Deep Learning to Improve Heart Disease Risk Prediction

Shelda Sajeev[1(✉)], Anthony Maeder[1], Stephanie Champion[1], Alline Beleigoli[1], Cheng Ton[2], Xianglong Kong[3], and Minglei Shu[3]

[1] Flinders Digital Health Research Centre, School of Nursing and Health Sciences, Flinders University, Adelaide, Australia
shelda.sajeev@flinders.edu.au
[2] Department of Big Data Engineering Technology Research Center of E-Government, Jinan, Shandong, China
[3] Shandong Computer Science Center, Shandong Provincial Key Laboratory of Computer Networks, Qilu University of Technology, Jinan, China

Abstract. Disease prediction based on modeling the correlations between compounded indicator factors is a widely used technique in high incidence chronic disease prevention diagnosis. Predictive models based on personal health information have been developed historically by using simple regression fitting over relatively few factors. Regression approaches have been favored in previous prediction modeling approaches because they are simplest and do not assume any non-linearity in the model for contributions of the chosen factors. In practice, many factors are correlated and have underlying non-linear relationships to the predicted outcome. Deep learning offers a means to construct a more complex modeling approach, along with automation and adaptation. The aim of this paper is to assess the ability of a deep learning model to predict the heart disease incidence using a common benchmark dataset (University of California, Irvine (UCI) dataset). The performance of deep learning model has been compared with four popular machine learning models (two linear and two nonlinear) in predicting the incidence of heart disease using data from 567 participants from two cohorts taken from UCI database. The deep learning model was able to achieve the best accuracy of 94% and an AUC score of 0.964 when compared to other models. The performance of deep learning and nonlinear machine learning models was significantly better compared to the linear machine learning models with increase in the dataset size.

Keywords: Cardiovascular disease · Risk factors · Risk prediction · Machine learning · Deep learning

This research was funded by the Government of South Australia and Shandong Provincial Government, China.

H. Liao et al. (Eds.): MLMECH 2019/CVII-STENT 2019, LNCS 11794, pp. 96–103, 2019.
https://doi.org/10.1007/978-3-030-33327-0_12

1 Introduction

Cardiovascular disease (CVD) is the leading cause of death worldwide (30%) and is regarded as highly preventable (90%) [14]. Coronary heart disease also known as heart disease is the most common form of CVD [1]. Primary prevention is thus, a high priority and requires screening for severity of the risk factors, and generally addressing these with medication or health behavior changing interventions. Likelihood of heart disease is conventionally assessed from known highly indicative risk factors using compound formulas based on underlying Cox regression analysis methods [8]. A major longitudinal study (Framingham) conducted in USA has provided evidence for risk factor effects contributing to these formulas [4]. Several CVD risk prediction models to estimate an individual's risk of a CVD event within a given period are available [11]. However, the existing models are limited to the use of clinical decision (or prediction) rules in the form of simple heuristics and scoring systems. These models use a small set of variables (risk factors) that are easily observable, known to be clinically relevant and therefore easily incorporated into calculations. In addition, the traditional models do not assume any non-linear relationships between the predictors and the outcome measure and suffer from generalization and lacks the ability to be updated as new information becomes available.

Deep learning/machine learning is an emerging computational technique that can address the issues of multiple and correlated predictors, nonlinear relationships and interactions between the predictors and outcome, better than the traditional approach [6]. A recent investigation within a UK population found machine learning approaches predicted cardiac events more accurately, compared to conventional models [13]. The aim of the work reported here was to investigate plausibility of using deep learning/machine learning approach, by demonstrating its ability to derive prediction models for heart disease. This study discusses variations that can arise in the performance of some typical linear and more sophisticated non-linear machine learning prediction methods on a case study for heart disease, using data from the well-known public domain UCI dataset. The effects of different underlying populations on predictive performance, and the impact of combining cohorts to mimic a more general population, are considered.

2 Materials and Methods

2.1 Dataset

The dataset used for this study was taken from the University of California, Irvine (UCI) machine learning repository. A detailed information of the database can be found in the literature [2]. As a result of the small sample sizes in the available datasets, two datasets (cohorts) with 13 common risk factors/variables and no overlap in data instances were combined for the purposes of the machine learning analysis, in addition to analyzing each cohort individually. The two datasets used were the Statlog heart dataset (270 participants) and Cleveland

heart disease dataset (303 participants). Six participants were excluded from Cleveland dataset due to missing values, reducing the total sample to 567. The risk factors and the outcome variable used in the machine learning analysis are listed in Table 1.

2.2 Multi-Layer Perceptron - A Deep Learning Model

Multi-Layer Perceptron (MLP) is a traditional deep learning architecture [7]. It uses supervised learning called back propagation to train the model. It is a feed forward network consists of three types of layers (input, hidden and output). There could be one input layer, multiple hidden layers and one output layer. Nodes in each layer connected to every node in the previous and following layer. Nodes are not connected with any other node in the same layer. These connections carry a weight which represents the strength of the connection, typically initialized randomly. Learning is summarized by an attempt to determine which network connection weights best reduce the difference between predicted and true outputs. Activation function used on the node describes the nonlinear relationship between input of the node to the node output.

A basic MLP approach with 4 layers was used in this study: input layer, 2 hidden layers and output layer with 12, 8, 4 and 1 hidden units respectively. ReLU was used as the activation function for input and hidden layers. Sigmoid was the activation function used for the output layer. Loss function used was *binary-cross entropy* and *Adam* as optimizer. Deep learning environment used includes Python (3.6.6), Anaconda (5.3.0), Keras (2.2.4) and Tensorflow (1.11.0).

3 Experimental Setup and Performance Measures

In addition to MLP, four popular machine learning models (logistic regression (LR) [9], linear discriminant analysis (LDA) [10], support vector machine (SVM) with RBF kernel [12], and random forest (RF) [3]) were used for comparison. LR and LDA are simple linear classifiers, while SVM and RF are more advanced machine learning models that support non-linear classification. All the machine learning algorithms code was implemented in Python using the Scikit-learn library.

After removing missing values, the data was randomly divided into training and testing data. The training data consisted of 454 samples (80% of total data) and the remaining 113 samples (20%) were used for testing. Before feeding the data to the machine learning algorithms, some preprocessing was necessary. The data was normalized to zero mean and unit variance, to have each variable same influence on the cost function in designing the classifier.

In machine learning, a confusion matrix calculates the actual and predicted classifications for each class, measuring the accuracy of the algorithm and identifying the type of errors being made by the classifier. In this study, a confusion matrix was used to review the performance of the classification

algorithm. The two-class confusion matrix reports four outcomes; true positives (TP) for subjects with heart disease, correctly classified as cases, false positives (FP) for healthy subjects incorrectly classified as cases, true negatives (TN) for healthy subjects correctly classified as healthy, and false negatives (FN) for subjects with heart disease incorrectly classified as healthy. The performance measures extracted from the confusion matrix were sensitivity, specificity, precision and accuracy and that are calculated as follows: $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{TN+FP}$, $Precision = \frac{TP}{TP+FP}$ and $Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$.

To visualize the performance of the classification algorithm, a receiver operating characteristic (ROC) curve was used. The curve is calculated by plotting the TP rate against the FP rate for every possible threshold. The area under the curve was used as a measure of the accuracy of the classification algorithm, an accepted approach for evaluating classification performance. Additionally, to ensure stable classification results, the overall process was repeated 50 times for each machine learning model. Performances results reported in Tables 2 and 3 are the average score from 50 iterations.

4 Results

4.1 Study Population Characteristics

The characteristics of the study population are reported in Table 1. The average age of the participants was 54 years. There were substantially fewer women than men (32% women, 68% men). Of the participants, 14% had diabetes and 52% had high cholesterol (above 240). In addition, 51% exhibited an abnormality in ECG results and 31% exhibited major vessel calcification in fluoroscopy, while 33% experienced exercise induced angina. There were 257 (45%) cases of heart disease, from 567 participants. In Statlog cohort, there were 120 cases out of 270 (44%) and in Cleveland 137 cases out of 297 (46%).

4.2 Prediction Accuracy

Tables 2 and 3 show the performance comparison of deep learning model and four machine learning models for predicting heart disease incidence for individual cohort and combined cohort respectively. As mentioned previously, the performance of the predictive models was accessed using sensitivity, specificity, precision, accuracy and AUC score. For individual cohort analysis, the machine learning model achieved an accuracy up to 0.838 and an AUC score up to 0.913 for Statlog cohort and an accuracy up to 0.840 and an AUC score up to 0.912 for Cleveland cohort. The results of the modeling indicated that the performance of the linear and nonlinear classifiers was similar in both cohorts.

For combined cohort analysis, deep learning model MLP obtained the highest scores (sensitivity = 0.932, specificity = 0.957, precision = 0.942, accuracy = 0.940 and an AUC score of 0.964). The next highest performance was achieved

Table 1. List of all 13 variables and the outcome variable that were used for machine learning analysis and their characteristics for combined cohort (Statlog and Cleveand).

Variables	Description	Values
Age	-	54.49 ± 9.06
Sex	Male	384 (68%)
	Female	183 (32%)
Cp	Chest pain type	
	Typical angina	43 (7.5%)
	Atypical angina	91 (16%)
	Non-anginal pain	162 (28.5%)
	Asymptomatic	271 (47.8%)
Trestbps	Resting blood pressure	131 ± 17.8
Chol	Serum cholesterol	248 ± 51.8
Fbs	Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	
	Diabetics	83 (14.6%)
	Non-diabetics	484 (85.4%)
Restecg	Resting electrocardiographic results	
	Normal	278 (49%)
	Having ST-T wave abnormality	6 (1%)
	Showing probable or definite left ventricular Hypertrophy by Estes criteria	283 (50%)
Thalach	Maximum heart rate achieved	149 ± 23
Exang	Exercise induced angina (1 = yes; 0 = no)	
	No	381 (67%)
	Yes	186 (33%)
Oldpeak	ST depression induced by exercise relative to rest (0–6.2)	1.05 ± 1.15
Slope	The slope of the peak exercise ST segment	
	Upsloping	269 (47.4%)
	Flat	259 (45.7%)
	Down sloping	39 (6.9%)
Ca	Number of major vessels (0–3) colored by fluoroscopy	
	0	334 (58.9%)
	1	123 (21.7%)
	2	71 (12.5%)
	3	39 (6.9%)
Thal	Thallium stress test	
	Normal	316 (55.7%)
	Fixed defect	32 (5.6%)
	Reversible defect	219 (38.7%)
Presence of heart disease		257 (45%)

by RF (sensitivity = 0.890, specificity = 0.955, precision = 0.943, accuracy = 0.933 and an AUC score of 0.963). It can be seen that deep learning approach gives the best results in all performance measures except precision, where it is

comparable with random forest. Further, the nonlinear models (MLP, RF and SVM) showed considerably superior results than the linear ones (LR and LDA).

Table 2. Comparison of the performance of deep learning and four machine learning models using thirteen risk factors predicting heart disease incidence for individual cohorts (Statlog and Cleveand). The reported values are the average of 50 iterations. DL represents deep learning.

Algorithms	Sensitivity	Specificity	Precision	Accuracy	AUC
Statlog heart dataset					
Logistic regression	**0.807**	0.859	0.821	0.836	0.910
Linear discriminant analysis	0.798	0.870	0.830	**0.838**	0.909
Support vector machine - RBF	**0.807**	0.849	0.849	0.830	0.907
Random Forest	0.788	0.879	0.838	0.836	**0.913**
DL - Multi-Layer Perceptron	0.701	**0.907**	**0.856**	0.814	0.881
Cleveland heart dataset					
Logistic regression	**0.794**	0.869	0.841	0.834	0.903
Linear discriminant analysis	0.789	**0.886**	0.858	**0.840**	0.904
Support vector machine - RBF	0.773	0.867	**0.867**	0.828	0.900
Random forest	0.778	0.883	0.853	0.832	**0.912**
DL - Multi-Layer Perceptron	0.780	0.879	0.850	0.833	0.861

Table 3. Comparison of the performance of deep learning and four machine learning models using thirteen risk factors predicting heart disease incidence for *combined* cohort (Statlog and Cleveand). The reported values are the average of 50 iterations. DL represents deep learning.

Algorithms	Sensitivity	Specificity	Precision	Accuracy	AUC
Logistic regression	0.817	0.873	0.844	0.848	0.913
Linear discriminant analysis	0.800	0.888	0.857	0.848	0.911
Support vector machine - RBF	0.866	0.906	0.885	0.888	0.943
Random forest	0.890	0.955	**0.943**	0.933	0.963
DL - Multi-Layer Perceptron	**0.932**	**0.957**	0.942	**0.940**	**0.964**

Figure 1 shows the ROC curves for all the five predictive models for combined cohort. The ROC curves have been drawn for one of the best cases of the 50 iterations. An AUC score 0f 0.988 was achieved using MLP. This indicates that the deep learning have the potential to build highly accurate prediction system that could give a second opinion in clinical decision making.

5 Discussion

In this study we presented deep learning and machine learning methodologies for predicting the presence of heart disease. Results for predictive accuracy obtained from deep learning model is compared with two popular linear (LR

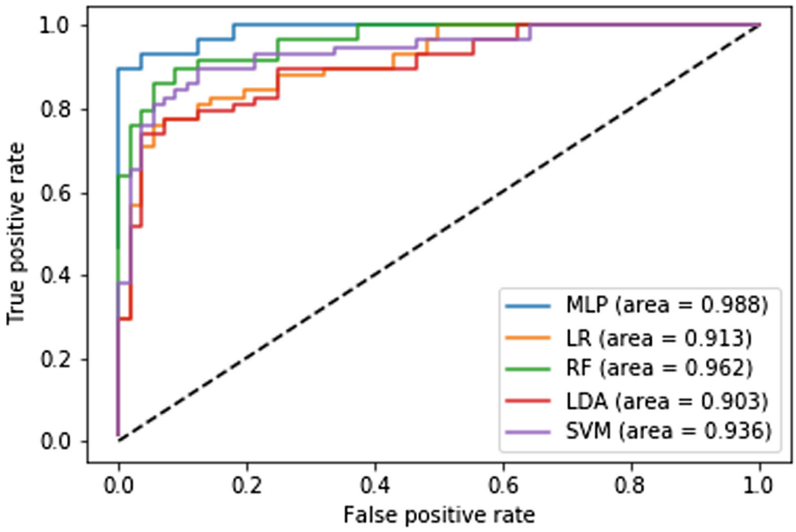

Fig. 1. ROC curves for MLP, LR, RF, LDA, SVM models for UCI study participants (Statlog and Cleveland cohorts combined). ROC is drawn for one of the 50 iterations.

and LDA) and non-linear machine learning models (SVM and RF). The models were applied on 13 highly indicative factors in the datasets, comparable with factors used in standard Framingham derived models. Evidence of heart disease diagnosis was available within the datasets through clinical history of chest pain, resting and exercise electrocardiogram, myocardial scintigraphy or angiogram tests (45% of cases). The results for application to two cohorts from different sources show that even for a small dataset, machine learning models can produce good results and variations in comparable cohorts do not affect this adversely. Furthermore, when the cohorts are combined, the overall non-linear model's performance increases significantly, while the results from linear models remain similar. The reason for superior performance could be due to its flexibility and non-linear function. Our train/test technique with 50 iterations assured the independence of the testing samples from training samples and validation of the model effectiveness.

As the deep learning model was created and tested on 2 small datasets, we have plans to validate the model in larger cohorts that will enable us to investigate the potential of deep learning with multiple layers and explore its suitability for general population heart disease risk prediction.

The availability of larger datasets from the electronic health records would allow deep learning/machine learning to discover unseen relationship and find new risk factors previously not identified as highly relevant. In addition, it could lead to the development of better cohort-based risk models and perhaps even individually tailored risk profiles. Finally, in this study we have not compared the proposed approach with the popular CVD risk prediction model: the American College of Cardiology/American Heart Association (ACC/AHA) model [5], as the information to compute the AHA model was not available in UCI dataset.

6 Conclusion

This work demonstrates value in considering deep learning method for disease prediction modeling, and the potential for modeling performance to improve as dataset size increases. This suggests that the deep learning approach may be more effective for maintaining prediction accuracy for datasets which change over time, as well as for specialized cohorts within the overall population, for which prediction may be less accurate due to deviation from the standard model. It provides an exciting prospect for achieving better and more specific disease risk assessment that may assist the drive towards personalised medicine.

References

1. AIHW: Cardiovascular disease: Australian facts 2011. Cardiovascular disease series. Cat. no. CVD 53. Canberra. Australian Institute of Health and Welfare (2011)
2. Bache, K., Lichman, M.: UCI Machine Learning Repository Irvine. University of California, School of Information and Computer Science, Oakland (2013). http://archive.ics.uci.edu/ml
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. D'Agostino, R.B., et al.: General cardiovascular risk profile for use in primary care. Circulation **117**(6), 743–753 (2008)
5. Goff, D.C., et al.: 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association task force on practice guidelines. J. Am. Coll. Cardiol. **63**(25 Part B), 2935–2959 (2014)
6. Goldstein, B.A., Navar, A.M., Carter, R.E.: Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur. Heart J. **38**(23), 1805–1814 (2016)
7. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
8. Hlatky, M.A., et al.: Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation **119**(17), 2408–2416 (2009)
9. Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, vol. 398. Wiley, Hoboken (2013)
10. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.R.: Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No. 98th8468), pp. 41–48. IEEE (1999)
11. Sajeev, S., Maeder, A.: Cardiovascular risk prediction models: a scoping review. In: Proceedings of the Australasian Computer Science Week Multiconference, p. 21. ACM (2019)
12. Van Gestel, T., et al.: Benchmarking least squares support vector machine classifiers. Mach. Learn. **54**(1), 5–32 (2004)
13. Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M., Qureshi, N.: Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE **12**(4), e0174944 (2017)
14. WHO: Prevention of cardiovascular disease : guidelines for assessment and management of total cardiovascular risk. World Health Organization (2007)