Health Informatics Journal

Original Article

Linear discriminant analysis and principal component analysis to predict coronary artery disease

Health Informatics Journal
2020, Vol. 26(3) 2181–2192

DOI: 10.1177/1460458219899210
journals.sagepub.com/home/jhi
SAGE

Carlo Ricciardi
University Hospital of Naples 'Federico II', Italy

Antonio Saverio Valente
University of Naples 'Federico II', Italy

Kyle Edmunds
Reykjavík University, Iceland; University of Oxford, UK

Valeria Cantoni
Roberta Green
Antonella Fiorillo
Ilaria Picone
University Hospital of Naples 'Federico II', Italy

Stefania Santini
Mario Cesarelli
University of Naples 'Federico II', Italy

Abstract

Coronary artery disease is one of the most prevalent chronic pathologies in the modern world, leading to the deaths of thousands of people, both in the United States and in Europe. This article reports the use of data mining techniques to analyse a population of 10,265 people who were evaluated by the Department of Advanced Biomedical Sciences for myocardial ischaemia. Overall, 22 features are extracted, and linear

Corresponding author:
Carlo Ricciardi, Department of Advanced Biomedical Sciences, University Hospital of Naples 'Federico II', Naples 80131, Italy.
Email: carloricciardi.93@gmail.com

discriminant analysis is implemented twice through both the Knime analytics platform and R statistical programming language to classify patients as either normal or pathological. The former of these analyses includes only classification, while the latter method includes principal component analysis before classification to create new features. The classification accuracies obtained for these methods were 84.5 and 86.0 per cent, respectively, with a specificity over 97 per cent and a sensitivity between 62 and 66 per cent. This article presents a practical implementation of traditional data mining techniques that can be used to help clinicians in decision-making; moreover, principal component analysis is used as an algorithm for feature reduction.

Keywords
cardiology, clinical decision-making, data mining, linear discriminant analysis, principal component analysis

Introduction

The prevalence of chronic diseases has increased due to poor eating habits, sedentary lifestyles, and the progression of ageing, thereby presenting an increasing burden upon individuals.[1] According to the Organization for Economic Cooperation and Development, chronic diseases account for major causes of deaths and health problems: 60 per cent (about 35 million) of the world's population is estimated to die due to chronic diseases. This epidemiological phenomenon has serious repercussions on life expectancy and its quality, thereby incurring an increasing economic burden upon healthcare systems and societies.[2] The rise of chronic diseases, alongside the need to improve the effectiveness, equity, and sustainability of healthcare systems, has required innovation among care services.[3–5] Many approaches have been proposed for the treatment and management of chronic diseases,[6,7] and innovative methodologies have been introduced to reduce waste in healthcare processes[8–10] and support decision-makers in the evaluation of technologies and in the choice of appropriate therapies.[11,12] Healthcare is rapidly generating big data (so-called both for its volume and its complexity and range) that can be exploited to derive new knowledge regarding patient care, compliance, and various regulatory requirements.[13] Researchers addressing these issues with data in the health sector belong to the field of health informatics: it is described as the union of the healthcare, computational, and information sciences in the study of healthcare information.[14] It includes acquisition, storage, manipulation, and retrieval of data to enable healthcare suppliers to deliver better results.[15] This set of tools and procedures is well-known in literature as big data analytics. This field is growing and has the potential to provide useful insights for healthcare systems. Currently, big data analytics is used to predict clinical decisions made by physicians and the outcomes of clinical activity for conditions based on patient features.[16] Data mining provides a variety of techniques for interpreting medical big data that can be used in patient care.[17] It helps in the analysis of clinical data stored in a shared information system and aims at identifying effective, useful, and generalizable relationships therein by combing massive datasets to elucidate patterns that may otherwise be too subtle or complex for traditional analytical methods to detect.[18]

There are several examples in the literature that encourage the application of data mining to different fields of medicine to discover hidden patterns or new information from clinical data.[19,20] Specific applications of machine learning are present in radiology; for example, Romeo et al.[21] used texture analysis to characterize adrenal lesions. In neurology, data mining techniques were applied to data collected from Parkinson's patients.[22–24] Other applications were developed for paediatric patients affected by developmental delays[25] and for orthopaedic patients to predict rehabilitation outcomes.[26] The support provided by data mining in healthcare must be taken into consideration not only for its assistance with clinical decision-making but also for its contribution to the assessment of physical and instrumental examinations performed for patient diagnoses. This evaluation allows health facilities to save money and avoid waste from the requisition of unnecessary patient examination.

Coronary artery disease

Coronary artery disease (CAD) is a multilevel process that originates from a combination of focal obstructive, diffuse, and microcirculatory alterations throughout the major coronary arteries supplying blood flow to the heart muscle.[27] CAD is one of the leading causes of death in the modern era, and according to the American Heart Association, 2300 Americans die because of cardiovascular diseases each day[28] – an average of one death every 38 s. Moreover, the European Heart Network stated in its report that CAD killed about 4 million people in Europe in 2017, constituting 45 per cent of all-cause mortality.[29] Due to the chronic nature of the disease and numerous consequential complications, there is a large volume of clinical data related not only to the number of patients but also to the number of variables per patient. Physicians have to interact with richer quantitative data that may contain a variety of instrumental parameters requisite for long-term data storage and biochemical data, such as total cholesterol, glycaemia, or *hemoglobina glicada*.[30] Cardiovascular disease management introduces a high number of variables per patient and necessitates the evaluation of large datasets to identify hidden patterns, compare information, and eventually make predictions.[31] Currently, there is enough awareness of both CAD pathology and data mining algorithms to strongly encourage the design and implementation of specific applications, as suggested by Chitra and Seenivasagam.[32]

Data mining was feasible within the daily process in institution in which clinical and instrumental variables are systematically registered to conform the body of information available for each subject at the time of diagnostic test. Therefore, we support the notion that data mining techniques can be applied in any context where systematic recording of patient-pertinent information is conducted. Considering the large amount of data, machine learning would be advantageous for clinicians who would benefit from systems trained with millions of data, especially for risk stratification of patients with suspected CAD, avoiding further unnecessary diagnostic tests.

In this article, we tested the use of a data mining application for the classification of patient with suspected or known CAD. As such, a linear discriminant analysis (LDA) algorithm was applied to patients with CAD exploiting features describing their state of health, and these results were compared to those obtained by using artificial features computed through principal component analysis (PCA). Different scores were computed: precision, recall, sensitivity, specificity, error, and accuracy. The data were acquired from anonymized information of patients who underwent stress single-photon emission computed tomography myocardial perfusion imaging (MPI) at the University Hospital 'Federico II' of Naples, Italy.

The rest of the article is organized as follows. Section 'Related work and aim' reviews the literature describing data mining in cardiology and describes our contribution to the literature. Section 'Materials' introduces the dataset. Section 'Methods' discusses the tools and algorithms used. Our results are presented in section 'Results', and finally, discussion and conclusion are found in section 'Discussion and conclusion'.

Related work and aim

We found different articles that exploited data mining to reveal in time the occurrence of heart disease in patients affected by CAD or to classify, according to diagnostic parameters, the level of clinical risk for patients. Some researchers focused only on the application of one algorithm, such as Shouman et al.,[33] who applied k-nearest neighbour on a benchmark dataset, and Chaurasia,[34] who used different types of decision trees. Conversely, Motwani et al.[35] investigated the feasibility and accuracy of machine learning to predict 5-year all-cause mortality in patients undergoing coronary computed tomographic angiography (CCTA) and compared the performances to the existing

clinical or CCTA metrics. Some classification procedures have been proposed by researchers that compare different techniques[36–39] or assess cardiovascular risk based on machine learning.[40–42] Another study investigated heart valve disease with the adaptive neuro-fuzzy inference system.[43] Weng et al.[40] illustrated a prospective cohort study in the United Kingdom related to the application of data mining on patients with cardiovascular pathology, while Rajkumar and Reena[44] conducted the classification of patients with heart diseases based on supervised machine learning algorithms, providing their accuracy, time taken to build the algorithm, and a comparison of the results. Soni et al.[45] compared different algorithms using artificial data of patients with heart disease made the same comparison using real data.

There are other studies in the literature that highlight the application of LDA: for example, Marcos et al.[46] used spectral features in nocturnal polysomnography; Luo et al.[47] included ultrasound (US) elastography features to detect thyroid nodules; and Yang et al.[48] predicted CAD through a combination of LDA and a fuzzy inference method that are described later in the 'Discussion and conclusion' section. Giri et al.[49] used different classifiers exploiting features extracted by PCA, yet they started from features related to heart rate signals. Here, the primary objective was to provide an automatic estimation of patient clinical outcomes to aid clinicians in their decision-making; moreover, the comparison detailed here enhances the importance of feature reduction, proving the equivalence of the two approaches.

As a result of these investigations, we found that many studies have been focused on the application of different data mining techniques in cardiology; moreover, PCA and LDA have been used for a wide range of disciplines and cardiac application.[46,47] Nevertheless, no studies were found describing the application of the LDA and PCA algorithms to CAD – particularly, no comparison exists between LDA and the combination of LDA and PCA regarding CAD diagnosis. Moreover, Lakshmi et al.[50] showed that LDA analysis is among the most accurate strategies that can be used in the heart disease classification, and Marcos et al.[46] implemented the combination of PCA and LDA successfully.

In this work, we used supervised data mining techniques to analyse a cohort of 10,265 patients who underwent stress single-photon emission computed tomography MPI at the Department of Advanced Biomedical Sciences of the public University Hospital 'Federico II' of Naples. Note that cardiopathic patients are always characterized by conventional risk factors, that is, the presence of diabetes, family history of CAD, hypertension, and so on.[51] Therefore, the aim of this study is to classify a large dataset of patients with CAD as either healthy or presenting this pathology through LDA by using conventional risk factors (age, gender, history of CAD, diabetes, etc.), as suggested by Kurt et al.,[52] who studied the classification of CAD by comparing different non-LDA algorithms. Here, the LDA algorithm was computed both on clinical features, provided and suggested by clinicians, and principal components (a linear combination of clinical features) obtained to reduce the previous number of features. These results were compared to highlight the effectiveness of the proposed methodology.

Materials

Population characteristics

The dataset includes 10,265 patients with suspected or known CAD that were evaluated for myocardial perfusion defect at the Department of Advanced Biomedical Sciences, University Hospital Federico II of Naples, between 2004 and 2017. As part of their initial check-up, clinicians collected information on traditional cardiovascular risk factors (age, gender, blood pressure, smoking history, serum cholesterol, family history of CAD, resting ECG characteristics, diabetes and its

Table 1. Population characteristics.

	Mean	Standard deviation
Age (years)	62.1	10.9
Diabetes (years)	11.6	9.8
Glycaemia (mg/dL)	114.0	46.2
LDL (mg/dL)	103.0	40.0
HDL (mg/dL)	47.8	15.4

LDL: low-density lipoprotein; HDL: high-density lipoprotein.

Table 2. Population characteristics: occurrences.

Variables	Categories	Occurrences	Variables	Categories	Occurrences
Gender	Men	7058	CAD	Suspected	5758
	Women	3207		Known	4507
Blood pressure therapy	Unknown	766	Diabetes	Unknown	13
	Yes	7627		Yes	6643
	No	1872		No	3609
Hypertension	Unknown	7	Family history of CAD	Unknown	27
	Yes	2448		Yes	6032
	No	7810		No	4206
Dyslipidaemia	Unknown	12	Angina	Unknown	23
	Yes	4042		Yes	6847
	No	6211		No	3395
Myocardial infarction	Unknown	17			
	Yes	6754			
	No	3494			

CAD: coronary artery disease.

complications, ECG stress testing). A known history of systolic blood pressure greater than 140 mmHg or the use of antihypertensive medication implicated hypertension. Hypercholesterolemia was defined based on either a known history of dyslipidaemia or treatment with cholesterol-lowering medication. A previous diagnosis of diabetes or treatment with oral hypoglycaemic drugs or insulin classified patients as diabetic. Only first-degree relatives were considered to assess a positive family history of CAD. Key population characteristics are presented in Table 1, with means and standard deviations. Risk factor occurrences are presented in Table 2.

Design of the dataset

The dataset was created using Excel by clinicians who have inserted patients one-by-one for 15 years. To enable the connection between the source of data and each possible software that can analyse data, we designed a database using PostgreSQL, a SQL Database Management System (DBMS) that allows to manage the relational database. It is well known that the integration between data mining systems and DBMS or spreadsheet visualization tools is fundamental in exploring databases. Holsheimer et al.[53] have showed that the use of an efficient relational DBMS can help with optimizing the identification of association rules.

Methods

The algorithm was implemented through the combination of the Knime analytics platform and R software.[54,55]

Tools

The extension of R programming language through packages was used: here, the MASS package was installed to use its 'lda' function.[56] In the literature, various tools are used for data analysis; our team used Knime analytics platform, whose utility is widely acknowledged in literature.[57] Knime Analytics Platform was created for the cheminformatics community, but it has been adapted over the years within other disciplines.[58] Moreover, the Weka data mining environment and additional R plugins offer access to a vast library of statistical routines.[59] It has been used in different research applications, such as in neurology, radiology, eye-related pathologies, and cardiotocography.[60–63]

PCA

PCA is a linear dimensionality reduction technique that can be used to reduce a large set of variables to a smaller set that still contains most of the original information.[64] It seeks a linear combination of variables such that the maximum variance can be extracted. PCA then removes this variance and seeks a second linear combination, which iteratively explains the maximum proportion of the remaining variance. This is called the principal axis method, which results in orthogonal (uncorrelated) factors. Moreover, it involves the computation of eigenvalues and eigenvectors of covariance matrices, followed by the sorting of these eigenvectors in the descending order of their eigenvalues and, finally, the projection of the actual data into the directions of sorted eigenvectors.

LDA

LDA was introduced for the first time by Fisher[65] in 1936; today, it remains a well-established statistical-based pattern classification method. Regarding two-class problems, the LDA method identifies a projection vector to maximize between-class scatter matrix while minimizing within-class scatter matrix in the feature space.[66] The aim here is to find a linear function

$$y = a_1 x_{i_1} + a_2 x_{i_2} + a_3 x_{i_3} + \cdots + a_q x_{i_q}$$

where

$$a^T = \left[\left\{ a_1, a_2, \ldots, a_q \right\} \right]$$

is a vector of coefficients that has to be determined, while

$$x_i = \left[x_{i_1}, x_{i_2}, \ldots, x_{i_q} \right]$$

are the patients and

$$x_j = \left[x_{j_1}, x_{j_2}, \ldots, x_{j_q} \right]$$

are the features.

To estimate the mean and variance in the dataset, the following assumptions, related to the multivariate analysis of variance, are necessary:

- *Normality*: independent variables are normal for each level of the grouping variable;
- *Independence*: the sampling is assumed to be random, and a sample's score on one variable is assumed to be independent of scores on that variable for all other participants;
- *Collinearity*: a high correlation between variables can decrease the predictive power.

Evaluation metrics

The following scores, identified as significant by the literature,[67,68] were used:

1. *Accuracy*: the ratio between correct predictions and total number of records;
2. *Error*: the ratio between wrong predictions and total number of records;
3. *Recall*: the number of positive patterns that are correctly classified;
4. *Precision*: the number of positive patterns correctly predicted from the total predicted patterns in a positive class;
5. *Sensitivity*: true positive rate, it assesses the effectiveness of the algorithm on a single class (the positive one);
6. *Specificity*: true negative rate, it assesses the effectiveness of the algorithm on a single class (the negative one).

Results

Data pre-processing

Evidently, our dataset underwent feature selection – a critical pre-processing step in machine learning that is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. To reconcile the high presence of missing values in the present work, empty fields were substituted with the mean values of those analysed through a node called 'missing values'. The selection of the most significant features was made through the calculation of a matrix of correlation among variables through the nodes 'linear correlation' and 'correlation filter': a threshold was chosen to exclude those that were too correlated and did not add information to the algorithms, as mentioned in the 'Methods' section. In conclusion, the algorithm included the class and the 22 variables.

Scores

The present dataset was divided into two parts: the model was fitted on a training set and the results were obtained on a test set. The variables were used to obtain the results presented in the first row of Table 3. Starting from the same variables, PCA was applied to identify 10 principal components. After excluding the 22 old features, principal components were used to perform the same previous classification, and the results are reported in the second row of Table 3.

Table 3. Scores.

	Accuracy (%)	Error (%)	Recall (%)	Precision (%)	Sensitivity (%)	Specificity (%)
LDA	84.5	15.5	62.8	94.2	62.8	97.7
LDA and PCA	86.0	14.0	65.4	96.2	65.4	98.4

LDA: linear discriminant analysis; PCA: principal component analysis.

The obtained classification accuracy was 84.5 per cent using only LDA and 86.0 per cent using the combination of LDA and PCA. Recall and sensitivity did not show high percentages (below 70%), while specificity and precision had better results – 94.2 and 97.7 per cent, respectively, from only LDA, and 96.2 and 98.4 per cent, respectively, from combined LDA and PCA.

Discussion and conclusion

We acquired clinical and instrumental data from patients admitted to the Department of Advanced Biomedical Sciences at the University Hospital of Naples 'Federico II'. Clinicians then identified some variables as useful, and the data were pre-processed as described. This analysis was implemented in two separate stages. LDA was first applied alone to classify the patients into two groups: healthy and pathological. Then, PCA and LDA were used together to reduce the number of features and reclassify the patients. Actually, there are too many variables involved when assessing the health status of a patient with cardiac issue; thus, the application of PCA allowed a summarization in a limited number of attributes. The key scores used (accuracy, error, precision, recall, sensitivity, and specificity) were found to improve under the combined PCA and LDA approach.

Extant studies in literature describe additional applications of LDA. Marcos et al.[46] showed an accuracy of 93 per cent using spectral features in their signal analysis (nocturnal polysomnography); Luo et al.[47] analysed US elastography features to classify thyroid nodules and obtained a discriminant score of 86 per cent; and Yang et al.[48] combined a fuzzy inference method and LDA to predict CAD with an accuracy of 80.2 per cent. Since a comparison between different applications is not always fair and possible, it is necessary to stress that our results are perfectly in accordance with literature describing the prospective applications of LDA. The power of this algorithm is also demonstrated when compared to the work of Kurt et al.,[50] who attempted to classify patients both with and without CAD. Their accuracy here was below 80 per cent for each trial, and their numbers of both patients and features were limited. The results from this analysis are strongly correlated with the quality of the data, with scores decreasing where data quality is poor. Improving the scores obtained with the LDA and PCA algorithms would encourage its use in health facilities to support clinicians with decision-making. Nevertheless, Jaarsma et al.[69] conducted a meta-analysis to determine the diagnostic accuracy of the three most commonly used non-invasive MPI modalities, single-photon emission computed tomography, cardiac magnetic resonance, and positron emission tomography perfusion imaging for the diagnosis of obstructive CAD. They, respectively, obtained a sensitivity of 88, 89, and 84 per cent and a specificity of 61, 76, and 81 per cent. Therefore, the machine learning would allow clinicians to obtain comparable to those obtained through the daily clinical practice, despite using only anamnestic variables and none of those obtained through instrumental exams. Not only is this method likely to apply to patients' diagnoses but it also supports clinicians in the formulation of prognoses through numerical predictive algorithms. The possibility to predict diagnosis from a few variables (perhaps restricted to anamnestic ones) would enable health facilities to spend less money on expensive exams, while nonetheless obtaining reliable diagnoses. Moreover, not only does the analytic platform provide simpler data analysis but it also

requires less time to perform this analysis, as there is little demand for programming experience. Clearly, one limitation of the study was the quality of data. Another limitation stems from the quality of the pre-processing pipeline. The addition of further pre-processing tools may benefit the reported methodology; a wrapper method could increase our scores, as could cross-validation (either leave-one-out or K-fold). Indeed, it may be possible in the future to use other methods to analyse these data and apply the same methodology to other clinical fields, thereby improving scores that could be obtained through the implementation of LDA combined with PCA. Following our implementation, PCA deserves further investigation: its use allows for a significant reduction of features while still obtaining an equal or higher set of scores. This notion implicates the method for better management of large quantities of data and faster analyses. PCA, as future development, could be applied in other healthcare contexts to reduce the number of features, while providing 'new ones' (principal components) that are as valid as original ones.

Acknowledgement

The authors wish to thank Alec Shawn for his contribute as regards "grammar and spell check". This work has been realized thanks to the collaboration of the Department of Advanced Biomedical Sciences of the University Hospital "Federico II" of Naples. The authors wish to thank Sabrina De Vita, Francesca D'Agostino, Giuseppina Toscano and Tania Di Monda for their valuable contribution to the implementation of data mining algorithm during their MS thesis. The work has been partially carried out under *TablHealth* [CUP B49J17000720008] project and *AK12 s.r.l.*"

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Carlo Ricciardi https://orcid.org/0000-0001-7290-6432

References

1. Park HS, Cho H and Kim HS. Development of a multiagent m-health application based on various protocols for chronic disease self-management. *J Med Syst* 2016; 40(1): 36.
2. Garcıa-Lizana F and Sarría-Santamera A. New technologies for chronic disease management and control: a systematic review. *J Telemed Telecare* 2007; 13(2): 62–68.
3. Montella E, Di Cicco MV, Ferraro A, et al. The application of Lean Six Sigma methodology to reduce the risk of healthcare associated infections in surgery departments. *J Eval Clin Pract* 2017; 23(3): 530–539.
4. Improta G, Balato G, Ricciardi C, et al. Lean Six Sigma in healthcare. *TQM J* 2019; 31(4): 526–540.
5. Ricciardi C, Fiorillo A, Valente A, et al. Lean Six Sigma approach to reduce LOS through a diagnostic-therapeutic-assistance path at A.O.R.N. A. Cardarelli. *TQM J* 2019; 31: 657–672.
6. Santini S, Pescapè A, Valente AS, et al. Using fuzzy logic for improving clinical daily-care of β-thalassemia patients. In: *2017 IEEE international conference on fuzzy systems (FUZZ-IEEE),* Naples, 9–12 July 2017, pp. 1–6. New York: IEEE.
7. Biondi M, Crispino M, Improta G, et al. The condroprotector role in the osteoarthritis of the knee. *Giorn Ital Ortoped Traumatol* 2013; 39: 44–47.

8. Improta G, Romano M, Di Cicco MV, et al. Lean thinking to improve emergency department throughput at AORN Cardarelli hospital. *BMC Health Serv Res* 2018; 18(1): 914.
9. Improta G, Balato G, Romano M, et al. Improving performances of the knee replacement surgery process by applying DMAIC principles. *J Eval Clin Pract* 2017; 23(6): 1401–1407.
10. Improta G, Ricciardi C, Borrelli A, et al. The application of six sigma to reduce the pre-operative length of hospital stay at the hospital Antonio Cardarelli. *Int J Lean Six Sigma* 2019; 11(3): 555–576.
11. Improta G, Russo MA, Triassi M, et al. Use of the AHP methodology in system dynamics: modelling and simulation for health technology assessments to determine the correct prosthesis choice for hernia diseases. *Math Biosci* 2018; 299: 19–27.
12. Improta G, Triassi M, Guizzi G, et al. An innovative contribution to health technology assessment. In: Ding W, Jiang H, Ali M, et al. (eds) *Modern advances in intelligent systems and tools*. Berlin: Springer, 2012, pp. 127–131.
13. ur Rehman MH, Chang V, Batool A, et al. Big data reduction framework for value creation in sustainable enterprises. *Int J Inform Manage* 2016; 36(6): 917–928.
14. Koh HC and Tan G. Data mining applications in healthcare. *J Healthc Inf Manag* 2011; 19(2): 64–72.
15. Revetria R, Catania A, Cassettari L, et al. Improving healthcare using cognitive computing-based software: an application in emergency situation. In: *International conference on industrial engineering and other applications of applied intelligent systems*, Syracuse, NY, 28 June–1 July 2012, pp. 477–490. Berlin: Springer.
16. Kumar S and Singh M. Big data analytics for healthcare industry: impact, applications, and tools. *Big Data Min Anal* 2019; 2(1): 48–57.
17. Youssef AE. A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments. *Int J Ambient Syst Appl* 2014; 2(2): 1–11.
18. Kreuze D. Debugging hospitals. *Technol Rev* 2001; 104(2): 32–32.
19. Bellazzi R, Ferrazzi F and Sacchi L. Predictive data mining in clinical medicine: a focus on selected methods and applications. *WIREs: Data Min Knowl* 2011; 1(5): 416–430.
20. Bellazzi R, Diomidous M, Sarkar IN, et al. Data analysis and data mining: current issues in biomedical informatics. *Methods Inf Med* 2011; 50(6): 536–544.
21. Romeo V, Maurea S, Cuocolo R, et al. Characterization of adrenal lesions on unenhanced MRI using texture analysis: a machine-learning approach. *J Magn Reson Imaging* 2018; 48: 198–204.
22. Amboni M, Barone P, Iuppariello L, et al. Gait patterns in Parkinsonian patients with or without mild cognitive impairment. *Movement Disord* 2012; 27(12): 1536–1543.
23. Nilashi M, bin Ibrahim O, Ahmadi H, et al. An analytical method for diseases prediction using machine learning techniques. *Comput Chem Eng* 2017; 106: 212–223.
24. Perumal SV and Sankar R. Gait monitoring system for patients with Parkinson's disease using wearable sensors. In: *Healthcare innovation point-of-care technologies conference (HI-POCT)*, Cancun, Mexico, 9–11 November 2016, pp. 21–24. New York: IEEE.
25. Chang CL. A study of applying data mining to early intervention for developmentally-delayed children. *Expert Syst Appl* 2007; 33(2): 407–412.
26. Chan F, Cheing G, Chan JYC, et al. Predicting employment outcomes of rehabilitation clients with orthopedic disabilities: a CHAID analysis. *Disabil Rehabil* 2006; 28(5): 257–270.
27. van de Hoef TP, Echavarria-Pinto M, van Lavieren MA, et al. Diagnostic and prognostic implications of coronary flow capacity: a comprehensive cross-modality physiological concept in ischemic heart disease. *JACC Cardiovasc Interv* 2015; 8(13): 1670–1680.
28. Mozaffarian D, Benjamin EJ, Go AS, et al. Heart Disease and Stroke Statistics-2016 Update. *Circulation* 2016; 133(4): e38–e360.
29. Wilkins E, Wilson L, Wickramasinghe K, et al. *European cardiovascular disease statistics 2017*. Brussels: European Heart Network, 2017.
30. Johnson KW, Soto JT, Glicksberg BS, et al. Artificial intelligence in cardiology. *J Am Coll Cardiol* 2018; 71(23): 2668–2679.
31. Alexander C and Wang L. Big data analytics in heart attack prediction. *J Nurs Care* 2017; 6(393): 2167–1168.

32. Chitra R and Seenivasagam V. Review of heart disease prediction system using data mining and hybrid intelligent techniques. *ICTACT J Soft Comput* 2013; 3(4): 605–609.
33. Shouman M, Turner T and Stocker R. Applying k nearest neighbour in diagnosing heart disease patients. *Int J Inform Educ Technol* 2012; 2(3): 220–223.
34. Chaurasia V. Early prediction of heart diseases using data mining techniques. *Caribbean J Sci Technol* 2013; 12(1): 208–217.
35. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J* 2016; 38(7): 500–507.
36. Ricciardi C, Amboni M, De Santis C, et al. Using gait analysis' parameters to classify Parkinsonism: a data mining approach. *Comput Methods Programs Biomed* 2019; 180: 105033.
37. Romeo V, Ricciardi C, Cuocolo R, et al. Machine learning analysis of MRI-derived texture features to predict placenta accreta spectrum in patients with placenta previa. *Magn Reson Imaging* 2019; 6464: 7171–7676.
38. Romano M, Iuppariello L, Ponsiglione AM, et al. Frequency and time domain analysis of foetal heart rate variability with traditional indexes: a critical survey. *Comput Math Method Med* 2016; 2016: 958543.
39. Improta G, Romano M, Ponsiglione A, et al. Computerized cardiotocography: a software to generate synthetic signals. *J Health Med Inform* 2014; 5(4): 162.
40. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* 2017; 12(4): e0174944.
41. Mannarino T, Assante R, Ricciardi C, et al. Head-to-head comparison of diagnostic accuracy of stress-only myocardial perfusion imaging with conventional and cadmium-zinc telluride single-photon emission computed tomography in women with suspected coronary artery disease. *J Nucl Cardiol* 2019: 1–10.
42. Ricciardi C, Cantoni V, Green R, et al. Is it possible to predict cardiac death? In: *Mediterranean conference on medical and biological engineering and computing*, Coimbra, 26–28 September 2019, pp. 847–854. Cham: Springer.
43. Sengur A. An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases. *Expert Syst Appl* 2008; 35(1–2): 214–222.
44. Rajkumar A and Reena GS. Diagnosis of heart disease using data mining algorithm. *Global J Comput Sci Technol* 2010; 10(10): 38–43.
45. Soni J, Ansari U, Sharma D, et al. Predictive data mining for medical diagnosis: an overview of heart disease prediction. *Int J Comput Appl* 2011; 17(8): 43–48.
46. Marcos JV, Hornero R, Álvarez D, et al. Automated detection of obstructive sleep apnoea syndrome from oxygen saturation recordings using linear discriminant analysis. *Med Biol Eng Comput* 2010; 48(9): 895–902.
47. Luo S, Kim EH, Dighe M, et al. Thyroid nodule classification using ultrasound elastography via linear discriminant analysis. *Ultrasonics* 2011; 51(4): 425–431.
48. Yang JG, Kim JK, Kang UG, et al. Coronary heart disease optimization system on adaptive-network based fuzzy inference system and linear discriminant analysis (ANFIS – LDA). *Pers Ubiquit Comput* 2014; 18(6): 1351–1362.
49. Giri D, Acharya UR, Martis RJ, et al. Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. *Knowl-Based Syst* 2013; 37: 274–282.
50. Lakshmi KR, Veera Krishna A and Prem Kumar S. Performance comparison of data mining techniques for predicting of heart disease survivability. *Int J Sci Res Pub* 2013; 3(6): 1–10.
51. Khot UN, Khot MB, Bajzer CT, et al. Prevalence of conventional risk factors in patients with coronary heart disease. *JAMA* 2003; 290(7): 898–904.
52. Kurt I, Ture M and Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst Appl* 2008; 34(1): 366–374.
53. Holsheimer M, Kersten ML, Mannila H, et al. A perspective on databases and data mining. In: *KDD-95*, Montreal, QC, Canada, 20–21 August 1995, pp. 150–155. Palo Alto, CA: AAAI.

54. Berthold MR, Cebron N, Dill F, et al. KNIME – the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explorat Newslett* 2009; 11(1): 26–31.
55. R Core Team. *R: A language and environment for statistical computing*. Vienna: R Core Team, 2013.
56. Venables WN and Ripley BD. *Modern applied statistics with S*. 4th ed. New York: Springer, 2002.
57. Sharma N and Bansal K. Comparative study of data mining tools. *J Adv Datab Manag Syst* 2015; 2(2): 35–41.
58. Warr WA. Scientific workflow systems: pipeline Pilot and KNIME. *J Comput Aided Mol Des* 2012; 26(7): 801–804.
59. Witten IH, Frank E, Hall MA, et al. *Data Mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann, 2016.
60. Ricciardi C, et al. Classifying different stages of Parkinson's disease through random forests. In: *IFMBE proceedings on XV Mediterranean conference on medical and biological engineering and computing – MEDICON 2019* (ed J Henriques, N Neves and P de Carvalho), vol. 76, Coimbra, 26–28 September 2019. Cham: Springer.
61. Ricciardi C, Cuocolo R, Cesarelli G, et al. Distinguishing functional from non-functional pituitary macroadenomas with a machine learning analysis. In: *IFMBE proceedings on XV Mediterranean conference on medical and biological engineering and computing – MEDICON 2019* (ed J Henriques, N Neves and P de Carvalho), vol. 76, Coimbra, 26–28 September 2019. Cham: Springer.
62. Improta G, Ricciardi C, Amato F, et al. Efficacy of machine learning in predicting the kind of delivery by cardiotocography. In: *IFMBE proceedings on XV Mediterranean conference on medical and biological engineering and computing – MEDICON 2019* (ed J Henriques, N Neves and P de Carvalho), vol. 76, Coimbra, 26–28 September 2019. Cham: Springer.
63. D'Addio G, Ricciardi C, Improta G, et al. Feasibility of machine learning in predicting features related to congenital nystagmus. In: *IFMBE proceedings on XV Mediterranean conference on medical and biological engineering and computing – MEDICON 2019* (ed J Henriques, N Neves and P de Carvalho), vol. 76, Coimbra, 26–28 September 2019. Cham: Springer.
64. Jolliffe I. Principal component analysis. In: Lovric M (ed.) *International encyclopedia of statistical science*. Berlin: Springer, 2011, pp. 1094–1096.
65. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugenic* 1936; 7(2): 179–188.
66. Duda RO, Hart PE and Stork DG. *Pattern classification*. New York: John Wiley & Sons, 2012.
67. Hossin M and Sulaiman M. A review on evaluation metrics for data classification evaluations. *Int J Data Min Knowl* 2015; 5(2): 1.
68. Sokolova M, Japkowicz N and Szpakowicz S. Beyond accuracy F-score and ROC: a family of discriminant measures for performance evaluation. In: *Australasian joint conference on artificial intelligence*, Hobart, TAS, Australia, 4–8 December 2006, pp. 1015–1021. Berlin: Springer.
69. Jaarsma C, Leiner T, Bekkers SC, et al. Diagnostic performance of noninvasive myocardial perfusion imaging using single-photon emission computed tomography, cardiac magnetic resonance, and positron emission tomography imaging for the detection of obstructive coronary artery disease: a meta-analysis. *J Am Coll Cardiol* 2012; 59(19): 1719–1728.