Knowledge on Heart Condition of Children based on Demographic and Physiological Features

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

Tiago T. V. Vinhoza, Ana Castro
Instituto de Telecomunicações, Porto, Portugal
tiago.vinhoza@ieee.org, ana.castro@dcc.fc.up.pt

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

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

Abstract

We evaluated a population of 7199 children between 2 and 19 years old to study the relations between the observed demographic and physiological features in the occurrence of a pathological/non-pathological heart condition. The data was collected at the Real Hospital Português, Pernambuco, Brazil. We performed a feature importance study, with the aim of categorizing the most relevant variables, indicative of abnormalities. Results show that second heart sound, weight, heart rate, height and secondary reason for consultation are important features, but not nearly as decisive as the presence of heart murmurs. Quantitatively speaking, systolic murmurs and a hyperphonetic second heart sound increase the odds of having a pathology by a factor of 320 and 6, respectively.

1 Introduction

Children are usually thought of as having healthy hearts. Therefore it maybe a surprise to many people to learn that, in the US, nine out of every 1000 babies are born with a congenital heart abnormality. It is estimated that one third of these babies require intervention to prevent death in the first year of life. Approximately 1.3 million people living in the US today were born with a congenital heart defect, and at least half of these individuals are under age 25[1]. In Portugal, the number of cardiac surgeries in children performed a year is around 500 and eight out of 1000 babies are born with a heart abnormality[2]. In Brazil, place where we collected our data it is estimated that between eight and ten children out of 1000 are born with a congenital cardiac disease [1].

Risk factors such as smoking, lack of exercise, and high cholesterol, that contribute to coronary artery disease and other cardiovascular diseases levels, often start at an early age. In the US, about 4.5 million children, ages 12 to 17, are already smokers. Nearly half of people aged 12 to 21 do not exercise on a daily basis, and an estimated 8.8 million (about 30 percent) US children (ages six to 19) are obese.

Because of the misperception that all children have healthy hearts, cardiac diseases can evolve, in a silent manner, to the point that it can be too late to revert the health deterioration by implementing an adequate treatment. Other factor that contributes to the non-detection of diseases is the lack of expert professionals available in certain regions. In particular, in certain regions in Brazil, this is a major concern.

We study a population of 7199 children between 2 and 19 years old, from the northeastern part of Brazil. Our goal is to identify variables that may be more indicative of normality or pathology and use this information to build classifiers that can, in an automatic fashion, distinguish between normal and cardiac pathological cases. Such clas-

[1] Source: Lucile Packard Children's Hospital at Stanford

[2] Source: Apifarma, Portuguese Association of the Pharmaceutical Industry.

978-1-4799-1053-3/13/$31.00 ©2013 IEEE

sifiers, and an increased understanding of the relations between physiological and demographic variables, may help on the decision making process, avoiding missing pathological patients when they are consulting with a less experienced professional. This would allow detecting and initiating the treatments earlier, improving patient outcome and reducing costs. Very few works in the literature report on prediction of heart diseases using machine learning techniques. The University of California at Irvine (UCI) repository has some datasets related to cardiology. The one most related to our dataset is the "Heart Disease" [2]. 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" attribute in this database 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 focused on attempting to distinguish presence of disease (values 1,2,3,4) from absence of disease (value 0). These experiments focused mainly on classification performance. Some of them use feature selection to build classifiers but do not focus on the individual feature importance as we do in this work. The results obtained with an instance-based learning algorithm (IB1) report an accuracy of 75.7% (± 0.8) [3]. Another experiment with the same dataset, also to diagnose disease, used a neural network algorithm and reported an accuracy of 87.5% [4]. More recently, the same Cleveland dataset was used in experiments that use Radial Basis Function Networks, that report an accuracy of 84% [5]. A more recent work [6] had as objective to model detection of heart failure more than 6 months before the actual date of clinical diagnosis using machine learning techniques to EHR data. They compared the performances of logistic regression, SVM and Boosting along with various variable selection methods in heart failure prediction. A value of 0.77 for the area under the ROC (AUROC) for the best classifier was reported. Previous work on a smaller dataset of 169 children [7], and performing an exhaustive search for the best classifier, produced an accuracy of 90.5% and AUROC of 0.83. For this dataset with 7199 children, our training accuracy is 93% when predicting cardiac pathological cases, using an SVM.

2 Materials and Methods

2.1 Dataset

The data used in this study was collected in the Real Hospital Português (RHP), Brazil, then anonymized and shipped to Portugal with the approval of the RHP Ethics Commitee. The Ethics Committee of the University of Porto, Portugal also approved this study.

In the original dataset containing 7603 instances and 33 features, we performed pre-processing tasks, namely, data cleaning, data transformation, data normalization, as well as removal of irrelevant features, like ID, name of doctor, etc. Since we want to use the body mass index (BMI) in our study and this feature is usually assessed for 2-year old or older children, we removed from our dataset all the children younger than 2 years old. Therefore, our dataset, after pre-processing, consists of 7199 cases and 21 features. Of these 7199 cases, 2507 (34.8%) are pathological, while 4692 (65.2%) are healthy. The description of the data with its variables is shown in Tables 8 and 9 in the appendix. We note that age, BMI, systolic blood pressure, and diastolic blood pressure are represented in numerical and categorical versions. We used only the categorical versions in our analysis, thereby reducing the number of analyzed features to 17. The main goal of this work is to apply feature importance metrics to rank the variables that are more predictive of cardiac pathologies.

2.2 Methodology

We now present in detail the methods applied in our dataset. We take two approaches to find the most relevant attributes to predict pathologies. First, we use a filter based approach, i.e., using model independent metrics such as mutual information and chi-squared tests. Then, we use model specific metrics, namely a variable importance measure by a random forest classifier and a logistic regression odds ratio analysis.

The mutual information approach to feature selection consists in computing the mutual information between each attribute and the class variable. This gives information on how correlated they are. Remembering the definition of mutual information $I(X;Y) = H(X) - H(X|Y)$, where X is our class variable and Y is the feature under analysis [8]. To have the results bounded between 0 and 1, we define a normalized mutual information measure as:

$$I_{\text{norm}}(X;Y) = I(X;Y)/H(X) = 1 - H(X|Y)/H(X).$$

The intuition this result gives us is the following: If X and Y are statistically independent random variables, then $H(X|Y) = H(X)$, that is, the knowledge of Y does not reduce the uncertainty about X. Therefore, $I_{\text{norm}}(X;Y) = 0$. If the knowledge of Y removes all the uncertainty about X, then $H(X|Y) = 0$, as a result $I_{\text{norm}}(X;Y) = 1$. So, the closer the result is to 1, the more important is the feature according to this measure.

The chi-squared test is used as a test of independence between two random variables. The first step is to calculate the chi-squared test statistic, χ^2, which resembles a normalized sum of squared deviations between observed and expected frequencies. The second step is to determine the degrees of

freedom, d, of that statistic, which is essentially the number of frequencies reduced by the number of parameters of the fitted distribution. In the third step, χ^2 is compared to the critical value of no significance from the χ^2_d distribution, which in many cases gives a good approximation of the distribution of χ^2.

We then analyze the specific importance metrics. First, we calculate the variable importance as measured by a random forest classifier. A random forest is an ensemble classifier that consists of many decision trees, and outputs the class that is the mode of the classes output by individual trees. The method combines Breiman's *bagging* idea [9] and the random selection of features, introduced independently by Ho [10] and Amit and Geman [11] to construct a collection of decision trees with controlled variation.

Finally, we apply a logistic regression and perform odds ratio analysis to infer the importance of each feature in the odds of having pathology. In a logistic regression, we can think of the class variable x as having a Bernoulli distribution with parameter p given by

$$p = P(x = 1|\mathbf{\Theta}^T\mathbf{y}) = h\left(\mathbf{\Theta}^T\mathbf{y}\right),$$

where $\mathbf{\Theta} = [\theta_1, \theta_2, \ldots, \theta_f]^T$ are the regression coefficients, f is the number of features, $\mathbf{y} = [y_1, y_2, \ldots, y_f]$ is the vector containing the features, y_i is the value the i-th feature, and $h(\cdot)$ is the logistic function. The log odds of the outcome is modelled as a linear combination of the predictor variables $\ln\left(\frac{p}{1-p}\right) = \mathbf{\Theta}^T\mathbf{y}$. The odds ratio for the i-th feature is simply given by exponentiating the i-th feature regression coefficient and can be seen as how an increase (presence) of a numerical (categorical) feature influence the probability of occurrence of the class variable [12].

3 Results and Discussion

In this section we present the results obtained by applying the feature importance metrics presented in the previous section to our database. The mutual information results, using the 17 features, place murmur as the feature that most reduces the uncertainty of the class variable (either pathological or not). In fact, murmur presents a score much higher than the remaining features (Table 1).

From the presented result, murmur plays an important role when predicting cardiac pathology, which is in accordance to clinical assessment [13]. Nonetheless, from all the cases in the database we observe that 5,000 patients have absent murmur, and of those, 404 have cardiac pathology (404 pathological, $\approx 8\%$ and 4,596 normal, $\approx 92\%$). Since murmur seems to be determinant in cardiac pathology detection, to study the impact of the absence of this characteristic, and to evaluate which other variables may aid in the detection of these pathological cases, we apply mutual information to the dataset containing only the patients with absent murmur. Results are presented in Table 2.

Table 1. Mutual Information applied to the 17 features: 7199 cases

Feature	Score
Murmur	0.61
Secondary Reason	0.10
Weight	0.09
Primary Reason	0.05
HR	0.05

Table 2. Mutual Information applied to the 17 features: 5000 cases with absent murmur

Feature	Score
Weight	0.17
HR	0.05
Height	0.04
S2	0.04
Secondary Reason	0.02

Although weight, heart rate, height, S2 and secondary reason, reach the top of the ranking (Table 2), their relevance is lower when compared to murmur (Table 1).

Chi-squared results using the 17 features from the 7199 instances reinforce also the idea that murmur is a key factor when assessing a cardiac disease (Table 3). When applying chi-squared test to the dataset with absent murmur (Table 4), the feature ranking obtained is very similar to the feature ranking in the mutual information approach, and as referred before.

Table 3. Chi-squared test applied to the 17 features: 7199 cases

Feature	Score
Murmur	5160.57
Weight	730.00
Secondary Reason	702.92
Primary Reason	501.63
HR	445.33

Moving to classifier-based approaches, we first compute the mean decrease Gini achieved by a random forest classifier to the 17 features from the 7199 instances. It focuses on measuring the total decrease in node impurities from split-

Table 4. Chi-squared test applied to the 17 features: 5000 cases with absent murmur

Feature	Score
Weight	577.86
S2	255.56
HR	149.01
Height	125.22
Secondary Reason	62.89

ting on the variable, averaged over all trees. The node impurity is measured by the Gini index. A variable that decreases the Gini index the most is responsible for a decreased node impurity, hence it is the most important in terms of separating the target classes. Analyzing the results with 17 features from the 7199-case dataset, murmur is the crucial feature in order to correctly separate the target classes, as it decreases the Gini index in an approximate score of 1976 (Table 5). When we apply a random forest classifier to the dataset with absent murmur (Table 6), we notice that the top 5 features present scores of mean decrease Gini that are less relevant when compared to the result achieved by murmur.

We noticed that the Secondary Reason is ranked in the top 2 twice and in the top 5 three times. This can be explained by the fact that one of the possible Secondary Reasons is "Presence of Murmurs" (see Table 9). From the 881 occurrences of that label, 635 ($\approx 72\%$) are associated with a cardiac disease. This reinforces the importance of murmur in the accurate detection/classification of cardiac pathology.

Table 5. Variable importance as measured by a Random Forest classifier applied to the 17 features: 7199 cases

Feature	Mean Decrease Gini
Murmur	1975.98
Secondary Reason	216.44
Weight	189.82
Height	172.65
HR	149.61

We then apply logistic regression and compute the odds-ratio of the features. The most important features according to the odds-ratio is the presence of systolic murmurs, which increases the probability of having a pathology by a factor of approximately 320. The results of the logistic regression also shows that an abnormal S2, such as having an hyperphonetic S2, increases the odds of having a cardiac pathology by six. This result contrasts with the ones obtained using the other importance metrics presented in the paper. Analyzing S2 in more detail, we may notice that if S2

Table 6. Variable importance as measured by a Random Forest classifier applied to the 17 features: 5000 cases with absent murmur

Feature	Mean Decrease Gini
Weight	131.59
Height	119.36
HR	84.64
CDH 1	48.05
Secondary Reason	46.45

is abnormal (i.e. unique, hyperphonetic or fixed split) it is possible to easily separate the pathological cases. There are 140 instances with abnormal S2. From these, 129 ($\approx 92\%$) are associated with a cardiac disease while 11 ($\approx 8\%$) refer to healthy children. Focusing on the 7005 instances with normal S2, the distribution of abnormal and normal cases is extremely similar to the *a priori* distribution (see Table 9). From the 7005 cases with normal S2, 2338 ($\approx 33\%$) are associated with a pathology, while the remaining 4667 ($\approx 67\%$) are from patients with no cardiac problems. As the normal S2 are in much higher number than the abnormal ones, knowing S2 does not clarify much in the prediction of cardiac pathologies, as proved by the mutual information and chi-squared tests.

Finally, we performed a small experiment using a SVM classifier applied to the 7199 cases to predict cardiac pathologies using/not using murmur as feature. The results, obtained with 10 times 10-fold cross-validation, are in Table 7 and are consistent with the results obtained in the feature importance analysis. As performance metrics, we report the average number of Correctly Classified Instances (CCI), sensitivity and specificity.

Table 7. Predicting Pathology

Metrics	Using murmur	Not using murmur
CCI (%)	93.21	73.12
Sensitivity	0.85	0.37
Specificity	0.98	0.92

4 Conclusions

In this study we present an exploratory analysis of various cardiac and demographic features collected from children (with and without cardiac pathology), in standard clinical practice. The most important result that is drawn from the several exploratory techniques presented, is the importance of the presence of murmur in the detection of cardiac pathology. Although this information is not new, and

is widely used in clinical practice to assess cardiovascular state in conjunction with other demographic data [13], it should be noticed that when this feature is not present, the remaining variables analyzed in this study do not contribute as decisively to the pathology detection. Hence it is crucial to have accurate information on murmur presence. Taking into consideration that some studies report a high detection error of mild murmurs, when evaluated by a less experienced clinician [14], this may present a motivation for the introduction of digital signal processing of the heart sound for feature extraction, aiding the less experienced clinician in the detection of such features. It should be noticed that information about S2, which is described in the literature [13] as being a decisive predictor of cardiopathies, did not yield significant performance gain in the presented exploratory analysis.

Acknowledgements

The authors are grateful to Nuno Marques (University of Porto, Portugal) and Bruno Lopes (Instituto de Telecomunicações, Porto, Portugal) for helpful and valuable discussions.

This work was partially funded by the DigiScope project (PTDC/EIA-CCO/100844/2008) and by the Fundação para a Ciência e a Tecnologia (FCT/Portugal). Pedro Ferreira is sponsored by a grant from INESC-TEC LA.

References

[1] V. C. Pinto Jr. *et al.*, "The situation of congenital heart surgeries in Brazil," *Revista Brasileira de Cirurgia Cardiovascular*, vol. 19, Jun. 2004.

[2] K. Bache and M. Lichman, "UCI machine learning repository," 2013.

[3] D. Aha and D. Kibler, "Instance-based prediction of heart-disease presence with the Cleveland database," tech. rep., University of California, Mar. 1988.

[4] 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)*, pp. 28–30, Dec. 2004.

[5] B. O'Hora, J. Perera, and A. Brabazon, "Designing radial basis function networks for classification using differential evolution," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, pp. 2932 – 2937, 2006.

[6] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches," *Medical Care*, vol. 48, pp. 106–113, Jun. 2010.

[7] P. Ferreira *et al.*, "Detecting cardiac pathologies from annotated auscultations," in *Proc. International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 1–6, IEEE, 2012.

[8] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley, 2006.

[9] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, Jan. 2001.

[10] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832–844, Aug. 1998.

[11] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural computation*, vol. 9, pp. 1545–1588, Oct. 1997.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer Series in Statistics, 2009.

[13] L. S. Bickley, *Bates' Guide to Physical Examination and History Taking*. Lippincott Williams & Wilkins, 2005.

[14] J. R. Carapetis *et al.*, "Evaluation of a screening protocol using auscultation and portable echocardiography to detect asymptomatic rheumatic heart disease in Tongan schoolchildren," *Nature Clinical Practice Cardiovascular Medicine*, vol. 5, pp. 411–417, Jul. 2008.

A Dataset Description Tables

Table 8. Numerical Features

Attribute	Range	Average ±Stdev	Missing
Age	2-19	8.6 ± 3.7	0
Height (cm)	51-198	130.2 ± 21.5	0
Weight (kg)	3.5-101.0	32.8 ± 15.0	0
BodyMassIndex	12.0-33.6	18.4 ± 3.6	0
HeartRate (bpm)	48-160	85.5 ± 11.0	310
SystolicPressure	70-170	101.0 ± 10.7	20
DiastolicPressure	35-120	62.1 ± 8.5	20

Table 9. Categorical features

Attribute	Values	Percent	Qty	Missing	Pathology	
					Yes	No
Sex	Female	41	2946	0	1023	1923
	Male	59	4253		1484	2769
Age Range	Pre-School 2-6	38.3	2757	0	1158	1599
	School 6-10	25.2	1817		636	1181
	Pre-Teen 10-14	28.3	2036		573	1463
	Teenager 14-19	8.2	589		140	449
Body Mass Index Percentile	Low Weight	4.5	324	0	148	176
	Normal	48.7	3509		1223	2286
	Overweight	17.1	1234		445	789
	Obese	29.6	2132		691	1441
Systolic Blood Pressure (SBP)	Normal	91.6	6574	20	2241	4333
	Limit	3.1	224		87	137
	Hypertense	5.3	381		172	209
Diastolic Blood Pressure (DBP)	Normal	90.0	6459	20	2234	4225
	Limit	5.7	409		149	260
	Hypertense	4.3	311		117	194
Result-SBP-DBP	Normal	86.2	6187	20	2113	4074
	Limit	6.6	472		175	297
	Hypertense	7.2	520		212	308
Murmur	Absent	69.5	5000	0	404	4596
	Systolic	30.4	2186		2093	93
	Diastolic	0.1	6		5	1
	Continuous	0.1	7		5	2
Second Heart Sound (S2)	Normal	98.0	7005	54	2338	4667
	Fixed Split	0.9	63		55	8
	Unique	0.1	9		8	1
	Hyperphonetic	1.0	68		66	2
Pulses	Normal	99.8	7168	17	2486	4682
	Diminished Femoral	0.1	7		7	0
	Ample	0.1	7		2	5
Current Disease History 1 (CDH 1)	Asymptomatic	72.3	3910	1789	1500	2410
	Cyanosis	1.0	54		22	32
	Precordial pain	9.7	527		176	351
	Dyspnea	6.1	332		135	197
	Palpitation	5.3	286		78	208
	Faint/Dizziness	3.2	172		37	135
	Weight Gain	2.4	129		50	79
Current Disease History 2 (CDH 2)	Cyanosis	8.4	26	6889	15	11
	Precordial pain	18.1	56		18	38
	Dyspnea	22.9	71		25	46
	Palpitation	29.7	92		28	64
	Faint/Dizziness	12.9	40		13	27
	Weight Gain	8.1	25		10	15
Primary Reason	Cardiopathy	5.7	408	60	258	150
	Routine check-up	7.2	513		119	394
	Others	2.5	178		67	111
	Cardiology Screening	53.1	3788		958	2830
	Possible Cardiopathy	31.5	2252		1082	1170
Secondary Reason	Physical Activity	13.1	701	1846	188	513
	Congenital Cardiopathy	6.1	324		222	102
	Surgery	34.2	1833		499	1334
	Risk factors	17.1	914		299	615
	Presence of Murmurs	16.5	881		635	246
	Others	13.1	700		241	459
Pathology			7199	0	2507 (34.8%)	4692 (65.2%)