APPLY MACHINE LEARNING TO PREDICT CARDIOVASCULAR RISK IN RURAL CLINICS FROM MEXICO

Misael Zambrano-de la Torre (1), Maximiliano Guzmán-Fernández (2), Claudia Sifuentes-Gallardo (3), Hamurabi Gamboa-Rosales (4), Huizilopoztli Luna-García (5), , Ernesto Sandoval-García (6), Ramiro Esquivel-Felix (7) and Héctor Durán-Muñoz (8)

Misael Zambrano-de la Torre

Unidad Académica de Ingeniería Eléctrica. Universidad Autónoma de Zacatecas, Zacatecas, Zac., México

4 - 5 AUGUST 2021

CONTENTS

1 INTRODUCTION

2 OBJECTIVES

3 LITERATURE REVIEW/JUSTIFICATIONS

4 METHODOLOGY

5 RESULTS AND DISCUSSIONS

6 CONCLUSIONS

Introduction

Currently, according to the World Health Organization, degenerative diseases cause the death of approximately 41 million people per year. From this total, cardiovascular diseases account for 17.9 million deaths worldwide. In the case of Mexico, according to the National Institute of Statistics and Geography (INEGI), in 2019, there are a total of 747, 784 deaths. Cardiovascular diseases occupy 23.5% (156, 041 people), from 88.8% of the total deaths in 2019. This is why artificial intelligence has been looking to develop solutions through the use of machine learning algorithms.

Objectives

The objective of this work is to propose a new stage in the methodology used in machine learning for the classification of cardiovascular risk in rural clinics in Mexico.

Classify patients using only attributes denoted as "non-invasive".

Verify the feasibility of the new stage through a previously validated database.

Obtain a classification without wasting a lot money, time and computational cost.

LITERATURE REVIEW/JUSTIFICATIONS

Some examples about the use of machine learning to classify cardiovascular disease were found.

Mezzatesta et al., In their work, it was reported that using the Support Vector Machine algorithm, a performance in terms of accuracy of 95.25% was obtained following a five-stage methodology.

Dimopoulos et al., They developed an algorithm capable of classifying with an accuracy of 95% following a five-stage methodology.

These stages are:

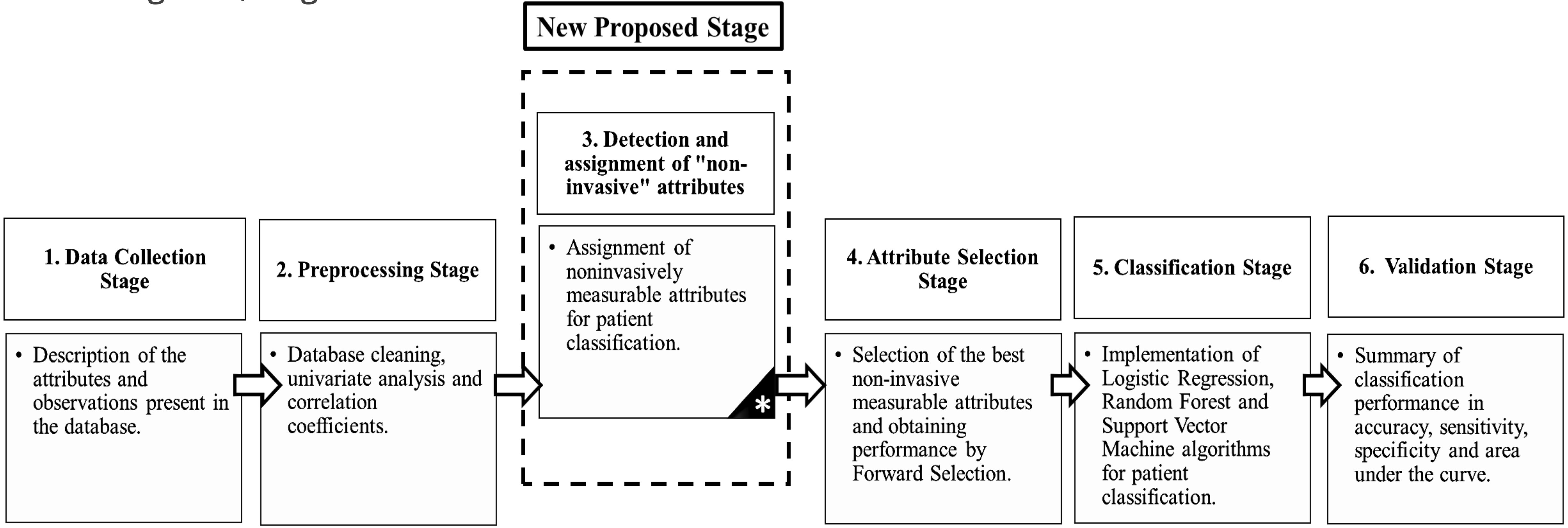
database description, pre-processing, attribute description, algorithm training and model validation.

S. Mezzatesta, C. Torino, P. De Meo, G. Fiumara, and A. Vilasi, “A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis,” *Comput. Methods Programs Biomed.*, vol. 177, pp. 9–15, Aug. 2019, doi: 10.1016/j.cmpb.2019.05.005.

A. C. Dimopoulos *et al.*,“Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk,” *BMC Med. Res. Methodol.*,vol. 18, no. 1, p. 179, Dec. 2018, doi: 10.1186/s12874-018-0644-1.

METHODOLOGY

The methodology used in this work is based in the literature. The proposed new stage is shown in Figure 1, stage three.



New Proposed Stage
1. Data Collection Stage
• Description of the attributes and observations present in the database.
2. Preprocessing Stage
• Database cleaning, univariate analysis and correlation coefficients.
3. Detection and assignment of "non-invasive" attributes
• Assignment of noninvasively measurable attributes for patient classification.
*
4. Attribute Selection Stage
• Selection of the best non-invasive measurable attributes and obtaining performance by Forward Selection.
5. Classification Stage
• Implementation of Logistic Regression, Random Forest and Support Vector Machine algorithms for patient classification.
6. Validation Stage
• Summary of classification performance in accuracy, sensitivity, specificity and area under the curve.

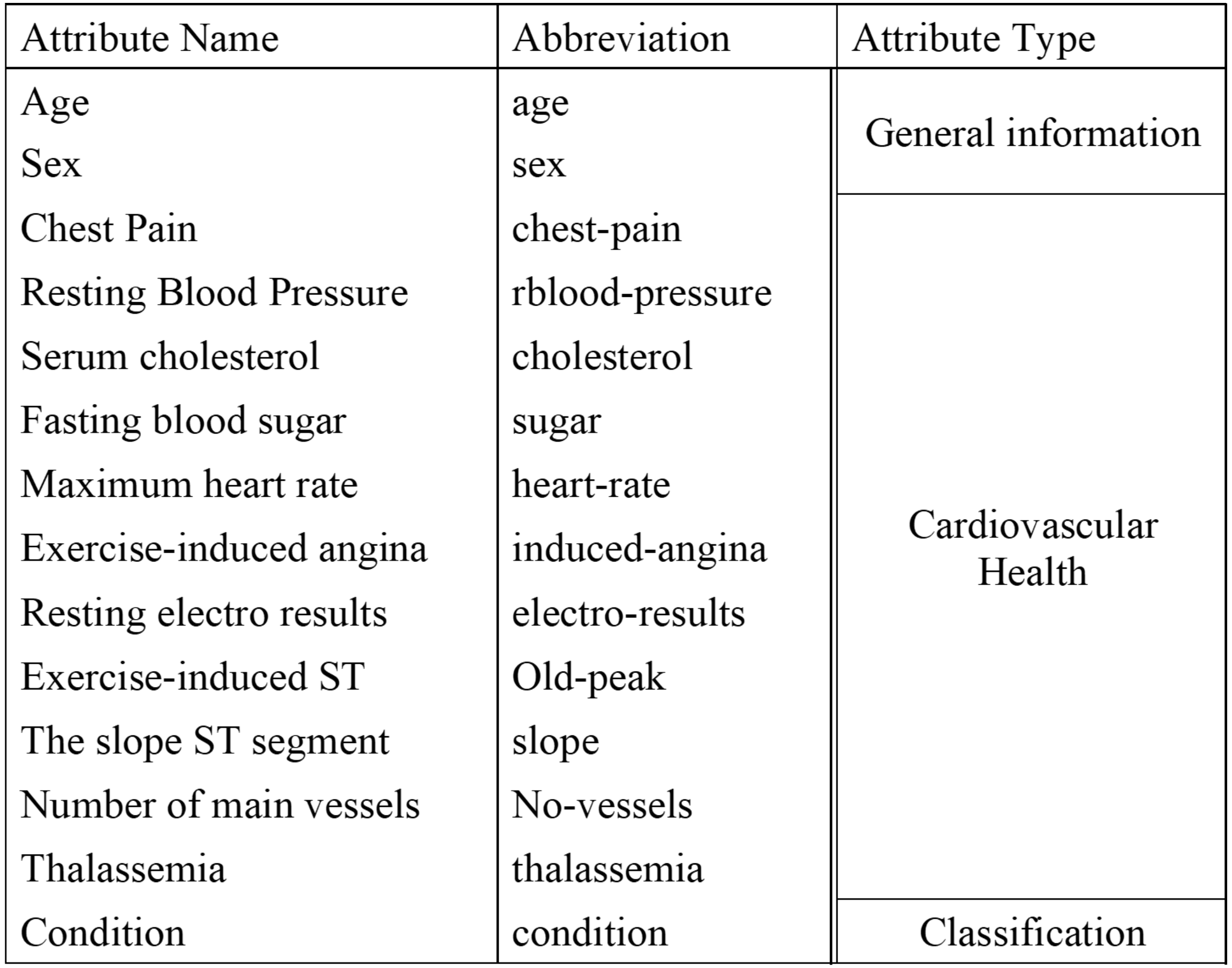
Figure 1: Methodology for classifying cardiovascular risk, incorporating the new stage 3: detection of "noninvasive" attributes

METHODOLOGY

"The Heart Disease Data Set" from the CAD repository from the University of California was used.

This database is previously validated data available for free use in research.

For practical purposes in this work, the file corresponding to the fourteen attributes was extracted directly from the repository.

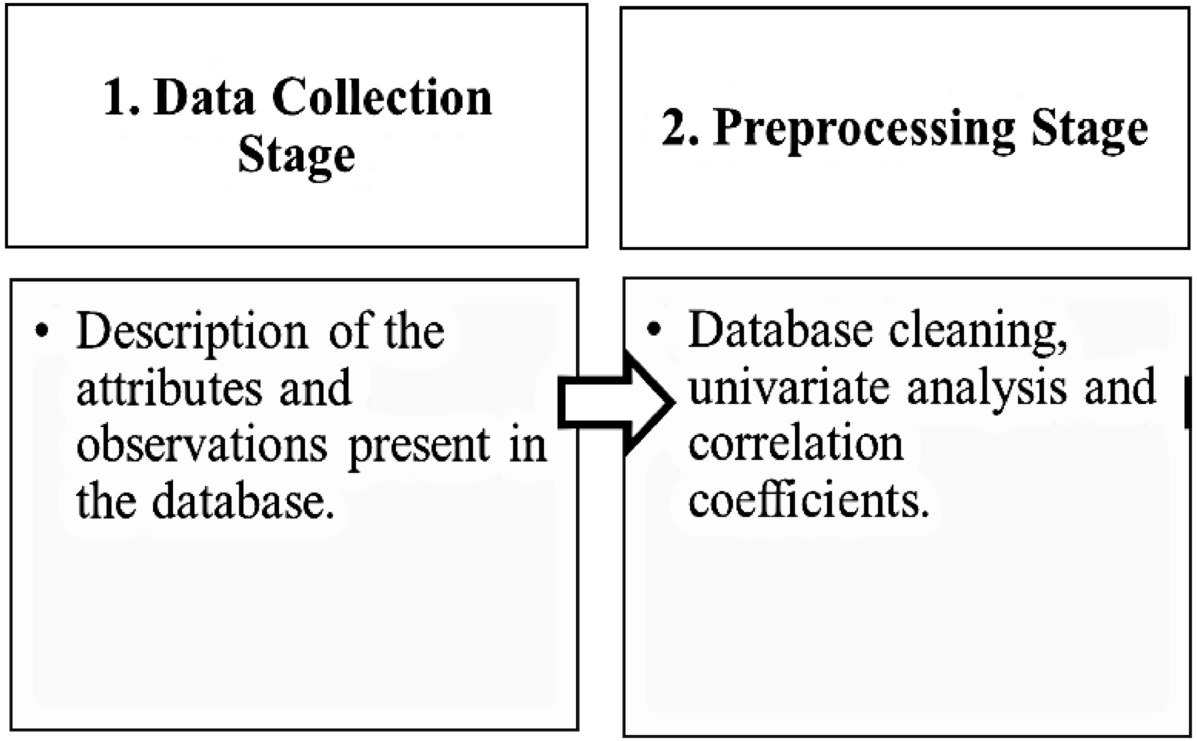


Attribute Name	Abbreviation	Attribute Type
Age	age	General information
Sex	sex	
Chest Pain	chest-pain	Cardiovascular Health
Resting Blood Pressure	rblood-pressure	
Serum cholesterol	cholesterol	
Fasting blood sugar	sugar	
Maximum heart rate	heart-rate	
Exercise-induced angina	induced-angina	
Resting electro results	electro-results	
Exercise-induced ST	Old-peak	
The slope ST segment	slope	
Number of main vessels	No-vessels	
Thalassemia	thalassemia	
Condition	condition	Classification

Figure 2: Attributes used in this work.

METHODOLOGY

Erroneous data in the database was cleaned and debugged. Univariate analysis was performed to see the distribution of the attributes.



1. Data Collection Stage

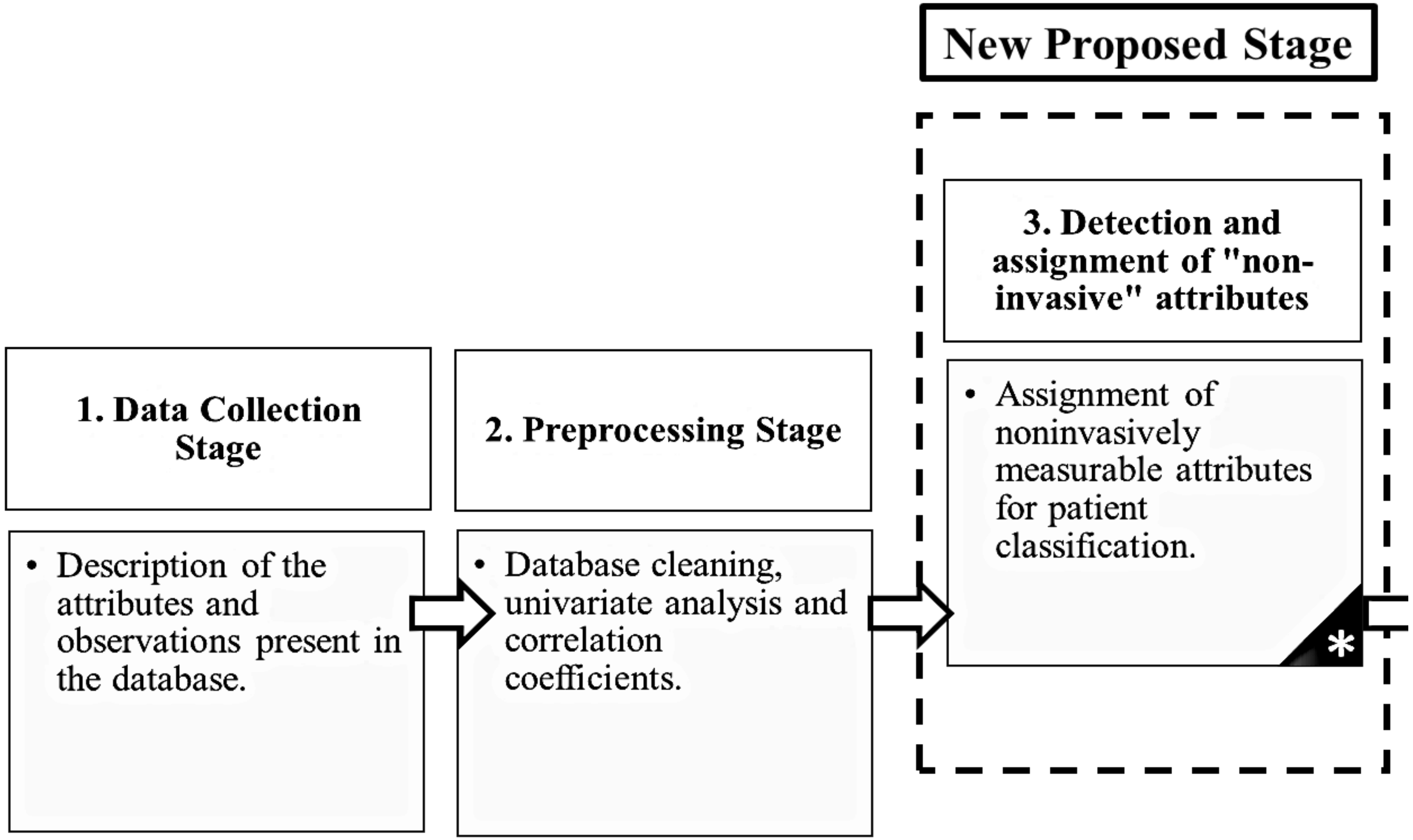
- Description of the attributes and observations present in the database.

2. Preprocessing Stage

- Database cleaning, univariate analysis and correlation coefficients.

Figure 3: Second Stage.

METHODOLOGY



New Proposed Stage

1. Data Collection Stage

- Description of the attributes and observations present in the database.

2. Preprocessing Stage

- Database cleaning, univariate analysis and correlation coefficients.

3. Detection and assignment of "non-invasive" attributes

- Assignment of noninvasively measurable attributes for patient classification.

Figure 4: New stage in the methodology.

Commonly in the literature an indiscriminate selection of attributes is made, without considering the complications to obtain such attributes. For this reason, a new third stage is proposed.

METHODOLOGY

At this stage, eight attributes were selected. In order that the patient does not require clinical analysis or interventions.

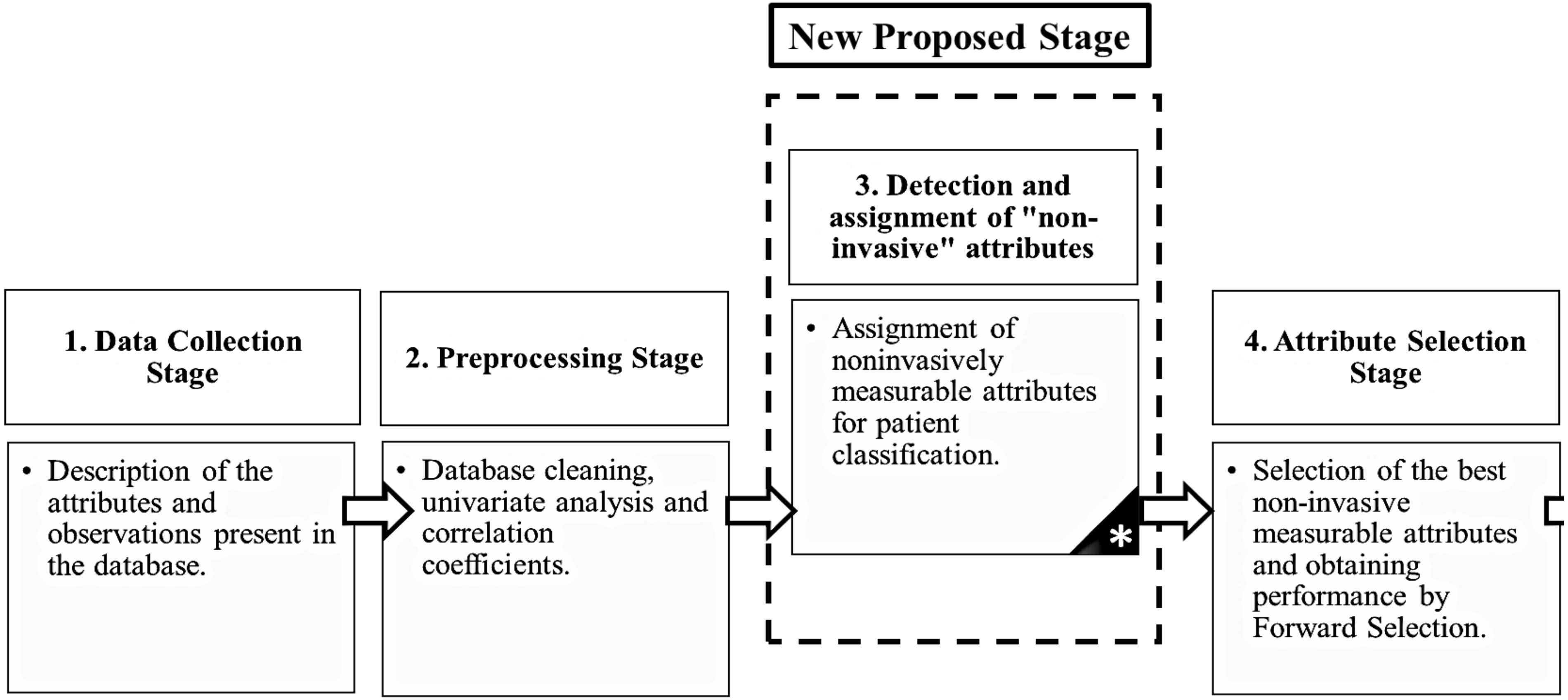
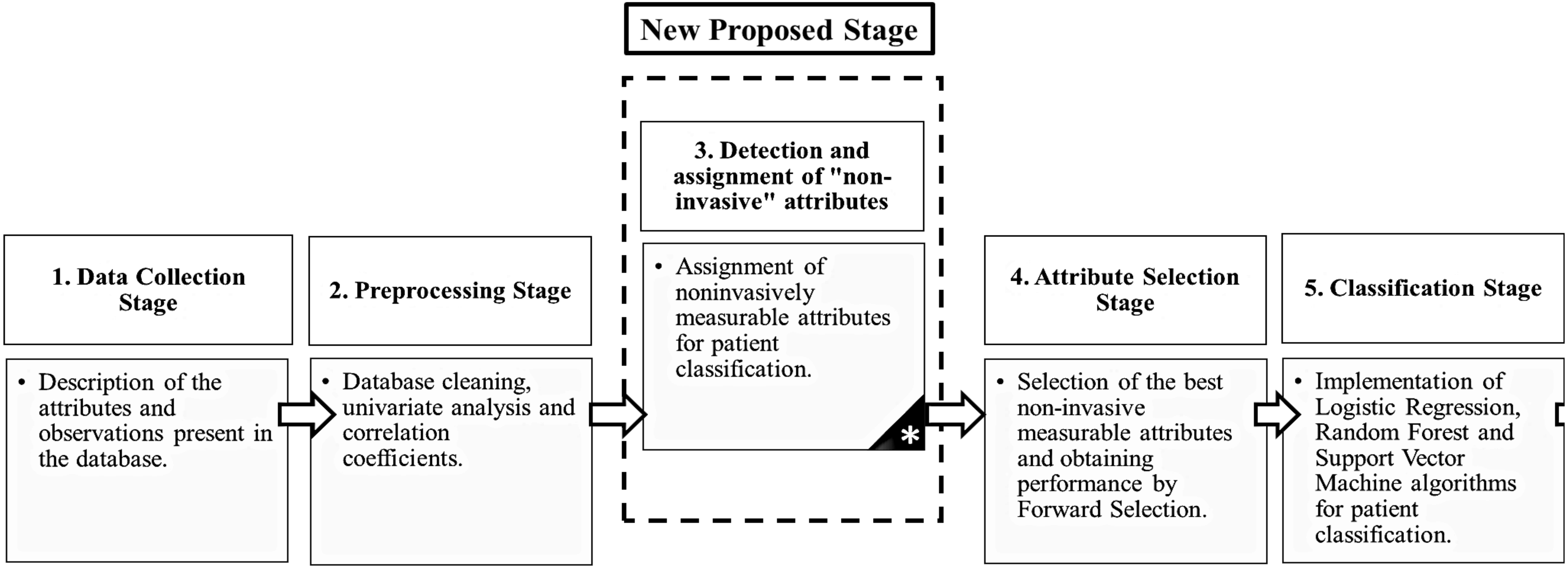

New Proposed Stage
1. Data Collection Stage
- Description of the attributes and observations present in the database.
2. Preprocessing Stage
- Database cleaning, univariate analysis and correlation coefficients.
3. Detection and assignment of "non-invasive" attributes
- Assignment of noninvasively measurable attributes for patient classification.
*
4. Attribute Selection Stage
- Selection of the best non-invasive measurable attributes and obtaining performance by Forward Selection.

Figure 5: fourth stage.

METHODOLOGY

After attribute selection, patient classification is applied using three different classification algorithms: multivariable Logistic Regression (LR), Random Forest (RF) and Support Vector Machines (SVM).



New Proposed Stage

1. Data Collection Stage
- Description of the attributes and observations present in the database.

2. Preprocessing Stage
- Database cleaning, univariate analysis and correlation coefficients.

3. Detection and assignment of "non-invasive" attributes
- Assignment of noninvasively measurable attributes for patient classification.

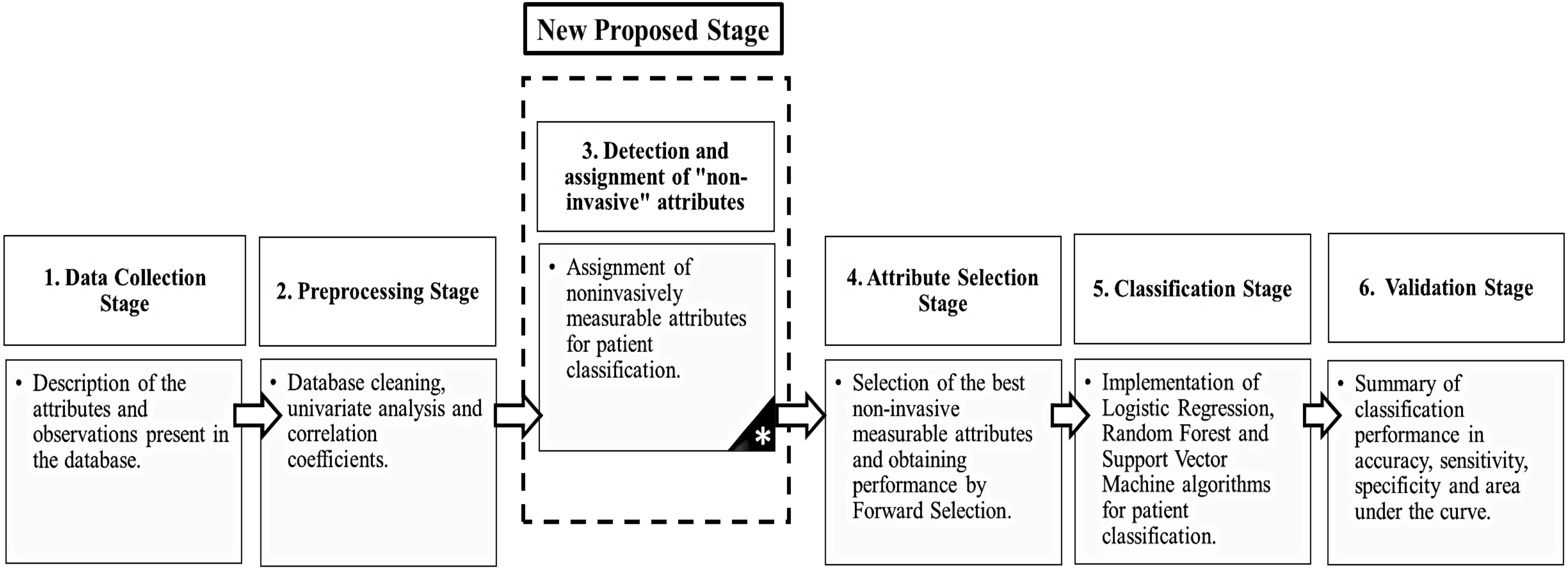
4. Attribute Selection Stage
- Selection of the best non-invasive measurable attributes and obtaining performance by Forward Selection.

5. Classification Stage
- Implementation of Logistic Regression, Random Forest and Support Vector Machine algorithms for patient classification.

Figure 6: fifth stage.

METHODOLOGY

In the sixth stage, the performance of the algorithms is compared.



New Proposed Stage

1. Data Collection Stage
- Description of the attributes and observations present in the database.

2. Preprocessing Stage
- Database cleaning, univariate analysis and correlation coefficients.

3. Detection and assignment of "non-invasive" attributes
- Assignment of noninvasively measurable attributes for patient classification.

*

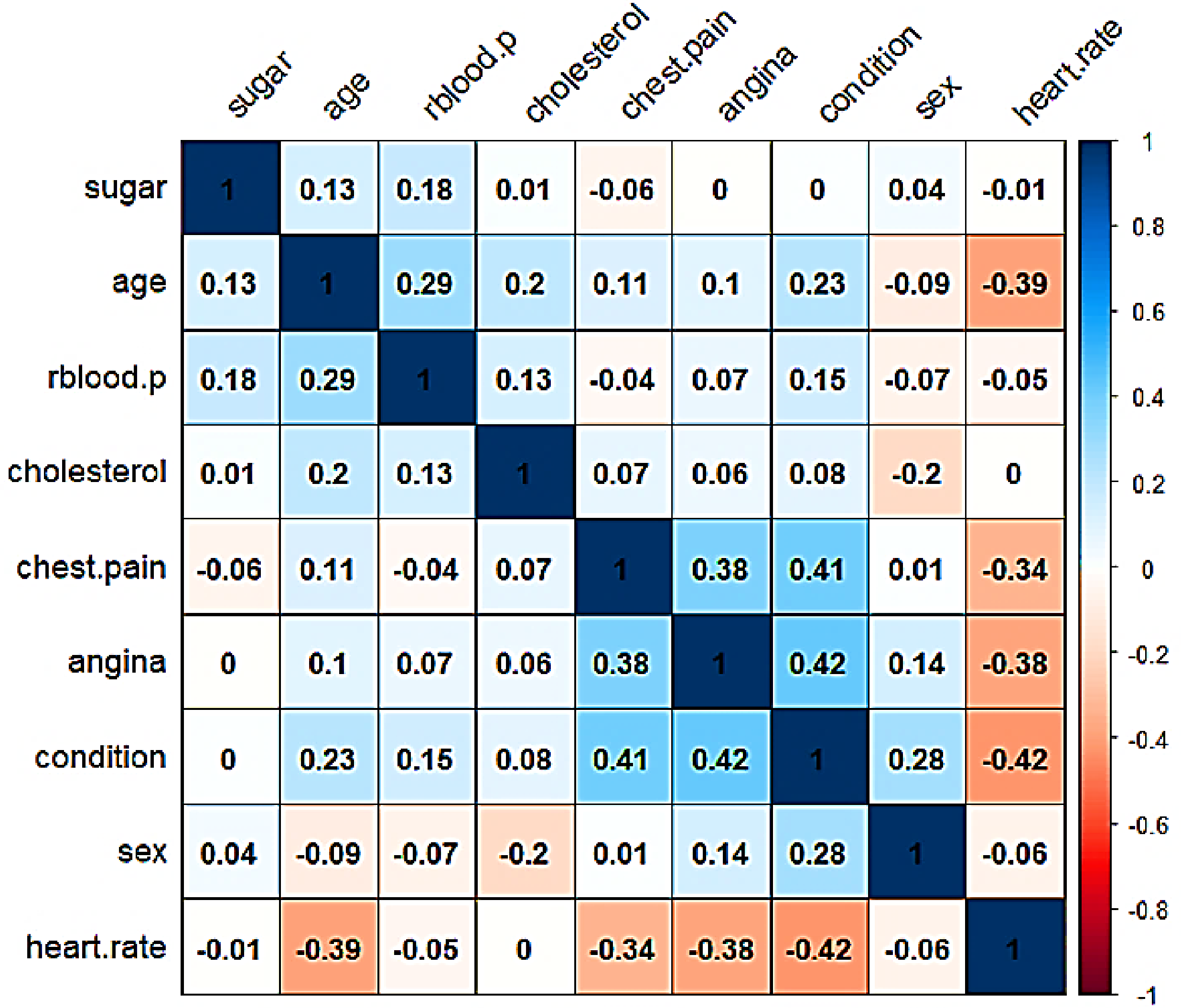
4. Attribute Selection Stage
- Selection of the best non-invasive measurable attributes and obtaining performance by Forward Selection.

5. Classification Stage
- Implementation of Logistic Regression, Random Forest and Support Vector Machine algorithms for patient classification.

6. Validation Stage
- Summary of classification performance in accuracy, sensitivity, specificity and area under the curve.

Figure 7: Final stage.

RESULTS AND DISCUSSIONS



	sugar	age	rblood.p	cholesterol	chest.pain	angina	condition	sex	heart.rate
sugar	1	0.13	0.18	0.01	-0.06	0	0	0.04	-0.01
age	0.13	1	0.29	0.2	0.11	0.1	0.23	-0.09	-0.39
rblood.p	0.18	0.29	1	0.13	-0.04	0.07	0.15	-0.07	-0.05
cholesterol	0.01	0.2	0.13	1	0.07	0.06	0.08	-0.2	0
chest.pain	-0.06	0.11	-0.04	0.07	1	0.38	0.41	0.01	-0.34
angina	0	0.1	0.07	0.06	0.38	1	0.42	0.14	-0.38
condition	0	0.23	0.15	0.08	0.41	0.42	1	0.28	-0.42
sex	0.04	-0.09	-0.07	-0.2	0.01	0.14	0.28	1	-0.06
heart.rate	-0.01	-0.39	-0.05	0	-0.34	-0.38	-0.42	-0.06	1

Figure 8: Heat map of the correlation matrix of eight non-invasively measurable attributes.

RESULTS AND DISCUSSIONS

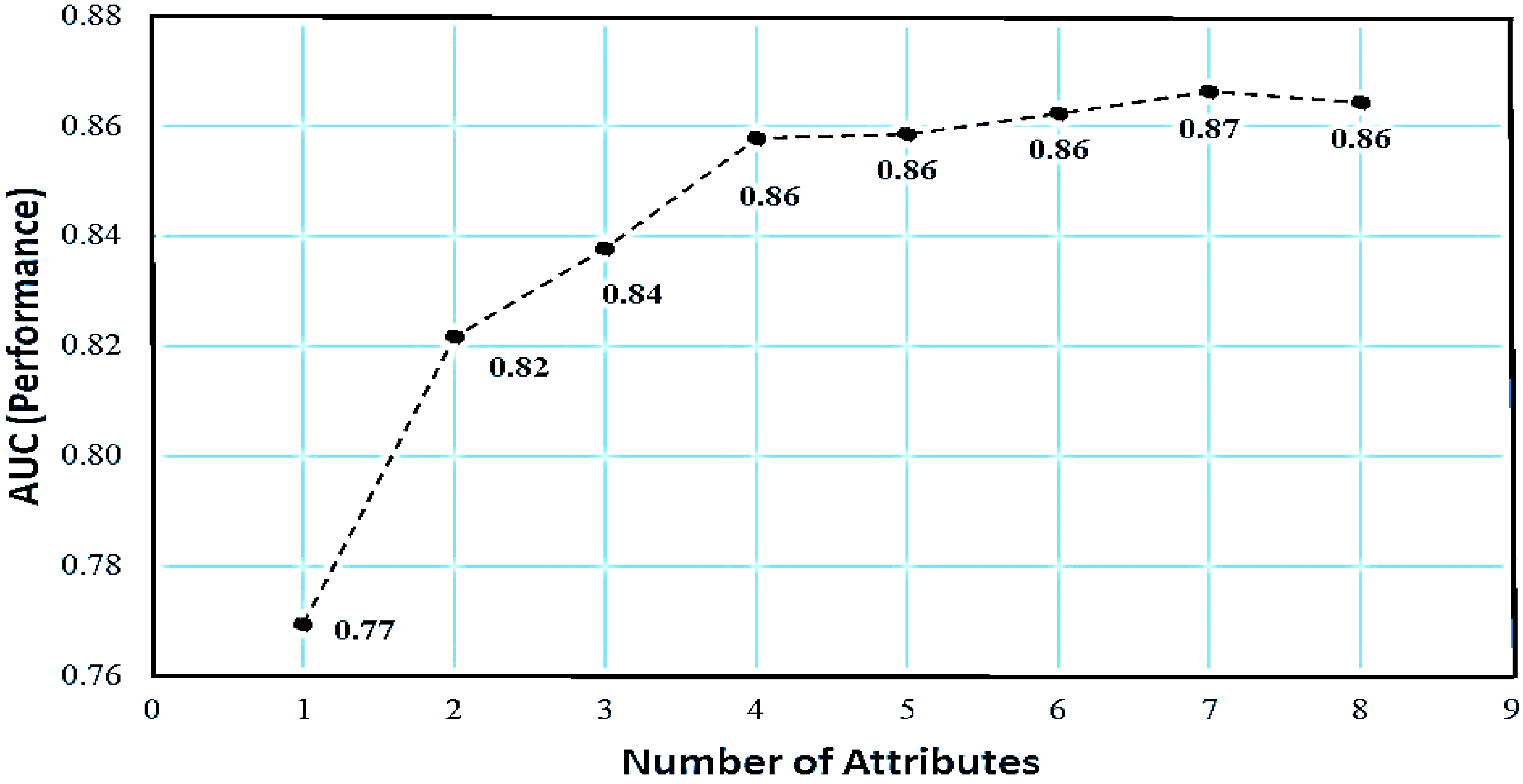
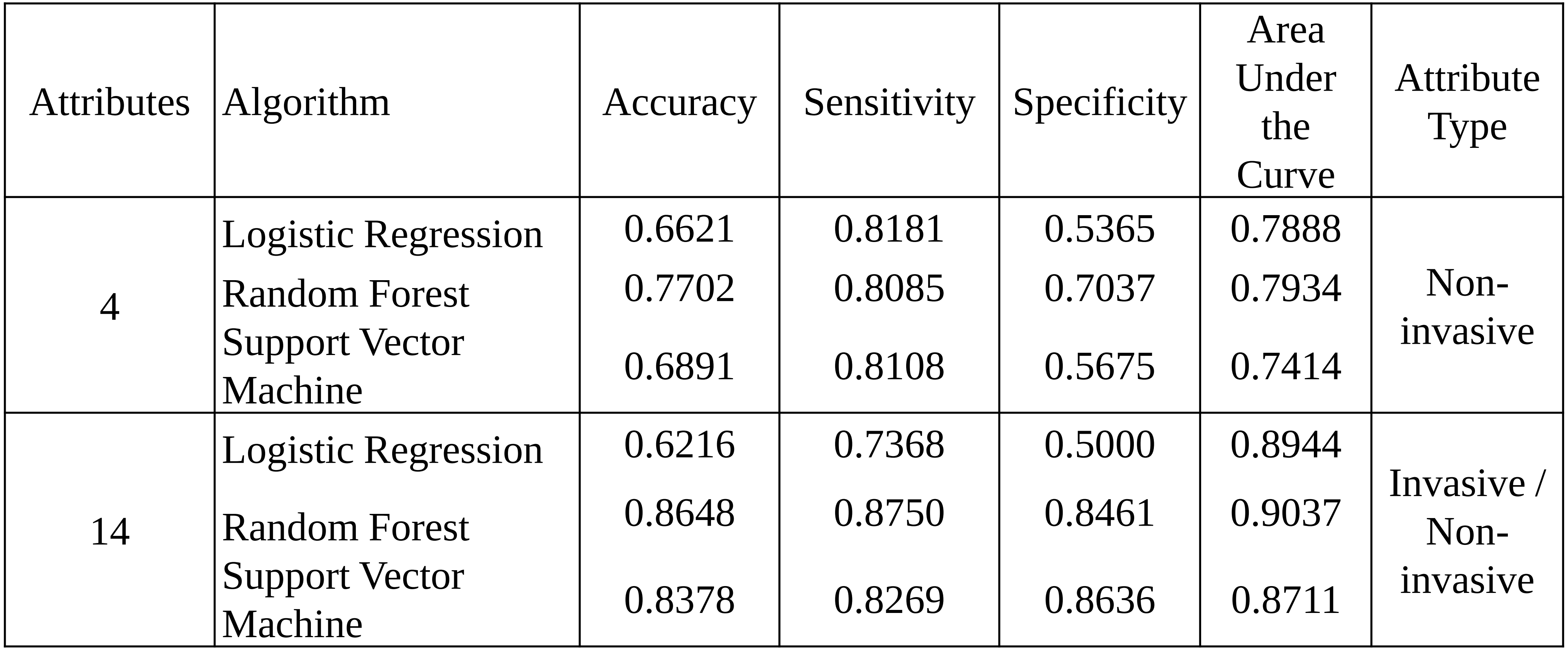


Figure 9: Forward selection in terms of the area under the curve of 8 attributes.

RESULTS AND DISCUSSIONS



Attributes	Algorithm	Accuracy	Sensitivity	Specificity	Area Under the Curve	Attribute Type
4	Logistic Regression	0.6621	0.8181	0.5365	0.7888	Non-invasive
	Random Forest	0.7702	0.8085	0.7037	0.7934	
	Support Vector Machine	0.6891	0.8108	0.5675	0.7414	
14	Logistic Regression	0.6216	0.7368	0.5000	0.8944	Invasive / Non-invasive
	Random Forest	0.8648	0.8750	0.8461	0.9037	
	Support Vector Machine	0.8378	0.8269	0.8636	0.8711	

Figure 10: Performance comparison.

CONCLUSIONS

It is concluded that the new proposed stage proves to be useful, acceptable, easy to implement and reproduce.

it is concluded that a reduction of approximately 70% of attributes was achieved in exchange for a loss in performance of 10%, in terms of the area under the curve.

In the future, this research is looking to generate a device that is low cost and easy to reproduce. This will allow classification to be performed even faster.

THANK YOU