimize any impacts on normal patient flows. An informational placard was also placed next to the sensor box to provide the public with further information and contacts regarding the study. Each of the waiting areas serves different patient sub-populations, as described below.

- **Walk-in clinic:** This is the general waiting room for all urgent care in the facility. Patients either present here after setting an appointment via a university-administered SMS system, or they just present and then wait for clinical staff to see them. One room has a wall which partially blocks the view of the thermal camera; consequently we deployed two separate units in this waiting room so that more of the room will be in the visual range of our system.
- **Checkup clinic:** This is the waiting room for patients who come for any routine (non-urgent) appointments. We deployed a single unit here.
- **Pediatric care:** This is the waiting room for pediatric care; one unit was deployed here.
- **Women's Health clinic:** This clinic provides gynecological and reproductive health services; again a single unit was sufficient for room-wide coverage.

Ground-truth Waiting Room Patient Count Data: To validate thermal camera-based patient counts, we extracted the ground-truth patient count data from the patient arrival and departure records within the walk-in clinic. For the walk-in clinic, a patient can make appointment via a Short Message Service (SMS)-based system. Alternatively, the patient can also make an appointment on arrival at the site. As the SMS-based appointments

Fig. 3. illustrates the deployment of the contactless sensing platform in different hospital waiting areas (i.e., women's clinic, pediatric clinic, and checkup clinic waiting areas from left to right).

typically result in less waiting time at the hospital, most patients generally utilize it. We used the records that capture whether the patient arrived with or without SMS appointment, their time of setting up the SMS appointment (If they make their appointment via SMS); their arrival time in the clinic; time when they are taken away from the main waiting room for meeting with the doctor and time when they left the clinic. In order to establish the daily patient count, we simply enumerated the total number of people who came to the walk-in clinic waiting room on any specific day.

Ground-Truth Influenza and ILI Data: For routine influenza testing, the clinical laboratory uses the "Alere (Abbott) ID NOW" influenza A & B rapid test. This instrument-based test platform allows for the detection of both influenza A and B from nasal or nasopharyngeal swabs [46]. Testing takes ca. 10 minutes and detects influenza with high sensitivity and specificity (>95% for both). This data source is widely utilized, well-validated and provides "gold standard" data for modeling. In our analysis, we utilize both total tests (an indicator of general ILI caseloads) and total positive tests (confirmed influenza). The total number of flu tests ordered per day is a proxy for influenza-like-illness cases, as the physicians and nurse practitioners run the tests on patients who show flu-like symptoms. Conversely, the total count of positive flu test cases per day is a direct indicator of diagnosed influenza cases.

5 LABORATORY DEVELOPMENT OF COUGH RECOGNITION ALGORITHMS FOR PUBLIC SPACES

5.1 Data Sources

In order to train and subsequently evaluate the cough and speech recognition system, we use several datasets including Google Audioset [18], ESTI [1], DEMAND [45], and TIMIT [17] (see table 2). Google Audioset is a collection of 10-second, partially annotated audio clips from Youtube that includes cough, sneeze, speech as well

Table 2. List of audio datasets that we used for training and testing our laboratory models

Dataset Name	*Class*	*Length (in seconds)*
Audioset [18] labeled	Cough, sneeze, speech, sniffle, gasp and background noise	45550
Audioset [18] unlabeled	Sound of other human activities and background noise	462220
ESTI [1]	Background noise	1403
DEMAND [45]	Background noise	86401
TIMIT [17]	Speech	19380

as sound of different human activities and background noise. While the audio annotations sufficiently described the sound type in each audio clip within Google Audioset, the start and stop times of these categories (especially cough and speech) are not marked. We recruited two human raters to manually and precisely label 45,550 seconds of Audioset dataset with the tags of speech, cough, sneeze, sniffle, gasp, silence, and background noise. The audio samples of the sneeze, sniffle, and gasp allow the cough recognition model to learn the difference between a cough and other abnormal respiratory noises. Lastly, we compiled a large set of audio files (approximately 462,220 seconds) from Audioset, MIT TIMIT [17], DEMAND [45] and ESTI [1] datasets to capture sounds of human activities, environmental noises, speech, and hospital noise.

5.2 Augmentation

We applied multiple complementary data augmentation techniques to maximize the accuracy and robustness of our model for use in challenging real-world scenarios.

5.2.1 Volume Augmentation. In order to simulate audio events at different distances, we increased and decreased the volume of our existing data by a factor of 0.25, 0.5, 2 and 4.

5.2.2 Background Noise Augmentation. To ensure our model would be robust to different background noise conditions (e.g., crowded public environment), we added different types of background noise from our noise dataset to the original unmixed training data. We randomly select the mixing ratio by generating a number between 0 and 0.75, where the mixing ratio close to 0 simulates audio recording in less noisy conditions and the mixing ratio close to 0.75 simulates highly noisy conditions.

5.2.3 Room Impulse Response Augmentation. The room acoustics (especially reverberation properties) can have important impacts on the audio recorded by the microphone array. In our deployment study, we had different waiting rooms with different geometry and acoustic properties. In order to ensure that our cough and speech recognition model is equally effective in these diverse spaces, we used the Aachen room impulse response dataset [22] to transform our original training data.

5.2.4 Combined Augmentation. We also applied different combinations of above-mentioned augmentations (e.g., room impulse response augmentation with background noise augmentation, volume augmentations with background noise augmentation or even all three augmentations combined). Together, these exercises helped ensure our training data represents realistic technical challenges in analysis.

After applying these augmentation processes we had a final dataset which was six times the original size. In these data, one subset was the original data, three subsets were generated by only applying volume augmentation, room impulse response augmentation and background noise augmentation respectively; one subset was generated by applying two augmentations on the same audio, and one subset was generated by applying combinations of all three augmentations.

5.3 Modeling

For modeling purposes, we randomly divided all of our labeled audio data from different sources using three fixed partitions: 80% of the audio files for training, 10% for validation with the remaining 10% for testing; this

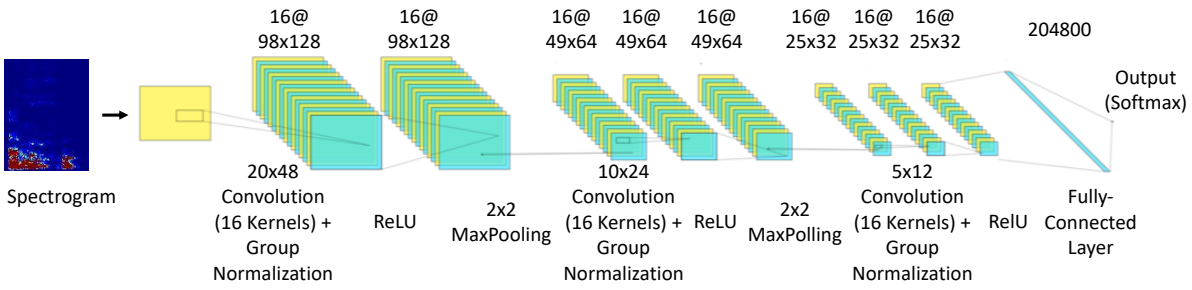

Fig. 4. illustrates the architecture of the CNN-based cough recognition algorithm.

same partition was used for all subsequent models. We passed our training partition data through the same augmentation pipeline. For all final testing in “Speech” and “Hospital Noise” conditions, we utilized only instances of "Speech" and "Hospital Noise" that had not been used for training.

For building the cough and speech recognition models, we started with the CNN based architecture “cnn-trad-fpool3” [40] which is optimized for resource-constrained settings. The continuous audio stream is divided into one second long audio snippets which are subsequently converted to spectrograms with fft bin size of 257, with window size and shift of 30ms and 10ms respectively. In the end, we have a 40 width and 98 height size image representation of each 1-second audio snippet. This image is then passed through a multilayer CNN model. We tuned different hyperparameters including the spectrogram frequency range, number of filters in different layers, filter size, size of maxpool filter, and total number of convolution layers. Table 3a shows different models that we developed by tuning the CNN architecture.

The best cough recognition performance was achieved by the model (model index 6 as shown in Table 3a) which incorporates a spectrogram having frequencies ranging from 0 to 4000 Hz. Figure 4 illustrates the internal architecture of another 3-layer CNN that was used to train our cough recognition model and to generate “cough” and “not cough” inferences. For reproducibility of our model, a detailed description of all training parameters is provided in Appendix B.

Table 3b shows the performance of the cough recognition models in different testing scenarios including no background noise; with speech; and with hospital noise conditions. The “no background” noise condition aims to evaluate the performance of the cough model in noise-free spaces. In contrast, the “with speech” and “with hospital noise” conditions aim to evaluate the cough model in the presence of complex environments with background speech and hospital noise. Lastly, the "all augmentation" testing condition passes the testing partition of our data through different augmentations (as described in section 5.2) and simulates a real-world hospital waiting room scenario where different conditions could happen in random order.

As shown in Table 3b, the 2-layer CNN with Mel Frequency Cepstral Coefficient (MFCC) performs with high accuracy (with an F1 score of 92.25%) in the "no background" scenario. However, in the presence of background speech and hospital noises, performance drops significantly. A similar trend can be observed with all augmentation testing condition where the F1 score of the binary classifier only reaches 50.6%, which is essentially a random classifier. This is not unexpected, as other studies have also found out that MFCC features are not robust against additive background noise [37]. Conversely, models based on spectrograms are relatively robust to noisy situations. Our studies have also found that using even half of the spectrogram in frequency axis (taking frequencies between 0-4 kHz instead of 0-8 kHz) and reducing the number of filters does not adversely affect the performance of the cough recognition model. Lastly, the 3 layer CNN with group normalization[51] (Model 6) yields the highest performance in all of the more challenging conditions including with speech, with hospital noise and with all

Table 3. (a) The list of different "Cough vs. Else" classification models. (b) The performance of these models under different testing conditions in terms of recall (R), precision (P) and F1 score.

Model Summary	Model Index
MFCC + 2 layer CNN (64)	1
0-8 KHz Spectrogram + 2 layer CNN (64)	2
0-4 KHz Spectrogram + 2 layer CNN (64)	3
0-4 KHz Spectrogram + 2 layer CNN (32)	4
0-4 KHz Spectrogram + 2 layer CNN (16)	5
0-4 KHz Spectrogram + 3 layer CNN(16) + Group normalization layers (8)	6

(a)

	No Background Noise			With Speech			With Hospital Noise			With All Augs		
Models	***R (%)***	***P (%)***	***F1 (%)***	***R (%)***	***P (%)***	***F1 (%)***	***R (%)***	***P (%)***	***F1 (%)***	***R (%)***	***P (%)***	***F1 (%)***
1	**92.26**	**92.26**	**92.25**	53.71	73.48	41.37	53.49	73.60	40.90	58.61	74.32	50.63
2	86.72	87.40	86.66	68.57	77.89	65.71	71.53	79.32	69.51	75.21	80.42	74.11
3	87.5	87.57	87.49	70.51	78.15	68.45	71.88	77.84	70.29	76.94	80.64	76.23
4	85.57	86.47	85.48	67.49	77.65	64.21	69.52	78.43	69.51	74.21	80.01	72.89
5	85.57	86.47	85.48	81.71	70.23	78.19	71.42	78.92	69.43	76.22	80.35	75.38
6	90.2	90.2	90.2	**82.45**	**82.38**	**82.41**	**84.55**	**85.43**	**84.45**	**87.02**	**87.3**	**86.99**

(b)

augmentations (e.g., an F1 score of 87.0% in all augmentations) while maintaining a high performance in the no background noise condition (an F1 score of 90.2%).

For the speech model, we followed the original approach proposed by [40] named "cnn-trad-fpool3". For this, we used the MIT TIMIT as speech corpus database in addition to using our own custom labeled dataset from AudioSet. We achieved 93% test accuracy for the speech detection tasks in the "all augmentation" condition.

5.4 Adaptation and Validation of Laboratory Cough Models with Real-World Waiting Room Data

To further evaluate the performance of our laboratory-developed cough model under routine waiting room conditions, we used audio snippets recorded within waiting rooms during our clinical study as described in section 3.1. From all the one-second audio snippets recorded from the four waiting areas in our clinical study, we have randomly sampled/selected 2500 one-second long audio snippets with two rules: (i) From each of the five FluSense platforms (deployed in the four waiting areas), we have sampled 500 one-second long audio snippets. (ii) The audio snippets were uniformly sampled from different probability ranges/bins as outputted by the cough model. We have divided the output probability ranges into 10 bins (e.g., 0.0-0.1, 0.1-0.2, etc.) and randomly sampled 250 samples from each bin. Lastly, a human rater manually labeled all the 2500 audio snippets into two categories: cough and non-cough.

Figure 5a shows the proportions of the cough and non-cough sounds manually labeled by a human rater in the ten different probability bins. These probabilities correspond to the fitness score from our cough model before any transfer learning (which corresponds to the first row of table 4). As can be seen in figure 5a, most of the audio snippets with low probability values (i.e., less than 0.5) belong to the non-cough category. However, with increases in probability, we can also observe an increase in the percentage of cough sounds detected by human rater. The majority of the audio snippets with probability values greater than 0.9 are judged to be cough by the human rater. In other words, the audio snippets that are more confidently classified to be a cough by the laboratory model are highly likely to contain actual cough sounds. However, this model also classified different

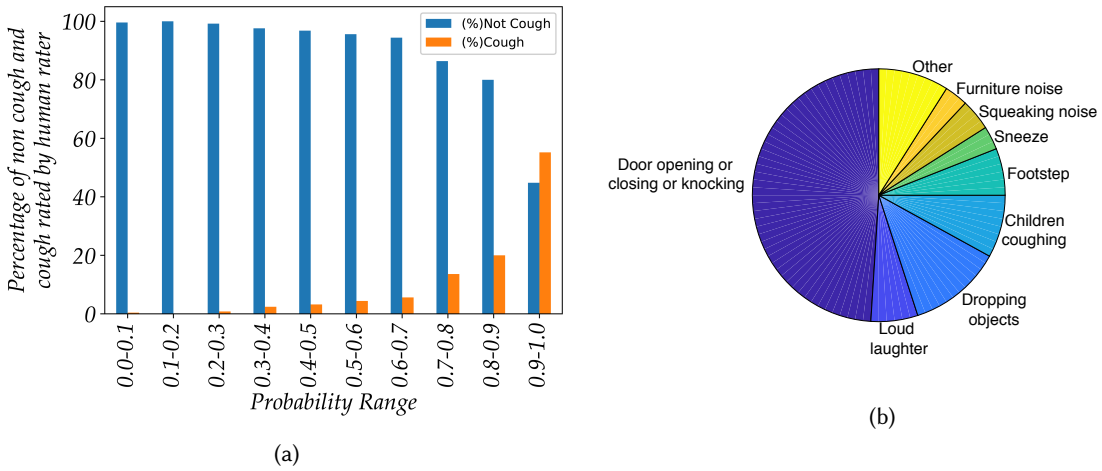

Fig. 5. (a) Overall proportion cough and non-cough sounds manually labeled by a human rater using decile probability bins from the cough model (Model 6) before transfer learning, (b) Distribution of audio events that produced false positive and false negative errors from Model 6 before transfer learning.

Table 4. Comparison of model performance before and after additional transfer-learning.

Model	Precision	Recall	F1
Model 6, prior to transfer learning	0.73	0.76	0.75
Model 6, after transfer learning	0.87	0.89	0.88

types of non-cough events under high probability range which include door openings/closings, footsteps, loud laughter, sneezes, squeaking, and furniture movement. As such, these sound types are falsely classified as cough by the model even with a fairly high probability threshold as a decision boundary (i.e., 0.88). Conversely, we also observed that some actual cough events were given relatively low probability values (between 0.7 and 0.9) which primarily includes coughing sounds from children. As our training data does not include many examples of children coughing, our model is under-trained on this domain. Figure 5b shows the distribution of different types of audio events that gave rise to either false positive or false negative errors, and by including more examples like these in the training process, the cough model could be improved even further.

To this point, our model has never been exposed to any cough or non-cough training data collected from our real-world clinical study at the waiting rooms, and was primarily trained using large pre-existing datasets (Table 2). In order to compensate for potential domain mismatch and to address any gaps in our training data, we used supervised transfer learning on our top model (model 6) with an additional 2,250 one-second long audio snippets collected from all the five FluSense platforms deployed in waiting rooms. We used approximately half of our labeled data for the transfer-learning. As can be seen in Table 4, the transfer learning resulted in a 12% increase in precision and a 11% increase in recall. In terms of F1 score, we achieved a 13% increase to 88%. Overall, this result is comparable to the F1 score of 84.5% that we achieved in the hospital noise testing condition (as can be seen in Table 3b). This transfer learned model was the final model deployed on all FluSense platforms.

Each waiting area has a different size and interior layout; consequently, the acoustic properties of each is quite different. Moreover, the cough sounds recorded from different waiting areas come from different group of patients. For example, the pediatric clinic sees cough sounds from children while a majority of the cough sounds

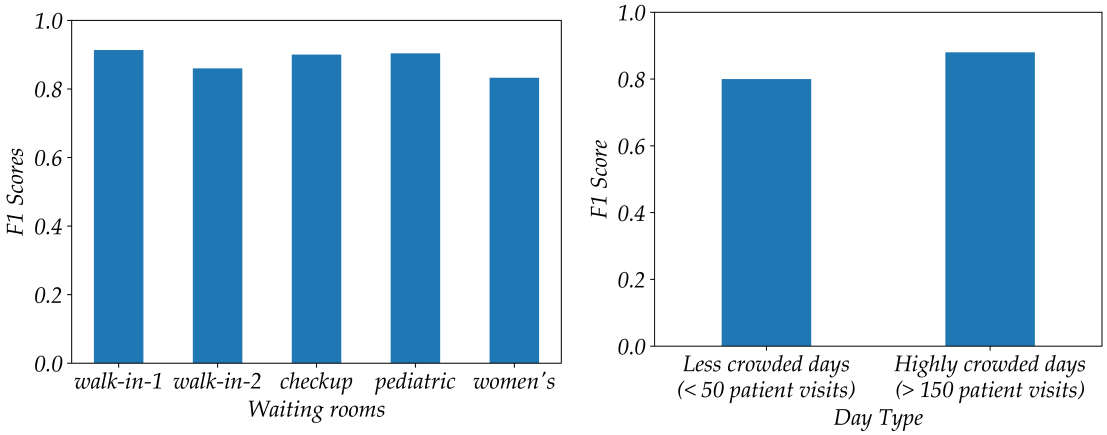

Fig. 6. (a) Performance of the cough model (after transfer learning) in different waiting areas in terms of F1 score. (b) Performance of the cough model for highly crowded vs less crowded days

captured in the women's clinic comes from an adult female population. By analyzing the performance of our cough model in these different waiting areas, we are able to explore whether and how the performance of the model varies across population and settings. Figure 6a shows how our transfer learned cough model performs in different waiting areas. The highest F1 score of 0.91 is achieved in the walk-in clinic, while the lowest F1 score 0.84 was found in the women's clinic. This result shows that our cough model achieves reasonably high performance in all the different waiting areas.

We also analyzed the performance of our model on crowded vs less-crowded days. For this study we chose four different days: two of which were weekends having less than 50 patient visits, plus two very busy days having more than 150 patient visits. We then randomly sampled 200 audio snippets from each of these days, and a human rater manually labeled all audio snippets. Model 6 after transfer learning, was then used for analysis. As shown in figure 6b on less-crowded days we achieved an F1 score of 0.80, while and on highly crowded-days we observed an F1 score of 0.88. Less crowded days appear to have had a low level of background noise when compared to more crowded days would with higher background noise. As a result, it is surprising that our cough model achieves a slightly less performance in less crowded days. However, this observation can be explained if we look at the type of sounds that cause errors (false positive and false negative) as shown in figure 5b. Sudden noises including door closing/opening, objects falling on the floor, and loud laughter are all falsely classified as coughs. In settings with less background noise, these sharp noises are more prominent which gives rise to the lowering of performance of the cough model in a less crowded setting. Our model trained with heavy background noise augmentation (as explained in section 5.2) can tolerate a relatively high level noise and attain a high level of performance during the highly crowded days.

Speech Model: Ethical review limitations prohibit us from recording raw speech audio data in public areas within the hospital due to privacy concerns and consideration of the US Federal Health Insurance Portability and Accountability Act (HIPAA) of 1996. However, in the pilot deployment at the hospital waiting rooms, we have recorded all non-speech audio snippets or segments (as detected by the speech detection model) which can be used to gauge the performance of our model. More specifically, we can investigate what proportion of nonspeech segments (as predicted by the speech model) actually contain intelligible speech (as rated by a human).

A human rater manually labeled a total of 2,500 1-second long audio snippets, with 500 audio snippets sampled from each of the 5 FluSense platforms. From this analysis we determined that only 79 out of 2500 audio snippets contain any speech. That is, 96.84% of the audio snippets classified as non-speech correctly contained only non-speech data. 3.16% of the audio snippets that have been classified as non-speech contained speech. While there are some cases where a full word could be understood, the vast majority of the snippets contain partial and unintelligible syllables. More importantly, the human rater was unable to understand the topic or context of the conversation from any of the audio snippets that were misclassifications (specifically the false negatives). In summary, these results clearly indicates that our system rarely misclassifies speech as non-speech, and in cases where misclassification does occur, there was no intelligible content. This underscores the privacy-preserving nature of the entire system.

6 DEVELOPMENT OF THERMAL IMAGING-BASED CROWD ESTIMATION ALGORITHM

6.1 Data

We have collected a total of 359,789 thermal images from the 5 FluSense units. We randomly selected four days during the clinical study and manually labeled all the thermal images collected by the five sensor platforms during these four days. We manually created bounding box labels for all the person-shapes inside these images which resulted in a thermal imaging dataset with 2,100 images for crowd density estimation.

Fig. 7. Sample thermal images collected in different waiting rooms (with one room per column column) and the inferred bounding boxes provided by our thermal imaging model. The last row illustrates situations where our thermal model performs sub-optimally.

6.2 Modeling

For thermal imaging-based crowd estimation, we considered two different model frameworks: Single Shot Detector [29] based upon Mobilenet [20] and Faster-RCNN [38] based on Inception-ResnetV2 [43]. We specifically used "ssd_mobilenet_v1_fpn_coco" and "faster_rcnn_inception_resnet_v2_atrous_coco" from the tensorflow object detection model zoo [21]. All of these models were pre-trained using the Microsoft COCO dataset[27]. We transfer-learn on top of the above mentioned models with our data to ensure that the models are well fitted to

the images taken by our thermal camera hardware. For training, we selected three days out of the four days of labeled data as our training set from each of the boxes (around 1700 images) and the other one day of data from each box (around 400 images) as the testing set data.

6.3 Performance

As shown in Table 5, we used different standard object detection measures including mAP (Mean-Average Precision) at IOU (Intersection over Union) 0.5 and mAP at IOU between 0.5 and 0.95. The mAP@0.5 measures the area under the Recall-Precision curve of an object detection model where the detection boxes must have an IOU of 0.5 or greater to be registered as successful detections. The mAP@0.5:0.95 is another similar Mean-Average Precision metric which is estimated by averaging mAP values across different IOU values from 0.5 to 0.95 with a 0.05 increment. We also used LAMR (Log Average Missing Rate) described in [11], which is considered a better metric for benchmarking object detection algorithms. This metric counts the average number of people missed in a single image for different levels of false positive values distributed in log-space between 1 and 10^{-2}. Lastly, we used MAE (Mean Average Error), RMSE (Root Mean Squared Error) between the actual crowd size and predicted crowd size (total number of bounding boxes) to evaluate crowd size estimation of the different models.

Table 5. The performance of different crowd estimation models.

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	LAMR (%)	MAE	RMSE
SSD MobileNet + COCO	27.4	12.8	61.92	1.54	1.73
SSD MobileNet + COCO + our data	85.6	58.7	17.4	0.31	0.59
Faster-RCNN + COCO	60.3	32.1	34.47	0.61	0.93
Faster RCNN + COCO + our data	90.5	65.87	9	0.23	0.54

We used Faster-RCNN and SSD-Mobilenet FPN pretrained on COCO as a baseline as these explicitly include "person" as a class label. The table 5 shows that our transfer learned model outperforms the baseline models while the transfer-learned Faster-RCNN model has the best performance as measured by all of the metrics in the table. For the baseline models even without transfer learning, Faster-RCNN performs reasonably well with a mAP@0.5 score of 60.3. After transfer learning the mAP@0.5 score becomes 90.3. On the other hand SSD-Mobilenet based model does not perform well for our thermal camera images. However, after transfer learning, there is a drastic performance improvement in all of the metrics. We can also see from the table that both of our fine-tuned models outperforms all the baseline models that were trained using only COCO with RGB images.

Figure 7 illustrates sample thermal images collected in different waiting rooms (each row represents a different waiting room) and the inferred bounding boxes determined by our thermal imaging algorithm. The first five columns represent the scenarios where our model detected all the persons correctly and the last column shows where our model failed to predict the correct number of people. From these results, we can see that our Faster-RCNN model performs accurately in a thermal scene with single and multiple individuals in different postures including sitting, standing with both front and back facing people. The model also performs well when the thermal camera is out of focus. It also performs well when there are multiple people, even in the case where there are occlusion or overlap between bounding boxes. However, from the last column, it can be seen that the model sometimes fails to detect the correct number of people for multiple person scenarios if the occlusion is high and/or if there are multiple people with different overlapping or heat signatures with different intensities.

7 FLUSENSE FEATURE ENGINEERING

In this section, we will discuss how different speech-, cough-, and crowd size-related features from raw speech, cough, and people count data streams. These multimodal FluSense features will then be used to model the total ILI or influenza burden within the college community.

7.1 Features

There are three groups of features that we extracted from the raw speech, cough, and crowd size inferences:
Crowd size-related feature (personTime): The *FluSense* platform captures a thermal image of its surroundings every minute. These are then passed to the crowd estimation models (as described in 6) to draw the bounding boxes around each each person and then to count the total number of bounding boxes in the thermal field. Here, if c_i is the total number of people/bounding boxes detected in i^{th} thermal scene in a certain day, then

$$personTime = \sum_{i=1}^{M} c_i$$

Here, M refers to the total number of minutes when the *FluSense* platform was operational during the day. A day with a high patient influx should therefore have a correspondingly high personTime value. Later, in section 7.2, we will demonstrate that the personTime feature has a statistically significant relationship with actual daily patient count (derived from hospital records).
Cough-related features (coughFeat): The cough-related feature subset consists of four cough related features that summarize the daily cough activity within a waiting room. We used the probability threshold of 0.88 to use only high probability cough events for this feature computation.

- **Total Adjusted Cough (tac)**: This feature captures the average number of cough events per hour in a certain day. As the hours of operation of the waiting rooms vary across different day type (i.e., weekday vs. weekend), the **tac** is adjusted based on the total operation hours per day.
- **Total Cough Epoch (tce)**: The total cough epoch is also computed to count the number of continuous cough segments per day. Here, we considered that two cough events belong to the same cough-epoch if they occur within 3 seconds.
- **Total Unique Cough DOAs (tucd)**: The tucd is a heuristic-based feature which was estimated using the direction of arrival information for different cough events with a preset timeout criteria. The DOA was estimated Generalized Cross Correlation with Phase Transform (GCCPHAT) method described in [23] to estimate the time delay and use the time delay and array geometry to localize the source. The main goal of this feature was to estimate the total unique direction of arrivals (DOAs) associated with different cough events during the day. While the total adjusted cough (tac) feature counts all the cough events with equal weight, the total unique cough DOAs (tucd) feature aims to categorize these events in different spatio-temporal clusters. In other words, several events coming from the same/similar DOAs within some particular period are counted as one unique cough DOA event. As different patients or visitors at the waiting rooms associated with coughing events at different parts of the day would fall under different spatio-temporal clusters and increase the tucd value, the tucd feature can be thought to be associated or indirectly linked with total number of people coughing. A detailed and intuitive explanation of the tucd feature extraction process can be found in Appendix C.
- **Cough to Speech Ratio (csr)**: The **csr** feature estimates the total adjusted cough (tac) to total adjusted speech (tas) ratio or each day.
- **Cough by People Time (cbypt)**: **cbypt** feature estimates the ratio between total adjusted cough count (tac) and personTime. cbypt is a symptom/incident rate feature that normalize the cough count information with a measure of crowd size/density and can potentially allow the ILI model to account for crowd size related variabilities in the cough count data.

Speech-related features (speechFeat): The speech-related features aim to summarize the daily speech activities in the waiting rooms. We used the probability threshold of 0.5 to detect speech snippets or segments. The decision boundary at 0.5 ensures that all speech segments in the public waiting areas can be reliably captured while no or

few speech segments are misclassified as non-speech. The misclassification leads to saving of the speech audio snippets which may give rise to privacy issues. Here is description of these features:

- **Total Adjusted Speech (tas)**: This feature captures the average number of speech events per hour in a certain day. Similar to the **tac** feature, the **tas** is also adjusted based on the total operation hours per day.
- **Total Speech Epoch (tse)**: The total speech epoch is also computed to count the number of continuous speech segments per day. Here, we considered that two cough events belong to the same cough-epoch if they occur within 3 seconds.

7.2 The PersonTime Feature is Closely Correlated with Daily Patient Counts

The personTime feature can be intuitively understood as a measure of total crowd density. However, due to the limited field of view of the thermal cameras (i.e., 32°), limited placement locations for mounting the sensor box (e.g., sensor locations might be constrained by the availability of power source/outlet), and occlusion by walls and furniture, we could only sample a portion of the waiting areas using thermal imaging. Does the personTime estimated from partially observed waiting areas capture daily patient count estimated from the hospital logs? If we can establish a statistically significant relationship between personTime feature and daily patient count, the feature can provide the denominator estimation of symptom incidence rate (e.g., total coughs per personTime) for ILI modeling.

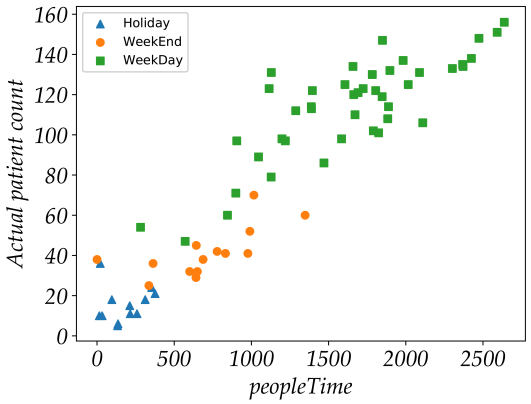

Fig. 8. Relationship between personTime and actual daily patient count.

In order to address this question, we developed regression models to predict actual daily patient count from different FluSense features including personTime. The actual daily patient count was derived from the patient arrival and departure log as described in section 4. To assess the performance of these models, two different cross-validation experiments were used: 70% training and 30% testing split (repeated 1000 times) and a Leave-One-Day-Out (LODO) cross-validation experiment. As shown in table 6, the model trained solely on personTime can achieve a Pearson correlation coefficient (ρ) of 0.91 and a Root Mean Squared Error (RMSE) of 16.38. Figure 8 shows that personTime has a strong linear relationship with the actual patient count across different day types. Relative to this personTime model, the performance of speechFeat and coughFeat model is significantly less. The speechFeat model achieves the lowest performance with the ρ of 0.07 and RMSE of 47.99 (with the 1000 fold random 70/30 split crossvalidation). Compared to the speechFeat model, coughFeat model achieves a much higher performance with the ρ of 0.62 and RMSE of 34.17 (with the 1000 fold random 70/30 split cross-validation). This indicates that the speech activity itself is a poor measure of waiting room patient count. However, cough

Table 6. The performance of Random Forest models trained on different groups of features for daily patient count prediction. The performance was compared using the Pearson correlation coefficient (ρ), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) from two types of cross-validation experiments.

	Random 70/30 split			**Leave One Day Out**		
Feature	ρ	*MAE*	*RMSE*	ρ	*MAE*	*RMSE*
personTime	0.91	12.87	16.38	0.93	12.30	16.08
speechFeat	0.07	41.04	47.99	0.22	37.67	44.57
coughFeat	0.62	28.04	34.17	0.69	25.29	32.70
isHoliday, dayType	0.86	14.61	19.53	0.88	14.85	21.34
isHoliday, dayType, personTime, speechFeat, coughFeat	0.93	11.96	15.10	0.95	9.32	12.95
isHoliday, dayType, personTime	0.94	9.52	12.58	0.95	8.91	13.62

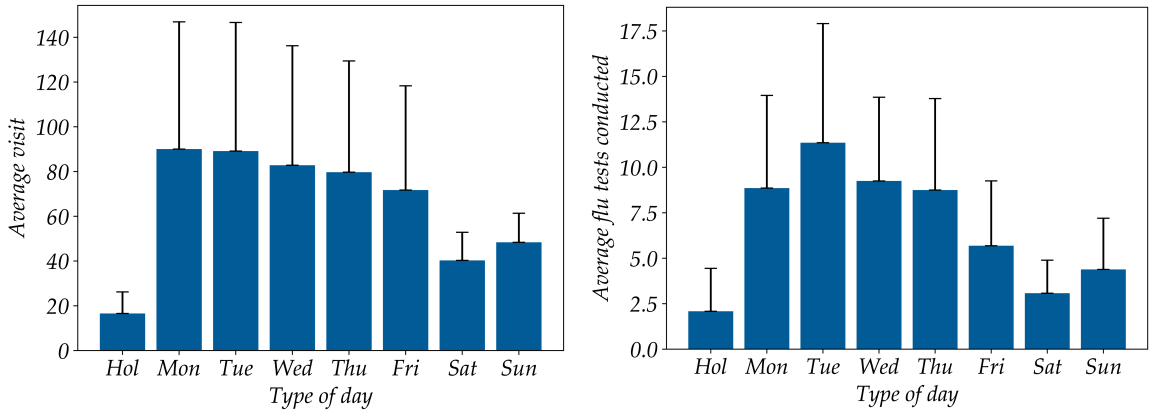

Fig. 9. Distribution of total daily patient visits (a) and total flu test counts (b) over holidays (Hol), weekdays (Mon-Fri), and weekends (Sat-Sun). Each bar represent the mean value, with the thin black line represents one standard deviation.

activity can be a weak indicator of waiting room patient count. These observations have a simple explanation. In a hospital waiting areas, patients themselves rarely speak, especially if they are very ill. Informal observations also suggest that the vast majority of talking that does occur comes from hospital staff, including administrators, nurses and doctors. The speech-related metrics are therefore closely related to only staffing levels and their activities. While the majority of the coughs typically originate from patients, not all patients cough. As a result, while cough-related features become an indirect and weak indicator of daily patient count, personTime is strongly related to daily patient count.

Figure 9a Natural variability of the daily patient count across types of days. Whether a specific day is a holiday ($isHoliday \in \{1, 0\}$) and the day type ($dayType \in \{Mon, Tue, Wed, Thu, Fri, Sat, Sun\}$) contains valuable information about the actual daily patient count. The combination of these two baseline features (i.e., isHoliday and dayType) provides a model with a ρ of 0.86 and 0.88 respectively in the 1000 fold random 70/30 split and LODO cross-validation experiments. The best regression model performance is obtained by the model with only three features: isHoliday, dayType, and personTime and achieves a ρ of 0.95 and an RMSE of 13.62. Considering the range of the ground truth daily person-count is between 5 and 169, the RMSE is about 8.3% of the total

range of daily patient count. This result also indicates that removing the speechFeat and coughFeat from the full feature set does not impact model performance. As shown in table 6, the feature of personTime alone is able to outperform the baseline model trained on isHoliday and dayType.

8 MODELING DAILY ILI AND CONFIRMED FLU CASE COUNTS

As described in section 4, the daily ILI and confirmed flu case counts were derived from the routine influenza testing in the clinical laboratory. To explore the predictive power of FluSense sensor data for modeling daily ILI and confirmed flu case counts, we trained different regression models including linear, ridge, Poisson, gradient boosted tree, and Random Forest regressions with data collected across different day-types including weekdays, weekend days and holidays (shown in table 7). As shown in figures 9b, there is a large natural variability in the total flu test across weekdays, weekend days, and holidays. An ANOVA analysis between the type of day and total influenza test reveals a statistically significant main effect at $p < 0.01$ for all eight day types $[F(7, 88), 0.000267]$. As a result, we trained a baseline model with dayType and isHoliday features which achieve a Pearson correlation coefficient (ρ) of 0.42 for daily total test count and 0.19 for daily total positive case count. We hypothesized that the regression model trained on both the FluSense sensor-based features and the baseline features could outperform the model trained only on the baseline features.

Table 7. The performance of baseline and different FluSense sensor-based models with different feature subset for total daily flu test count and total daily test positive count prediction. The performance was measured with respect to Pearson correlation coefficient (ρ) and Root Mean Squared Error (RMSE) with Leave-One-Day-Out cross-validation experiment.

		Leave-One-Day-Out			
		Total Test		**Total Positive**	
Model	**Feature**	ρ	*RMSE*	ρ	*RMSE*
Baseline	Baseline ∈ {dayType, isHoliday}	0.42	5.02	0.19	2.28
Linear Regression	Baseline + top 3	0.53	4.58	0.33	2.07
Ridge Regression	Baseline + top 3	0.53	4.57	0.33	2.07
Poisson Regression	Baseline + top 3	0.43	5.48	0.25	2.24
Gradient Boosted Tree	Baseline + top 3	0.58	4.44	0.45	2.01
Random Forest	Baseline + top 3	0.65	4.28	0.61	1.68
Random Forest	Baseline + tac	0.59	4.34	0.40	2.00
Random Forest	Baseline + csr	0.58	4.39	0.38	2.03
Random Forest	Baseline + cbypt	0.56	4.44	0.36	2.05

In order to train the FluSense sensor-based regression models, we selected a highly informative feature subset consisting of three FluSense sensor-based features (described in section 7.1) selected from a sequential forward feature selection algorithm. For both the daily total flu test count and the total positive test count, the top selected feature subset includes total adjusted cough count (tac), cough by person time (cbypt), and cough to speech ratio (csr). This highly informative feature subset is then concatenated with the two baseline features to train different regression models. In order to train and test the models presented in table 7, we use data collected from the FluSense platforms deployed in the walk-in clinic waiting area as the walk-in clinic waiting room gets a substantial portion of the total daily patient traffic which can be as high as 160 patient per day. Later, in this section, we will demonstrate how the sensor data captured from different waiting rooms compares.

All the FluSense sensor-based regression models outperform the baseline model by a large margin (Table 7). The Random Forest regression model achieves the top performance for both total flu test (ILI count) and total test positive (confirmed Flu case count) modeling. For total flu test, it achieved a ρ of 0.65 and an RMSE

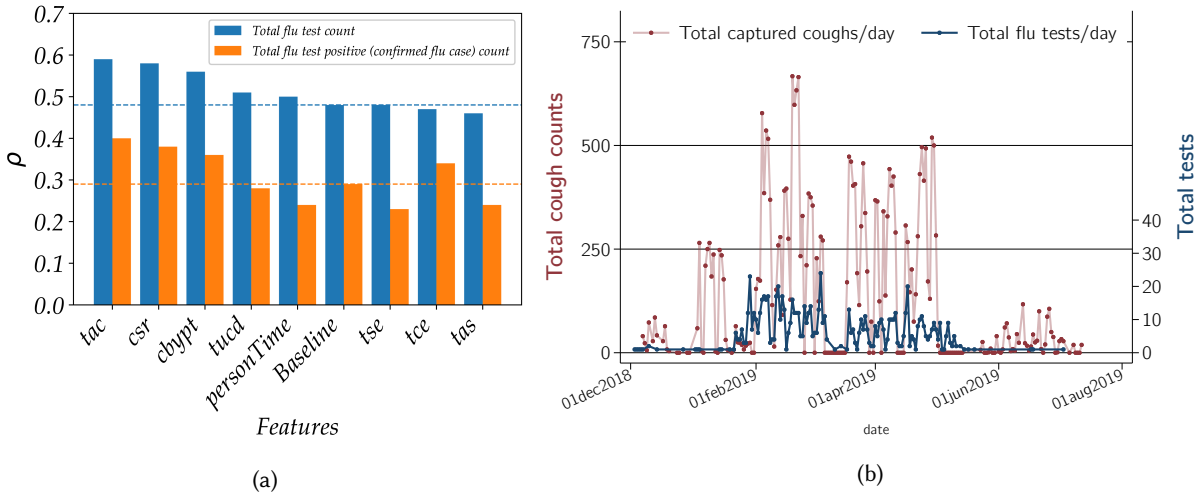

Fig. 10. Depiction of (a) the performance (measured via Pearson correlation coefficient ρ) of different daily total flu test and total test positive count regression models trained on different feature groups and (b) comparison of total cough counts captured by our device vs the total patient load with ILI, by day.

of 4.28 while for total test positives it achieved a ρ of 0.61 and an RMSE of 1.68. The RF model with FluSense sensor data yields an improvement of 0.23 and 0.42 in terms of ρ for respectively daily total flu test and total test positives. From these results, we can conclude that our sensor-based features contain information that can help to predict daily ILI and confirmed influenza case counts. ILI count by definition measures the "syndromic" aspects of influenza-like illness which is directly used in public health epidemic forecasting. Our results demonstrate that the physical symptom information passively captured by the contactless sensor directly from the hospital waiting room crowd can be a rich information source for epidemic modeling of influenza-like illness.

8.1 Feature Importance Analysis

To assess the relative importance of individual features, we trained Random Forest regression models with the baseline features and a single sensor-based feature and estimated the performance of these models with a leave-one-day-out cross-validation experiment. As shown in figure 10a, for both the total daily flu test and total daily positive test counts, the three most-informative features are the total adjusted cough count (tac), cough to speech ratio (csr), and coughs per person time (cbypt). The total cough count (tac) feature along with the two baseline features (i.e., isHoliday and dayType) achieves the top performance with *rho* values of 0.59 and 0.40 respectively for the total flu test and total flu test positive count. All cough-related features are generally ranked at the top while the speech-related features (i.e., tas, tse) are poorly ranked. The model trained on baseline features and total adjusted speech (tas) had the weakest performance with *rho* of 0.46 for the total flu test count prediction while the total speech epoch (tse) yields the lowest performance (*rho* of 0.23) for total positive test count prediction. During the deployment and data labeling, we have observed that most of the talking was by the hospital staff while the vast majority of patients were silent during their stay at the waiting room. As a result, the total speech feature fails to capture information about the patient with high signal-to-noise ratio and thus was poorly correlated with disease state variable of the patient including ILI or confirmed flu case counts. Another important observation was that although personTime was not found to be a strong feature for ILI and confirmed flu case count prediction, the symptom incident rate feature estimated with personTime, the total cough by

people time (cbypt) feature, is found to be one of the top informative features. The model trained on cbypt and the baseline features achieve a *rho* of 0.56 an RMSE of 4.44 for ILI count prediction. Lastly, two additional cough related features, the total unique cough DOAs (tucd) and total cough epoch (tce), are also not found to be highly linked to influenza variables, although they are also not found at the bottom of the ranked list. Patient movements and other cough-like sounds in the waiting areas can decrease the quality of the tucd and tce features. The total adjusted cough count is significantly more informative about the two influenza-related outcomes than the tucd or tce features and likely contains most of the information contained by these two features.

Figure 10b shows how total captured cough by the FluSense device corresponds to total ILI tests. These results show that at the beginning of data capture (December and January) the total number of patients with ILI symptoms was very low. However from February onward, a large influx of patients with ILI symptoms gives rise to correspondingly rapid increase in the flu testing which marks the start of the flu season. Although there is some within week variability, we continue to observe a high number of flu patient visits until mid of May when a sharp decline of ILI patients can be observed. The total captured coughs by FluSense also shows a similar pattern. It is also noteworthy that total coughs starts to increase slightly earlier than the total flu testing. It is possible that the waiting room patient population presents with physical symptom of coughing slightly early, which is captured by the FluSense platform.

8.2 Performance Across Different Waiting Rooms

The waiting rooms are not only different in their sizes and configurations, but also serve different patient populations. For example, the checkup and women's clinic waiting room primarily sees persons with chronic and gynecological health problems. Although some patients in these waiting rooms may also have influenza or ILI symptoms, the sensor data captured in the checkup and women's clinic waiting rooms could contain a much-attenuated level of ILI signal in theory. Conversely, the walk-in and pediatric clinic waiting rooms should receive a much higher number of acutely patients with ILI- related symptoms, and can be considered as primary triage areas for respiratory illness. As a result, we hypothesized that the FluSense sensor data collected at the walk-in and pediatric clinic waiting rooms should contain more information about ILI trends, and will be able to capture these changes with a much higher level of accuracy. These results would provide a secondary validation of the sensor platforms, as it would show the signals being captured are directly related to population-level influenza-like illness.

Table 8. Individual univariable random-effects negative binomial models for the relationship between daily total captured cough counts (Flusense platform) and total clinical influenza tests by waiting area. All models adjusted for total population in waiting areas, and for day-of- week effects. (Stata 16; College Station TX). Note: IRR= Incidence rate ratio.

Waiting area	*IRR*	*p-value*	95% CI	
Walk-in 1	1.00024	<0.001	1.0014	1.0033
Walk-in 2	1.00297	<0.001	1.0016	1.0043
Checkup	1.014	0.055	0.99968	1.029
Pediatric	1.0193	0.003	1.0064	1.032
Womenś clinic	1.0071	0.354	0.9922	1.022

To assess the relationship between captured cough counts as measured by FluSense and total influenza testing per day at the clinic (a proxy for the total ILI burden), negative-binomial random-effects models were used. In univariable comparisons of this relationship for waiting area, the two walk-in clinic units and the pediatric clinic all showed a statistically significant relationship ($p < 0.05$), while the checkup and women's clinic showed no evidence of a relationship as shown in table 8.

To further quantify these relationships for the waiting areas with significant univariable coefficients (i.e., with walk-in 1, walk-in 2 and pediatric clinic only) and to provide effect sizes, the same models were used, and now included covariate adjustments for day-of-week, and total captured population denominator, with robust clustering of errors. In this analysis with the Walk-In clinic 1 as the reference, the Walk-In clinic 2 and Pediatrics both showed a positive association between influenza testing and the FluSense-captured coughing rate. These incidence rate ratios imply that for each unit increase in cough count per person-time the Walk-In Clinic 2 had a 62% (95% CI: 28 to 103%; $p < 0.0001$), and the Pediatric clinic had an 84% (95%CI: 35 to 144%; $p < 0.0001$) greater likelihood of a flu test being ordered relative to the testing rate in the Walk-In clinic 1. (see Table 9). This observed biological gradient (with high ILI burden in pediatric populations) is consistent with prior state-level surveillance studies [15] and with studies showing very high prevalence of cough specifically in pediatric influenza cases [39], which has been identified as the strongest independent predictor of influenza in pediatric populations [16]. Together these results provide strong statistical evidence that *FluSense* accurately captures true disease-related events by showing strong signals in more-likely-to-be-ill populations, and a null signal in less-likely-to-be-ill populations during the influenza season.

We can also observe a similar trend when we train ILI and confirmed flu count regression models for each location separately with the baseline and top 3 sensor-based features (as described in table 7). It should be highlighted that we have deployed two FluSense platforms in the walk-in clinic waiting room as it is the largest waiting room with a more complex structure (i.e., two subunits divided by a barrier/wall). The daily total flu test and total positive test count can be predicted at a significantly higher level of accuracy with the sensor data collected by the two platforms deployed in the walk-in clinic waiting room. The model trained on walk-in clinic 2 achieves the highest performance with a ρ of 0.65 with a RMSE of 4.28 for total test prediction, while it achieves the ρ of 0.61 and the RMSE of 1.68 for total positive test count prediction. The performance of the walk-in platform 1 (with a ρ of 0.64 with a RMSE of 4.23 for total test) is comparable to that of the walk-in platform 2. Overall, these results support our hypothesis that walk-in clinic is the most informative location for the hospital influenza modeling in this study. On the other hand, the checkup and women's clinic waiting room models perform at a significantly lower at a ρ of respectively 0.53 and 0.52 for total test prediction. Overall, these results show that determining the sensor deployment location is highly critical.

9 DISCUSSION

In this discussion section we first aim to summarize the novel insights generated in the course of this research. We expand discussion of the technical design of the system, and describe how our design could generalize across different real-world scenarios, and in different populations. Finally, we will discuss the limitations and our plans for future work.

Table 9. Multivariable random-effects negative binomial models for relationship between daily total captured cough counts (in 3 waiting areas only) and total clinical influenza tests. Model adjusted for total population in waiting areas, day-of-week effects, and clustering by waiting area and calendar date. Note: IRR = Incidence rate ratio.

Total Influenza tests	*IRR*	*p-value*	95% CI
Daily cough count	1.0026	<0.001	1.0017 - 1.0034
Waiting area			
Walk-In 1	(reference)	-	-
Walk-In 2	1.62	<0.001	1.28 - 2.03
Pediatric	1.84	<0.001	1.38 - 2.44

9.1 New Insights and Key Takeaways

There are several major findings from this work. We have demonstrated that a non-speech body sound, cough, captured directly and anonymously in hospital waiting areas, provides important epidemiological information about influenza trends. This validates the premise of the FluSense platform for routine public health surveillance. We also provide results showing that features like cough count, number of coughs by number of speech activity and number of cough count by number of person-time are better predictors of total of ILI and influenza patients compared to just a simple patient count. Secondly, we have also demonstrated the feasibility and utility of a low-cost and efficient edge-computing platform to capture cough and underlying population-counts in a noisy environment while ensuring personal privacy. Together, these findings illustrate one way this type of edge computing sensor platform can be used to improve the timeliness and effectiveness of current influenza-like illness prediction models.

9.2 Generalizability Across Different Populations and in Real-World Settings and Privacy Considerations

In this work, our contactless sensing platform for influenza-like-illness patient count and influenza positive patient count has been validated in the context of several waiting areas of a university health clinic/hospital. However, we argue that our technique is potentially transferable to diverse public locations. In section 5, we demonstrated that the audio-based cough model performs well in different noisy settings. For example, with different augmentation techniques simulating different types of difficult real-world scenario, we have demonstrated that our cough detector (i.e., model 6 in Table 3b) can achieve an F1 score of 82.4% even when a significant level of speech exists in the background. The model also achieves an F1 score of 87.0% with different mixed augmentations representing a highly noisy and dynamically variable setting. Overall, these results indicate that our cough classification model has the potential to achieve high performance in a crowded public locations including restaurants, large classrooms in schools, waiting areas in public offices, train or bus stations. These types of settings are currently not part of any of the influenza surveillance systems, and represent a important missed opportunity for epidemic early-warning signals.

In the planning stages of the project, the researchers canvassed opinions from clinical care staff and the university ethical review committee to ensure the sensor platform was acceptable and well-aligned with patient protection considerations. All persons discussed major hesitations about collection any high-resolution visual imagery in patient areas, especially in consideration of HIPAA (US Federal Health Insurance Portability and Accountability Act of 1996). High-resolution visual imagery has no additional utility for platform performance beyond IR-based sensing, but is a major data protection liability within a patient care setting. Data breach of any identifiable patient data would incur major financial and legal liabilities for the clinic, researchers and university. Given the aims and scope of the research, this was determined to be an unacceptable risk, and was not pursued. This is the rationale for the system's speech classifier, as we wanted to ensure that we didn't collect any speech data for our system validation purpose and we built a speech classifier that does not collect any speech data.

9.3 Limitations and Future Opportunities

This FluSense sensor array is not without limitations. The system is designed to run all machine learning computations on the edge, so the current platform is limited by the computation power/memory/power requirements for edge computing devices. However, as the capabilities of edge computing devices are rapidly increasing, this situation will only improve, and we anticipate running more complex models at point of sound capture. The thermal camera that we used also has limitations as it is a low-resolution camera with a limited field of view. During our initial planning phase, we explored other thermal cameras with higher-resolution, wide-angle and very accurate skin temperature measurement capabilities; but we found that they are highly expensive

(approximately $10K to $20K USD) and not suitable for a low-cost mobile deployment setting. In this paper, we have demonstrated that even with a low cost thermal camera, the total waiting room patient count could be accurately estimated from the personTime feature estimated from thermal images. We also found during the deployment that there were several sensor array outages; however second versions of the units are expected to be more robust. While there were several time periods when no data was collected due to hardware failures, these data were randomly missing and do therefore not impact the overall findings.

Optimization of the locations for deployment of the FluSense sensor boxes is a critical next step. The device locations should carefully chosen to capture large, and diverse populations with a high likelihood likelihood of ILI symptoms. The sensor boxes also currently require infrastructure for deployment, and administrative/governmental approvals for public use. There may also be concerns about public perception of such devices and some people might find the deployment of such devices as an invasion of their privacy. The IRB process was important in this regard, and clarified several important aspects of data capture. While we have collected data from a single health facility during one influenza transmission season, our results highlight the feasibility and utility of this platform for routine syndromic surveillance. Longer-term studies across multiple transmissions seasons, and in more acoustically and epidemiologically diverse settings will be required to fully validate it. Additionally, a well-defined network of sensors deployed in diverse settings may provide additional insight into the spatio-temporal dynamics of ILI in public settings [10].

10 CONCLUSION

One of the major challenge of deploying ubiquitous computing platforms in real world environment is to effectively analyze diverse sets of noisy signals within the constraints of computational power, size, budget, and ease of deployment. With our FluSense platform, we have developed a system that can gather representative and actionable public health data using a low-cost and privacy sensitive edge computing platform. For this FluSense platform we have developed audio and image recognition models that were subsequently validated in real world settings and can be deployed on edge computing devices. Furthermore, we have shown that based on our sensor data it is possible to predict total patient ILI volumes with a 0.65 correlation coefficient, while predicting the total flu positive patients (correlation = 0.61), illustrating that FluSense provides a novel and useful signal to inform seasonal influenza surveillance and forecasting.

ACKNOWLEDGMENTS

We sincerely thank reviewers for their insightful comments and suggestions that helped improve the paper. We thank the staff of the University Health Services (UHS) at the University of Massachusetts Amherst for helping us during the data collection. This work is partially funded by the Center for Data Science (CDS), the College of Information and Computer Sciences and the Institute for Applied Life Sciences at the University of Massachusetts Amherst. This work was also partially funded by the National Institute of General Medical Sciences (NIGMS) Grant R35GM119582. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of the NIH or NIGMS.

REFERENCES

[1] [n. d.]. ETSI Binaural sound database. https://docbox.etsi.org/stq/Open/TS103224BackgroundNoiseDatabase/Binaural. ([n. d.]).

[2] [n. d.]. *Incidence rate*. https://wiki.ecdc.europa.eu/fem/w/wiki/incidence-rate.

[3] [n. d.]. Intel Neural Compute Stick 2. https://software.intel.com/en-us/neural-compute-stick. ([n. d.]).

[4] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using twitter data. In *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*. IEEE, 702–707.

[5] Benjamin M Althouse, Yih Yng Ng, and Derek AT Cummings. 2011. Prediction of dengue incidence using search query surveillance. *PLoS neglected tropical diseases* 5, 8 (2011), e1258.

[6] J. Amoh and K. Odame. 2016. Deep Neural Networks for Identifying Cough Sounds. *IEEE Transactions on Biomedical Circuits and Systems* 10, 5 (Oct 2016), 1003–1011. https://doi.org/10.1109/TBCAS.2016.2598794

[7] Samantha J. Barry, Adrie D. Dane, Alyn H. Morice, and Anthony D. Walmsley. 2006. The automatic recognition and counting of cough. *Cough* 2 (28 Sep 2006), 8–8. https://doi.org/10.1186/1745-9974-2-8 1745-9974-2-8[PII].

[8] Matthew Biggerstaff, Krista Kniss, Daniel B Jernigan, Lynnette Brammer, Joseph Bresee, Shikha Garg, Erin Burns, and Carrie Reed. 2017. Systematic assessment of multiple routine and near real-time indicators to classify the severity of influenza seasons and pandemics in the United States, 2003–2004 through 2015–2016. *American journal of epidemiology* 187, 5 (2017), 1040–1050.

[9] David A Broniatowski, Michael J Paul, and Mark Dredze. 2013. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one* 8, 12 (2013), e83672.

[10] Vivek Charu, Scott Zeger, Julia Gog, Ottar N Bjørnstad, Stephen Kissler, Lone Simonsen, Bryan T Grenfell, and Cécile Viboud. 2017. Human mobility and the spatial transmission of influenza in the United States. *PLoS computational biology* 13, 2 (2017), e1005382.

[11] P. Dollar, C. Wojek, B. Schiele, and P. Perona. 2009. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 304–311. https://doi.org/10.1109/CVPR.2009.5206631

[12] Wen Dong, Tong Guan, Bruno Lepri, and Chunming Qiao. 2019. PocketCare: Tracking the Flu with Mobile Phones using Partial Observations of Proximity and Symptoms. *arXiv preprint arXiv:1905.02607* (2019).

[13] Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, and Richard E Rothman. 2013. Influenza forecasting with Google flu trends. *PloS one* 8, 2 (2013), e56176.

[14] Gunther Eysenbach. 2006. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium Proceedings*, Vol. 2006. American Medical Informatics Association, 244.

[15] Ashley Fowlkes, Sharoda Dasgupta, Edward Chao, Jennifer Lemmings, Kate Goodin, Meghan Harris, Karen Martin, Michelle Feist, Winfred Wu, Rachelle Boulton, et al. 2013. Estimating influenza incidence and rates of influenza-like illness in the outpatient setting. *Influenza and other respiratory viruses* 7, 5 (2013), 694–700.

[16] Marla J Friedman and Magdy W Attia. 2004. Clinical predictors of influenza in children. *Archives of pediatrics & adolescent medicine* 158, 4 (2004), 391–394.

[17] L.; Fisher W.; Fiscus J.; Pallett D.; Dahlgren N. Garofolo, J.; Lamel. 1990. TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. (1990).

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[19] Jack M Gwaltney Jr. 2002. Clinical significance and pathogenesis of viral respiratory infections. *The American journal of medicine* 112, 6 (2002), 13–18.

[20] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR* abs/1704.04861 (2017).

[21] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara Balan, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. 2017. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 3296–3297.

[22] M. Jeub, M. Schafer, and P. Vary. 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *2009 16th International Conference on Digital Signal Processing*. 1–5. https://doi.org/10.1109/ICDSP.2009.5201259

[23] C. H. Knapp and G. Clifford Carter. 1976. The generalized correlation method for estimation of time delay.

[24] E.C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S.N. Patel. 2011. Accurate and Privacy Preserving Cough Sensing Using a Low-cost Microphone *(UbiComp '11)*. 375–384.

[25] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[26] S. Le and W. Hu. 2013. Cough sound recognition based on Hilbert marginal spectrum. 03 (Dec 2013), 1346–1350. https://doi.org/10.1109/CISP.2013.6743882

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

[28] J. Liu, M. You, G. Li, Z. Wang, X. Xu, Z. Qiu, W. Xie, C. An, and S. Chen. 2013. Cough signal recognition with Gammatone Cepstral Coefficients. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. 160–164. https://doi.org/10.1109/ChinaSIP.2013.6625319

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *ECCV*.

[30] S. Matos, S. S. Birring, I. D. Pavord, and H. Evans. 2006. Detection of cough signals in continuous audio recordings using hidden Markov models. *IEEE Transactions on Biomedical Engineering* 53, 6 (June 2006), 1078–1083. https://doi.org/10.1109/TBME.2006.873548

[31] Noelle-Angelique M Molinari, Ismael R Ortega-Sanchez, Mark L Messonnier, William W Thompson, Pascale M Wortley, Eric Weintraub, and Carolyn B Bridges. 2007. The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine* 25, 27 (2007), 5086–5096.

[32] Arnold S Monto, Stefan Gravenstein, Michael Elliott, Michael Colopy, and Jo Schweinle. 2000. Clinical signs and symptoms predicting influenza infection. *Archives of internal medicine* 160, 21 (2000), 3243–3247.

[33] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10)*. Omnipress, USA, 807–814. http://dl.acm.org/citation.cfm?id=3104322.3104425

[34] Elaine O Nsoesie, Patrick Butler, Naren Ramakrishnan, Sumiko R Mekaru, and John S Brownstein. 2015. Monitoring disease trends using hospital traffic data from high resolution satellite imagery: A feasibility study. *Scientific reports* 5 (2015), 9112.

[35] Dave Osthus, Ashlynn R Daughton, and Reid Priedhorsky. 2019. Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLoS computational biology* 15, 2 (2019), e1006599.

[36] Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. 2008. Using internet searches for influenza surveillance. *Clinical infectious diseases* 47, 11 (2008), 1443–1448.

[37] Sourabh Ravindran, David V. Anderson, and Malcolm Slaney. 2006. Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing. In *SAPA@INTERSPEECH*.

[38] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), 1137–1149.

[39] Bernhard R Ruf and Markus Knuf. 2014. The burden of seasonal and pandemic influenza in infants and children. *European journal of pediatrics* 173, 3 (2014), 265–276.

[40] Tara N. Sainath and Carolina Parada. 2015. Convolutional neural networks for small-footprint keyword spotting. In *INTERSPEECH*.

[41] Jeffrey Shaman, Sasikiran Kandula, Wan Yang, and Alicia Karspeck. 2017. The use of ambient humidity conditions to improve influenza forecast. *PLoS computational biology* 13, 11 (2017), e1005844.

[42] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one* 6, 5 (2011), e19467.

[43] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI*.

[44] Shinya Takahashi, Tsuyoshi Morimoto, Sakashi Maeda, and Naoyuki Tsuruta. 2004. Cough detection in spoken dialogue system for home health care. In *INTERSPEECH*.

[45] J. Thiemann, N. Ito, and E. Vincent. 2013. The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. *Acoustical Society of America Journal* 133 (2013), 3591. https://doi.org/10.1121/1.4806631

[46] https://www.alere.com/en/home/product-details/id-now.html. [n. d.]. ([n. d.]). Abbott ID NOW Rapit Influenza Test.

[47] https://www.thermal.com/compact-series.html/. [n. d.]. ([n. d.]). Seek CompactPRO: Thermal Camera.

[48] http://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/. [n. d.]. ([n. d.]). ReSpeaker: Microphone Array V2.

[49] US Centers for Disease Control and Prevention. 2018. Overview of Influenza Surveillance in the United States. (2018). https://www.cdc.gov/flu/weekly/overview.htm.

[50] US Centers for Disease Control and Prevention. 2018. Past Seasons Estimated Influenza Disease Burden. (2018). https://www.cdc.gov/flu/about/burden/past-seasons.html.

[51] Yuxin Wu and Kaiming He. 2018. Group Normalization. In *ECCV*.

[52] C. Zhu, L. Tian, X. Li, H. Mo, and Z. Zheng. 2013. Recognition of cough using features improved by sub-band energy transformation. In *2013 6th International Conference on Biomedical Engineering and Informatics*. 251–255. https://doi.org/10.1109/BMEI.2013.6746943

A SYSTEM BENCHMARKING: THROUGHPUT, POWER AND MEMORY

Figure 11a shows the data throughput in terms of duration per frame of the audio and thermal image processing pipelines. To be realtime, our system needs to operate below the horizontal dash line (as shown in figure 11a). To process one second of audio data including sampling, Fourier transformation, speech and cough inference, our system requires about 0.3 seconds. Our system takes about 0.97 seconds to process one thermal image which includes thermal image capture, preprocessing, and crowd size estimation inference. As the system needs to process a thermal image every minute in addition to the one audio frame per second, the total execution time becomes 0.32 seconds, and thus our system can operate in a real-time manner. Power benchmarking results are shown in figure 11b. The base raspberry pi-system without running our algorithms consumes about 2.9 W. The full thermal imaging and audio processing pipeline on the raspberry pi and Intel neural stick increase the power consumption by about 1 W and the complete system operates at 4.06 W. The memory consumption of the FluSense audio and thermal image processing are fairly minimal. While the audio model consumes about 14 MB of the Raspberry Pi's memory, the image model (i.e., SSD MobileNet) alone consumes about 148 MB. When we run both the audio and image models together, the total required memory on the raspberry pi is about 149 MB. This memory requirement of the system on the raspberry pi is fairly minimal as both the audio and image models are executed by the neural computing engine. The neural computing engine can run the audio model with 700k parameters and the image model (based on SSD mobilenet_fpn) with 12 million parameters in real-time. During the clinical deployment we stored the raw thermal images with encryption in the hard drive as we needed the raw data for modeling and evaluation. However, this system benchmarking analysis clearly shows that the FluSense platform can handle all the required audio and thermal image computation on the edge in real time.

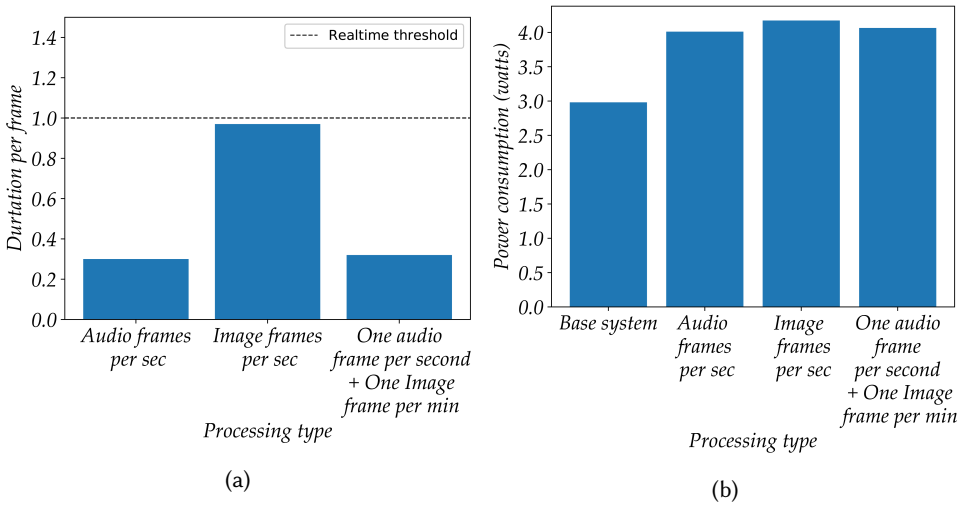

Fig. 11. shows the system benchmarking in terms of (a) throughput and (b) power consumption for the proposed platform.

B SYSTEM BENCHMARKING: THROUGHPUT, POWER AND MEMORY

APPENDIX B: Cough Model Training Parameters

The Cough model uses 3 convolutional layers with 16 filters in each layer. The first layer consists of CNN filters with 64 filters each having 8 size width and 20 size height with stride 1. The output of the filters are then passed through a ReLU[33] and a maxpool layer with filter dimensions $2x2$ with stride value 2. A group normalization

layer is situated between two convolutional layers and helps to converge faster. The fully connected layer takes the output activation maps from the third layer of CNN and projects it down to a two-class score or binary classifier (cough vs not cough). These scores are then passed to a softmax layer, which converts this logit scores to probability for a "cough" and "not cough" events.

C SYSTEM BENCHMARKING: THROUGHPUT, POWER AND MEMORY

APPENDIX C: Extracting Total Unique Cough DOAs (tucd) Feature

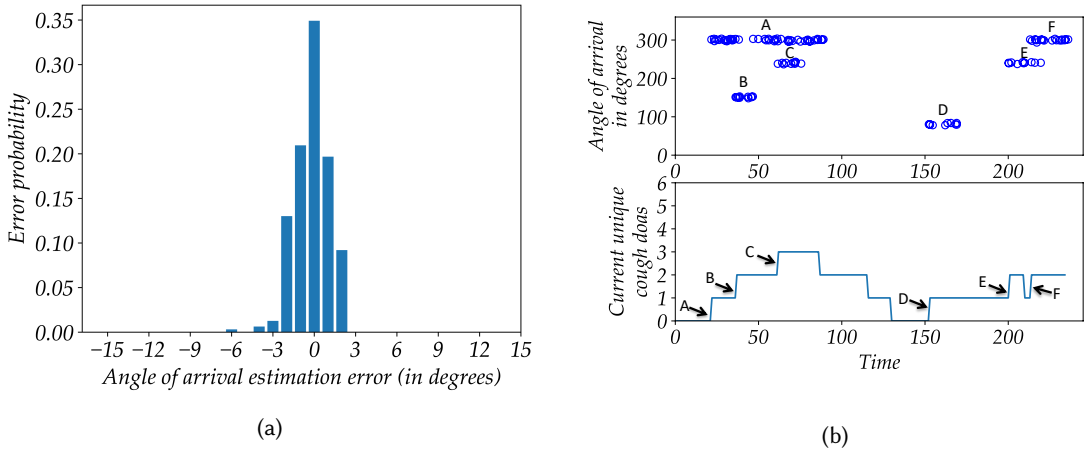

Fig. 12. (a) shows the DOA estimation error by the microphone array. (b) shows the Total Unique Cough DOAs (tucd) feature estimation when six persons are coughing (referred as A-F) over a period of time.

The heuristic feature tucd is estimated in the following way: Every time a new cough event is inferred, we examine all the past cough events that took place in the last 6 minutes within 9°of the direction of arrival (DOA). If such an event is found, the tucd estimation algorithm considers that the new cough event is coming from the same spatio-temporal cluster which has already been registered. As a result, the tucd count is not increased. On the other hand, if no cough event is found from the same spatio-temporal cluster, the newly inferred cough event is registered under a new spatio-temporal cluster and the tucd count is increased. Here, we have used two parameters to define the spatio-temporal cluster including a timeout criteria of 6 minutes and a DOA range bin of 9°. In our deployment within the waiting rooms, the distance between the seats and the FluSense platforms are at least 20 feet while the seat width is approximately 3 feet. As a result, we define the spatial range of the spatio-temporal cluster to be an angular region of 9 degrees (approximately, 2 arctan 1.5/20 degrees or 8.57°). Also, as can be seen in figure 12a, the standard deviation of the microphone array estimation error is 1.26 which is significantly less than the DOA angular region of 9 degrees. The timeout hyper-parameter of 6 minutes defines the temporal range of the spatio-temporal cluster and was chosen based on overall ILI model performance.

Figure 12b illustrates how we can estimate tucd feature when 6 persons (A-F) coughs over a period of time in a room. On the top figure, the blue hollow circles represent the direction of arrivals (DOAs) of different inferred coughing events from unique individuals. As A, B, and C start to cough from different unique DOAs, the current unique cough DOAs feature increases the counter to account for 3 new unique DOAs. At time=90, the maximum

inactivity period of B crosses the timeout period, B is unregistered and the current unique DOA count is decreased by one point. Similarly, both the C and A clusters are unregistered after a period of inactivity, which allows new spatio-temporal clusters to form at the same DOA at a future time. Later, both E and F are again registered as two new unique cough DOAs which overlaps with respectively C and A spatially. To estimate the total unique cough DOAs, we sum the number of times the current unique DOA increases (indicated by the arrow).