Identifying Key Factors in Right Ventricular Involvement in Ischaemic and Non-ischaemic Cardiomyopathy

Carlos Barroso-Moreno
Faculty of Biomedical and Health Sciences
Universidad Europea de Madrid
Madrid, Spain
email: 22014885@live.uem.es
0000-0002-1609-2267

Hector Espinos Morato
i3M Molecular Imaging
Universidad Europea de Valencia
Valencia, Spain
0000-0002-4089-1368

Enrique Puertas
School of Architecture, Engieering and Design
Universidad Europea de Madrid
Madrid, Spain
0000-0002-5115-1226

Juan José Beunza Nuin
Faculty of Biomedical and Health Sciences
Universidad Europea de Madrid
Madrid, Spain
0000-0001-8192-2952

José Vicente Monmeneu
Cardiac Magnetic Resonance
Exploraciones Radiológicas Especiales (ERESA)
Valencia, Spain
email: jmonmeneu@eresa.com

David Moratal
Center for Biomaterials and Tissue Engineering
Universitat Politècnica de Valencia
Valencia, Spain
0000-0002-2825-3646

María P. López-Lereu
Cardiac Magnetic Resonance
Exploraciones Radiológicas Especiales (ERESA)
Valencia, Spain
email: mplopezl@ascires.com

***Abstract*—Cardiomyopathy is a disease of the heart muscle that makes it harder for the heart to pump blood. Previous studies have focused on the left ventricle, but in recent years the relevance of the right ventricle has been the focus of current research. The aim is to determine those clinical and cardiac parameters that influence right ventricular involvement in ischaemic and non-ischaemic cardiomyopathy. The used database is composed of 56,447 subjects collected from 2008 to 2020 by ASCIRES Biomedical Group. The methodology is divided into two blocks: in the clinical aspect, decision trees are used to gain interpretability and in the technical aspect, Machine Learning (ML) is used for a greater degree of prediction. The results show the influence of the difference in aortic artery beat volume and vascular pulmonary volume as key factors, reaching an Area Under the Curve (AUC) of 92.3% using RapidMiner tool with decision trees algorithm. The conclusions demonstrate the ability to identify clinical variables of right ventricular involvement and consequently reduce the number of diagnostic tests and associated times in a situation of cardiomyopathy.**

***Index Terms*—Machine Learning; right ventricular involvement; Pulmonary Vascular Resistance; Cardiomyopathy.**

I. INTRODUCTION

Among cardiovascular diseases, ischaemic heart disease accounts for 16% of all deaths worldwide, rising from over 2 million deaths in 2000 to 8.9 million in 2019, and has become the disease attributed with the largest increase in deaths since 2000 [1]. Other cardiac conditions, such as non-ischaemic cardiomyopathy, arrhythmia, valvular heart disease, and heart failure are highly prevalent in developed countries and also cause high morbidity and mortality [2]. Due to the complexity and high prevalence of these diseases, a better understanding of the pathophysiology, as well as earlier diagnosis is of vital importance to increase the success rate of therapies, which is reflected in a reduced level of disability and lower mortality. To this end, for decades, all attention has been directed to the study of the left ventricle, making the right ventricle the "forgotten side of the heart". On the other hand, a direct extrapolation of the knowledge acquired about the physiology of the left side of the heart to the right side is not possible, as the normal right ventricle is anatomically and functionally different from the left ventricle. However, in recent years, advances in non-invasive cardiac imaging techniques have made it possible to discover the importance of the right ventricle in different cardiac diseases [3]. Therefore, there is a need for a better understanding of those factors that influence right ventricular dysfunction, given the accumulating evidence of their clinical relevance from both a symptomatic or diagnostic and prognostic perspective.

Numerous researchers have demonstrated the feasibility of applying Machine Learning (ML) algorithms in health studies to predict strokes [4], ICU patients with Covid-19 [5], prostate cancer [6] or acute coronary diseases syndrome [7] and others. The joint use of this type of algorithms with visualisation tools, such as Power BI is used in numerous areas [8] [9]

 ISBN: 978-1-68558-105-3

[10]. However, the diagnostic use of decision-making tools is still rare in the health sector. In recent years, these types of diagnostic aid tools have become popular. For example, a success story is the application of this type of tools in the private health sector in Finland, which has based its health system's decisions on data, identifying key factors [11].

Previous studies have identified influential variables in right ventricular compromise, such as: pulmonary arterial hypertension (PAH) associated with pressure overload [12]; diabetes, dyslipidemia [13], blood flow [14] and habits, such as smoking [15] and others.

The main objective of this project is to determine those clinical and cardiac parameters that influence the involvement of the right ventricle in ischemic and non-ischemic cardiomyopathy using ML techniques. To this end, predictive models capable of identifying patients with right ventricular dysfunction will be developed and the key parameters used by the models will be studied from the point of view of their clinical implication.

This paper is organized as follows. In Section 2 presents the details of dataset and it describes the methodology for Power BI and RapidMinner. Section 3 describes the results focusing on significant variables, distributions and decision trees . Finally, Section 4 presents the conclusions and directions for future work.

II. MATERIALS AND METHODS

The methodology used makes a comparison of the most common supervised classification algorithms in ML: Support Vector Machines (SVM) [16], decision tree [17], Random Forest [18] and neural networks [19]. To compare results, the following metrics have been used: precision, sensitivity, specificity and Area Under the Curve (AUC). This last variable has been used to evaluate the performance of binary classification models.

The data used for the study comes from the ASCIRES Biomedical Group database. This database has 56,447 records of variables collected from 2008 to 2020.

A. Software used

The research uses two software tools: Power BI and RapidMinner.

- Power BI is a data analytic service from Microsoft that provides interactive graphs focused on analytic intelligence to generate reports [20]. In the present research it allows easy visualisation of the database to automate the analysis.
- RapidMiner is a software for data analysis and data mining by chaining operations in a graphical environment [21]. Version 9.10.013 is used to obtain the results of the ML models.

B. Data preparation

At this point, clinical filters and patient labels are made according to age, gender, systole and diastole of the right ventricle. The volume of data cleaning by means of the filters means, that the initial database has 56,447 patients and 1,815 variables; after applying the filters, these are reduced to 12,083 patients and 120 variables. Of the latter subgroup, 7153 are labelled as unaffected and the target group is 4944 with right ventricular involvement.

The process applies logical cleaning, such as: (i) Removal of inconsistent data, such as the presence of letters in numerical values. (ii) Elimination of erroneous data, heights < 1 metre and > 2.3 metre or ages < 0 and > 120 years. (iii) Checking whether the numeric value zero represents such a value or is a null value (NULL). (iv) For having the same content as other variables, but with a different name, e.g., Vol.eyec.Ao for vol.lat. (v) For having all data set to 0; (vi) Transformation into international units of certain variables, such as wood units. (vii) For containing inconsistent data from the clinical perspective (outlayer).

An additional step is the elimination of variables with higher correlations, in order to avoid multicollinearity in our database. These steps are the following:

(i) Frac.reg.Ao.por.vol.lat to Ao.reg.Vol.beat.vol.dif with $\rho = 0,91$ (ii) Ao.reg.Vol.beat.vol.dif to Vol.reg.Pulm.dif.vol.lat with $\rho = -1$ (iii) Frac.reg.Ao.por.vol.lat to Vol.reg.Pulm.dif.vol.lat with $\rho = -0,92$ (iv) IMVI to MVI with $\rho = 0,93$ (v) NLVEDV to NLVESV with $\rho = 0,94$ (vi) Weight (kg) to S.Corp with $\rho = 0,95$ (vii) NRVSV to RVSV with $\rho = 0,95$ (viii) NLVSV to LVSV with $\rho = 0,95$ (ix) NRVEDV to RVEDV with $\rho = 0,95$ (x) RWT...relative.wall.thickness to RWT.2...relative.wall.thickness.pwd.sd with $\rho = 0,90$ (xi) LVEDV to LVESV with $\rho = 0,93$ (xii) NLVEDV to LVEDV with $\rho = 0,95$ (xiii) NLVEDv to LVESV with $\rho = 0,90$ (xiv) IVTSVD to RVESV with $\rho = 0,97$ (xv) NLVESV to LVESV with $\rho = 0,97$

C. Data labels

Aligned with the main objective of the research, the database is labelled to assess whether the patient has right ventricular involvement. For the identification of these patients, the consensus tables specified by the European Society of Cardiology [22] establishes ranges of variables (age, gender, systoles, diastole and others) to identify RV involvement. These values are adapted in Table I, The normal values of RV systolic and diastolic parameters vary according to age and gender. The standard of normal values used for the recognition of impairment is used with a similar range in current papers, such as that of the researchers Petersen et al. (2019) [23].

The absolute values of end-systolic volume (ESV), end-diastolic volume (EDV) and body mass have been used, whose values are provided automatically within the framework of clinical tests. Clinics automatically provide based on clinical evidence frameworks.

Systolic volume (SV) is calculated by the difference between EDV and ESV; Additionally, ejection fraction (EF) is calculated as SV/VDE. Sex, body surface area (BSA), and age are independent predictors of several RV parameters, as suggested by previous well-established studies [24]. The

 ISBN: 978-1-68558-105-3

TABLE I
RIGHT VENTRICLE LABELS. STANDARD RANGES BY RV VOLUMES, SYSTOLIC FUNCTION AND MASS BY AGE INTERVAL (95% CONFIDENCE INTERVAL). ADAPTED FROM MACEIRA ET AL. (2006) [22].

Right Ventricule labels						
Age (years)	***20-29***	***30-39***	***40-49***	***50-59***	***60-69***	***70-79***
Males						
Absolute values						
1-1 EDV (mL) SD 25.4	(127,227)	(121,221)	(116,216)	(111,210)	(105,205)	(100,200)
ESV (mL) SD 15.2	(38,98)	(34,94)	(29,89)	(25,85)	(20,80)	(16,76)
SV (mL) SD 17.4	(74,143)	(74,142)	(73,141)	(72,140)	(71,139)	(70,138)
EF (%) SD 6.5	(48,74)	(50,76)	(52,77)	(53,79)	(55,81)	(57,83)
Mass (g) SD 14.4	(42,99)	(40,97)	(39,95)	(37,94)	(35,92)	(33,90)
Normalized to BSA						
EDV/BSA (mL//mˆ(2)) SD 11.7	(68,114)	(65,111)	(62,108)	(59,105)	(56,101)	(52,98)
ESV/BSA (mL//mˆ(2)) SD 7.4	(21,50)	(18,47)	(16,45)	(13,42)	(11,40)	(8,37)
Females						
Absolute values						
1-1 EDV (mL) SD 21.6	(100,184)	(94,178)	(87,172)	(81,166)	(75,160)	(69,153)
ESV (mL) SD 13.3	(29,82)	(25,77)	(20,72)	(15,68)	(11,63)	(6,58)
SV (mL) SD 13.1	(61,112)	(59,111)	(58,109)	(56,108)	(55,106)	(53,105)
EF (%) SD 6	(49,73)	(51,75)	(53,77)	(55,79)	(57,81)	(59,83)
Mass (g) SD 10.6	(33,74)	(31,72)	(28,70)	(26,68)	(24,66)	(22,63)
Normalized to BSA						
1-1 EDV/BSA (mL//mˆ(2)) SD 9.4	(65,102)	(61,98)	(57,94)	(53,90)	(49,86)	(45,82)
ESV/BSA (mL//mˆ(2)) SD 6.6	(20,45)	(17,43)	(14,40)	(11,37)	(8,34)	(6,32)

standardised values of EDV/BSA and ESV/BSA are obtained from these variables.

The filtered database uses functions to apply binary labeling of patients: Normal or Abnormal (RV involvement). As an example, a 62-year-old male patient with an EDV of 211 EDV (mL) is labeled as abnormal (RV impairment) because he is not within the range of (105,205) set in the parameters of Table I. If the same patient has an EDV of 204 EDV (mL), the patient is considered to have no RV involvement (normal), if the patient also meets the other variables in their corresponding ranges.

The database input and output variables are described in Table II. The $\overline{\mathbf{X}}$ is the average; SD is the standard deviation; Max is the maximum value and Min is the minimum value. The database contains 12,083 patients, of which 76.6% ($n = 9260$) are men and 23.4% ($n = 2823$) are women. The mean age is 62.49 years with a standard deviation of 14.1. The average body mass index (BMI) is 27.87 ($SD = 4.69$), the formula is $BMI = Weight(kg)/[Height(m)]^2$. According to the BSA, the average is 18.04 ($SD = 2.8$).

The output result corresponds to the classification of RV involvement with patients without RV involvement (normal) in a percentage of 59.1% ($n = 7139$) and with RV involvement (abnormal) in a percentage of 40.9% ($n = 4944$). Missing data is 0 because these records are removed in the preprocessing step so as not to distort the output of the ML algorithms in later steps.

In the first tests, the variables in the table are eliminated due to their direct relationship in the table calculations. These are discarded, as detailed in Table III: right ventricular systolic volume (RVSV), left ventricular systolic volume (LVSV), right ventricular end-diastolic volume (RVEDV), left ventricular end-diastolic volume (LVEDV), right ventricular end-systolic volume (RVESV), left ventricular end-systolic volume (LVESV). However, they are reintroduced in the prediction tests, as there is no clear correlation with the output variable and they are not detected as primary variables for predicting RV involvement.

TABLE II
INPUT VARIABLES FOR OUTPUT VARIABLE LABEL.

Input				
Variable	***Categories***	***n***	***%***	***Missing***
Gender	Males	9260	76,6	0
	Females	2823	23,4	0
	$\overline{X}$	***SD***	***Min***	***Max***
Age	62,49	14,1	20	95
BMI	27,87	4,69	13,1	71,4
BSA	18,04	2,8	4,22	53,13
EDV/BSA	70,94	28,34	12,0	403,3
ESV/BSA	33,86	21,03	0,9	251,9
Output				
Variable	***Categories***	***n***	***%***	***Missing***
DV involvement	Normal	7139	59,1	0
	Abnormal	4944	40,9	0

III. RESULTS

After data preparation with a single model, the samples of the training set (n-train) and test sets (n-test) are 8458 and 3625, respectively. These patients are changed by cross-validation in the tests.

The results of the algorithms are shown in the table IV, which are Accuracy (ACC), Sensitivity (SE), Specificity (SP), Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Area Under the Curve (AUC).

 ISBN: 978-1-68558-105-3

TABLE III
VARIABLES ELIMINATED BY INDIRECT USE IN THE OUTPUT VARIABLE.

Elimnated variables				
	$\overline{X}$	*SD*	*Min*	*Max*
RV.DTD	32,78	7,33	6	100
RVSV	70,73	27,24	4	250
LVSV	75,05	24,16	8	200
RVEDV	135,5	55,69	23	500
LVEDV	196,89	75,97	33	600
RVESV	64,68	40,56	2	437
LVESV	122,08	69,62	15	500

The algorithm with the highest AUC in RapidMiner is performed with XGBoost ($AUC = 0.87$), although the difference is small with the neural network algorithm ($AUC = 0.85$), determining a predominance of neural networks as predictors RV involvement. Although Random Forest algorithms have a ($AUC = 0.82$), and the lowest of the results is from Support Vector Machine (SVM) with a AUC =0.79. The results are quite promising, as it means that we can predict almost with an accuracy of 9 out of 10 patients performing the clinical tests whether they have RV involvement.

However, direct comparisons of the algorithms should be treated with caution. On the one hand, XGBoost provides the best results of the study but the complexity of ML algorithms makes them ”black boxes”. On the other hand, decision trees obtain inferior results but allow for great interpretability of the results gaining a value-added clinical perspective. The size of the database and the optimisation of the algorithms used allow low processing times of less than 3 minutes with a MacBook Pro 2.6 GHz Intel Core i7 2.6 GHz 6-core computer with 6 GB 2400 MHz DDR4 memory.

TABLE IV
COMPARISON AND EVALUATION OF DIFFERENT ALGORITHMS IN RAPIDMINNER.

Algorithms	*ACC*	*SE*	*SP*	*NPV*	*AUC*
SVM	0.76	0.76	0.76	0.77	0.79
Random Forest	0.78	0.78	0.79	0.78	0.82
Neural Network	0.78	0.78	0.79	0.77	0.85
XGBoost	0.82	0.78	0.80	0.77	0.87

The Table V shows the main variables with the greatest weight detect in the ML algorithms, whose selection is automatic as key risk predictors. The descriptive analysis of the variables are gender in male with 76.6% ($n = 9260$) and female 23,4% ($n = 2823$), This is logical as it is a necessary variable for patient labelling. Dyslipidemia is present in 51.6% ($n = 6231$) of patients, hypertension has a 59.0%($n = 7133$), diabetes has a low presence with a prevalence rate of 0.7% ($n = 85$), however, type II diabetes (low involvement) has a high incidence with 67.3% ($n = 8131$). The rest of the patients of each typology do not present the pathology. Smoking patients are high with 72.8% ($n = 8800$). Stent implantation is high with 82.2% ($n = 9937$) because the database is made up of patients who go to the cardiologist and suffer from some kind of affectation or signs of affectation of the heart. The stress study as a medical test is performed in 48.5% of patients ($n = 5857$), although the use of this aggressive test is decreasing every year.

In reference to the numerical variables, the most relevant is RVP [Wood] with $\overline{X} = 14,99$ ($SD = 1,27$). El aortic arch is $\overline{X} = 25,71$ ($SD = 3,68$), height is $\overline{X} = 167,86$ ($SD = 9,13$), Weight (kg) is $\overline{X} = 78,63$ ($SD = 14,96$). The Regurgitate Volume of the Aorta Artery (Reg.Vol.Ao.) is available for $\overline{X} = 4,31$ ($SD = 24,27$). The Diameter of the Aorta with Pulmonary Artery (DoA.PA) is $\overline{X} = 0,96$ ($SD = 0,25$). The Descending Thoracic Aorta (Desc.T.Ao) es de $\overline{X} = 24,43$ ($SD = 3,85$). Sinus pressure (Sinus.P) is $\overline{X} = 34,1$ ($SD = 4,77$). Posterior Wall in Diastole (PWD) has an average of $\overline{X} = 8,63$ ($SD = 2,75$). Finally, the ratio of diastolic volumes (diastolic RV) are of $\overline{X} = 1,59$ ($SD = 0,71$).

TABLE V
ANALYSIS OF THE VARIABLES WITH THE GREATEST WEIGHT IN THE ML ALGORITHM IN THE ASCIRES BIOMEDICAL GROUP DATABASE.

Main variables				
Variable	*Categories*	*n*	*%*	*Miss.*
Gender	Males	9260	76,6	0
	Females	2823	23,4	0
Dyslipidemia	Yes	6231	51,6	0
	No	5852	48,4	0
Hypertension	Yes	7133	59,0	0
	No	4950	41,0	0
Diabetes	Yes	85	0,7	0
	No	11998	99,3	0
Diabetes II	Yes	8131	67,3	0
	No	3952	32,7	0
Smoker	Yes	8800	72,8	0
	No	3283	27,2	0
Stent	Yes	9937	82,2	0
	No	2146	17,8	0
Stress study	Yes	5857	48,5	0
	No	6226	51,5	0
	$\overline{X}$	*SD*	*Min*	*Max*
PVR [wood]	14,99	1,27	0,49	18,97
Aortic.Arch	25,71	3,68	2	60
Height (cm)	167,86	9,13	131	205
Weight (kg)	78,63	14,96	35	187
Vol.reg.Ao. dif.vol.lat	4,31	24,27	-247	305
DoA.PA	0,96	0,25	0,01	5,01
Desc.T.Ao	24,43	3,85	2	63
Sinus.P	34,1	4,77	9	83
PWD	8,63	2,75	1	125
diastolic RV	1,59	0,71	0,17	10,88

A. Pulmonary Vascular Resistance (PVR)

Pulmonary Vascular Resistance (PVR) is the mean pressure drop from the main pulmonary artery to the divided left atrium. The units of measurement are Wood's units, which arise from the equivalence one Wood's unit = 80·s·cm^{-5}.

PVR is defined by the Swan-Ganz catheter from a central vein [25], the formula is:

$$PVR = 80 * (PAP - CEP) * CO, \quad (1)$$

therefore depend on Pulmonary Arterial Pressure (PAP) in mmHg units; the Capillary Locking Pressure (CEP) in mmHg units; and Cardiac Output (CO) in l/min units. The normal value for a subject is 1-2 [Wood], but this increases with age, as determined by the correlation table and Vizza et al. (2022) [26], increasing by 0.2 Wood in subjects over 50 years of age.

The database provided uses an estimation model [27]:

$$PVR[Wood] = 19.38 - (4.62 * Ln(PAAV) - (0.08 * RVEF)) \quad (2)$$

where PAAV are in centimeter per second and RVEF in percentage.

B. Distributions

The raw database model not shown in the results of the paper has an AUC value of 98%, the reason is that most of the patients in the database do not have RV involvement. This caused the algorithms to estimate a normal value by default and thus be right in most cases (sensitivity almost 0). But after data filtering, the database is balanced, preventing this negative event in the analysis of the observations. Another noteworthy aspect of the distributions is the elimination of data in the filtering that are empty in any of the variables, which may cause a bias by eliminating patients who intrinsically did not complete the data due to some particularity. As the volume is high, this is not an impediment to the analysis and the algorithms yield better results.

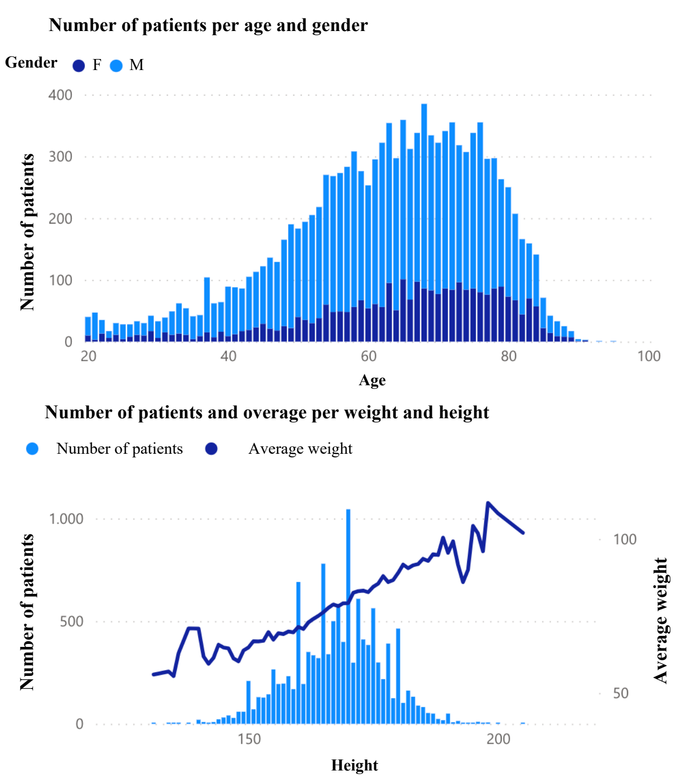

Fig. 1. Multivariate histograms of number of patients, (top) age as a function of gender, (bottom) height as a function of mean weight.

The use of Power BI makes it possible to visualise these and many other particularities quickly, avoiding bias or inconsistencies in the data. The Figure 1 shows the histogram of the number of patients by age and gender, quickly showing how an increase in age means a higher incidence, as well as a predominantly male gender. The 60 to 80 year age range accounts for almost half of the RV affectations. In reference to height and weight, there is a clear correlation between the increase in height and weight, which is why BMI is used as a predictive variable in the labelling of the database.

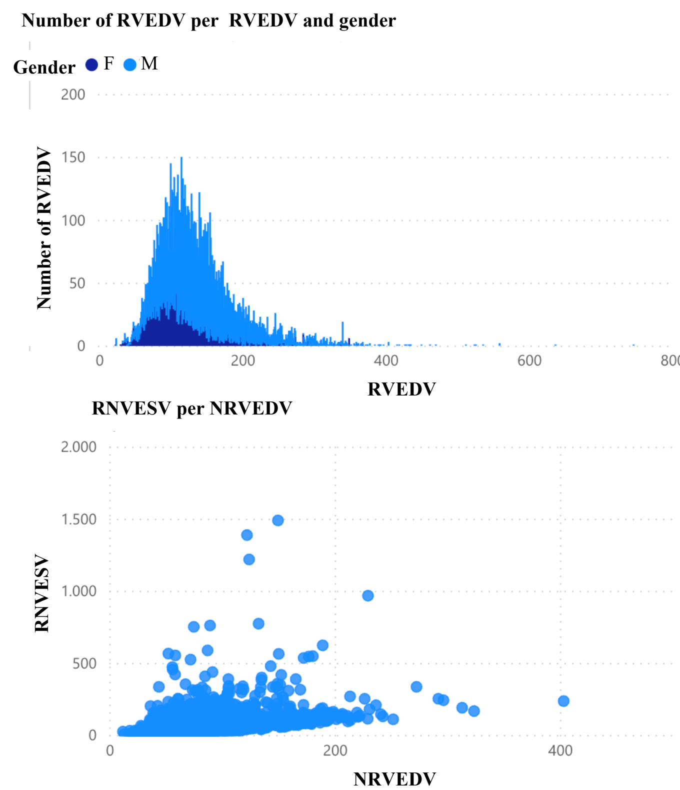

Fig. 2. Multivariate histogram of the number of patients, (top) RVEDV by gender, and scatter plot of NRVEDV with RNVESV.

Figure 2 shows the multivariate analysis of RVEDV according to gender and number of patients, in the analysis it can be seen that the female anatomy has a smaller volume than the male anatomy. This situation is contemplated in the labelling of RV involvement, so this differentiation from previous studies is correct. Regarding the lower figure, it represents the dispersion in relation to NRVEDV with ICTSVD, this analysis allows the detection of outliers to eliminate patients with highly distorted values. These graphs allowed to establish minima and maxima in the variables without interfering negatively in the algorithms. The interactive graphs allow for easy filtering of the outliers in order to reach a consensus with the doctors on their discarding and clinical criteria with a range limit for each variable. This process is repeated for each of the variables used in the labelling of the data.

According to PVR [Wood], Figure 3, which represents the

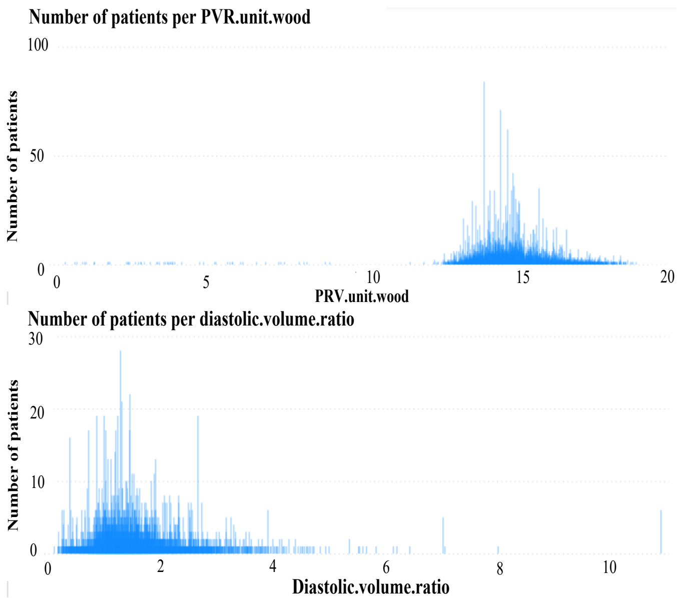

Fig. 3. Multivariate histograms of the number of patients, (top) as a function of PVR [Wood], (bottom) as a function of diastolic volume ratio.

data distributions concentrated in the ranges of 12 to 18 [Wood], establishing a valid variable quality to be considered relevant in the research. This concentration allows us to discard outlayer patients (higher than 6) to avoid confusion of the algorithms in the predictions.

C. Decision trees

The decision tree has a couple of advantages, the first, as already mentioned, have and easily interpretability for clinicians in decision making, the second is easy data manipulation, and the third is speed of execution and design. The disadvantage is an algorithm with low predictive power.

A pruning is performed on the best decision tree, however, if it is too high we can produce an overfitting. The hyperparameters used are adjusted for avoiding overfitting. Overfitting is not observed.

According to the algorithm, the patient can be classified with an AUC (91.2%) for RV involvement, based on the interpretation of the provided tree, Figure 4:

- RV involvement (Abnormal): If PVR.en.units.wood $>$ 15.221 + Height $>$ 1.40m + DoA.PA $>$ 0.087 + Aortic.arch $>$ 14.
- No RV involvement (normal): If PVR.en.units.wood ¡ 15.221 and Ao.reg.Vol.beat.vol.dif $>$-90.500.

With this philosophy, decision trees are generated to generate visualisations that follow the branches to generate those visualisations on the medical side. On the technical side we use more complex but less visual algorithms to fit higher quality predictor data. Another important variable in the decision tree is Aortic Regurgitation Volume by Beat-Volume Difference (Ao.reg.Vol.beat.vol.dif).

The following tree would be a similar interpretation, if we remove the variable PVR.in.wood.units, the variable reason.of.diastolic.volumes stands out, but the accuracy is reduced to 66.6% in the decision tree and 78.03% in the Gradient Boosted Trees.

If we include the variables gender and age, which were not initially included because they are used to classify the DV affectations, the results do not change. However, if we prune the tree, the variables age and gender are included, but they are not decisive, and a test with grouping by decades of age is carried out to ensure that they win.

The left ventricular involvement variables have been included; the algorithm does not identify them as a key element for right ventricular involvement. Although it is true that there is joint involvement and influence.

Previous studies with ML algorithms in cardiology achieve predictions greater than 90%, as they focus on achieving the best results, not on their interpretability [28]. It is interesting to see that our study achieves close values with interpretability by combining both tools (RapidMiner and Power BI). The results obtained are in line with recent research, which states that neural networks are the most predictive of cardiac parameters [29] [30].

This investigation has some relevant points. The main one is the creation of a tool to support clinical diagnosis that cardiologists can use for the prescription of new tests and a more detailed follow-up, similar to previous experiences already carried out [4] [31]. A second point is the creation of a visual interface, which allows dynamic monitoring, which facilitates dynamic interpretation, generating reports of high statistical value. It is worth exploring new algorithms to improve the interpretation of those key factors in the involvement of the right ventricle, thus improving the possible diagnosis. Tests could be carried out with different databases to create an algorithm that is robust enough to be able to limit any bias that the database used may contain. On the other hand, the authors aim to create a standardized protocol of measurements and tests that is carried out in daily clinical practice. The benefits of data analysis in cardiology using these types of techniques are evident, allowing them to increase the quality of diagnosis, prognosis and therapy.

IV. CONCLUSION

The ML algorithms and decision trees described, demonstrate the ability to identify the variables of greatest weight in right ventricular involvement with a reduced number of parameters. As a result, the number of diagnostic tests and their associated times can be reduced, allowing faster intervention in ischaemic and non-ischaemic cardiomyopathy. Therefore, the main objective of determining the clinical parameters that influence the involvement of the right ventricle in ischaemic and non-ischaemic cardiomyopathy is fulfilled.

The results show the influence of pulmonary vascular resistance and aortic artery beat volume difference as key factors, reaching an AUC of 87,7% with XGBoost in decision trees. Other variables with height (BMI), DoA.PA and Ao.reg.Vol.beat.vol.dif stand out as variables with a high weight in the predictions.

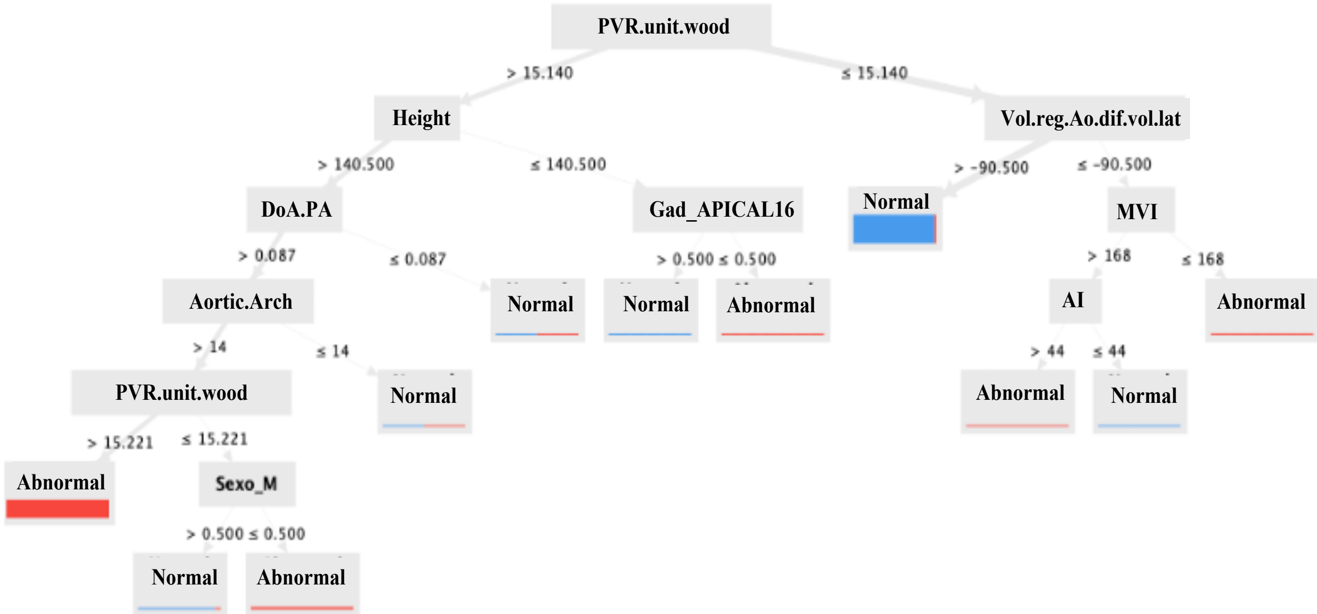

Fig. 4. Decision tree pruned in RapidMiner.

.

Author contributions

CB designed the article, performed the analysis in Power BI and drafted the text. HE co-managed the development of the article and obtained the database permissions. EP performed the ML analysis with RapidMiner. JJB engineering and medical design consultancy with descriptive analysis of the database. JVM interpretation of medical results. DM engineering design consultancy. MPL interpretation of medical results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research is funded by the European University through the project "Identification of factors influencing right ventricular involvement in ischaemic and non-ischaemic cardiomyopathy" through the project 2022/UEM18. The authors of this study wish to express their gratitude to ASCIRES Biomedical Group for the confidentiality agreement reached with the European University; Silvia Ruiz-España (Universitat Politècnica de València) for the documentation compiled from the database; to the advice of the interdisciplinary working group of Machine Learning Health-UEM; as well as to many other researchers for being a source of inspiration in the convergence of technology applied to health.

References

[1] O. W. Health, "The top 10 causes of death in the world." OWH, 2020. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] E. Boot, M. S. Ekker, J. Putaala, S. Kittner, F. E. De Leeuw, and A. M. Tuladhar, "Ischaemic stroke in young adults: a global perspective," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 91, no. 4, pp. 411–417, 2020.

[3] L. Cavigli, M. Focardi, M. Cameli, G. E. Mandoli, S. Mondillo, and F. D'Ascenzi, "The right ventricle in "left-sided" cardiomyopathies: the dark side of the moon," *Trends in Cardiovascular Medicine*, vol. 31, no. 8, pp. 476–484, 2021.

[4] J. J. Beunza, E. Puertas, E. García-Ovejero, F. Villalba, E. Condes, G. Koleva, C. Hurtado, and M. F. Landecho, "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)," *Journal of biomedical informatics*, vol. 97, p. 103257, 2019.

[5] S. Martínez-Agüero, A. G. Marques, I. Mora-Jiménez, J. Álvarez Rodríguez, and C. Soguero-Ruiz, "Data and network analytics for covid-19 icu patients: A case study for a spanish hospital," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 12, pp. 4340–4353, 2021.

[6] R. Cuocolo, M. B. Cipullo, A. Stanzione, L. Ugga, V. Romeo, L. Radice, A. Brunetti, and M. Imbriaco, "Machine learning applications in prostate cancer magnetic resonance imaging," *European radiology experimental*, vol. 3, no. 1, pp. 1–8, 2019.

[7] A. García-García, I. Prieto-Egido, A. Guerrero-Curieses, J. R. Feijoo-Martínez, S. Muñoz-Romero, S. Fernández-Manzano, P. J. Flores-Blanco, J. L. Rojo-Álvarez, and A. Martínez-Fernández, "Data science analysis and profile representation applied to secondary prevention of acute coronary syndrome," *IEEE Access*, vol. 9, pp. 78 607–78 620, 2021.

[8] D. Andriansyah and L. Nulhakim, "The application of power business intelligence in analyzing the availability of rental units," in *Journal of Physics: Conference Series*, vol. 1641, no. 1. IOP Publishing, 2020, p. 012019.

[9] M. Mariani, R. Baggio, M. Fuchs, and W. Höepken, "Business intelligence and big data in hospitality and tourism: a systematic literature review," *International Journal of Contemporary Hospitality Management*, vol. 30, no. 12, pp. 3514–3554, 2018.

[10] A. Sánchez-Ferrer, H. Pérez-Mendoza, and P. Shiguihara-Juárez, “Data visualization in dashboards through virtual try-on technology in fashion industry,” in *2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*. IEEE, 2019, pp. 1–6.
[11] M. Ratia, J. Myllärniemi, and N. Helander, “The new era of business intelligence: Big data potential in the private health care value creation,” *Meditari Accountancy Research*, vol. 26, no. 3, pp. 531–546, 2018.
[12] C. R. Greyson, “Ventrículo derecho y circulación pulmonar: conceptos básicos,” *Revista española de cardiología*, vol. 63, no. 1, pp. 81–95, 2010.
[13] J. Sanz, D. Sánchez-Quintana, E. Bossone, H. J. Bogaard, and R. Naeije, “Anatomy, function, and dysfunction of the right ventricle: Jacc state-of-the-art review,” *Journal of the American College of Cardiology*, vol. 73, no. 12, pp. 1463–1482, 2019.
[14] S. Wang, H. Wang, M. Ng, Y. Tada, G. Pontone, J. Urmeneta, I. Saeed, H. Patel, C. Mariager, J. V. Monmeneu-Menadas *et al.*, “Quantification of myocardial blood flow using stress cardiac magnetic resonance for the detection of coronary artery disease,” *European Heart Journal-Cardiovascular Imaging*, vol. 24, no. Supplement_1, pp. jead119–375, 2023.
[15] J. M. Oakes, J. Xu, T. M. Morris, N. D. Fried, C. S. Pearson, T. D. Lobell, N. W. Gilpin, E. Lazartigues, J. Gardner, and X. Yue, “Effects of chronic nicotine inhalation on systemic and pulmonary blood pressure and right ventricular remodeling in mice,” *Hypertension*, vol. 75, no. 5, pp. 1305–1314, 2020.
[16] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, pp. 273–297, 1995.
[17] S. K. Murthy, “Automatic construction of decision trees from data: A multi-disciplinary survey,” *Data mining and knowledge discovery*, vol. 2, pp. 345–389, 1998.
[18] T. Ho-Kam, “Random decision forests,” in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.
[19] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[20] T. Lachev and E. Price, *Applied Microsoft Power BI Bring your data to life!* Prologika Press, 2018.
[21] V. Kotu and B. Deshpande, *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann, 2014.
[22] A. M. Maceira, S. K. Prasad, M. Khan, and D. J. Pennell, “Reference right ventricular systolic and diastolic function normalized to age, gender and body surface area from steady-state free precession cardiovascular magnetic resonance,” *European heart journal*, vol. 27, no. 23, pp. 2879–2888, 2006.
[23] S. E. Petersen, M. Y. Khanji, S. Plein, P. Lancellotti, and C. Bucciarelli-Ducci, “European association of cardiovascular imaging expert consensus paper: a comprehensive review of cardiovascular magnetic resonance normal values of cardiac chamber size and aortic root in adults and recommendations for grading severity,” *European Heart Journal-Cardiovascular Imaging*, vol. 20, no. 12, pp. 1321–1331, 2019.
[24] J. M. Gardin, M. K. Rohan, D. Davidson, A. Dabestani, M. Sklansky, R. Garcia, M. L. Knoll, D. White, S. K. Gardin, and W. L. Henry, “Doppler transmitral flow velocity parameters: relationship between age, body surface area, blood pressure and gender in normal subjects,” *American journal of noninvasive cardiology*, vol. 1, no. 1, pp. 3–10, 1987.
[25] J. Swan-Harold, W. Ganz, J. Forrester, H. Marcus, G. Diamond, and D. Chonette, “Catheterization of the heart in man with use of a flow-directed balloon-tipped catheter,” *New England Journal of Medicine*, vol. 283, no. 9, pp. 447–451, 1970.
[26] C. D. Vizza, I. M. Lang, R. Badagliacca, R. L. Benza, S. Rosenkranz, R. J. White, Y. Adir, A. K. Andreassen, V. Balasubramanian, S. Bartolome *et al.*, “Aggressive afterload lowering to improve the right ventricle: a new target for medical therapy in pulmonary arterial hypertension?” *American journal of respiratory and critical care medicine*, vol. 205, no. 7, pp. 751–760, 2022.
[27] A. García-Alvarez, L. Fernandez-Friera, J. G. Mirelis, S. Sawit, A. Nair, J. Kallman, V. Fuster, and J. Sanz, “Non-invasive estimation of pulmonary vascular resistance with cardiac magnetic resonance,” *European heart journal*, vol. 32, no. 19, pp. 2438–2445, 2011.
[28] T. Smole, B. Žunkovič, M. Pičulin, E. Kokalj, M. Robnik-Šikonja, M. Kukar, D. I. Fotiadis, V. Pezoulas, N. S. Tachos, F. Barlocco *et al.*, “A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy,” *Computers in biology and medicine*, vol. 135, p. 104648, 2021.
[29] P. Revuelta-Zamorano, A. Sánchez, J. L. Rojo-Álvarez, J. Álvarez-Rodríguez, J. Ramos-López, and C. Soguero-Ruiz, “Prediction of healthcare associated infections in an intensive care unit using machine learning and big data tools,” in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016: MEDICON 2016, March 31st-April 2nd 2016, Paphos, Cyprus*. Springer, 2016, pp. 840–845.
[30] R. Garcia Carretero, L. Vigil-Medina, O. Barquero-Perez, I. Mora-Jimenez, C. Soguero-Ruiz, and J. Ramos-Lopez, “Machine learning approaches to constructing predictive models of vitamin d deficiency in a hypertensive population: a comparative study,” *Informatics for Health and Social Care*, vol. 46, no. 4, pp. 355–369, 2021.
[31] A. Hernández-Casillas, . Del-Canto, S. Ruiz-España, M. P. López-Lereu, J. V. Monmeneu, and D. Moratal, “Detection and classification of myocardial infarction transmurality using cardiac mr image analysis and machine learning algorithms,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 1686–1689.