Detecting Cardiac Pathologies from Annotated Auscultations

Pedro Ferreira
CRACS-INESC TEC, Porto, Portugal
pedroferreira@dcc.fc.up.pt

Daniel Pereira
CINTESIS, Porto, Portugal
danielclaudiopereira@gmail.com

Felipe Mourato and Sandra Mattos
Real Hospital Português, Recife, Pernambuco, Brazil
felipe.a.mourato@gmail.com, ssmattos@cardiol.br

Ricardo Cruz-Correia
CINTESIS, Porto, Portugal
Faculdade de Medicina, University of Porto, Porto, Portugal
rcorreia@med.up.pt

Miguel Coimbra and Inês Dutra
IT-Porto and CRACS-INESC TEC
Department of Computer Science, Faculdade de Ciências
University of Porto, Porto, Portugal
mcoimbra@dcc.fc.up.pt, ines@dcc.fc.up.pt

Abstract

The DigiScope project aims at developing a digitally enhanced stethoscope capable of using state of the art technology in order to help physicians in their daily medical routine. One of the main tasks of DigiScope is to build a repository of auscultations (sound and medical related data). In this work, we present a preliminary analysis and study of the first auscultations performed on children of a Brazilian hospital. Results indicate that classifiers can be obtained that distinguish reasonably well patients with cardiac pathologies from those that do not have pathologies.

1. Introduction

Since the invention of the first stethoscope, by the French physician René Laënnec in 1816, the auscultation of the heart and lungs, using a stethoscope, is often conducted on patients thought to have cardiac or pulmonary disease before recommending additional diagnostic procedures, treatment, or no further action. Because this process is simple, cheap, and quick to detect diseases, the stethoscope still maintains a key position in medicine in the modern era. Auscultation is a subjective process that depends on the experience and hearing capability of the individual, which may lead to a large variability in findings. Auscultation, however, is a hard skill to master. Physically, a stethoscope covers a broad sound spectrum and the average frequency depends on the point of auscultation. It requires significant practice for a human ear to distinguish between them.

Digital stethoscopes are medical devices that can collect, store and sometimes transmit acoustic auscultation signals in a digital format. These can then be replayed, sent to a colleague for a second opinion, studied in detail after an auscultation, used for training or, as we envision it, can be used as a cheap powerful tool for screening cardiac pathologies. DigiScope [16, 15] is one of the enhanced stethoscopes that aims at using state of the art technology in order to help physicians in their daily medical routine. DigiScope aims to be a prototype of a digitally enhanced stethoscope, capable of automatically extracting clinical features from the collected data, as well as providing a clinical second opinion on specific heart pathologies. Several other electronically enhanced and digital stethoscopes have been developed and described in the literature [10, 18, 2], including models such as the iStethoscope Pro (application developed for the iPhone) or products such as the stethoscope developed by Zargis Medical Corp. While all provide some interesting ideas for digitally enhanced stethoscopes, their primary focus is on the technological development of the apparatus itself. A published review [5] argues that the key to robust solutions lies in a stronger interaction with the clinical community, both for understanding the needs of cardiologists and for robust clinical validation of not only the methods but also the final prototype.

In this work, we use data collected by the DigiScope Data Collector, which is being used in two hospitals, one in Portugal and another one in Brazil [16]. The data collected in Portugal is from adults while the data collected from Brazil is from children. In this work, we concentrate on analyzing the Brazilian children data. We discuss DigiScope's data model, give some statistics about the data being collected in Brazil, and report preliminary results on the correlation among the data attributes that may lead to valuable recommendations to the doctors and their patients. This work was approved by the Ethical Review Board of the Real Hospital Português.

This work has two goals:

1. automatically learn classifiers that distinguish between normal patients and those with any cardiac pathology. Our classifiers rely only on the cardiologist provided annotation and not on the raw sound data itself.

2. automatically extract new and relevant knowledge from the dataset.

Very few works in the literature report on prediction of heart diseases using machine learning techniques. The University of California at Irvine (UCI) repository (`http://archive.ics.uci.edu/ml/`) has some datasets related to cardiology. The one most related to our dataset is the "Heart Disease". According to the UCI website, this database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by machine learning researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is an integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). The results obtained with an instance-base learning algorithm (IB1) report an accuracy of 75.7% (±0.8). Another experiment with the same dataset, also to diagnose disease, used a neural network algorithm and reported an accuracy of 87.5% [12]. More recently, the same dataset was used in experiments that use Radial Basis Function Networks, that report an accuracy of 84% [14]. A more recent work [19] had as objective to model the detection of heart failure more than 6 months before the actual date of clinical diagnosis using machine learning techniques to Electronic Health Record (EHR) data. They compared the performances of logistic regression, SVM and Boosting along with various variable selection methods in heart failure prediction. They reported a value of 0.77 for the Area Under the ROC (AUROC) for the best classifier. With our dataset and performing an exhaustive search for the best classifier, we obtained an accuracy of 90.5% and AUROC of 0.83 on unseen cases.

2. The DigiScope Data Model

Given the significant amount of data that DigiScope can capture in a day, it is important to think on how to describe the data in a meaningful way. Medical sub-specialties often use a specific standardized lexicon to describe findings in a patient (for example, in the area of breast cancer, the terminology is based on the BIRADS - Breast Imaging Reporting and Data System - lexicon [13]).

We defined metadata that is interesting for pathology screening, and that is feasible to be annotated in this context. Contributions to this definition also came from the HL7 standard [11] and openEHR publicly available archetypes [3]. We emphasized collecting data that could be helpful for machine learning research for cardiac pathology detection.

In particular, we focused on the following types of information:

- The exam history of a patient. This information can be useful for detecting temporal relationships between diseases.

- Attributes that can be used to distinguish between normal and abnormal cases. In particular, the second heart sound (S2) is very important for this task.

- Attributes that can differentiate between multiple diseases for the same patient (multi-labeling [8]).

- Relationships between patients exams.

- Relationships between medications and a patient's health status.

The resulting data model defines both physician annotated attributes as well as other attributes that will result from future work involving signal processing (e.g. A2_intensity).

3. Materials and Methods

3.1. Data

The DigiScope application has been used for four months at two hospitals. During this period, June to September 2011, around 200 patients (children) were auscultated at the Real Hospital Português, in Pernambuco, Brazil. Each auscultation was recorded, and a XML file containing the physician annotated data was associated with each sound file. In this work, we focus only on the annotated nominal data available in the XML files. For each patient, we have 40 attributes, but we only use the variables that were annotated for the majority of the annotations (i.e., those that have few missing values). These attributes are shown in Table 1 along with their possible values. For each numerical attribute, the table reports its average and standard deviation. For each categorical attribute, the table reports the raw number of occurrences for each value along with the percentage. We also show the number of occurrences according to the CardiacPathology values (last two columns of Table 1). The average age for the children in the study is 7.3 years.

We worked with the 169 cases that had almost complete annotations and were not missing the class label CardiacPathology. 40 cases were labeled as having a CardiacPathology while the remaining 129 were annotated as being normal. Besides being auscultated, every patient also had an ecocardiogram. The final label for CardiacPathology was determined during auscultation and possibly modified after the ecocardiogram.

In Table 1, we highlight in bold 14 attributes that have multiple values and are rarely missing.

3.2. Methodology

Besides the original attributes, we derived the following additional ones: body mass index (BMI, calculated as the person's weight in kilograms divided by the square of the height in meters), and categorized versions of SystolicSystemicPressure-mmHg and DiastolicSystemicPressure-mmHg, according to the children's sex, age and respective height percentiles. Table 2 shows statistics about the derived attributes.

For training, we omitted the attributes related to weight, height, and age, since these are captured by the derived attributes. Attributes like S1Status, PressurePosition, S3Exist and S4Exist are not predictive and therefore are omitted in our experiments. StatusForm was not used either because it only indicates if the patient information is complete. We used the categorized versions of the attributes (the ones with suffix "_def") in order to ensure that the discretized versions are medically relevant. No weighing was used to compensate the class size imbalance.

Our main goals with this work are: (1) to distinguish between abnormality (patients that have a cardiopathy) and normality (patients that do not have any cardiopathy), and (2) explore the data for new relevant knowledge.

For the first goal we focused on learning a classifier to predict CardiacPathology using the other attributes. According to the specialists, the attribute Murmur is considered to be very important to predict a cardiac pathology. In order to study the influence of the attribute Murmur, we also performed experiments removing the attribute Murmur from our dataset. These experiments were performed using the WEKA tool [9]. We compared several classifiers with a variety of parameters: ZeroR (reference), PART, OneR, DTNB, SimpleCart, RandomForest, NBTree, J48, DecisionStump, SMO, NaiveBayes, BayesNet(TAN) and 5 more meta-learning algorithms: AdaBoostM1, Bagging, Dagging, Grading, Stacking and Vote (with 3 different kinds of combination rules: Majority Voting, Average of Probabilities and Maximum Probability). Most meta-learning algorithms use DecisionStump as the default base classifier, with the exception of Bagging and Dagging, that use REPTree and SMO, respectively. For all of these experiments, we used stratified 10-fold cross-validation with 10 iterations, with tuning sets. We compared the results using a two-tailed corrected paired t-test, with p=0.05. The best models found with the internal tuning were then applied to the test sets. We report the average number of Correctly Classified Instances (CCI), sensitivity and specificity, calculated according to what is discussed by Forman and Scholz [7].

For the second goal, we focused on exploring the data trying to discover new knowledge. In WEKA, we used association rule mining, feature selection and used classifiers that produce interpretable results (e.g., J48). When trying to find relations among attributes, in the WEKA system, we tested all possible combinations of "Attribute Evaluator" and "Search Method". The most frequent ranked attributes were used again to further filter and select the best attributes. Interestingly, the attribute Murmur was highly ranked. This step was done using 10-fold cross-validation and in each run, we selected the attributes that were most frequently selected in the 10 folds. We also used Aleph [17], an inductive logic programming system that produces human-readable first-order rules. Experiments with Aleph were performed over the entire dataset.

4. Results

The best models found in the internal 10-fold cross-validation (tuning) were applied to the test sets. When predicting CardiacPathology, with the attribute Murmur, in seven folds, the best classifier was Grading, and in the remaining folds, SMO. The overall performance on the tuning and test sets are shown in Table 3. The results on the tuning sets are statistically better than ZeroR, which we use as the reference classifier.

It is interesting to note that results in the literature using meta-learning algorithms (more specifically, AdaBoostM1), report that boosting yields minor to modest performance on the classification of heart diseases [1]. In our work, the results of AdaBoostM1 using DecisionStump as the base clas-

Attribute	Range/	Percent/	Stdev/	Missing	Pathology	
	Values	Average	Qty		Yes	No
Age	0–19	7.31	±4.17	2	6.80	7.46
Height-cm	49–183	122.22	±26.85	4	116.03	124.02
Weight-kg	3–97	29.36	±16.87	1	25.98	30.42
SystolicSystemicPressure-mmHg	90–145	100.79	±7.51	36	104.78	99.95
DiastolicSystemicPressure-mmHg	50–90	61.44	±5.70	36	63.48	61.01
Sex	Female	35.50	60	0	18	42
	Male	64.50	109		22	87
AuscultationPosition	Sit	12.65	21	3	3	18
	Lying downwards	87.35	145		36	109
SystemicPressureMethod	Manometry	96.38	133	31	23	110
	Unknown	3.62	5		2	3
Murmur	No	80.47	136	0	11	125
	Systolic	18.93	32		28	4
	Diastolic	0.60	1		1	0
S2Status	Normal	98.22	166	0	37	129
	Abnormal	1.78	3		3	0
IfAbnormal	NA	98.22	166	0	37	129
	Single	1.18	2		2	0
	Fixed split	0.59	1		1	0
PulmonaryComponent	Normal	99.41	168	0	39	129
	Hypophonetic	0	0		0	0
	Hyperphonetic	0.59	1		1	0
CardiacPathology	Yes	23.67	40	0		
	No	76.33	129			
CardiacPathologyType	PulmonaryHypertension (PH)	0	0		0	0
	ArterialHypertension (AH)	1.18	2		2	0
	ValvularAorticDisease (VAD)	0.59	1		1	0
	IntraventricularCommunication (IC)	3.55	6		6	0
	OtherCardiacPathology (OCP)	13.61	23		23	0
	AH and VAD	1.18	2		2	0
	IC and OCP	2.37	4		4	0
	VAD and OCP	0.59	1		1	0
	None	76.92	130		1	129
S1Status	Normal	100	169	0	40	129
PressurePosition	Sit	99.41	168	1	40	128
S3Exist	No	100	169	0	40	129
S4Exist	No	100	169	0	40	129
StatusForm	Complete	56.80	96	0	25	71
	Incomplete	43.20	73		15	58

Table 1. Attributes associated with the auscultations and their values

Attribute	Values	Qty	Percentage
BMI	Normal	88	53.33
	Overweight	24	14.55
	Underweight	23	13.94
	Obese	30	18.18
SystolicSystemicPressure-mmHgDisc	High	2	1.53
	Normal	129	98.47
DiastolicSystemicPressure-mmHgDisc	High	2	1.53
	Normal	129	98.47

Table 2. Extra derived attributes

Metrics	Tuning	Test
CCI (%)	91.56	90.53
Sensitivity	0.72	0.70
Specificity	0.98	0.97

Table 3. Results with Murmur

sifier (the same as the one reported in the literature) produces results that are statistically close to the ones reported in Table 3.

Metrics	Tuning	Test
CCI (%)	79.37	79.29
Sensitivity	0.28	0.28
Specificity	0.95	0.95

Table 4. Results without Murmur

We repeated the same experiment, removing the attribute Murmur from our dataset. For this experiment, the best results were always obtained with a Naive Bayesian network classifier in all folds. The accuracy (CCI) is statistically worse than the accuracy achieved by the classifier that uses Murmur, and not statistically different from the reference classifier ZeroR. Not every murmur is related to a pathology, but in our dataset, the attribute Murmur seems to be very important to predict it. These results suggest that, if possible, the attribute Murmur should always be annotated. If not, with only the attributes we have, our classifier would have a poor performance. One alternative would be to extract the attribute Murmur from the wave sound through signal processing, but this is an open research issue and could be a challenging task. Concluding, the attribute Murmur needs to be annotated and, according to these experiments, is crucial to obtain a classifier that can predict cardiac pathology with good sensitivity and specificity.

When doing feature selection, the relevant attributes chosen by all algorithms were: BMI_def, Age_def, Sex, SystolicSystemicPressure_def, DiastolicSystemicPressure_def, Hypertension, Murmur, Grading, S2Status, IfAbnormal, PulmonaryComponent, CardiacPathology and CardiacPathologyType, which coincide with the attributes we used for classification.

The HotSpot algorithm correlated the CardiacPathology attribute (class variable) with BMI (these results were also obtained using a reduced set of patients [6], where the relationship was for height and weight, which are the basic attributes used for computing the BMI). Removing the attribute Murmur maintains this relationship, but BMI is replaced by Sex. Similarly, when trying to discover the best attributes to predict the class variable (CardiacPathology), all algorithms select CardiacPathologyType and Murmur. In the absence of either or both of these attributes, S2Status, IfAbnormal and SystolicSystemicPressure_def are selected. These are all clinically relevant variables related to cardiopathies.

When learning first-order rules, we found an intriguing rule shown in Figure 1. This rule says that if a child has a systolic murmur and a high BMI (Body Mass Index), it is very likely that the child has a pathology. BMI in children is rarely related to cardiac pathologies according to most specialists. This rule may open a new stream of research into this relationship in clinical practice. This rule does not contain the sex and age of the child and this omission needs to be further investigated. The rule holds for 6 out of the 40 patients with a cardiac pathology, and does not apply to any healthy patient (129). This finding has been discussed in other work. Daniels *et al.* [4] mention that classic signs and symptoms of heart failure are not always present in obese patients, whose body habitus may mask signs of edema and may muffle the heart and lung sounds during auscultation. In their study, patients with high BMI were less likely to have documented murmurs. In our dataset, the opposite seems to be true, since we had 6 patients with annotated murmurs and a high BMI.

```
'CardiacPathology'(A) if
        bmi(A,obese) and
        'Murmur'(A,'Systolic').
```

Figure 1. Example of Aleph rule.

5 Conclusions and Future Work

In this work, we studied the first data collected in the context of the DigiScope project. An application is being routinely used in 2 hospitals, and already recorded auscultations and medical information for 200 patients. We studied the data annotated for the patients with the intention of producing classifiers to predict cardiac pathologies. Our results indicate that the attribute Murmur is relevant to predict a cardiac pathology. In the absence of this attribute, a classifier has a very poor sensitivity. When learning rules, we uncovered an intriguing one that relates BMI with Murmur and CardiacPathology. Usually, BMI is not considered relevant to predict cardiac pathologies in children.

When learning classifiers, results show that we can train a classifier close to a specialist with a performance of 90.5%, sensitivity of 0.70 and specificity of 0.97 to predict pathologies on unseen cases. The area under the ROC curve was 0.83. We consider these results very promising and have been working to acquire more data to improve the classification task.

The work developed in the context of DigiScope has been paving the way to a new vision of cardiology. One important stream to follow is that of education in cardiology. The sounds recorded by DigiScope are being used to train novices to identify basic and more challenging cardiopathies. A second and also challenging path is to process the signal and extract important indicators such as amplitude and distance between signals (A1, P1, A2, P2 etc). With these indicators, we believe it will be possible to learn more effective classifiers. Our final goal is to have an integrated tool, capable of online predicting the cardiac pathologies and recommending additional screening.

Ongoing work is being done to acquire data from more patients. We also have been working with data from adults and from pregnant women, which poses new challenges on the auscultations and annotations of the sounds.

Acknowledgments

This work has been supported by the projects DigiScope (PTDC/EIA-CCO/100844/2008), HORUS (PTDC/EIA-EIA/100897/2008) and ADE (PTDC/EIA-EIA/121686/2010) and by the Fundação para a Ciência e a Tecnologia (FCT/Portugal).

References

[1] P. C. Austin and D. S. Lee. Boosted classification trees result in minor to modest improvement in the accuracy in classifying cardiovascular outcomes compared to conventional classification trees. *Am J Cardiovasc Dis.*, 1:1–15, April 2011.

[2] M. Brusco and H. Nazeran. Development of an intelligent pda-based wearable digital phonocardiograph. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 2005.

[3] CKM. Openehr clinical knowledge manager, April 2011. visited April 2011.

[4] L. B. Daniels, P. Clopton, V. Bhalla, P. Krishnaswamy, R. M. Nowak, J. McCord, J. E. Hollander, P. Duc, T. Omland, A. B. Storrow, W. T. Abraham, A. H. Wu, P. G. Steg, A. Westheim, C. W. Knudsen, A. Perez, R. Kazanegra, H. C. Herrmann, P. A. McCullough, and A. S. Maisel. How obesity affects the cut-points for bnp in acute hf diagnosis. *American Heart Journal*, 151:999–1005, 2006.

[5] F. de Lima Hedayioglu, M. T. Coimbra, and S. da Silva Mattos. A survey of audio processing algorithms for digital stethoscopes. In *HEALTHINF*, pages 425–429, 2009.

[6] P. Ferreira, D. Pereira, I. Dutra, F. Hedayioglu, and M. Coimbra. The digiscope auscultation data: First explorations. In *17th Portuguese Conference on Pattern Recognition, RecPad 2011*, Oct 2011.

[7] G. Forman and M. Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57, Nov. 2010.

[8] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 195–200, New York, NY, USA, 2005. ACM.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11:263–286, 2009.

[10] F. L. Hedayioglu, S. S. Mattos, L. Moser, and M. E. de Lima. Development of a tele-stethoscope and its application in pediatric cardiology. *Indian Journal of Experimental Biology*, 45, 2007.

[11] H. L. S. I. HL7. Hl7 - cardiology, 2011.

[12] S. M. Kamruzzaman, A. R. Hasan, A. B. Siddiquee, and M. E. H. Mazumder. Medical diagnosis using neural network. In *3rd International Conference on Electrical & Computer Engineering (ICECE)*, pages 28–30, Dec 2004.

[13] A. C. o. Radiology. Breast imaging reporting and data system (biradstm), 1998.

[14] B. O'Hora, J. Perera, and A. Brabazon. Designing radial basis function networks for classification using differential evolution. In *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, pages 2932 –2937, 0-0 2006.

[15] D. Pereira, F. Hedayioglu, R. Correia, I. Dutra, F. Almeida, S. Mattos, and M. Coimbra. Digiscope - unobtrusive collection and annotating of auscultations in real hospital environments. In *17th Portuguese Conference on Pattern Recognition, RecPad 2011*, Oct 2011.

[16] D. Pereira, F. Hedayioglu, R. Correia, T. Silva, I. Dutra, S. Mattos, F. Almeida, and M. Coimbra. Digiscope - unobtrusive collection and annotating of auscultations in real hospital environments. In *Proceedimgs of the 33rd IEEE International Conference on Engineering in Medicine and Biology*. IEEE, 2011.

[17] A. Srinivasan. *The Aleph Manual*, 2001.

[18] M. E. Tavel, D. D. Brown, and D. Shander. Enhanced auscultation with a new graphic display system. *Arch. Intern. Med.*, 154:893, 1994.

[19] J. Wu, J. Roy, and W. F. Stewart. Prediction modeling using ehr data: Challenges, strategies, and a comparison of machine learning approaches. *Medical Care*, 48:106–113, June 2010.