Journal of Biomedical Informatics 97 (2019) 103257

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Special Communication

Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)

Juan-Jose Beunza[a,b,*], Enrique Puertas[a,c], Ester García-Ovejero[a,d], Gema Villalba[a,e], Emilia Condes[a], Gergana Koleva[a], Cristian Hurtado[a,f], Manuel F. Landecho[a,g]

[a] Machine Learning Health Working Group, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Madrid, Spain
[b] Department of Medicine, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Madrid, Spain
[c] Department of Computer Science and Technology, School of Architecture, Engineering and Design, Universidad Europea de Madrid, Madrid, Spain
[d] Department of Nursing and Psychology, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Madrid, Spain
[e] Indra, Madrid, Spain
[f] Department of Pharmacy and Biotechnology, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Madrid, Spain
[g] Departament of Internal Medicine, Clinica Universidad de Navarra, Pamplona, Spain

ARTICLE INFO

Keywords:
Machine learning
Supervised machine learning
Support vector machines
Research techniques
Area under curve
Diagnostic techniques and procedures

ABSTRACT

Aim: The aim of this study is to compare the utility of several supervised machine learning (ML) algorithms for predicting clinical events in terms of their internal validity and accuracy. The results, which were obtained using two statistical software platforms, were also compared.
Materials and methods: The data used in this research come from the open database of the Framingham Heart Study, which originated in 1948 in Framingham, Massachusetts as a prospective study of risk factors for cardiovascular disease. Through data mining processes, three data models were elaborated and a comparative methodological study between the different ML algorithms – decision tree, random forest, support vector machines, neural networks, and logistic regression – was carried out. The global selection criterium for choosing the right set of hyperparameters and the type of data manipulation was the area under a curve (AUC). The software tools used to analyze the data were R-Studio® and RapidMiner®.
Results: The Framingham study open database contains 4240 observations. The algorithm that yielded the greatest AUC when analyzing the data in R-Studio was neural network applied to a model that excluded all observations in which there was at least one missing value (AUC = 0.71); when analyzing the data in RapidMiner and applying the same model, the best algorithm was support vector machines (AUC = 0.75).
Conclusions: ML algorithms can reinforce the diagnostic and prognostic capacity of traditional regression techniques. Differences between the applicability of those algorithms and the results obtained with them were a function of the software platforms used in the data analysis.

1. Introduction

The algorithms and techniques deployed in machine learning (ML) can be framed within a more general process known as *knowledge discovery in databases* or simply *data mining*. Some of these techniques were described more than 50 years ago [1], however in recent years interest in and about them has surged dramatically, driven in part by major advances in algorithmic programming, increasing processing capacity of modern computers (Graphics Processor Unit for video and graphics cards and Tensor Processing Unit for neural learning), and growth of data availability (Big Data).

In the last three years these types of algorithms and techniques have begun to be applied to clinical environments, including within the fields of diagnostic radiology [2–4], cardiac electrophysiology [5], diabetes [6], dermatology [7] and psychiatry [8,9]. Given their practicality and accessibility, and the impressive results obtained so far, we expect an explosion in ML applications in healthcare settings in 2019.

Recent publications of the Ministry of Science, Innovation and Universities of the Government of Spain identify the incorporation of artificial intelligence in healthcare as a priority [10]. Official

Abbreviations: ML, machine learning; ACC, accuracy; AUC, area under curve; SE, sensitivity; SP, specificity; PPV, positive predictive value; NPV, negative predictive value; NA, not applicable

[*] Corresponding author at: Universidad Europea de Madrid, Calle Tajo, s/n, 28670 Villaviciosa de Odón, Madrid, Spain.
E-mail address: juanjo@juanjobeunza.com (J.-J. Beunza).

https://doi.org/10.1016/j.jbi.2019.103257
Received 27 May 2019; Received in revised form 21 July 2019; Accepted 22 July 2019
Available online 30 July 2019
1532-0464/ © 2019 Elsevier Inc. All rights reserved.

documents of the European Commission and strategic plans of other European countries point in the same direction [11,12].

The Machine Learning development group Salud-UEM was created in November 2018 at University Europea de Madrid with the aim of applying ML techniques to the study of health-related problems, generating evidence of their potential utility, and eventually incorporating them in the teaching curriculum of students and health professionals. It brings together a heterogeneous mix of health professionals (doctors, nurses, psychologists, pharmacists and biotechnologists), computer specialists specialized in big data, and eHealth consultants. As a first test of the possibilities that ML can offer in the development of predictive models in health, it was decided to test various algorithms to see whether and to what extent their prediction scores approximate or improve upon the results obtained in the original Framingham model [13], one of the most important cardiovascular risk prediction tables from the point of view of clinical practice [14].

Although other researchers have tried to improve upon the Framingham model's predictive value [15], the focus of the present study is purely methodological, given our overarching goal of exploring the applicability of ML models in health. Specifically, the primary objective was to compare in terms of internal validity and accuracy several supervised ML algorithms applied to the prediction of clinical events, using structured data; the secondary objective was to compare the utility, usability and results obtained by means of two software tools, R-studio (script code) and RapidMiner (graphic interface).

2. Materials and methods

A comparative methodological study was carried out between the most commonly used supervised classification algorithms in ML: decision tree, random forest, support vector machines, and neural networks, in addition to traditional logistic regression.

The following outcome variables were selected for comparison: accuracy, sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, and area under the curve (AUC). The global criterion for selecting the right set of hyperparameters and the type of data manipulation was the AUC.

The data used came from the Framingham database available on Kaggle [16], a Google-owned platform for data scientists that published ML-based prediction competitions using publicly available databases.

2.1. Data preparation

A descriptive analysis of the database was performed, attending to the nature of the variables: mean, standard deviation and extreme values were described for quantitative variables; absolute and relative frequencies for qualitative variables. The number of missing values was identified for each variable, since for some of the algorithms to work it was anticipated that it may be necessary to eliminate observations with missing values.

To guide the selection of independent variables to be included in the prediction model for coronary risk at 10 years, an automatic "stepwise" technique was applied, using multiple logistic regression (stepAIC function of the MASS library in R-studio) and establishing as a cut-off point the significance level of $p < 0.10$. Nonetheless, the algorithms identified as "best" used $p < 0.05$ as the significance level.

For the comparison of the algorithms, three different data models were prepared. Model A included the original variables in the Framingham Heart Study database without any modification. Model B included the same variables but excluded all observations in which there was at least one missing value in any of the variables (analysis of complete cases). Finally, in model C missing values in any of the continuous variables were imputed using the average obtained from the rest of the non-missing values present in that variable so as to avoid reducing the number of observations, notwithstanding that some minimal alterations introduced by the imputation.

In all three models the data were randomly divided into two subgroups: training set (*train*), which contained 80 percent of the observations, for the training of the algorithm, and test set (*test*), comprising the remaining 20 percent of observations, for the evaluation of the algorithm's capacity for prediction. Different base models were elaborated by normalizing (subtracting the mean and dividing by the standard deviation, which resulted in means of zero in all the variables) and standardizing (subtracting the minimum value of the variable and dividing by the range, which resulted in minimum values of 0 and maximums of 1 in all the variables) the values in both the *train* and the *test* sets in order to homogenize their range, an important condition for some algorithms, (e.g. the neural network algorithm).

Finally, the number of positive events or results, referred to as *labels* in informatics terminology, was balanced using the ROSE library and the functions *over* (duplicating subjects with positive cardiovascular event), *under* (eliminating subjects with negative cardiovascular event), *both* (combining both techniques) and *rose* (artificially generating completely new subjects based on the distribution of variables of the original database). This resulted in a change from the original prevalence of coronary events of 15% to a prevalence of 50%. The purpose of the balancing, as commented on in the discussion section, is to improve the algorithm's prediction capability. In addition, each algorithm was optimized by way of adjusting its hyperparameters (the internal values of the algorithm that determine its learning function and therefore directly impact on the final result it produces).

The participants were not stratified by sex, as the original authors of the Framingham study had done, so as not to reduce the sample size for training the algorithm, something to which ML algorithms are very sensitive. As a workaround, the variable *sex* was included in the model, because it is strongly associated with the event under study. The accuracy values for both sexes from the original Framingham publication were calculated using the formula [(sensitivity * prevalence) + (specificity * (1 − prevalence))].

2.2. Software used

Two software tools were used in carrying out this research:

(a) R-Studio open source, version 1.1.463 with R open source version 3.5.2 (2018-12-20, "Eggshell Igloo") for data preparation and algorithm training and evaluation. The libraries and functions used in training the different algorithms were the following: for the decision tree and the boosted decision tree (the most powerful version of the decision tree) algorithms, the *C50* library and the *C5.0* function; for random forests, the *randomForest* library and the *randomForest* function; for the support vector machines, the *kernlab* library and the *ksvm* function; for the neural network, the *neuralnet* library and the *neuralnet* function. The complete code is available on Github [17] and will be published as open source in *markdown* format (html) with explanatory comments with teaching objectives.
(b) RapidMiner version 9.2.0 to compare it with the results obtained with RStudio. Rapid Miner code is available on Github [18].

3. Results and discussion

The Framingham open database consisted of 4240 observations, of which 57.1% corresponded to women (n = 2420) and 42.9% to men (n = 1820). The mean age was 49.6 years (standard deviation (SD) = 8.6). The percentage of patients who were smokers was 49.4% (n = 2095), 31.1% had hypertension (n = 1317) and 2.6% had diabetes (n = 109). The mean total cholesterol was 236.7 (SD = 44.6), the mean systolic blood pressure was 132.4 mm Hg (SD = 22), and the mean diastolic blood pressure was 82.9 mm Hg (SD = 11.9). Discordant extreme values were seen in blood glucose, whose mean was 82 mg/dl (SD = 24), with a minimum of 40 and a maximum of 394. In total, 15.2% of participants suffered coronary events at 10 years (n = 644).

Table 1
Exploratory descriptive analysis of the Framingham Heart Study open database [16].

Variable	Categories	n	%			*Missing*
Sex	Women	2420	57.1			0
	Men	1820	42.9			
Educational level *(n = 4135)*	*Some High School*	1720	41.6			105
	High School or GED	1253	30.3			
	Some College or Vocational School	689	16.7			
	college	473	11.4			
Current smoker	No	2145	50.6			0
	Yes	2095	49.4			
Antihypertensive treatment *(n = 4187)*	No	4063	97			53
	Yes	124	3			
Prevalent stroke	No	4215	99.4			0
	Yes	25	0.6			
Hypertension	No	2923	68.9			0
	Yes	1317	31.1			
Diabetes	No	4131	97.4			0
	Yes	109	2.6			
Coronary events at 10 years	No	4131	84.8			0
	Yes	644	15.2			
		$\bar{x}$	SD	Max	Min	
Age		49.6	8.6	70	32	0
Daily cigarettes *(n = 4211)*		9.01	11.9	70	0	29
Cholesterol *(n = 4190)*		236.7	44.6	696	107	50
Systolic blood pressure		132.4	22.0	83.5	295	0
Diastolic blood pressure		82.9	11.9	48	142.5	0
Body mass index		25.8	4.1	56.8	15.5	19
Heart rate *(n = 4239)*		75.9	12.0	143	44	1
Glycaemia *(n = 3852)*		82.0	24.0	394	40	388

$\bar{x}$: mean; SD: standard deviation; Max: maximum value; Min: minimum value.

The variables with missing data were: educational level, antihypertensive treatment, daily cigarettes, body mass index, heart rate, and glycaemia (Table 1).

By sex, the 10-year prevalence of coronary heart disease was calculated as 19% (n = 343) in men and 12% (n = 301) in women. Using the formula already described, accuracy values of 0.78 for men and 0.81 for women were obtained.

The automatic "stepwise" selection of variables to include in the models determined the following features to be key predictors for coronary risk at 10 years: sex, age, cigarettes per day, prevalent cerebral infarction, prevalent hypertension, total serum cholesterol, arterial systolic pressure, and glucose serum ($p < 0.10$).

After data preparation for each model, the sizes of the training sets (n-train) and test sets (n-test) were as follows: model A, gross (n-train: 3392; n-test: 848); model B, analysis of complete cases (n-train: 3053; n-test: 764); model C, imputation of the mean (n-train: 3392; n-test: 848). The results in Table 2 were calculated with these numbers.

The algorithm that showed the highest AUC when performing the analysis with R-studio was neural network applied to model B (AUC = 0.71). Among the results obtained with RapidMiner, the algorithm that behaved best was support vector machines applied to model B (AUC = 0.75). Table 2 also summarizes the best outcomes in accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and area under the curve obtained for each prediction algorithm model, considering the three data models and the two computer tools R-studio and RapidMiner.

The ML algorithms applied to the Framingham database have obtained comparable, and in some cases superior, results to those described in the original article of the Framingham Coronary Risk Scale [13]. The sensitivity values described for each sex in the original study were 0.49 for men and 0.6 for women, and the specificity values were 0.85 and 0.84, respectively. However, direct comparisons between results should be approached with caution, given that the original article used Cox regression (time to event), calculated the probability of being in the fifth risk quintile while using more of the available predictor variables, and differed in the number of observations and variables used compared to the open database.

Working with the Framingham database was selected in the first place because it is freely accessible and is easy to use for teaching objectives. Secondly, because due to its relatively small size for ML studies, as a result of which the obtained results were probably not optimal, it allows for the algorithms to be executed in almost any personal computer with reasonable processing times (for example, it does not exceed 10 min using an Intel Core i7 2.5 GHz processor and 16 GB 1600 mHz DDR3 memory). In this way it was possible to avoid having to manage clusters of local computers or to use a cloud, with the added technical complications that implies.

Reviewing the algorithms one by one, the decision tree had several advantages. Firstly, it could manage variables with missing values and therefore required less manipulation of the original data. In addition, it offered a prioritization of the variables, something that is of value for clinicians in their decision-making. Interestingly, the variable most frequently used (in 100% of cases) in the design of the trees was glycemia. Finally, it was a fast algorithm in its execution and design, requiring minimal processing times (almost immediate). The disadvantage was that the accuracy values obtained were low and that it is an algorithm generally considered as having low prediction power. The boosting, which adds a greater number of trees (around 10 in our case) made almost no difference. A very aggressive pruning process in single-tree algorithms may incur in overfitting, but this problem is mitigated when using decision tree ensembles like Random Forest or Boosted Trees. When using default hyperparameters of the models, we did not observe symptoms of overfitting or underfitting. Different performance results may be obtained using full or pruned trees.

The random forest algorithm improved the values of the decision tree, although its interpretation was more complex (hence it being called a "black box") and the processing time was prolonged beyond five minutes in some configurations of the hyperparameters. The support vector machines algorithm improved the AUC a little more, but again extended the processing times and code complexity. The neural network was the algorithm that required the most programming time – around 1250 lines of code – and it ran 73 different architecture models, each comprising between one and three layers and between one and nine nodes per layer; some processing times exceeded 10 min; however, the neural network also offered the best result. Lastly, the traditional logistic regression offered similar results to the previously described algorithms.

In general, all the results of the algorithms improved after normalization/standardization of the data, an almost mandatory requirement in some cases such as support vector machines and neural networks. On the other hand, data balancing was another key factor that led to improved results; however, it is rarely done in traditional statistical analyses.

In Model A, acceptable accuracy values (85%) were almost universally obtained, which could indicate that the algorithm did a good job because it correctly guessed the classification of 85 out of 100 participants. However, what the algorithms were doing in almost all cases was to classify all subjects as non-events. In this way the accuracy was good, but the prediction of "positive events" was almost null (sensitivity close to 0, specificity close to 100). This happens usually when the incidence of the event in studies is not balanced, that is, when it is far from 50%, which is something very frequent in the study of medical or healthcare events. In our case, the incidence was 15%. The way to solve this was to balance the base by increasing the number of observations with positive events, which can be achieved by cloning observations or creating them again by artificial methods, while respecting the distribution of variables; decreasing the number of

Table 2
Comparison and evaluation of the different algorithms with R-studio and RapidMiner.

	R-Studio						RapidMiner					
Algorithms	ACC	SE	SP	PPV	NPV	AUC	ACC	SE	SP	PPV	NPV	AUC
Decision tree												
Model A	84	8	98	48	85	0.53	85	3	99	44	85	0.53
Model B	67	39	72	21	86	0.55	54	4	99	83	55	0.5
Model C	84	8	98	44	85	0.53	62	4	99	83	61	0.5
Boosted decision tree												
Model A	85	6	99	57	85	0.53	63	73	53	61	66	0.67
Model B	81	28	91	37	87	0.60	64	54	70	51	72	0.7
Model C	84	5	99	44	85	0.52	62	53	67	51	69	0.69
Random forests												
Model A	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Model B	79	35	88	39	86	0.63	65	9	97	65	64	0.71
Model C	78	30	87	31	87	0.59	63	14	95	62	63	0.69
Support vector machines												
Model A	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Model B	69	67	69	29	92	0.68	69	42	84	61	71	0.75
Model C	68	69	68	28	92	0.68	68	49	81	62	71	0.71
Neural network												
Model A	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Model B	67	70	66	28	92	0.71	69	36	90	67	70	0.73
Model C	71	64	72	29	92	0.68	68	56	77	61	73	0.72
Logistic regression												
Model A	84	5	99	50	85	0.5	63	47	79	69	60	0.68
Model B	68	69	67	29	92	0.68	68	43	83	59	71	0.73
Model C	66	69	66	27	92	0.68	67	46	81	61	70	0.73

ACC: accuracy; SE: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value; AUC: area under curve; NA: not applicable (model did not converge).

negative events by eliminating observations; or a combination of both.

This type of data manipulation almost certainly introduces significant biases in causality studies. For example, the imputation of missing values with the mean of the variable is something frequently done in prediction work yet is viewed with suspicion in traditional cause-and-effect analysis because of the bias it introduces in case the participants with missing values are intrinsically different from those without missing values.

Model B, which is the one that behaved the best in terms of predictive capacity, is nevertheless the one that follows the most discouraged strategy from the point of view of analytical power. The best approach to handling of missing data remains a problem without a universally agreed-upon solution in epidemiology so, at a minimum, it needs to be included in the analysis plan and communicated in the results [19]. It is a consideration to consider because in risk prediction studies and ML studies in general, the prediction objective allows us to manipulate the data more freely if the objective (prediction in a different database, evaluation) is satisfactory. It implies a change of mentality for health professionals trained in cause-and-effect studies.

Model C handled missing values by imputating mean values. Other classic imputation options are to use the most frequent value for that variable, zero or constant imputation, k nearest neighbors (k-NN) through feature similarity, multivariate imputation by chained equation (MICE), deep learning, stochastic regression imputation, extrapolation and interpolation and hot-deck imputation [20]. Regardless of the chosen method, there is no perfect way to compensate for missing values in a data set and it is up to the knowledge and intuition of the analyst to select the best method for that specific dataset and context. In our case, we considered that imputation by means – which is important in the analysis of relatively small data sets because it “prevents” discarding observations due to missing values – offered comparison value of algorithms without losing participants.

Finally, differences were found among the results produced by the software programs used in this research. R-studio proved to be more flexible and powerful a priori than RapidMiner, although the programming time it required was much higher. However, its ability to keep track of all the changes in a reproducible file (script) was more replicable and amenable to correction, which provided a sense of assurance. RapidMiner, on the other hand, was more intuitive and simpler, which is something very useful in simple or preliminary analyses; however, when the process becomes more complex, it seems easier to make mistakes. The analysis with RapidMiner is more difficult to replicate by an external person and more difficult to correct. It seems therefore an excellent tool for teaching purposes and for selecting algorithms (piloting), but it does not replace code-based programs from our point of view. The differences with respect to the results obtained with R-Studio may be due to the different implementations of the algorithms and the default hyperparameters used by the libraries of both testing environments. Therefore, comparing tools and algorithms often requires a judgment call, given that a good hyperparameter tuning is more an art than a mechanical process and should be based on both the characteristics of the dataset and the knowledge domain.

The present study has several strengths. The principal one is that it was managed by an interprofessional team that included clinical and informatics experts, which has proven to be extremely enriching. Further, a large amount of sensitivity analysis (multiple models of variables management, feature engineering) and model and hyperparameters design was carried out. The fact that the data are public and open and that the code has been released also allowed us to review and replicate our results, as well as to use the present study as a gateway to the world of health ML in particular and to the programming of prediction models in general. Finally, since the analysis drew upon a relatively small database (4240 observations with 16 variables), it lends itself to being conducted on any personal computer.

In terms of weaknesses, the main drawback of this research is that when using data presented as part of an ML competition, neither its clinical reliability nor the quality of the results can be guaranteed. In addition, as mentioned earlier, the limited number of observations likely diminishes the prediction capacity of the trained models, designed to be working with much larger numbers of observations.

Further and somewhat unexpectedly, some variables that are clearly predictive of coronary disease, such as blood levels of LDL-cholesterol, were not included in the dataset; a much more precise model could be obtained by using more information of risk factors of the participants. The possibility of requesting such information from those responsible for the Framingham study was considered but ruled out because of the likelihood that readers would be excluded from access to the additional data. Finally, the fact that stepwise selection was done through logistic regression – and the results applied to all models for the sake of simplicity – could have led to a missed opportunity to obtain “better” results by using other models. It would have been beneficial to do feature selection for each algorithm.

We obtained low levels of AUC values (close for 0.5 for decision tree and 0.6–0.73 for others). We did not find other published results on the same dataset (available from Kaggle) for comparison. However, similar approaches with other larger data sets (378,256 patients from UK family practices) yielded similar results [21]. Failure to obtain better results could be due to the low number of subjects or because the available variables do not have a higher predictive capacity. It could also be due to the fact that we may not have been able to adequately refine the hyperparameters; however, this is unlikely, given that one of the researchers in the group is an expert in big data and has won several national ML awards [22,23]. Nonetheless, it is still possible to refine the algorithms used, as we must not forget that, like medicine, ML is both an art and a science.

The next steps in the development of ML algorithms in healthcare settings is to apply them more widely to different environments and populations. A potential handicap is that as the adoption of advanced data analytics and machine learning in healthcare accelerates and databases grow in volume, the ever-rising number of observations makes it increasingly likely that researchers may need to use a special computer, a cluster, or the cloud to host and process the data; however, the upside to working with more data is the hope to obtain much more accurate results with clinical applicability. In our experience, although today it is difficult to obtain large volumes of health data in some countries, it is possible to apply ML to smaller data volumes (small data) without compromising validity or applicability, especially when the data quality is high. The benefits for the health system can be significant, beginning with the potential improvements of diagnostic, prognostic, and therapeutic tools. In addition, analyzing and learning from smaller datasets now will allow our clinical research teams to become familiar with these techniques and be prepared for when big data containing clinical and healthcare information truly becomes a reality. However, it is very important to highlight that each method has its own merits and may be preferable to others under certain conditions or assumptions. Therefore, it is not possible to compare them in general terms. Results will depend on the data, context, user, processes of imputation, variable selection, parameter tuning, re-balancing, and data partitioning. And although one method may outperform others in some metrics, different scenarios will probably change comparisons again.

Another one of our research group's projects is to prepare a digital manual of ML algorithms with applicability to healthcare, which will be published in late 2019 so that health professionals can start their ML learning journey in a simple and friendly way. Meanwhile, all our code for this study is available on Github [17].

4. Conclusions

The use of ML algorithms when working with small databases (around 4000 participants) is relatively simple. These algorithms can enhance the diagnostic and prognostic capacity of more traditional regression techniques.

R-Studio is a powerful tool for conducting complex ML analytics with high reliability in creating a record of all changes. RapidMiner runs and visualizes ML algorithms using a very simple and intuitive graphic interface, although its capacity for manipulating the parameters can be smaller and less reliable in the case of complex analyses.

Mixed research teams, comprising healthcare professionals and computer scientists or mathematicians, are optimal for the conceptualization and development of ML projects.

Author contributions

JJB designed the article, obtained the database, performed the main analysis and drafted the text; EP performed the analysis with Rapid Miner and co-directed the main analysis with R-Studio; EG performed the descriptive analysis of the Framingham database and adapted the article to the requirements of the journal; ML collaborated in the overall design of the study. All the authors have reviewed and contributed revisions up to the final draft.

Financial support received

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

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

Acknowledgements

The authors of this study wish to express their gratitude to the team behind the Framingham Heart Study for providing open access to its data; to Elena Gazapo, Dean of the Faculty of Biomedical Sciences and Health at Universidad Europea de Madrid, for her support for the creation and development of the Machine Learning Health-UEM working group; to Andrew NG for making the machine learning algorithms he has developed accessible and comprehensible; to Brett Lantz and Hadley Wickham for their creativity and empowering us, as well as countless other researchers, through R; and to R-Ladies, Group R of Spain, and Machine Learning Spain, for being sources of constant inspiration for the work done by our group dedicated to teaching and research in machine learning applied to health.

References

[1] A. Samuel, Some studies in machine learning using the game of checkers, IBM J. Res. Dev. 3 (1959) 210–229, https://doi.org/10.1147/rd.33.0210.

[2] P. Rajpurkar, J. Irvin, K. Zhu, et al., CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, 2017, < https://arxiv.org/abs/1711.05225 > (accessed 20 mar 2018).

[3] M. Grewal, M. Muktabh, P. Kumar, et al., RADNET: Radiologist Level Accuracy using Deep Learning for HEMORRHAGE detection in CT Scans, 2018, < https://arxiv.org/abs/1710.04934 > (accessed 20 mar 2018).

[4] Z. Li, C. Wang, M. Han, et al., Thoracic Disease Identification and Localization with Limited Supervision, 2018. < https://arxiv.org/abs/1711.06373 > (accessed 20 mar 2018).

[5] P. Rajpurkar, A.Y. Hannun, M. Haghpanahi, et al., Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks, 2017, < https://arxiv.org/abs/1707.01836 > (accessed 20 mar 2018).

[6] D.S.W. Ting, C.Y. Cheung, G. Lim, et al., Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes, JAMA 318 (2017) 2211–2223, https://doi.org/10.1001/jama.2017.18152.

[7] A. Esteva, B. Kuprel, R.A. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118, https://doi.org/10.1038/nature21056.

[8] T. Alhanai, M. Ghassemi, J. Glass, Detecting depression with audio/text sequence modeling of interviews, Interspeech 2522 (2018) 1716–1720.

[9] Y.H. Huang, L.H. Wei, Y.S. Chen, Detection of the prodromal phase of bipolar disorder from psychological and phonological aspects in social media, 2017, < https://arxiv.org/pdf/1712.09183.pdf > (accessed 20 mar 2018).

[10] Secretaría General de Coordinación de Política Científica del Ministerio de Ciencia, Innovación y Universidades y al Grupo de Trabajo en Inteligencia Artificial GTIA. Estrategia Española de I + D + I en inteligencia artificial. Secretaría General Técnica del Ministerio de Ciencia, Innovación y Universidades, 2019, p. 48, < http://www.ciencia.gob.es/stfls/MICINN/Ciencia/Ficheros/Estrategia_Inteligencia_Artificial-

IDI.pdf > (accessed 13 mar 2019).

[11] M. Craglia, A. Annoni, P. Benczur, et al., Artificial Intelligence – A; Luxembourg; European Perspective. Joint Research Centre (JRC), the European Commission's. Publications Office of the European Union, 2018, 140p, doi: 10.2760/11251.

[12] Dictamen de iniciativa 526° pleno del cese de 31 de mayo y 1 de junio de 2017. Dictamen del Comité Económico y Social Europeo sobre la «Inteligencia artificial: las consecuencias de la inteligencia artificial para el mercado único (digital), la producción, el consumo, el empleo y la sociedad» (2017/C 288/01). Diario Oficial de la Unión Europea. Comité Económico y Social Europeo.

[13] R. D'Agostino, R.S. Vasan, M.J. Pencina, et al., General cardiovascular risk profile for use in primary care the Framingham heart study, Circulation 117 (2008) 743–753, https://doi.org/10.1161/CIRCULATIONAHA.107.699579.

[14] C. Brotons Cuixart, J.J. Alemán Sánchez, J.R. Banegas Banegas, et al., Grupo de Prevención Cardiovascular del PAPPS, Recomendaciones preventivas cardiovasculares, Actualización PAPPS 2018, Aten Primaria 50(Suppl 1) (2018) 4–28, doi: 10.1016/S0212-6567(18)30360-3.

[15] J. Marrugat, I. Subirana, E. Comin, et al., Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA study, J. Epidemiol. Commun. Health Rev. 61 (2007) 40–47.

[16] Aman Ajmera, Framingham Heart study dataset [online dataset]. Kaggle Inc; Publicado y actualizado 7 nov 2017. URL < https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset > (accessed 8 mar 2019), Versión 1.

[17] J.J. Beunza, R-studio code for Machine Learning algorithms applied to 10-year coronary risk in the Framingham Heart Study database, GitHub Inc. URL < https://github.com/Juanjobeunza/Aprendizaje-Automatico-FRAMINGHAM > (published on March 28, 2019), Updated and accessed on March 28, 2019.

[18] E. Puertas, Comparison of machine learning algorithms for the prediction of coronary heart disease by using the Framingham data set, GitHub Inc. URL < https://github.com/epuertas/framingham_Rapidminer > (published on March 22, 2019), Updated on March 22, 2019, (accessed on March 22, 2019).

[19] K.M. Lang, T.D. Little, Principled missing data treatments, Prev. Sci. 19 (2018) 284–294.

[20] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, Inc., Hoboken, NJ, 2019.

[21] S.F. Weng, J. Reps, J. Kai, et al., Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 12 (2017) 1–14.

[22] E. Puertas, Enrique Puertas Ganador del Premio FUJITSU OPEN DATA, URL < http://projectbasedschool.universidadeuropea.es/Enrique+Puertas+Ganador+del+Premio+FUJITSU+OPEN+DA > (published 9 Jun 2015), (accessed 28 mar 2019).

[23] Ganadores del II hackathon de tecnologías del lenguaje, URL < http://projectbasedschool.universidadeuropea.es/II_HackathonTecnologiaLenguaje > (accessed 28 mar 2019).