medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Validity of deep learning algorithms for detecting wheezes and crackles from lung sound recordings in adults

Hasse Melbye[1], Johan Ravn[2], Mikolaj Pabiszczak[2], Lars Ailo Bongo[3], Juan Carlos Aviles Solis[1]

[1] General Practice Research Unit, UIT The Arctic University of Norway

[2] Medsensio AS, Tromsø, Norway

[3]Department of Computer Science, UIT The Arctic University of Norway

Correspondence: hasse.melbye@uit.no

Abstract

We aimed at evaluating deep learning algorithms for detecting wheezes and crackles developed based on sound files from 4033 adults in two samples of sound files not used in the algorithm development**.** In sample A, ground truth was established by experienced raters in 615 files from the Tromsø population study. Sample B contained 120 sound files from a previous interobserver study with ground truth determined by four experts. The algorithms' probability scores for wheezes and crackles were evaluated against the ground truth by calculating Area Under Curve (AUC). Agreements between the algorithm and the human annotations were also assessed by Kappa statistics. In sample A the AUC was 0.88 (95% CI 0.84 – 0.92) for wheezes and 0.88 (95% CI 0.84 – 0.92) for crackles. The kappa agreement between dichotomized labelling and the ground truth was 0.63 (95% CI 0.56 – 0.71) for wheezes and 0.68 (95% CI 0.60 – 0.75) for crackles. The corresponding kappa agreements between the human raters were 0.47. In sample B, an AUC of 0.99 (95% CI 0.98 – 1.0) was reached for wheezes and 0.95 (95% CI 0.89 – 1.0) for crackles with corresponding kappas of 0.78 (95% CI 0.58 – 0.99) and 0.75 (95% CI 0.59 – 0.92). The corresponding mean kappas between the ground truth and 24 observers were 0.68 and 0.55. The algorithm agreed substantially with ground truth, and with higher kappa agreements than observed between human annotators.

Keywords: Deep learning, algorithms, crackles, wheezes

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

1. Introduction

The abnormal lung sounds, wheezes and crackles, are present with increased frequency in patients with chronic obstructive pulmonary disease [1, 2], heart failure [3] pneumonia [4], and pulmonary fibrosis [5], and are regarded to be useful in the diagnosis of these diseases. Their identification during chest auscultation with a traditional stethoscope is hampered by subjectivity and interobserver variation [6, 7] This problem has been met by letting computers classify sounds recorded by electronic stethoscopes [8]. In recent years machine learning based algorithms for detecting adventitious lung sound, mainly wheezes and crackles, have been developed and evaluated[9-12], and discussions on opportunities and pitfalls are going on [12, 13]. There has also been attempts to go one step further and use lung sounds for direct diagnosing lung diseases [14-18]. In this study we stick to the identification of specific adventitious lung sounds and evaluate a state-of-the-art deep learning algorithm for detecting wheezes and crackles developed from recordings in a general population. We have evaluated the algorithm against human ratings of wheezes and crackles in sound files not been involved in the development of the algorithm.

2. Methods

2.1 The lung sound recordings

Two samples of lung sound recordings were used. The first sample, sample A, was based on 12090 sound files from 2015 participants of the 7th Tromsø Study, previously not annotated by humans and not used in the development of the algorithm. The 7th Tromsø Study was a population-based health survey performed between May 2015 and October 2016. Main features of the methodology and study design have previously been described [19, 20]. All Tromsø residents 40 years and older (n=32 591) were invited to participate and a random sample was selected for a second visit where lung sound recording was included. Lung sounds from 6048 participants were recorded at six locations of the chest (Fig 1), 15 seconds at each site, with a Sennheiser microphone inserted in the tube of a Littmann Classic II stethoscope. No preprocessing or filtering was done. Of the 12090 files not previously annotated, 615 were selected for this study.

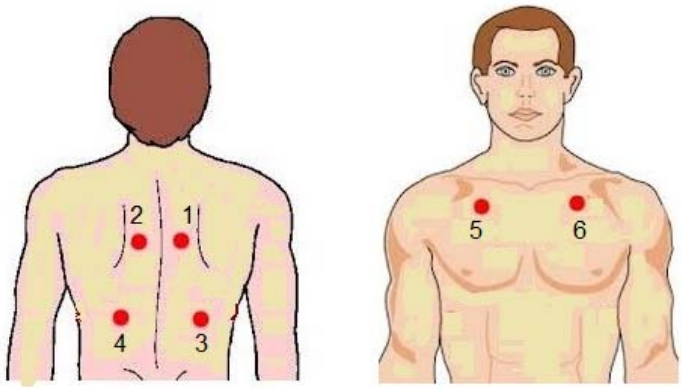


Figure 1. The six recording sites used

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

In sample B, the same stethoscope was used and lung sounds were recorded among volunteers aged 40 years or more, most of them patients at a rehabilitation center. Lung sounds from six chest locations recorded in 20 of the volunteers were used, 120 in total. The sound files had originally a duration of 10 -15 seconds but had been shortened to avoid sections with noise. These recordings were first used in an interobserver study, in which four lung sound researchers, 20 medical doctors and four medical students took part [21].

2.2 Algorithm development

The algorithm was developed using deep learning based on 24198 lung sound recordings from 4033 participants of the 7th Tromsø study. These files did not overlap with the samples A and B described above.

<u>The ground truth</u>

Presence of wheezes and crackles during inspiration and expiration had been determined through a rigorous classification process[19]. At first, two observers (clinicians) independently classified the recordings blinded for other information about the participants. When the observers disagreed, they discussed the actual recordings with a third observer. The recordings judged to contain certain or likely crackles or wheezes, were evaluated in a second round, where experienced lung sound researchers were among the observers. Based on the observers' annotations, a final decision was made on whether wheezes or crackles were present or not [19], which we used as "ground truth". When listening to and classifying the lung sounds, the observers watched spectrograms of the recordings generated by Adobe Audition©.

<u>The algorithm</u>

An architecture based on inception V3 was selected to build the deep learning algorithm[22]. The raw audio data was converted to mel spectrograms, thus turning the audio classification task into an image classification task.

For a single raw audio signal three such spectrograms were extracted, each using different size of the Fourier transform window. Subsequently, the three spectrograms were stacked together forming an RGB format-like image that was used as an input to the machine learning models. The last layer of the architecture is a *sigmoid* activation function that predicts each class with a probability of between 0 and 1. As this is partially a multi-label classification task, a sample can be both wheeze and crackle at the same time (although it cannot be wheeze and bad quality at the same time), we use binary cross-entropy as the loss function for the model.

The models were trained in the 5-fold cross validation procedure. In each fold a model was trained on the fold-specific training set and evaluated on the fold-specific validation set. For each fold, the model that obtained the highest ROC AUC (calculated on the fold-specific validation dataset) was selected as the result model. After selecting the best model in each fold, the next step was the selection of thresholds for each label. Thresholds are used for deciding whether the probability for a given label is high enough, so that the given label should be assigned (encoded). Distinct sets of thresholds are selected for each of 5 models and the selection of thresholds is performed with the use of the fold-specific validation set (the selection is performed in such a way as to maximize label-specific F1-score). The result model from each of the folds was used to form an ensemble of 5 models which together yielded the final label by the means of majority voting.

The development of the algorithm started at the Department of Computer Science at UIT the Arctic University of Norway. Based on funds from the Norwegian Research Council, a start-up company,

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Medsensio AS, was established, where the algorithm has been further developed. We evaluate the most recent version.

2.3 Evaluating the algorithm in sample A

Selection of study sample

To reduce the annotation workload, we selected 615 files from the 12090 sound files in sample A. To get a selection with close to equal numbers of files with normal sounds, wheezes, and crackles, we applied another algorithm to classify the recordings. It was a single model developed on the same subset from the Tromsø study that was used for the development of the ensemble of models evaluated in this paper, with sparser input features. To further reduce possible bias in the evaluation results favoring the algorithmic solution that could arise by this method of selecting a subset, the following procedure was followed. Firstly, prediction with the use of the algorithm was performed for all 12090 files. For each class (normal, wheeze, crackle) assigned by the algorithm, 200 files were selected in such a way, that – conditional on the given label being assigned – a quarter of files came from the 1st quartile, a quarter from the 2nd etc. and the division into quartiles for a given label was performed based on values of the scores (probabilities) for a given label assigned by the deep learning model. Additionally, 17 files classified as bad quality were incorporated into selection. Within the selection 2 files were predicted by the algorithm to have both wheeze and crackles present, hence the total of 615 files was selected.

Establishing "ground truth" in the study samples

The 615 sound files were annotated independently by two medical doctors who were experienced raters (HM and JCAS). The raters watched spectrograms of the recordings and were blinded for clinical information. Sound files on which the two raters disagreed were annotated again by both of them together and consensus was reached, which was used as ground truth.

In Sample B the ratings of the 120 sound files done by the four lung sound researchers were used when establishing ground truth. These raters had watched spectrograms of the recordings and were blinded for clinical information. The criteria for presence of wheezes and crackles were fulfilled when at least three out of the four expert raters had annotated their presence.

2.4 Statistical analysis

The ability of the algorithm to detect wheezes and crackles was assessed by calculating area under the curve (AUC) through Receiver operating characteristics (ROC) curve analysis and by calculating Kappa-agreement with ground truth. Sensitivity (also called "recall"), specificity, and positive predictive value (PPV, also called "precision") for detecting wheezes and crackles were also calculated). For comparing the algorithm with human annotations, we also calculated the kappa-agreement between the two annotators in sample A, and between each of 24 raters and the ground truth in sample B. We chose to use this many statistical methods to make the results more easily comparable with other studies. SPSS statistical software was used in most of the analyses. The 95% confidence intervals (CI) of sensitivities, specificities, and PPVs were obtained by use of MedCalc® statistical software, the 95% CI of kappa-agreements were calculated by use of Vassarstats®. F1-scores were calculated based on the confusion matrices as F1-score = $\frac{TP}{TP+\frac{1}{2}(FP+FN)}$, where TP, FP, and FN denote the numbers of true positives, false positives and false negatives, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Table 1. 2x2 tables showing sensitivity and specificity positive predictive value (PPV) of the algorithm for detecting wheezes and crackles (ground truth) in lung sound recordings in to separate samples, A (n=616) and B (n=120).

Sample A

Wheezes

		Ground truth		
		Yes	No	
Algorithm	Yes	96	56	152
	No	22	441	463
		118	497	615

Sensitivity: 81% (95% CI 73% - 88%),
Specificity: 89% (95% CI 86% - 91%)
PPV: 63% (95% CI 57% - 69%)
F1-score: 71%

Crackles

		Ground truth		
		Yes	No	
Algorithm	Yes	85	20	105
	No	41	469	510
		126	489	615

Sensitivity: 67% (95% CI 59% - 76%),
Specificity: 96% (95% CI 94% - 97%)
PPV: 81% (73% - 87%)
F1-score: 74%

Sample B

Wheezes

		Ground truth		
		Yes	No	
Algorithm	Yes	8	4	12
	No	0	108	108
		8	112	120

Sensitivity: 100% (95% CI 63% - 100%),
Specificity: 96% (95% CI 91% - 99%)
PPV: 67% (95% CI 43% - 84%)
F1-score 80%

Crackles

		Ground truth		
		Yes	No	
Algorithm	Yes	15	4	150
	No	4	97	446
		19	101	120

Sensitivity: 79% (95% CI 54% - 94%),
Specificity: 96% (95% CI 90% - 99%)
PPV: 79% (95% CI 58% - 91%)
F1-score 79%

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

3. Results

3.1 Evaluation in Sample A

Among the 12090 sound files in sample A, the algorithm identified wheezes in 1060 (8.8%) and crackles in 592 (4.9%) Within the 615 selected files the ensemble of models being evaluated considered only 252 as containing any abnormality, 152 files with wheezes and 105 files with crackles. Among the 17 files selected due to bad quality, 16 were also found to be of too bad quality to be annotated by the human raters, no lung sound was heard in 13, while four were too noisy to be annotated.

Table 2. Kappa-agreement with 95% confidence interval between algorithm and human annotations in sample A (596 recordings from Tromsø 7)

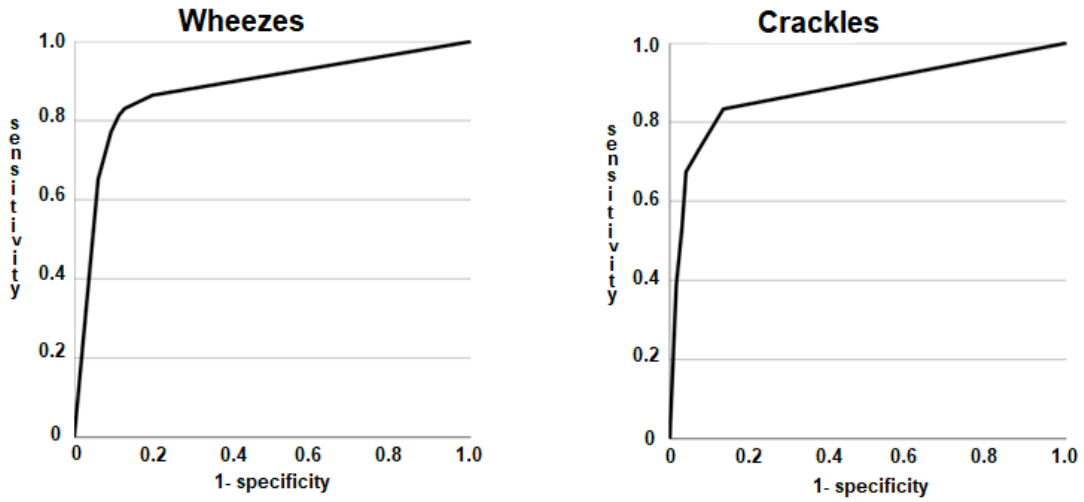
	Kappa	**(95% CI)**
Wheezes		
Algorithm versus rater A	0.576	(0.501 – 0.650)
Algorithm versus rater B	0.524	(0.443 – 0.606)
Algorithm versus ground truth (annotator consensus)	**0.631**	**(0.558 – 0.705)**
Annotator A versus Annotator B	0.465	(0.383 – 0.547)
Crackles		
Algorithm versus rater A	0.609	(0.530 – 0.689)
Algorithm versus rater B	0.602	(0.519 – 0.685)
Algorithm versus ground truth (annotator consensus)	**0.676**	**(0.600 – 0.751)**
Rater A versus rater B	0.469	(0.382 – 0.555)

Compared to the ground truth, the algorithms' prediction scores for wheezes and crackles had an AUC of 0.88 (95% CI 0.84 – 0.92) for wheezes and 0.88 (95% CI 0.84 – 0.92) for crackles (Fig. 2). When the scores were dichotomized, the algorithm detected wheezes with a sensitivity of 81% and a specificity of 89% and crackles with a sensitivity of 67% and a specificity of 97% (Table 1). The kappa agreement between the algorithm and the ground truth was 0.631 (95% CI 0.558 – 0.705) for wheezes and 0.68 (95% CI 0.60 – 0.75 for crackles, which was better agreement than with each of the annotators (Table 2). For comparison, the corresponding kappa agreements between the two annotators were 0.47 and 0.47.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Figure 2. ROC-curves showing the predictive value of the algorithms' wheeze and crackle scores for ground truth wheezes (n=118) and crackles (n=126) in sample A (615 lung sound recordings)



AUC: 0.878 (95% CI 0.838 – 0.919)

AUC: 0.880 (95% CI 0.839 – 0.921)

3.2 Evaluation in Sample B

Among the 120 files in sample B, the algorithms' prediction scores had an AUC of 0.991 (95% CI 0.976 – 1.0) for wheezes and 0.949 (95% CI 0.885 – 1.0) for crackles. Dichotomized, the algorithm detected wheezes with a sensitivity of 100 % and specificity of 96% and crackles with a sensitivity of 79% and specificity of 96% (Table 1), The kappa agreement between the algorithm and the ground truth was 0.783 (95% CI 0.578 – 0.987) for wheezes and 0.749 (95% CI 0.585 – 0.915) for crackles. The kappa agreements between the ground truth and the 24 observers varied between 0.130 and 1.000 for wheezes with a mean of 0.674 and between 0.211 and 0.820 for crackles with a mean of 0.548 (Fig 3).

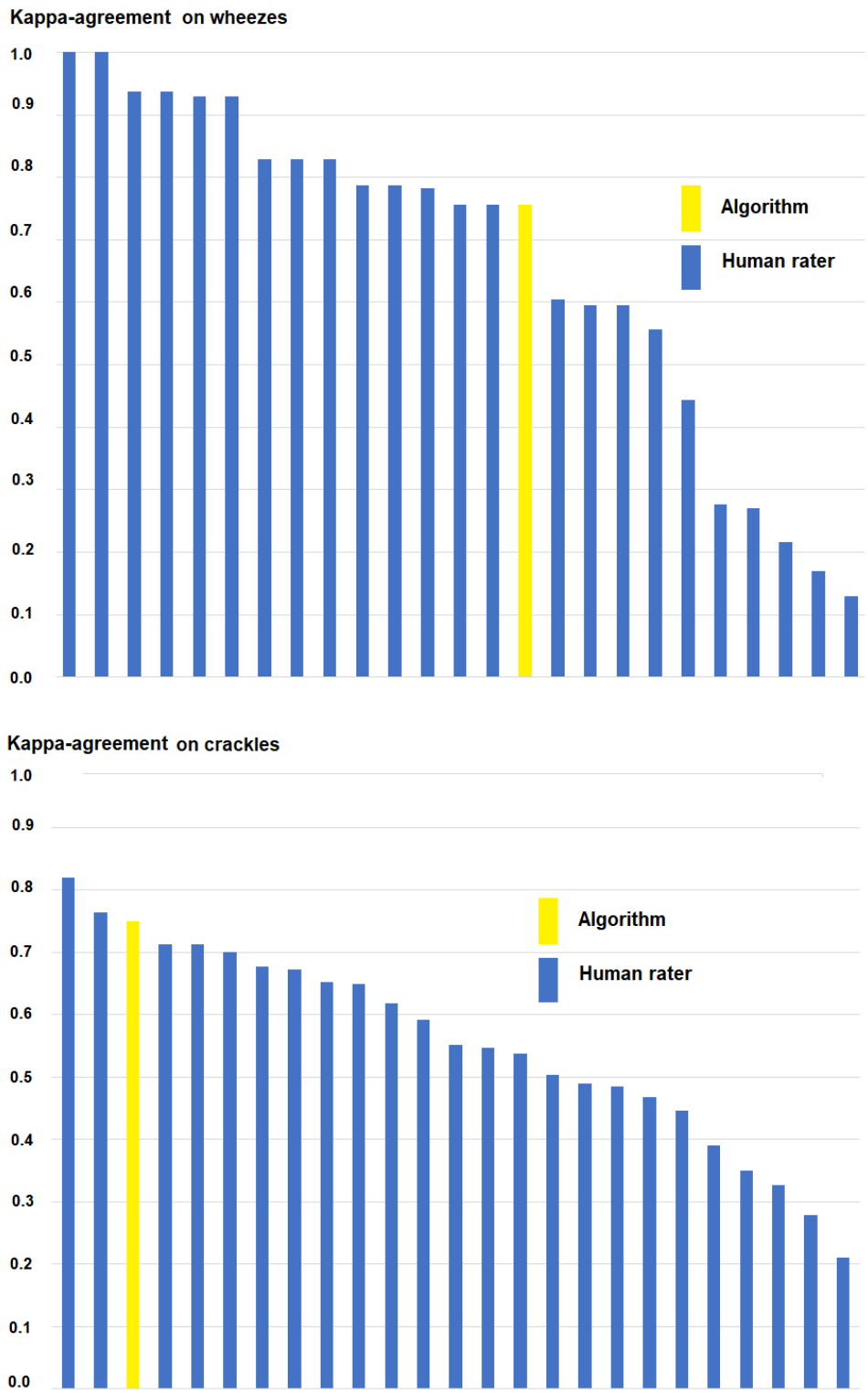
4. Discussion

The algorithm showed agreements with the ground truth that surpassed the agreements between human annotators rating the same files. Accordingly, lung sound classification done by a computer seems to be more reliable than when done by an average physician. The lung sound files in Sample A were recorded by the same method as the data set that trained the algorithm. The recordings in sample B were done with the same electronic stethoscope, but in a different setting. It was still not surprising that the best agreement was found in sample B, since those files were specially selected to be of good quality.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Figure 3. Kappa-agreements between 24 human raters and the ground truth on the presence of wheezes and crackles in sample B (120 lung sound recordings).



medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Comparison with previous studies

We found Kappa agreements in the range 0.6 to 0.8 which is regarded to be substantial [23]. Studies of agreement between human annotators often show lower values. McCollum and coworkers found kappa-agreements of 0.45 for wheezes and 0.41 for crackles [24]. Melbye and coworkers found kappa-agreements between experienced observers to be 0.59 for wheezes and 0.62 for crackles[6]. Ferreira-Cardoso and coworkers found a kappa-agreement on any abnormal sound in good-quality recordings of 0.66 [25]. Three pediatric pulmonologists identified wheezes with a kappa of 0.76 in 55 good-quality recordings from six patients [26]. In a clinical study of 115 hospitalized patients, similar kappa-agreements as in our study was found for wheezes and fine crackles, but lower values for rhonchi and coarse crackles [27].

A few comparable studies have evaluated algorithms for detection of wheezes and crackles. Kim and coworkers found higher concordance between algorithms and the reference classification, but the evaluated was done in sound files which had also been used for the development of the algorithms [9]. Grzywalski and coworkers found that the algorithm (neural network) detected wheezes and crackles in children with sensitivities ranging from 56% to 88% and with specificities from 79% to 88% [10]. Compared to our results, these values were generally somewhat lower, but the subdivision of wheezes into wheezes and rhonchi and crackles into fine and coarse may have contributed to lower concordance [6]. Anyhow, the algorithm obtained higher sensitivity than pediatricians, and similar specificity. Kevat and coworkers found a concordance between the algorithm and the expert classification, also in recordings from children, to be in a similar range as in our study. The sensitivity and specificity for detecting wheezes were 76% - 90% and 95% - 97%, respectively, and for crackles 60%-86% and 96%-99%. They concluded that the algorithm was promising and with at least similar diagnostic accuracy to that of many clinicians [11].

Strengths and limitations

Some strengths are related to the development of the algorithm, others to how the evaluation has been carried out. The development of the algorithm was based on a large number of recordings, 24090 in total, and rigorous human classifications without access to clinical information. We could evaluate the algorithm in two samples of lung sound recording, one principally identical to the sample that trained the algorithm and one from a different setting. However, the same stethoscope was used for both samples. This might limit the validity of the algorithm when it comes to recordings with different devices. The algorithm we evaluated only tell whether wheezes and crackles are present or not and give no detailed information on quality and timing of the adventitious sounds, which could increase the algorithms' diagnostic usefulness [12]. The algorithm would probably have performed worse if the evaluation sets had been recorded in more noisy settings [13].

5. Conclusion

The algorithm evaluated in this study agreed substantially with the ground truth on the presence of wheezes and crackles in lung sound recordings. The algorithm reached higher kappa agreements than what was observed between human raters. However, we cannot be sure that similar good results can be obtained in different settings or when other recording devices are used. These results may enable novel use of lung sounds in clinical and research applications.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Author Contributions: Conceptualization HM and JR; Sound file annotations HM and JCAS; Algorithm development JR, MP, and LAB; Statistical and formal analysis HM and MP; Writing original draft HM and MP; Writing review and editing HM, JR, MP, LAB and JCAS.

Funding: This research received no external funding

Data availability: Researchers can apply for access to the The Tromsø Study data at: https://uit.no/research/tromsostudy

Acknowledgments: The Tromsø study was approved by the Norwegian Data Inspectorate and the Regional Ethical Committee of North Norway (REK). Only the sound files and variables classifying the sounds were used, and identification of the participants was not possible. We have used the TRIPODreporting guidelines [28].

Conflicts of Interest: The results of this work will be used by the Medsensio AS company where JR is CTO. JR and LAB have shares in Medsensio AS. MP is an employee in Medsensio AS. HM and JCAS have done paid work for Medsensio AS.

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

References

1. Bohadana A, Izbicki G, Kraman SS: **Fundamentals of lung auscultation**. *N Engl J Med* 2014, **370**(8):744-751.
2. Melbye H, Aviles Solis JC, Jacome C, Pasterkamp H: **Inspiratory crackles-early and late-revisited: identifying COPD by crackle characteristics**. *BMJ Open Respir Res* 2021, **8**(1).
3. Melbye H, Stylidis M, Solis JCA, Averina M, Schirmer H: **Prediction of chronic heart failure and chronic obstructive pulmonary disease in a general population: the Tromso study**. *ESC Heart Fail* 2020.
4. van Vugt SF, Broekhuizen BD, Lammens C, Zuithoff NP, de Jong PA, Coenen S, Ieven M, Butler CC, Goossens H, Little P *et al*: **Use of serum C reactive protein and procalcitonin concentrations in addition to symptoms and signs to predict pneumonia in patients presenting to primary care with acute cough: diagnostic study**. *BMJ* 2013, **346**:f2450.
5. Vyshedskiy A, Murphy R: **Crackle Pitch Rises Progressively during Inspiration in Pneumonia, CHF, and IPF Patients**. *Pulm Med* 2012, **2012**:240160.
6. Melbye H, Garcia-Marcos L, Brand P, Everard M, Priftis K, Pasterkamp H: **Wheezes, crackles and rhonchi: simplifying description of lung sounds increases the agreement on their classification: a study of 12 physicians' classification of lung sounds from video recordings**. *BMJ Open Respir Res* 2016, **3**(1):e000136.
7. Benbassat J, Baumal R: **Narrative review: should teaching of the respiratory physical examination be restricted only to signs with proven reliability and validity?** *J Gen Intern Med* 2010, **25**(8):865-872.
8. Pramono RXA, Bowyer S, Rodriguez-Villegas E: **Automatic adventitious respiratory sound analysis: A systematic review**. *PLoS One* 2017, **12**(5):e0177926.
9. Kim Y, Hyon Y, Jung SS, Lee S, Yoo G, Chung C, Ha T: **Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning**. *Sci Rep* 2021, **11**(1):17186.
10. Grzywalski T, Piecuch M, Szajek M, Breborowicz A, Hafke-Dys H, Kocinski J, Pastusiak A, Belluzzo R: **Practical implementation of artificial intelligence algorithms in pulmonary auscultation examination**. *Eur J Pediatr* 2019, **178**(6):883-890.
11. Kevat A, Kalirajah A, Roseby R: **Artificial intelligence accuracy in detecting pathological breath sounds in children using digital stethoscopes**. *Respir Res* 2020, **21**(1):253.
12. Kim Y, Hyon Y, Lee S, Woo SD, Ha T, Chung C: **The coming era of a new auscultation system for analyzing respiratory sounds**. *BMC Pulm Med* 2022, **22**(1):119.
13. Rocha BM, Pessoa D, Marques A, Carvalho P, Paiva RP: **Automatic Classification of Adventitious Respiratory Sounds: A (Un)Solved Problem?** *Sensors (Basel)* 2020, **21**(1).
14. Yu H, Zhao J, Liu D, Chen Z, Sun J, Zhao X: **Multi-channel lung sounds intelligent diagnosis of chronic obstructive pulmonary disease**. *BMC Pulm Med* 2021, **21**(1):321.
15. Hafke-Dys H, Kuznar-Kaminska B, Grzywalski T, Maciaszek A, Szarzynski K, Kocinski J: **Artificial Intelligence Approach to the Monitoring of Respiratory Sounds in Asthmatic Patients**. *Front Physiol* 2021, **12**:745635.
16. Vidhya B, Nikhil Madhav M, Suresh Kumar M, Kalanandini S: **AI Based Diagnosis of Pneumonia**. *Wirel Pers Commun* 2022:1-16.
17. Tariq Z, Shah SK, Lee Y: **Feature-Based Fusion Using CNN for Lung and Heart Sound Classification**. *Sensors (Basel)* 2022, **22**(4).
18. Fukumitsu T, Obase Y, Ishimatsu Y, Nakashima S, Ishimoto H, Sakamoto N, Nishitsuji K, Shiwa S, Sakai T, Miyahara S *et al*: **The acoustic characteristics of fine crackles predict honeycombing on high-resolution computed tomography**. *BMC Pulm Med* 2019, **19**(1):153.
19. Aviles-Solis JC, Jacome C, Davidsen A, Einarsen R, Vanbelle S, Pasterkamp H, Melbye H: **Prevalence and clinical associations of wheezes and crackles in the general population: the Tromso study**. *BMC Pulm Med* 2019, **19**(1):173.
20. Hopstock LA, Grimsgaard S, Johansen H, Kanstad K, Wilsgaard T, Eggen AE: **The seventh survey of the Tromso Study (Tromso7) 2015-2016: study design, data collection,**

medRxiv preprint doi: https://doi.org/10.1101/2022.11.18.22282442; this version posted November 18, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

attendance, and prevalence of risk factors and disease in a multipurpose population-based health survey. *Scand J Public Health* 2022:14034948221092294.

21. Aviles-Solis JC, Vanbelle S, Halvorsen PA, Francis N, Cals JWL, Andreeva EA, Marques A, Piirila P, Pasterkamp H, Melbye H: **International perception of lung sounds: a comparison of classification across some European borders**. *BMJ Open Respir Res* 2017, **4**(1):e000250.
22. Szegedy C VV, Ioffe S, Shlens J, Wojna Z: **Rethinking the inception architecture for computer vision**. *ArXiv* 2015:1512.00567v00563.
23. Landis JR, Koch GG: **The measurement of observer agreement for categorical data**. *Biometrics* 1977, **33**(1):159-174.
24. McCollum ED, Park DE, Watson NL, Buck WC, Bunthi C, Devendra A, Ebruke BE, Elhilali M, Emmanouilidou D, Garcia-Prats AJ *et al*: **Listening panel agreement and characteristics of lung sounds digitally recorded from children aged 1-59 months enrolled in the Pneumonia Etiology Research for Child Health (PERCH) case-control study**. *BMJ Open Respir Res* 2017, **4**(1):e000193.
25. Ferreira-Cardoso H, Jacome C, Silva S, Amorim A, Redondo MT, Fontoura-Matias J, Vicente-Ferreira M, Vieira-Marques P, Valente J, Almeida R *et al*: **Lung Auscultation Using the Smartphone-Feasibility Study in Real-World Clinical Practice**. *Sensors (Basel)* 2021, **21**(14).
26. Bae W, Kim K, Yoon JS: **Interrater reliability of spectrogram for detecting wheezing in children**. *Pediatr Int* 2021, **64**(1):e15003.
27. Ramos-Hernandez C, Botana-Rial M, Nunez-Fernandez M, Lojo-Rodriguez I, Mouronte-Roibas C, Salgado-Barreira A, Ruano-Ravina A, Fernandez-Villar A: **Validity of Lung Ultrasound: Is an Image Worth More Than a Thousand Sounds?** *J Clin Med* 2021, **10**(11).
28. Collins, G.S., Reitsma, J.B., Altman, D.G. *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* **13**, 1 (2015). https://doi.org/10.1186/s12916-014-0241-z