Decoding Key Variables Contributing to Right Ventricular Involvement in Ischaemic and Non-ischaemic Cardiomyopathy

Carlos Barroso-Moreno
Faculty of Biomedical and Health Sciences
Universidad Europea de Madrid
Madrid, Spain
22014885@live.uem.es

Hector Espinos Morato
Faculty of Engieering
Universidad Europea de Valencia
Valencia, Spain
hector.espinos@universidadeuropea.es

Enrique Puertas
S. Architecture, Engieering and Design
Universidad Europea de Madrid
Madrid, Spain
enrique.puertas@universidadeuropea.es

Juan José Beunza Nuin
Faculty of Biomedical and Health Sciences
Universidad Europea de Madrid
Madrid, Spain
juanjose.beunza@universidadeuropea.es

José Vicente Monmeneu
Cardiac Magnetic Resonance
ASCIRES Biomedical Group
Valencia, Spain
jmonmeneu@ascires.com

María P. López-Lereu
Cardiac Magnetic Resonance
ASCIRES Biomedical Group
Valencia, Spain
mplopezl@ascires.com

David Moratal
Center for Biomaterials and Tissue Engineering
Universitat Politècnica de València
Valencia, Spain
dmoratal@eln.upv.es
,

***Abstract*—Cardiomyopathy is a condition affecting the heart muscle, poses challenges to effective blood pumping by the heart. While prior research predominantly concentrated on the left ventricle, recent investigations underscore the significance of the right ventricle. This study aims to ascertain the clinical and cardiac parameters influencing right ventricular engagement in both ischaemic and non-ischaemic cardiomyopathy. A database comprising 56,447 subjects, collected between 2008 and 2020 by the ASCIRES Biomedical Group, forms the basis of this investigation. The methodology encompasses two main blocks: the clinical aspect utilizes decision trees for enhanced interpretability, while the technical aspect employs Machine Learning to achieve a higher degree of prediction accuracy. Power Business Intelligence and RapidMiner are the main software tools enabling data transformation and deep data analysis within the EU legal framework for health data. The outcomes reveal the pivotal influence of disparities in aortic artery beat volume, pulmonary vascular volume and aortic arch as key factors. Remarkably, the RapidMiner tool, employing the decision trees algorithm, attains an impressive Area Under the Curve (AUC) of 0.873 with XGBoost and decreases to AUC= 0.791 with SVM. In conclusion, the study underscores the ability to identify crucial clinical variables associated with right ventricular involvement, offering the potential to streamline diagnostic procedures and reduce associated timeframes in cardiomyopathy scenarios.**

***Keywords*—Cardiomyopathy; Machine Learning Algorithms; right ventricular; Pulmonary Vascular Resistance.**

I. INTRODUCTION

This paper is an extended and updated presentation of the research based on the publication and presentation at the HEALTHINFO 2023 conference in Valencia, Spain [1]. The extension allows to show details of the methodology used and a greater number of results.

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 [2]. 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 [3]. 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 [4]. 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 [5], ICU patients with Covid-19 [6], prostate cancer [7] 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: restaurant marketing [8], rental car [9], tourism [10] or fashion industry [11]. 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 [12].

The RapidMiner tool is a data analytics platform that offers manipulation and predictive analytics with a graphical interface. The platform is used to analyse large datasets, especially using decision trees in the health domain. Beyond healthcare, RapidMiner showcases its adaptability in diverse domains, exemplified by its utilization in social network analysis as demonstrated by Anand *et al.* [13], tourism research highlighted by Seovcanac *et al.* [14], and the burgeoning field of electric vehicles, as explored by Fernandes *et al.* [15]. The tool is expansive utility underscores its significance as a multifaceted instrument contributing to advancements in analytics across various sectors [16].

Previous studies have identified influential variables in right ventricular compromise, such as: pulmonary arterial hypertension (PAH) associated with pressure overload [17]; diabetes, dyslipidemia [18], blood flow [19] and habits, such as smoking [20], family history of disease [21] and others. These multifaceted insights from previous research collectively contribute to a comprehensive understanding of the diverse variables influencing right ventricular health.

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. Section II presents the details of dataset and it describes the methodology for Power BI and RapidMiner. Section III describes the results focusing on significant variables, distributions and decision trees. Finally, Section IV presents the conclusions and directions for future work.

II. MATERIALS AND METHODS

The employed methodology involves a comprehensive evaluation of the prevalent supervised classification algorithms in ML: Support Vector Machines (SVM) [22], decision trees [23], Random Forest [24], and neural networks [25]. The assessment of results encompasses the utilization of key metrics such as precision, sensitivity, specificity, and the Area Under the Curve (AUC). This last variable is a pivotal measure to evaluate the performance of binary classification models.

The dataset is sourced from the ASCIRES Biomedical Group database, comprising 56,447 records encompassing variables collected meticulously between 2008 and 2020.

A. Overall methodology flow

The flow of methodology in the research project is visually presented in Figure 1. The first step corresponds to the extraction of data from the ASCIRES servers. The application of the ETL process allows managing server extractions, customised transformations and loads for a first exploratory analysis [26]. The transformations seek a pre-processing of the data to adapt it to the legal framework with secure encryption for variables such as names, surnames, addresses or removal of the date of birth by the variable 'age', among others. The next point of the process focuses on improving the quality of the data, considering aspects such as the consistency of dates of birth, age, adaptation of formats or grouping of similar categories that are differentiated by orthographic criteria (such as, for example, Fusion or Fusion). Finally, validations of duplicate patients are performed using the unique patient ID. For the exploratory analysis, statistical mathematics such as mean, median or histograms, among others, are applied. Likewise, the multivariate algorithms of Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) allow us to detect the variables that provide the most information.

The second step is the division of the resulting database into an 80% for training and a 20% for testing the algorithm. This division is not static since the algorithms randomly select patients for training and testing.

The third step corresponds to the development and integration of the ML algorithms, which are: Linear binary SVM, binary classification tree, neural networks, Naive Bayes and discriminant analysis. The choice of these algorithms is based on medical and technical criteria, with simple algorithms (SVM) seeking to explain the results, and complex neural network algorithms achieving the highest scores in the evaluation. This situation allows the creation of feature vectors.

Fourth to model cross-validation is a technique for evaluating ML models by training various subsets of available input data and evaluating them with other complementary subsets of data. The operation of cross-validation is divided into four steps: division into subsets of data, the default is 10 folds; 1 fold is reserved for validation and the remaining folds for training the model; this process is repeated as many times as folds are available, evaluating the accuracy; finally, a set of performance metrics and results are generated from all the data. This technique allows to assess the reliability of any

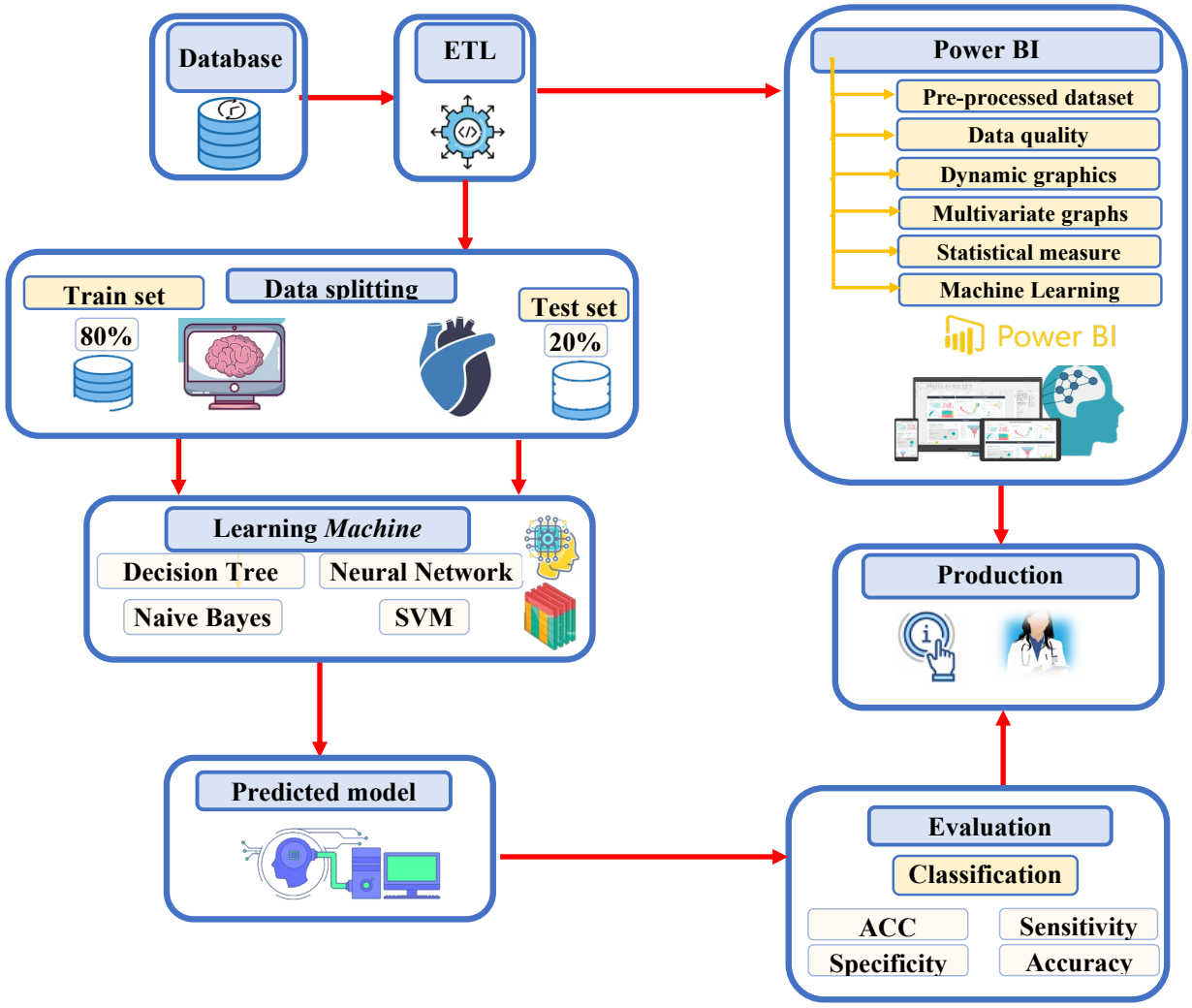

Figure 1. Overview of the flow of methodology in the research project.

trained model by testing the variability of the dataset, thus obtaining the final values for the evaluation. This requires a high computational cost due to random partitioning and use of a larger number of data, by training and evaluating the model several times [27].

The fifth step corresponds to the prediction model generated. The generation of the model requires a high computational cost of more than 1 hour of computation with the equipment used. The advantage lies in the speed of the result when introducing a new patient for prediction, in tenths of a second. Therefore, the models require a high computational load to be generated, but their medical application is instantaneous, fulfilling the medical requirement of minimising waiting times.

The sixth and seventh steps are the model scoring results and the model evaluation criteria, respectively. Each model has its own peculiarities; as a general rule, the higher the complexity of the algorithm, the higher the prediction score, but also the lower the control of the understanding of the learning process, becoming black boxes for the researcher and clinician. The evaluation measures for the algorithms are mainly: ACC, SE, SP, NPV and AUC [28]. At this point it is necessary to take into account different results for each algorithm applied and also the different models of the database as input parameter. All this is done in RapidMiner.

The last step corresponds to the entry into production of the algorithms, providing health professionals with the results of the predictor in an informative way. The advantage of the methodological scheme used is the speed of the results, cost savings and the ability to diagnose ischaemic and non-ischaemic cardiomyopathy.

Another complementary methodology employed in the research is the Power BI business tool developed by Microsoft, which allows the integration of various data analysis and results gathering techniques. The first step is the loading of data into the Power BI tool with connectors already integrated in the tool. In this loading, data pre-processing is carried out, such as the formatting of variables, or the elimination of duplicates with the selection of options, among others. This

automation makes it possible to maintain high data quality. Subsequently, various types of graphs are created: column, linear, combined, circular, ring, card, scatter, bubble, meter, etc. All the graphs constructed are dynamic and interactive, an advantage of Power BI, allowing the researcher to filter the graphs according to their research needs. In addition, there are a series of ML algorithms implemented in Power BI that allow quantifying relationships of variables with exitus. Finally, all this information is accessible to healthcare professionals. In Figure 1 a dashed line is drawn to production because this analysis is investigative and time-consuming, unlike ML algorithms designed to provide real-time feedback on patient survival.

B. Software used

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

- Power BI is a data analytic service from Microsoft that provides interactive graphs focused on analytic intelligence to generate reports [29]. 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 [30]. Version 9.10.013 is used to obtain the results of the ML models.

C. 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 1815 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 $\rho = 0.97$

D. 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 [31] 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) [32].

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 [33]. The 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 6 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

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) [31].

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)

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

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

step so as not to distort the output of the ML algorithms in later steps.

In the first tests, the variables in table are eliminated due to their direct relationship in 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.

E. Legal framework and data security

The General Data Protection Regulation (GDPR) is a European regulation that regulates the protection of natural persons in the processing of personal data and the free movement of such data, which includes the law Regulation (EU) 2016/679 of the European Parliament and of the Council [34]. The Ley Orgánica de Protección de Datos Persoonales y Garantía de los Derechos Digitales (LOPD-GDD) is the Spanish Organic Law 3/2018, of 5 December, which aims to guarantee and protect the processing of personal data according to the guidelines set out in the RGPD [35]. In Spain, the public body in charge of ensuring compliance with this law is the Spanish Data Protection Agency (AEPD). Both European and national legislation are present during the research, and some highlights are mentioned below. The consent of the patients according to the legal basis for the treatment must be explicit for which the data are used, at the point that affects the university the research purpose is preserved, so the authorisation of the patients is not needed again. Furthermore, there is an action protocol for cases of loss or theft of data, which is developed

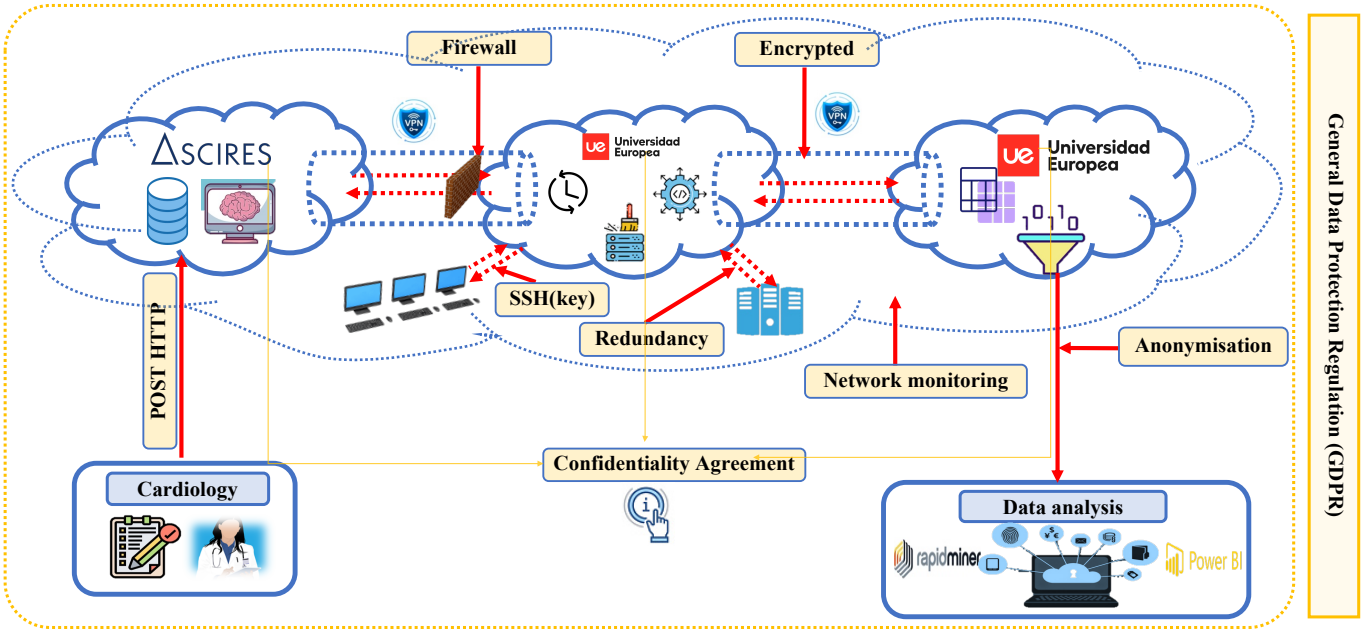

Figure 2. Overview of the legal framework and data security in research project.

by the Data Protection Delegate.

After mentioning the legislation in force and their respective protocols for action, Figure 2 compiles the techniques used both to comply with the legislation and the proactive part of security. The description in the figure follows the data flow in the process. This process starts after a new stroke patient is admitted and the physician enters the data in a form that is sent with the HTTPS protocol to the ASCIRES systems. This protocol prevents other users from intercepting the confidential information between the client and the server.

Once the data has been uploaded to the system, ASCIRES generates Virtual Private Network (VPN) to the EMU systems to upload new data on a regular basis. During this connection, a Firewall is set up to ensure that only the right person and files receive the data, and to block access to unauthorised users [36]. Researchers accessing such data need to have the SSH key, as this protocol establishes secure communications between two systems using client-server architecture to connect to the machine remotely. In addition, regular backups and checks allow for the creation of a backup system, the so-called redundancy. The company periodically carries out a vulnerability analysis to identify possible loopholes and to address them.

Different public and private institutions are involved in data traffic, which is why the confidentiality and non-disclosure agreement covers issues related to the database.

The data security with the various protocols ensures high security of digital privacy to prevent unauthorised access to the data, so it is worth mentioning that the description contains the protocols and techniques used, which can be detailed. A part of the process is omitted due to the confidentiality and privacy agreements of all parties.

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 Table IV, which are Accuracy (ACC), Sensitivity (SE), Specificity (SP), Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Area Under the Curve (AUC).

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.

Nevertheless, exercising caution is paramount when directly comparing the algorithms. While XGBoost yields the most favorable outcomes in this study, the inherent complexity of ML algorithms often renders them as ”black boxes,” challenging the interpretation of their inner workings. Conversely, decision trees, while achieving comparatively lower results, offer a distinct advantage in terms of interpretability, providing a valuable clinical perspective. The efficiency of the algorithms, coupled with the optimization techniques and the size of the database, ensures swift processing times, consistently under three minutes. These computations were conducted on a MacBook Pro 2.6 GHz Intel Core i7 6-core computer with 6 GB 2400 MHz DDR4 memory.

TABLE IV
COMPARISON AND EVALUATION OF DIFFERENT ALGORITHMS IN RAPIDMINER.

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

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

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 $\overline{X} = 1.59$ ($SD = 0.71$).

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 [37], 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) [38], increasing by 0.2 Wood in subjects over 50 years of age.

The database provided uses an estimation model to obtain the PVR value, described in equation 2 [39]:

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

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

B. Distributions

The original database model, which is not illustrated in the document, records an AUC value of 98% primarily due to the high proportion of patients in the database who do not show right ventricular (RV) involvement, causing imbalance in the data analysis. Consequently, the algorithms default to estimating a normal value, leading to a high accuracy rate (with sensitivity nearly at 0). However, through meticulous data filtering, the database achieves a balanced representation, mitigating this potential issue in the analysis of observations. Another notable aspect of the distributions involves excluding data during filtering if any of the variables are empty. This exclusion addresses the risk of bias introduced by eliminating patients who inherently did not complete data due to specific circumstances. Given the substantial volume of data, this

does not impede the analysis; instead, it enhances algorithmic performance, yielding more robust and accurate results.

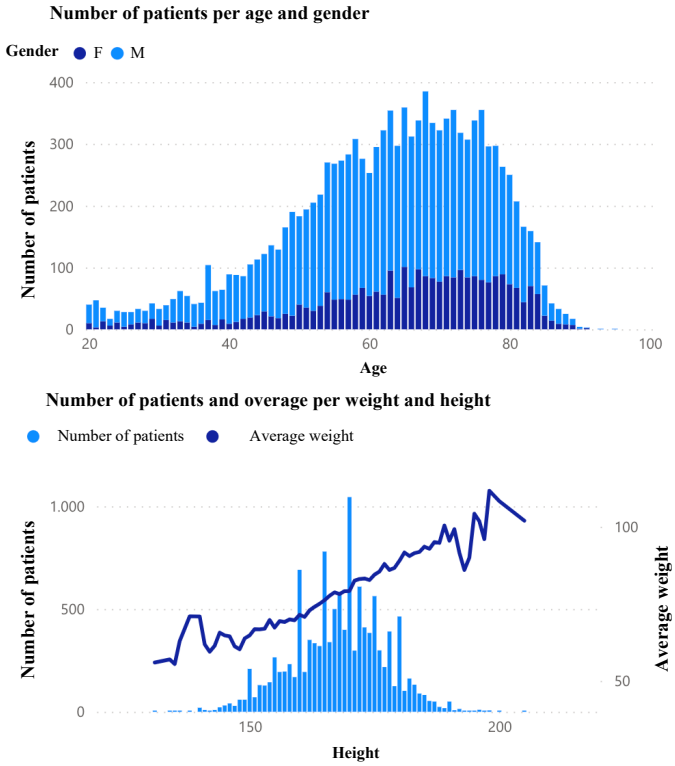

Figure 3. Multivariate histograms of number of patients, (top) age as a function of gender, (bottom) height as a function of mean weight.

The use of Power BI allows for quick visualisation of these and many other complexities, mitigating the risk of bias or inconsistencies in the data. Illustrated in figure 40, the histogram shows the distribution of patients by age and gender, highlighting the correlation between older age and higher incidence, especially among men. It should be noted that the age range between 60 and 80 years comprises almost half of the cases of right ventricular (RV) involvement. For height and weight, a discernible correlation emerges, prompting the use of body mass index (BMI) as a predictive variable in database labelling.

Figure 4 is the multivariate exploration of Right Ventricular End-Diastolic Volume (RVEDV) based on gender and patient count. The analysis distinctly reveals that the female anatomy tends to exhibit a smaller volume compared to male anatomy. This nuanced understanding is pivotal in appropriately labeling Right Ventricular (RV) involvement and aligns with findings from prior studies. The lower section of the figure depicts the scatter plot representing the relationship between Non-Right Ventricular End-Diastolic Volume (NRVEDV) and Indexed Cardiothoracic Surface Volume Difference (ICTSVD). This analysis serves to identify outliers, aiding in the exclusion of patients with significantly distorted values. The use of these graphs facilitates the establishment of acceptable ranges for each variable without adversely impacting the algorithms.

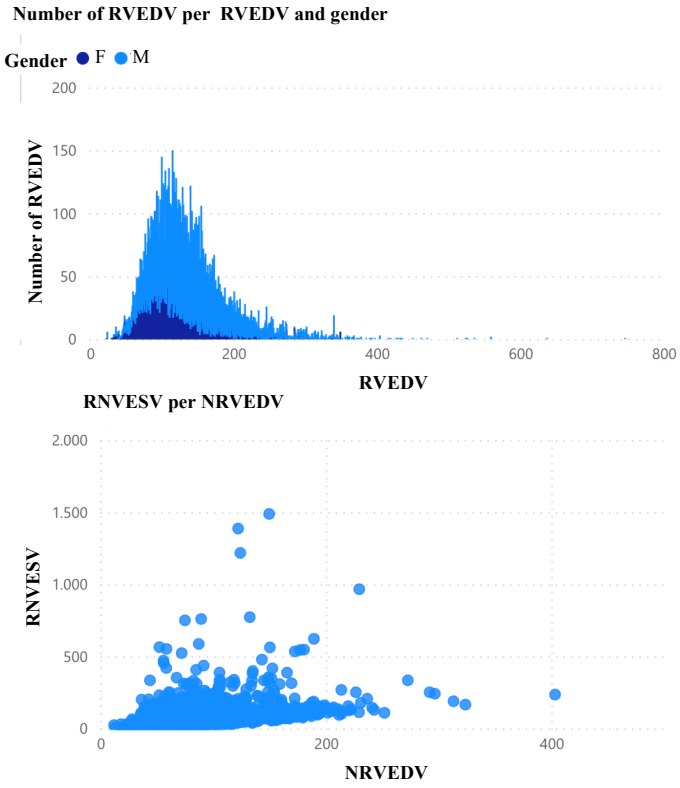

Figure 4. Multivariate histogram of the number of patients, (top) RVEDV by gender, and scatter plot of NRVEDV with RNVESV.

Furthermore, the interactive nature of the graphs streamlines the process of filtering out outliers, allowing for consensus-building with medical professionals regarding their exclusion based on clinical criteria and predefined ranges for each variable. This iterative procedure is replicated for all variables employed in the data labeling process.

In accordance with PVR [Wood], as illustrated in Figure 5, the data distributions are notably concentrated within the ranges of 12 to 18 [Wood]. This concentration validates PVR as a pertinent variable meriting consideration in the research. Such focus within a specific range facilitates the exclusion of outlier patients, particularly those with values exceeding 6. This deliberate exclusion is crucial for preventing algorithmic confusion during predictions and ensuring the reliability of the research findings.

C. Key influencer in Power BI

Power BI services have powerful features such as key influencer visualisation that allow you to detect and understand the factors that drive results from the database [40]. Figure 6 provides the relationships of the two most influential variables to understand the key concepts, in this case corresponding to the PVR variable in wood units and the Aortic Regurgitation Volume by Beat-Volume Difference (Ao.reg.Vol.beat.vol.diff) or also known as ARVBVD. Technically, Power BI uses decision trees that evaluate the impact of variables on the target metric to identify and rank key factors. Determining

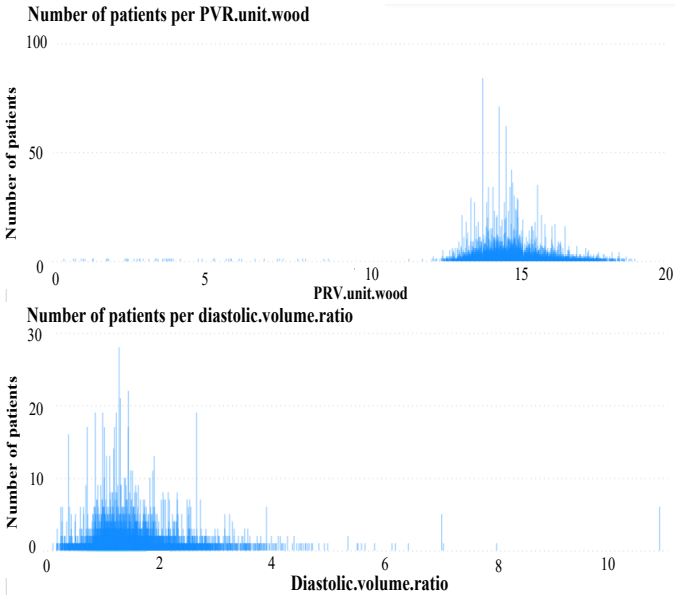

Figure 5. Multivariate histograms of the number of patients, (top) as a function of PVR [Wood], (bottom) as a function of diastolic volume ratio.

that the behaviour of a patient with Aortic Regurgitation Volume greater than 33 applies a factor of 0.59 on each PVR unit, if it is less than -34 it affects the same. Finally, Aortic Regurgitation Volume less than -46 applies a factor of 0.34 in each PVR unit. Both variables, when calculated in tree are similar to those determined in RapidMiner in Figure 9.

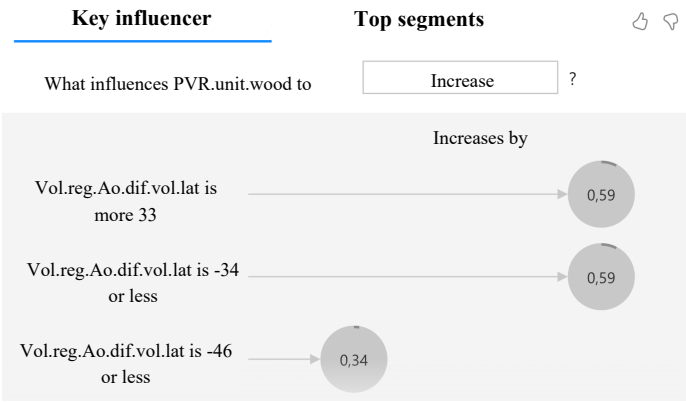

Figure 6. Influencing factors as a function of two key variables

Similarly, there is a cluster analysis tool that allows the identification of natural groups or clusters within the dataset providing a clearer structure of the internal organisation [41]. Figure 7 represents three clusters detected on the two most influential variables detected in Figure 9. The first cluster shows how ARVBVD values in the upper ranges $\pm 100 - 40$ with PVR values below 10 [Wood] are quite dispersed. If we decrease the ARVBVD range to $\pm 40 - 20$ and values between 10-30 [Wood] of PVR it starts to converge. The third cluster converges, finding the ARVBVD range at $\pm 20 - 0$ and values above 30 [Wood] PVR, this means a severity in patients with such values because it means an elevated pulmonary pressure with a low volume difference.

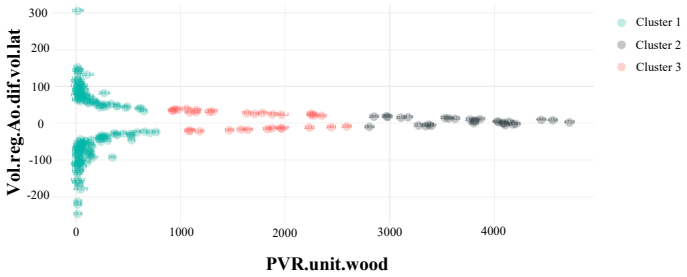

Figure 7. Automatic cluster detector

Another technique for pattern identification is Principal Component Analysis (PCA). The technique allows to reduce the dimensionality of the dataset by identifying the main directions of variability in the data. The 8 shows the two main components, which contribute more than half of the information, assuming PC1 = 37.98% and PC2= 26.10% of the information contribution. These results indicate that the information is concentrated in a few variables, as determined above.

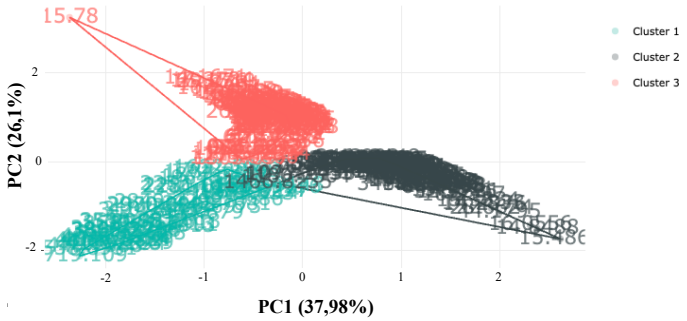

Figure 8. Cluster-based PCA analysis

D. Decision trees in RapidMiner

The decision tree has several advantages. First, as noted above, it is easy to interpret, which facilitates the decision-making process for clinicians. Second, it allows for easy manipulation of the data. Third, it excels in both speed of execution and efficiency of design. However, a notable drawback is its relatively low predictive power.

The solution to this problem is to perform pruning to optimise the best decision tree. However, care must be taken to avoid overfitting, as excessive pruning can lead to this problem. The hyperparameters are carefully adjusted to mitigate the risk of overfitting, and the analysis reveals no observable cases of overfitting in the decision tree model.

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 9:

- RV involvement (Abnormal): If PVR.en.units.wood $>$ 15.221 + Height $>$ 1.40m + DoA.PA $>$ 0.087 + Aortic.arch $>$ 14.

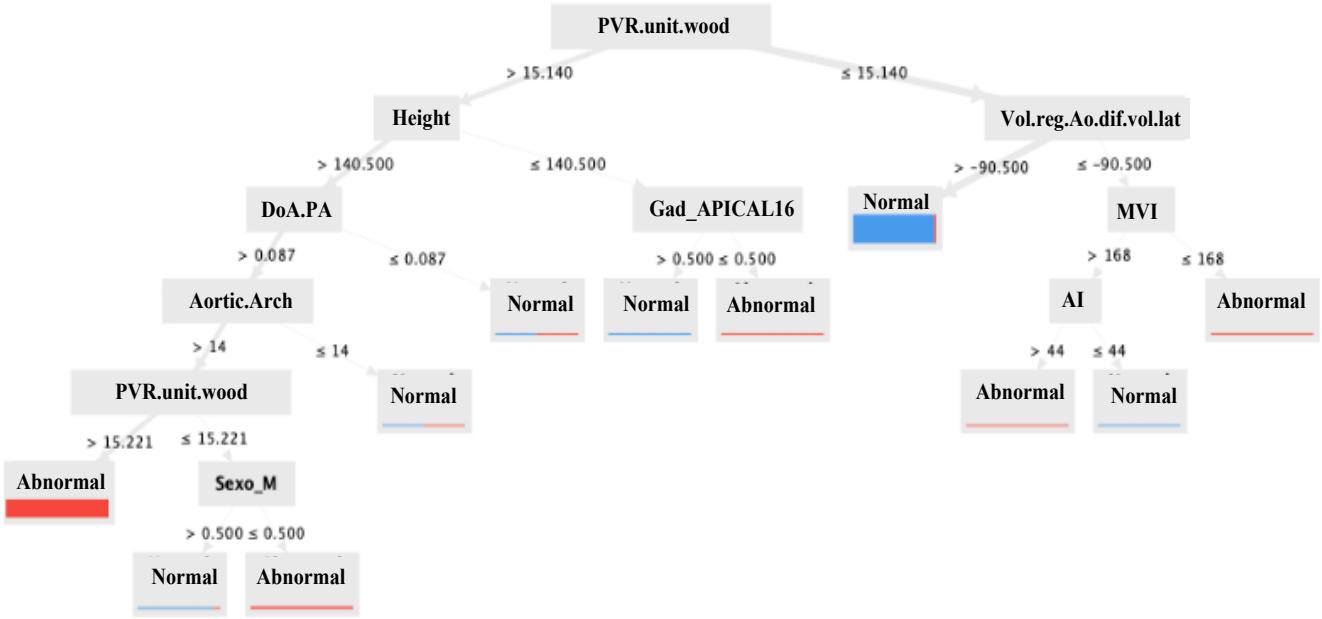

Figure 9. Decision tree pruned in RapidMiner.

- 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 variables related to left ventricular involvement have been incorporated into the analysis. However, the algorithm does not recognize them as primary determinants for right ventricular involvement. It is essential to acknowledge the interplay and mutual influence between these variables, even though the algorithm does not designate them as key elements specifically for right ventricular involvement. This underscores the complexity of the interactions between left and right ventricular aspects, warranting further investigation into their collective impact on cardiac health.

Previous studies with ML algorithms in cardiology achieve predictions greater than 90%, as they focus on achieving the best results, not on their interpretability [42]. 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 [43] [44].

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 [28] [45]. 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 presented in this research show a remarkable ability to discern the most influential variables associated with right ventricular involvement, performing this task with a concise set of parameters. This

efficiency translates into the potential reduction of both the number of diagnostic tests and their corresponding durations, facilitating expeditious interventions in cases of ischaemic and non-ischaemic cardiomyopathy. Consequently, the primary objective of identifying clinical parameters that influence right ventricular involvement in these specific cardiomyopathies is successfully met.

The findings highlight the substantial influence of pulmonary vascular resistance and the difference in aortic artery beat volume, serving as pivotal factors. Notably, the XGBoost algorithm within decision trees attains an AUC of 87.3%. Additionally, variables such as height (BMI), DoA.PA, and Ao.reg.Vol.beat.vol.dif emerge as influential factors, contributing significantly to the predictive of the model.

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] C. Barroso-Moreno, H. Espinos-Morato, E. Puertas, J. J. Beunza-Nuin, J. V. Monmeneu, D. Moratal, and M. P. López-Lereu, "Identifying key factors in right ventricular involvement in ischaemic and non-ischaemic cardiomyopathies," in *HEALTHINFO 2023, The Eighth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*. IARIA, 2023, pp. 1–8.

[2] 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

[3] 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.

[4] 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.

[5] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthcare Analytics*, vol. 2, p. 100032, 2022.

[6] 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.

[7] S. L. Goldenberg, G. Nir, and S. E. Salcudean, "A new era: artificial intelligence and machine learning in prostate cancer," *Nature Reviews Urology*, vol. 16, no. 7, pp. 391–403, 2019.

[8] K. K. Halim, S. Halim *et al.*, "Business intelligence for designing restaurant marketing strategy: A case study," *Procedia Computer Science*, vol. 161, pp. 615–622, 2019.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] V. Anand and S. Ghosh, "Sentiment analysis of social network for information professionals: A case study of lislinks discussion forum," *International Journal of Information Science and Management (IJISM)*, pp. 63–74, 2024.

[14] M. Seočanac and D. Dimitrovski, "Factors influencing tourists' nightlife experience in belgrade," *Consumer Behavior in Tourism and Hospitality*, vol. 17, no. 4, pp. 436–452, 2022.

[15] J. Fernandes Andry, J. Gunadi, G. Dwinoor Rembulan, and H. Tannady, "Big data implementation in tesla using classification with rapid miner," *International Journal of Nonlinear Analysis and Applications*, vol. 12, no. Special Issue, pp. 2057–2066, 2021.

[16] B. R. Pagliaro, F. Cannata, G. G. Stefanini, and L. Bolognese, "Myocardial ischemia and coronary disease in heart failure," *Heart Failure Reviews*, vol. 25, no. 1, pp. 53–65, 2020.

[17] C. R. Greyson, "Ventrículo derecho y circulación pulmonar: conceptos básicos," *Revista Española de Cardiología*, vol. 63, no. 1, pp. 81–95, 2010.

[18] 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.

[19] 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. 119–375, 2023.

[20] 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.

[21] J. Ingles, C. Burns, R. D. Bagnall, L. Lam, L. Yeates, T. Sarina, R. Puranik, T. Briffa, J. J. Atherton, T. Driscoll *et al.*, "Nonfamilial hypertrophic cardiomyopathy: prevalence, natural history, and clinical implications," *Circulation: Cardiovascular Genetics*, vol. 10, no. 2, p. e001620, 2017.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[23] 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.

[24] T. Ho-Kam, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[25] 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.

[26] V. Manickam and M. Rajasekaran Indra, "Dynamic multi-variant relational scheme-based intelligent etl framework for healthcare management," *Soft Computing*, vol. 27, no. 1, pp. 605–614, 2023.

[27] M. W. Browne, "Cross-validation methods," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 108–132, 2000.

[28] 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.

[29] T. Lachev and E. Price, *Applied Microsoft Power BI Bring your data to life!* Prologika Press, 2018.

[30] V. Kotu and B. Deshpande, *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann, 2014.

[31] 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.

[32] 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.

[33] 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.

[34] E. Commission, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016," European Commission, Tech. Rep., 2021.

[35] J. Calvo-López and B. Varela-Adsuara, *La adaptación al nuevo marco de protección de datos tras el RGPD y la LOPDGDD*. Wolters Kluwer España ISBN: 978-84-9090-345-2, 2019.

[36] P. Nyakomitta and S. Abeka, "Security investigation on remote access methods of virtual private network," *Global Journal of Computer Science and Technology*, 2020.

[37] 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.

[38] 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.

[39] 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.

[40] M. Ehrenmueller-Jensen and M. Ehrenmueller-Jensen, "Discovering key influencers," *Self-Service AI with Power BI Desktop: Machine Learning Insights for Business*, pp. 53–80, 2020.

[41] B. Powell, *Mastering Microsoft Power BI: expert techniques for effective data analytics and business intelligence*. Packt Publishing Ltd, 2018.

[42] 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.

[43] 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.

[44] 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.

[45] A. Hernández-Casillas, S. 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.