Classification of Adventitious Respiratory Sound Events: A Stratified Analysis

Tiago Fernandes
Univ Coimbra
CISUC, DEI
Coimbra, Portugal
tfernandes@student.dei.uc.pt

Bruno M. Rocha
Univ Coimbra
CISUC, DEI
Coimbra, Portugal
bmrocha@dei.uc.pt

Diogo Pessoa
Univ Coimbra
CISUC, DEI
Coimbra, Portugal
dpessoa@dei.uc.pt

Paulo de Carvalho
Univ Coimbra
CISUC, DEI
Coimbra, Portugal
carvalho@dei.uc.pt

Rui Pedro Paiva
Univ Coimbra
CISUC, DEI
Coimbra, Portugal
ruipedro@dei.uc.pt

2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-8791-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/BHI56158.2022.9926841

***Abstract*—Respiratory diseases are among the deadliest in the world. Adventitious respiratory sounds, such as wheezes and crackles, are commonly present in these pathologies. Automating the analysis of adventitious respiratory sounds can help health professionals monitor patients suffering from respiratory conditions. The ICBHI Respiratory Sound Database, a benchmark dataset in respiratory sound analysis, has large and diverse data available publicly. Given its diversity in data, a stratified analysis by recording equipment, age, sex, body-mass index (BMI), and clinical diagnosis is proposed in this article. Regarding the experiments, three machine learning algorithms (Support Vector Machine - SVM, Random Undersampling Boosting - RUSBoost, and Convolutional Neural Network - CNN) were employed in three tasks: 2-class crackles (crackles vs. others), 2-class wheezes (wheezes vs. others), and 3-class (crackles vs. wheezes vs. others). Overall, the CNNs achieved the best results in almost every category, except when the equipment was Littmann3200 or Meditron, where RUSBoost achieved better results. In terms of stratification categories, we observed significant differences in classification performance, namely in terms of equipment, where the Littmann3200 underperformed the other equipment analysed. In addition, in the 3-class task, the CNNs achieved better results in Male subjects than Female subjects. In terms of BMI, the CNN of the Overweight class in the 2-class wheeze task achieved worse results than the other two BMI classes (Normal and Obese).**

***Index Terms*—Adventitious Respiratory Sounds, Crackles, Wheezes, Deep Learning, Machine Learning, Stratification**

I. Introduction

The number of deaths caused by respiratory diseases such as chronic obstructive pulmonary disease (COPD), lower respiratory tract infections (LRTI) and trachea, bronchus, and lung cancers is increasing every year [1]. At the moment, physicians use stethoscopes to auscultate patients and try to identify any respiratory disorder. However, the results are not accurate (since they depend on the level of hearing accuracy and expertise of the physician), and continuous monitoring is impossible to provide [2].

Respiratory sounds are produced by airflow in the respiratory tract, during the inspiration and expiration phases, and can be recorded on the thorax, trachea or mouth [3]. Adventitious respiratory sounds (ARS) are abnormal respiratory sounds that are superimposed on breathing sounds. There are 2 types of ARS: continuous ARS, e.g., wheezes, lasting more than 80-100 ms, with a frequency range between 100-1000 Hz, and discontinuous ARS, e.g., crackles, lasting less than 20 ms, with a frequency range between 60-2000 Hz [4]. Depending on their duration, intensity and location on the respiratory cycle, ARS can assist in the diagnosis of respiratory conditions [4]. Figure 1 shows the spectrogram representation of the normal breath sounds, crackles and wheezes.

Over the last decades, several studies have addressed the classification of ARS events [5]–[9]. In these articles, various datasets were used, such as the Respiration Acoustics Laboratory Environment (RALE) dataset, which is used to teach students; private datasets; and the one used in this study [10], [11]. Even though several studies have tackled the classification of ARS events, a stratified analysis is missing in those works.

Hence, the main contribution of this work is the stratified analysis of ARS event classification in the RSD. To the best of our knowledge, it is the first stratified analysis of this task.

This work is structured as follows: in section II, an overview of the dataset, a description of the stratification process and a summary of the models used are provided; in section III, the obtained results are analysed; and the conclusion is presented in section IV.

II. Materials & Methods

A. Dataset

The ICBHI Respiratory Sound Database (RSD)[1] [10], [11] was created by two independent teams from two countries (Portugal and Greece). It contains 5.5 hours of respiratory sounds (920 annotated audio samples) from 126 patients, with a total of 8877 annotated crackles and 1898 annotated wheezes. To standardise the comparison between results among other studies that use this dataset, the authors partitioned it into training and test sets.

The respiratory sounds were collected from patients of different ages and with various respiratory conditions, such as upper and lower respiratory tract infection (URTI and LRTI), COPD, asthma, pneumonia, bronchiectasis, and bronchiolitis. These sounds were recorded with stethoscopes from different manufacturers, i.e. WelchAllyn Meditron (Meditron), 3M Littmann Classic II SE (LittC2SE), 3M Littmann 3200

[1]https://bhichallenge.med.auth.gr/ICBHI_2017_Challenge

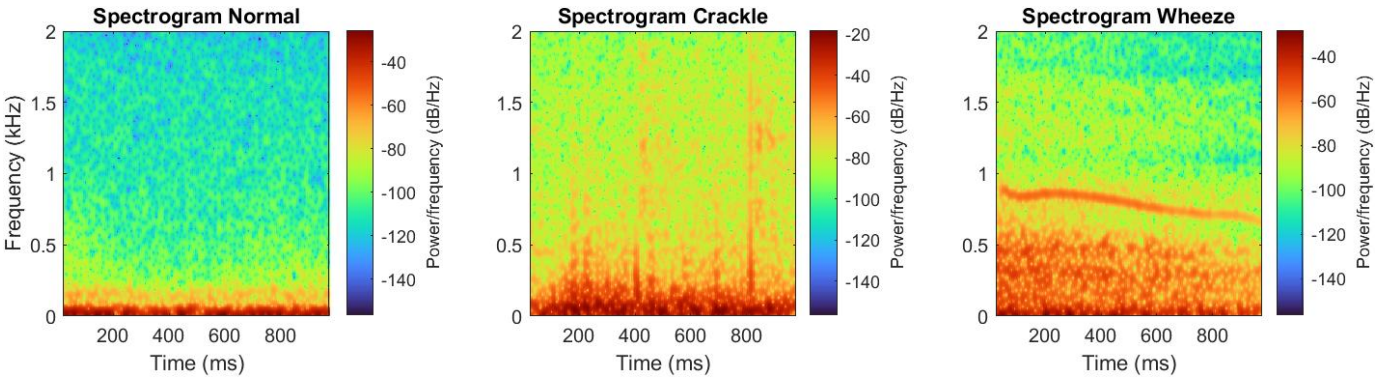

Fig. 1. Spectrogram representation of the normal breath, crackles and wheezes sounds

(Litt3200)), as well as microphones (AKG C417L) with different sampling rates and background noises.

Randomly generated events were also added to increase realism to the challenge for the models [9] with a variable duration following two Burr distributions according to the type of ARS: events shorter than 100 ms as the negative class against which the crackle events would be classified, or events between 100 ms and 2 s as the negative class against which the wheeze events would be classified.

B. RSD Stratification

As the sounds in this dataset were collected from subjects of different ages, with various diseases, body-mass index (BMI), and sex and recorded with different types of equipment, a stratified analysis of the results of the models was performed to understand in more detail the behaviour of the models for each subpopulation.

In the categorization by age, all patients under 18 were considered children, the others were considered adults. Regarding the BMI categories, they were defined according to the World Health Organization guidelines [12] and since there were only three underweight patients, they were included in the normal weight category. Concerning the diagnosis category, patients with COPD, asthma, and bronchiectasis were considered chronic; patients with LRTI, URTI, bronchiolitis, or pneumonia were considered non-chronic, whereas participants with no diseases were considered healthy. Table I shows the number of events per class in each category in the data.

From Table I we can conclude that:

- The test set contains no audio files recorded using the LittC2SE stethoscope
- Unknown Age subjects do not have any files in the training set; a similar situation is observed for patients with Unknown Sex
- For Healthy subjects there are no cases with annotated wheezes and only 9 cases with annotated crackles identified in the test set
- The number of events in each stratification category is not balanced between classes since the goal of the original splitting was to guarantee a 60/40 partition of the data according to the number of respiratory cycles (per class), number of patients and number of files

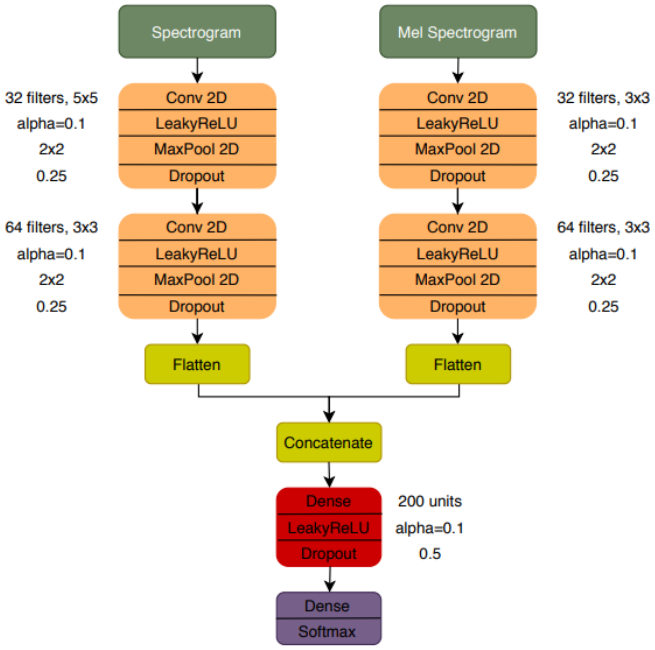

Fig. 2. Dual Input CNN architecture.

We used three machine learning algorithms to classify the test set: Random Undersampling Boosted Trees (RUSBoost), Support Vector Machine with radial basis function (SVM-rbf) and Convolutional Neural Network (CNN) (Figure 2). The same training set was used for all models. Following [9], handcrafted features were extracted for all the models except the CNN. Namely, the following set of features was extracted: spectral (e.g., centroid, spread, skewness, kurtosis), Mel-frequency cepstral coefficients (MFCCs), and melodic features (e.g., pitch, inharmonicity, voicing). Feature ranking using the minimum redundancy maximum relevance (MRMR) algorithm was employed to select the first 100 features.

Given the above stratification and the models already trained, the test data were divided into the aforementioned cat-

TABLE I
DISTRIBUTION OF EVENTS IN THE TRAIN AND TEST SETS PER EQUIPMENT, AGE (RANGE, MEAN±STANDARD DEVIATION), SEX, BMI (RANGE), AND DIAGNOSIS [F: FILES, C: ANNOTATED CRACKLES, W: ANNOTATED WHEEZES, OC: ANNOTATED OTHER CRACKLES, OW: ANNOTATED OTHER WHEEZES]

Category	Elements	F Train	F Test	C Train	C Test	W Train	W Test	OC Train	OC Test	OW Train	OW Test
Equipment	AKG C417L	361	285	5387	2682	749	482	1170	1115	274	257
	Littmann 3200	5	5	14	55	28	162	26	297	7	65
	Meditron	87	41	273	144	174	81	884	268	198	66
	Littmann C2SE	86	0	322	0	222	0	398	0	96	0
Age (years)	Adults (19-93, 67.7±11.6)	493	345	5927	2810	1110	676	2204	1441	510	329
	Children (0-18, 4.9.±4.6)	46	30	69	52	63	36	274	180	65	48
	Unknown	0	6	0	19	0	13	0	59	0	11
Sex	Male	272	319	2189	2741	728	682	1510	1256	348	292
	Female	267	56	3807	121	445	30	968	365	227	85
	Unknown	0	6	0	19	0	13	0	59	0	11
BMI	Normal (below 25)	235	91	3913	925	567	115	721	360	179	84
	Overweight (25-29.9)	171	189	1216	908	460	437	910	862	207	190
	Obese (above 30)	84	65	784	977	76	124	501	219	107	55
	Unknown	49	36	83	71	70	49	346	239	82	59
Diagnosis	Chronic (64 COPD, 7 Bronchiectasis, 1 Asthma)	459	351	5899	2829	1085	689	1966	1500	455	340
	Non-Chronic (14 URTI, 2 LRTI, 6 Bronchiolitis, 6 Pneumonia)	62	13	77	43	85	36	385	52	90	15
	Healthy	18	17	20	9	3	0	127	128	30	33

egories, applied to two binary classification problems (crackles vs. others, and wheezes vs. others) and one 3-class problem (crackles vs. wheezes vs. others). Healthy subjects were ignored in the analysis of the results, as their test set did not have annotated wheezes and only contained 9 crackles. The files with missing data (6 files with no information regarding age or sex) were also discarded.

III. RESULTS & DISCUSSION

Four metrics were used to evaluate the performance of the classification models: accuracy, area under the curve (AUC), F1-Score, and Matthews Correlation Coefficient (MCC). For the binary classification tasks, we calculated the accuracy, AUC, MCC, and F1 of the positive class (crackles or wheezes) and macro-averaged F1 (F1 Macro), considering that the dataset is unbalanced. For the 3-class problem, we computed the accuracy, the F1 for each class, and F1 Macro.

Table II shows the results for three of the analysed models: SVMrbf with 100 selected features, RUSBoost with all the features, and CNN with dual input, i.e., with a combination of spectrogram and Mel spectrogram inputs. Henceforth, these models will be called SVM, Boost, and CNN, respectively.

In all the performed comparisons (discussed in the following paragraphs), statistical significance tests were conducted. When comparing the results for different subpopulations, unpaired tests were performed, namely the unpaired t-test (when the distributions are Gaussian) or the Wilcoxon rank sum test (when the distributions are non-Gaussian). When comparing the results of different algorithms in the same subpopulations, paired tests were performed, namely, the paired T-test (Gaussian distributions) or the Wilcoxon signed rank test (non-Gaussian distributions). In all cases, the Kolmogorov-Smirnov test was employed to test for Gaussianity and the threshold for statistical significance was set to $p < 0.01$. Unless otherwise stated, all the results compared in the paragraphs below are statistically significant.

Looking at Table II, we can observe that, for the three types of classification problems, the model that obtained the best results overall was the CNN, except for four cases: Children in wheezes classification, where the Boost performed better; Litt3200 stethoscope in wheezes classification, where the SVM outperformed the CNN; Meditron stethoscope in crackles classification, where the SVM model also outperformed the CNN; and finally, in the 3-class classification problem, where the SVM also performed better than the CNN in the Litt3200 and Meditron.

For the Equipment category, in wheezes classification, the results are quite similar between all 3 stethoscopes/microphones, with a slight advantage for the AKGC417L microphone (with F1 wheezes of 83.8% in the AKGC417L microphone, maximum F1 wheezes of 80.7% in the Litt3200, and F1 wheezes of 80.3% in the Meditron). In crackles classification, the results are higher for the AKGC417L microphone and Meditron, while in the Litt3200, the results are worse (F1 crackles of 90.9% in the AKGC417L microphone, F1 crackles of 80.4% in the Litt3200, and a maximum F1 crackles of 83.5% in the Meditron). In the 3-class problem, the AKGC417L microphone also achieved better results than the other two (F1 macro of 81.8% in the AKGC417L microphone, a maximum F1 macro of 50.4% in the Litt3200, and F1 macro of 67.4% in the Meditron). Overall, the AKGC417L microphone achieved better results, since this microphone is more sensitive, it has no filters, and its training and test sets are larger than the sets for the other types of equipment.

When we look at the Age of the subjects, in wheezes classification, Children' achieved better results than Adults (with F1 wheezes of 84.9% in Children using the Boost model and F1 wheezes of 77.0% in Adults using the CNN model). Even though the Boost achieved better results than the CNN in wheezes classification in Children, the difference was not statistically significant ($p > 0.01$). In crackles classification, the reverse occurs, with Adults outperforming Children (with F1 crackles of 90.5% in Adults and F1 crackles of 70.9% in Children). Regarding the 3-class problem, Adults once again outperformed Children (with F1 macro of 78.9% and F1 macro of 78.2%, respectively).

Regarding the Sex category, Male subjects achieved superior results in wheezes classification than Female subjects (F1

TABLE II
RESULTS (ACC: ACCURACY, C: CRACKLE, W: WHEEZE, O: OTHER, SVM: SVMRBF_100MRMR, BOOST: RUSBOOST_FULL, CNN: CNN_DUALINPUT, M: MACRO)

			2 class crackles			2 class wheezes				3 class		
			SVM	Boost	CNN	SVM	Boost	CNN		SVM	Boost	CNN
Equipment	AKGC417L	Acc	70.3 ± 0.5	68.7 ± 0.6	**86.4 ± 0.8**	64.2 ± 1.3	60.7 ± 1.3	**78.2 ± 0.9**	Acc	69.1 ± 0.6	68.4 ± 0.7	**84.5 ± 1.0**
		AUC	62.7 ± 2.5	64.0 ± 1.0	**79.5 ± 2.3**	59.3 ± 1.5	59.4 ± 1.5	**74.5 ± 2.2**	F1 C	80.5 ± 0.5	77.9 ± 0.8	**90.4 ± 0.8**
		F1 C/W	79.4 ± 0.8	77.3 ± 0.8	**90.9 ± 0.5**	73.3 ± 1.9	67.8 ± 2.3	**83.8 ± 1.0**	F1 W	72.2 ± 2.2	76.8 ± 0.7	**83.0 ± 1.5**
		F1 M	62.8 ± 3.3	63.5 ± 1.2	**81.9 ± 1.8**	59.3 ± 3.3	58.5 ± 2.6	**75.1 ± 2.2**	F1 O	38.8 ± 4.3	45.3 ± 1.3	**71.9 ± 1.9**
		MCC M	26.3 ± 3.6	27.2 ± 1.5	**66.2 ± 1.8**	19.4 ± 2.3	18.1 ± 2.8	**51.2 ± 2.2**	F1 M	63.8 ± 2.3	66.7 ± 0.9	**81.8 ± 1.4**
	Litt3200	Acc	77.3 ± 6.7	78.2 ± 1.9	**90.9 ± 1.3**	**71.5 ± 3.4**	65.2 ± 3.0	51.4 ± 1.8	Acc	55.7 ± 5.9	**68.9 ± 1.4**	65.1 ± 1.8
		AUC	71.2 ± 3.1	67.0 ± 3.4	**86.3 ± 0.7**	62.4 ± 1.7	**62.5 ± 2.2**	57.7 ± 3.3	F1 C	9.6 ± 4.8	2.0 ± 3.0	**16.1 ± 7.1**
		F1 C/W	46.7 ± 4.0	42.0 ± 4.8	**80.4 ± 2.1**	**80.7 ± 2.9**	73.7 ± 3.0	57.1 ± 2.0	F1 W	62.7 ± 6.3	**71.2 ± 2.1**	50.1 ± 6.0
		F1 M	66.0 ± 4.8	64.3 ± 3.1	**87.3 ± 1.5**	**63.1 ± 2.3**	60.9 ± 2.8	50.5 ± 2.7	F1 O	65.0 ± 5.4	78.1 ± 1.4	**78.2 ± 1.7**
		MCC M	36.1 ± 5.1	29.8 ± 5.9	**74.7 ± 3.1**	**27.2 ± 5.5**	23.4 ± 4.3	14.0 ± 6.0	F1 M	45.8 ± 5.5	**50.4 ± 2.2**	48.1 ± 4.9
	Meditron	Acc	86.9 ± 1.0	**87.7 ± 1.4**	85.2 ± 1.5	72.2 ± 3.0	77.4 ± 2.4	**79.6 ± 1.8**	Acc	70.7 ± 2.5	**73.6 ± 1.7**	71.6 ± 1.8
		AUC	**88.6 ± 0.9**	86.5 ± 1.8	86.3 ± 0.8	72.6 ± 2.9	78.4 ± 2.2	**80.4 ± 1.8**	F1 C	56.3 ± 2.5	**58.5 ± 5.8**	57.5 ± 3.8
		F1 C/W	**83.5 ± 1.2**	82.4 ± 2.2	81.3 ± 1.2	73.0 ± 3.8	77.0 ± 3.1	**80.3 ± 2.2**	F1 W	59.6 ± 3.3	59.5 ± 1.5	**61.5 ± 3.3**
		F1 M	86.4 ± 1.1	**86.5 ± 1.6**	84.5 ± 1.4	72.1 ± 3.3	77.4 ± 2.5	**79.6 ± 2.0**	F1 O	80.8 ± 2.9	**84.2 ± 1.3**	80.7 ± 1.2
		MCC M	**74.2 ± 1.9**	73.0 ± 3.2	70.4 ± 1.9	45.1 ± 5.5	57.1 ± 3.9	**60.5 ± 3.5**	F1 M	65.6 ± 2.9	**67.4 ± 2.9**	66.6 ± 2.8
Age	Adults	Acc	71.5 ± 0.9	69.9 ± 0.6	**86.7 ± 0.8**	65.6 ± 0.6	61.4 ± 1.6	**70.5 ± 1.0**	Acc	67.2 ± 0.9	68.1 ± 0.5	**81.9 ± 0.8**
		AUC	66.9 ± 2.3	67.4 ± 0.7	**82.5 ± 1.9**	59.8 ± 1.6	60.0 ± 1.3	**68.7 ± 1.9**	F1 C	77.8 ± 0.4	76.0 ± 0.7	**88.7 ± 0.8**
		F1 C/W	79.0 ± 0.8	76.8 ± 0.7	**90.5 ± 0.5**	75.0 ± 0.7	69.0 ± 2.4	**77.0 ± 1.6**	F1 W	68.7 ± 2.7	73.9 ± 0.9	**74.4 ± 1.9**
		F1 M	67.1 ± 2.8	67.1 ± 1.0	**84.2 ± 1.5**	60.0 ± 2.2	58.8 ± 2.3	**67.7 ± 2.3**	F1 O	45.2 ± 3.6	52.4 ± 0.9	**73.5 ± 1.1**
		MCC M	34.9 ± 3.2	34.3 ± 1.2	**70.0 ± 1.7**	20.2 ± 2.7	19.0 ± 2.3	**36.3 ± 2.9**	F1 M	63.9 ± 2.2	67.4 ± 0.8	**78.9 ± 1.3**
	Child	Acc	80.6 ± 1.5	**84.4 ± 1.8**	82.9 ± 2.1	81.2 ± 3.9	**86.9 ± 2.1**	86.0 ± 3.2	Acc	77.2 ± 4.2	78.2 ± 1.5	**80.6 ± 3.0**
		AUC	84.2 ± 1.5	80.8 ± 3.4	**86.4 ± 1.4**	81.6 ± 3.8	**86.9 ± 2.5**	86.1 ± 3.6	F1 C	65.3 ± 5.1	64.1 ± 4.0	**67.1 ± 2.3**
		F1 C/W	67.7 ± 2.0	68.0 ± 4.3	**70.9 ± 2.0**	79.4 ± 4.1	**84.9 ± 2.9**	84.0 ± 4.2	F1 W	72.7 ± 4.9	71.0 ± 2.0	**82.8 ± 3.2**
		F1 M	76.9 ± 1.6	78.9 ± 2.8	**79.4 ± 1.9**	81.1 ± 4.0	**86.7 ± 2.3**	85.7 ± 3.6	F1 O	82.4 ± 3.8	83.7 ± 1.4	**84.8 ± 3.0**
		MCC M	58.9 ± 2.8	58.2 ± 5.6	**63.4 ± 2.4**	62.8 ± 7.6	**73.6 ± 4.5**	72.3 ± 6.7	F1 M	73.5 ± 4.6	72.9 ± 2.5	**78.2 ± 2.8**
Sex	Male	Acc	72.0 ± 0.8	70.4 ± 0.7	**87.3 ± 0.7**	68.7 ± 1.0	64.4 ± 1.9	**72.4 ± 1.3**	Acc	69.2 ± 0.7	69.4 ± 0.7	**82.7 ± 1.0**
		AUC	66.6 ± 1.8	67.8 ± 0.8	**82.4 ± 1.9**	62.2 ± 1.6	62.8 ± 1.1	**70.4 ± 1.9**	F1 C	79.4 ± 0.4	77.1 ± 0.8	**89.4 ± 0.9**
		F1 C/W	79.8 ± 1.0	77.6 ± 0.8	**91.2 ± 0.4**	77.8 ± 1.0	72.4 ± 2.4	**79.2 ± 1.8**	F1 W	71.6 ± 2.7	**76.9 ± 0.8**	76.8 ± 2.1
		F1 M	66.8 ± 2.3	67.0 ± 1.0	**84.2 ± 1.4**	62.2 ± 2.2	61.0 ± 2.1	**68.9 ± 2.2**	F1 O	46.1 ± 2.9	51.9 ± 0.9	**72.7 ± 1.1**
		MCC M	34.0 ± 2.2	34.4 ± 1.4	**69.9 ± 1.6**	24.7 ± 2.7	23.9 ± 2.0	**38.9 ± 2.6**	F1 M	65.7 ± 2.0	68.6 ± 0.8	**79.6 ± 1.4**
	Female	Acc	72.0 ± 5.4	72.9 ± 1.7	**80.3 ± 3.1**	50.8 ± 4.2	54.6 ± 3.1	**66.0 ± 3.4**	Acc	55.7 ± 4.5	63.1 ± 1.2	**74.6 ± 2.3**
		AUC	78.3 ± 3.2	76.3 ± 2.3	**84.1 ± 1.1**	48.0 ± 5.7	46.4 ± 3.3	**62.3 ± 1.7**	F1 C	49.2 ± 2.4	53.8 ± 2.4	**64.2 ± 1.9**
		F1 C/W	62.0 ± 3.7	60.4 ± 2.3	**70.0 ± 2.7**	30.8 ± 6.6	24.9 ± 5.6	**45.5 ± 2.3**	F1 W	29.0 ± 4.9	34.4 ± 3.0	**41.7 ± 4.7**
		F1 M	69.9 ± 4.8	69.9 ± 1.9	**77.7 ± 2.8**	46.3 ± 5.2	46.1 ± 4.4	**60.3 ± 2.9**	F1 O	62.9 ± 5.6	72.1 ± 1.3	**81.9 ± 2.1**
		MCC M	49.3 ± 5.3	45.8 ± 3.8	**60.5 ± 3.0**	-3.5 ± 10.1	-6.7 ± 6.2	**22.8 ± 3.1**	F1 M	47.0 ± 4.3	53.4 ± 2.2	**62.6 ± 2.9**
BMI	Normal	Acc	79.0 ± 1.7	80.0 ± 1.4	**88.1 ± 1.3**	63.0 ± 2.7	63.1 ± 1.7	**79.4 ± 2.5**	Acc	75.0 ± 0.8	76.9 ± 1.0	**87.2 ± 0.9**
		AUC	65.9 ± 3.7	70.4 ± 1.5	**80.2 ± 3.0**	61.1 ± 2.8	62.4 ± 1.8	**78.1 ± 2.6**	F1 C	85.9 ± 0.3	86.7 ± 0.9	**92.6 ± 0.6**
		F1 C/W	86.8 ± 0.8	86.9 ± 1.1	**92.3 ± 0.8**	69.5 ± 3.0	67.6 ± 1.9	**82.9 ± 2.2**	F1 W	69.2 ± 2.8	72.2 ± 1.0	**82.2 ± 1.7**
		F1 M	67.7 ± 4.6	72.3 ± 1.8	**83.3 ± 2.6**	60.9 ± 4.4	62.3 ± 2.3	**78.4 ± 2.9**	F1 O	41.5 ± 5.0	51.4 ± 2.2	**75.1 ± 2.4**
		MCC M	41.8 ± 6.4	46.6 ± 3.4	**69.7 ± 3.4**	23.1 ± 5.5	24.8 ± 3.6	**57.6 ± 5.1**	F1 M	65.5 ± 2.7	70.1 ± 1.4	**83.3 ± 1.6**
	Overweight	Acc	74.2 ± 2.8	75.1 ± 1.2	**85.5 ± 1.6**	**68.3 ± 2.0**	64.2 ± 2.4	67.9 ± 2.2	Acc	64.9 ± 1.9	70.3 ± 0.8	**78.0 ± 0.7**
		AUC	73.7 ± 2.9	74.7 ± 1.3	**85.3 ± 1.7**	59.5 ± 2.6	60.6 ± 2.2	**66.9 ± 1.2**	F1 C	72.0 ± 0.9	75.5 ± 0.6	**83.4 ± 0.9**
		F1 C/W	78.5 ± 1.4	78.7 ± 0.7	**87.0 ± 1.0**	**78.2 ± 1.4**	73.0 ± 2.5	75.0 ± 2.8	F1 W	69.9 ± 1.8	**74.0 ± 1.1**	70.4 ± 1.7
		F1 M	73.0 ± 3.7	74.4 ± 1.5	**85.3 ± 1.8**	59.9 ± 2.9	59.7 ± 2.9	**65.0 ± 2.2**	F1 O	52.2 ± 4.3	61.9 ± 1.7	**76.0 ± 1.3**
		MCC M	51.2 ± 4.0	52.0 ± 2.0	**72.0 ± 2.5**	20.5 ± 5.5	20.3 ± 4.3	**31.8 ± 2.2**	F1 M	64.7 ± 2.3	70.5 ± 1.1	**76.6 ± 1.3**
	Obese	Acc	59.5 ± 5.3	51.6 ± 3.3	**88.1 ± 2.5**	59.3 ± 3.3	50.0 ± 4.0	**75.0 ± 3.8**	Acc	62.9 ± 1.7	54.8 ± 2.2	**83.4 ± 2.6**
		AUC	62.5 ± 3.4	61.7 ± 2.2	**79.8 ± 0.8**	57.9 ± 1.1	54.7 ± 2.9	**72.6 ± 3.7**	F1 C	75.5 ± 1.7	62.7 ± 2.8	**90.7 ± 1.8**
		F1 C/W	69.7 ± 5.2	60.6 ± 3.8	**92.6 ± 1.8**	67.4 ± 4.1	53.6 ± 7.4	**81.3 ± 3.3**	F1 W	62.8 ± 9.2	74.8 ± 1.9	**76.6 ± 4.0**
		F1 M	53.8 ± 4.2	48.8 ± 2.8	**80.1 ± 2.4**	56.2 ± 3.0	49.3 ± 5.2	**71.7 ± 4.1**	F1 O	33.0 ± 2.2	36.0 ± 1.1	**61.8 ± 2.6**
		MCC M	19.6 ± 5.2	18.3 ± 3.4	**61.1 ± 4.9**	15.0 ± 2.6	9.0 ± 5.5	**43.9 ± 7.4**	F1 M	57.1 ± 4.4	57.8 ± 1.9	**76.4 ± 2.8**
Diagnosis	Chronic	Acc	71.9 ± 0.9	70.5 ± 0.6	**86.6 ± 0.9**	65.6 ± 0.7	61.9 ± 1.6	**70.9 ± 1.0**	Acc	67.4 ± 0.9	68.4 ± 0.5	**81.8 ± 0.9**
		AUC	67.8 ± 2.1	68.3 ± 0.7	**82.7 ± 1.9**	60.2 ± 1.6	60.7 ± 1.2	**69.3 ± 1.8**	F1 C	77.6 ± 0.4	75.9 ± 0.8	**88.3 ± 0.9**
		F1 C/W	79.1 ± 0.8	76.9 ± 0.7	**90.4 ± 0.5**	74.8 ± 0.7	69.2 ± 2.5	**77.3 ± 1.6**	F1 W	68.5 ± 2.7	73.6 ± 0.8	**74.2 ± 1.9**
		F1 M	68.1 ± 2.6	67.9 ± 0.9	**84.3 ± 1.5**	60.3 ± 2.1	59.5 ± 2.3	**68.3 ± 2.2**	F1 O	47.1 ± 3.4	54.1 ± 0.9	**73.9 ± 1.2**
		MCC M	36.7 ± 3.0	36.0 ± 1.2	**70.2 ± 1.8**	20.8 ± 2.7	20.4 ± 2.2	**37.4 ± 2.7**	F1 M	64.4 ± 2.2	67.9 ± 0.8	**78.8 ± 1.3**
	Non-Chronic	Acc	81.6 ± 2.5	80.5 ± 2.6	**85.3 ± 3.3**	82.4 ± 3.7	84.1 ± 3.6	**85.5 ± 4.4**	Acc	77.8 ± 3.5	78.2 ± 2.2	**84.7 ± 1.7**
		AUC	82.4 ± 2.5	80.7 ± 2.8	**86.0 ± 2.8**	80.9 ± 4.0	82.3 ± 2.5	**83.5 ± 5.1**	F1 C	78.7 ± 3.6	75.4 ± 4.1	**84.0 ± 1.7**
		F1 C/W	81.8 ± 2.4	79.2 ± 3.2	**85.4 ± 2.6**	87.0 ± 3.0	88.4 ± 3.0	**89.6 ± 3.2**	F1 W	83.4 ± 4.7	85.8 ± 2.0	**90.4 ± 3.0**
		F1 M	81.6 ± 2.6	80.4 ± 2.8	**85.3 ± 3.3**	79.5 ± 4.2	81.5 ± 3.5	**82.9 ± 5.1**	F1 O	73.6 ± 4.1	75.0 ± 2.5	**81.6 ± 2.6**
		MCC M	65.1 ± 4.9	61.3 ± 5.5	**72.4 ± 5.0**	60.1 ± 7.4	63.5 ± 6.2	**66.0 ± 10.1**	F1 M	78.6 ± 4.1	78.7 ± 2.9	**85.3 ± 2.4**

wheezes of 79.2% in Males and F1 wheezes of 45.5% in Females). In the classification of crackles, the same occurs, Male subjects also achieved superior results than Female Subjects (F1 crackles of 91.2% in Males and F1 crackles of 70.0% in Females). As for the 3-class classification problem, even though there is still an advantage for the Male subjects, the difference is lower (F1 macro of 79.6% in Males and F1 macro of 62.6% in Females). Even though Female subjects have a large number of annotated crackles in the training set, these differences can be explained by the unbalanced data in both crackles and wheezes in Female subjects between train and test sets. Overall, the results on the classification of crackles were superior to the results on the classification of wheezes, as there are more annotated crackles in both training and test sets than wheezes.

Regarding the BMI category, in the classification of wheezes, Obese and Normal BMI subjects achieved better results than Overweight BMI subjects (with F1 wheezes of 81.3% in Obese BMI subjects, F1 wheezes of 82.9% in Normal BMI subjects, and maximum F1 wheezes of 78.2% in Overweight BMI subjects). In the crackles classification, Obese and Normal BMI subjects achieved better results than

Overweight BMI subjects (with F1 crackles of 92.6% in Obese BMI subjects, F1 crackles of 92.3% in Normal BMI subjects, and F1 crackles of 87.0% in Overweight BMI subjects - except for the Obese and Normal CNNs where $p > 0.01$). In the 3-class problem, Obese and Overweight BMI subjects achieved similar results (in terms of F1 macro, with 76.4% and 76.6%, respectively), while the Normal BMI subjects achieved superior results (with F1 macro of 83.3% - except for the Obese and Normal CNNs, where $p > 0.01$). Overall, crackles classification achieved better results than wheezes classification, maybe due to having more annotated crackles than wheezes in the training set, which benefits the CNN model.

In the Diagnosis category, the subjects with a Non-Chronic diagnosis achieved better results in the wheezes classification than the subjects with a Chronic diagnosis (F1 wheezes of 89.5% and F1 wheezes of 77.3%, respectively). In the classification of crackles, the reverse occurs and the subjects with Chronic diagnosis surpassed the subjects with Non-Chronic diagnosis (F1 crackles of 90.4% and F1 crackles of 85.4%, respectively). As for the 3-class classification problem, the same as the wheezes classification happened: the subjects with Non-Chronic diagnosis once again achieved better results than those with Chronic diagnosis (F1 macro of 85.3% and F1 macro of 78.8%, respectively). Regarding wheezes and crackles classification, that difference may be explained by the fact that the AKGC417L microphone has more sensibility than any other equipment, and in the training set of the subjects with Non-Chronic diagnosis, there are only files where the equipment used were the LittC2SE and Meditron (less sensibility in general), while in the training set of the subjects with Chronic diagnosis, most of the files were recorded using the AKGC417L microphone.

IV. Conclusion

We have presented a stratified analysis of ARS event classification in the RSD. As discussed, we have observed several significant differences in the analysed stratification categories such as the fact that sounds recorded using the AKGC417L microphone achieve better results, as well as in Male subjects and Normal BMI subjects. Regarding the employed machine learning models, the CNNs attained in general the best results, except in some situations where the data were scarcer; in those cases, the SVM or the Boost models achieved better results.

Since the splitting currently available was done according to the number of respiratory cycles and the number of events on each set, for a possible future work, a new splitting can be created based on the demographic information to try to balance the number of events between categories and sets to achieve a more balanced partition of the data in both sets.

If a new partitioning is proposed, probably some models can achieve better results, since the data are not always well-balanced, making it a harder task for the models to learn the characteristics of the less represented stratification categories. The best model overall here presented (CNN) requires a large amount of data to achieve better results. The other two models, SVM and Boost, sometimes can achieve better results with fewer data but require a feature extraction stage.

Further studies on this topic require a more detailed analysis of Adults and Children results in all 3 classification problems.

Acknowledgements

This research is partially supported by Fundação para a Ciência e a Tecnologia (FCT) Ph.D. scholarships (2020.04927.BD and SFRH/BD/135686/2018), by the Horizon 2020 Framework Programme of the European Union project WELMO (under grant agreement number 825572) and project CISUC - UID/CEC/00326/2020 with funds from the European Social Fund, through the Regional Operational Program Centro 2020.

References

[1] WHO, "The top 10 causes of death," https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2020.

[2] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of respiratory sounds: state of the art," *Clinical medicine. Circulatory, respiratory and pulmonary medicine*, vol. 2, pp. CCRPM–S530, 2008.

[3] A. Sovijarvi, F. Dalmasso, J. Vanderschoot, L. Malmberg, G. Righini, and S. Stoneman, "Definition of terms for applications of respiratory sounds," *European Respiratory Review*, vol. 10, no. 77, pp. 597–610, 2000.

[4] A. Marques and A. Oliveira, "Normal versus adventitious respiratory sounds," in *Breath Sounds*. Springer, 2018, pp. 181–206.

[5] L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, I. Chouvarda, N. Maglaveras, V. Tsara, C. Teixeira, P. Carvalho, J. Henriques *et al.*, "Detection of wheezes using their signature in the spectrogram space and musical features," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 5581–5584.

[6] C. Pinho, A. Oliveira, C. Jácome, J. Rodrigues, and A. Marques, "Automatic crackle detection algorithm based on fractal dimension and box filtering," *Procedia Computer Science*, vol. 64, pp. 705–712, 2015.

[7] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artificial intelligence in medicine*, vol. 88, pp. 58–69, 2018.

[8] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–9, 2017.

[9] B. M. Rocha, D. Pessoa, A. Marques, P. Carvalho, and R. P. Paiva, "Automatic classification of adventitious respiratory sounds: A (un) solved problem?" *Sensors*, vol. 21, no. 1, p. 57, 2020.

[10] B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques *et al.*, "A respiratory sound database for the development of automated classification," in *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 33–37.

[11] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, p. 035001, 2019.

[12] WHO, "A healthy lifestyle - who recommendations," https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle—who-recommendations, 2010.