

Long-term Cholesterol Risk Prediction using Machine Learning Techniques in ELSA Database

Nikos Fazakis^a, Elias Dritsas^b, Otilia Kocsis^c, Nikos Fakotakis and Konstantinos Moustakas^d

*Department of Electrical and Computer Engineering, University of Patras, 26504 Rion, Greece
{fazakis, okocsis, moustakas}@ece.upatras.gr, dritsase@ceid.upatras.gr, fakotakis@upatras.gr*

Keywords: Cholesterol, Long-term Prediction, Machine Learning.

Abstract: Cholesterol is a crucial risk factor for cardiovascular diseases (CVDs) which in their turn are among the main causes of death worldwide and public health concern, with heart diseases being the most prevalent ones. For cholesterol control, the early prediction is considered one of the most effective ways. Utilizing the English Longitudinal Study of Ageing (ELSA), a large-scale database of ageing participants, a dataset is derived to evaluate the long-term cholesterol risk of elderly men and women using Machine Learning (ML) techniques. Several ML prediction models were assessed concerning Accuracy and Recall where the Logistic model tree was the best performer. The ultimate goal of this study is to identify individuals at risk and facilitate earlier intervention to prevent the future development of cholesterol.

1 INTRODUCTION

Cholesterol is a waxy substance found in many of the consumed foods and also in the body cells. There are three main types of cholesterol in the body cells. High-density lipoprotein (HDL), also called the "good" cholesterol, it helps removing excess cholesterol from the body. In contrast, Low-density lipoprotein (LDL), the bad or "lousy" cholesterol, it can lead to a buildup of plaque in the arteries. In addition, Very low-density lipoprotein (VLDL) exists, which also tends to promote plaque buildup.

Quantifying an individual's risk for developing non-contiguous chronic conditions such as high cholesterol, which are linked to controllable lifestyle behaviours and attitudes, is an important goal of prediction analysis in healthcare, as it is linked on one side with the long-term well-being on the individual and active and independent ageing, and on the other side with important potential savings for the social care systems. Recent research has demonstrated that it is possible to use ML tools to predict individual risk of hospitalization by only using data related to socio-economic features (age group, gender and race) and behavioural data, without requiring clinical risk factors

(Chen et al., 2020). In this context, the SmartWork (Kocsis et al., 2019) system has been developed with the aim to provide the ageing sedentary population with the right tools in order to promote healthy and active ageing and improve their workability. During this process the monitoring of the individual's health and the prognosis of several chronic diseases, i.e. high cholesterol, is considered as a vital step.

Cholesterol levels are tested through blood cholesterol or lipid tests, and although high levels are asymptomatic, the relation between the different forms of cholesterol provides an indication of risks of developing heart diseases (Group et al., 2000). Thus, preventing or lowering high cholesterol levels is directly linked to lowering risks of cardiovascular diseases. The main risk factors (Anagnostis et al., 2015), (Crouse et al., 1995), (Wakabayashi and Groschner, 2009) potentially impacting on the high cholesterol levels include:

- Gender: after menopause, a woman's LDL cholesterol level usually goes up.
- Age: men aged 45 years or older and women aged 55 years or older are at increased risk of high cholesterol and heart disease.
- Family history of heart diseases: the risk of high cholesterol may increase if a father/ brother was affected by heart disease before age 55 or a mother/ sister was affected by heart disease before age 65.

^a <https://orcid.org/0000-0001-7687-2380>

^b <https://orcid.org/0000-0001-5647-2929>

^c <https://orcid.org/0000-0001-6937-6442>

^d <https://orcid.org/0000-0001-7617-227X>

- Diet: high fat, high sodium, or frequent consumption of food from animal sources (red meat, eggs, cheese, etc.) increases total and LDL cholesterol levels.
- Obesity/ High BMI: a BMI higher than 25 is a high-risk factor for increased LDL cholesterol levels.
- Physical activity: increased physical activity and regular exercising helps to lower LDL and raise HDL cholesterol levels.
- Alcohol: although moderate alcohol intake is linked to increasing HDL-cholesterol, excessive alcohol intake may increase LDL-cholesterol and triglyceride levels.
- Smoking: past or present smoking, or exposure to tobacco smoke (passive smoking) increases cholesterol levels.
- Diabetes: persons suffering of diabetes have increased cholesterol levels.

The contribution of this paper is twofold: i) the engineering of a meaningful dataset that can facilitate the prognosis of high cholesterol regarding the elder workers, ii) a comparative assessment of different ML models families in order to spot the most efficient for the long-term cholesterol risk prediction of older people aged at least 50 years. At this point, it should be noted that these models will be integrated into the predictive AI tools of the SmartWork system, which aims to sustain workability of older office workers (Kocsis et al., 2019) based on personalized models acquired from the integration of the health condition of specific patient models. Cholesterol is one of the chronic conditions that will be considered in SmartWork (Fazakis et al., 2021).

The rest of this paper is organized as follows. Related work regarding the prediction of high cholesterol as chronic disease is discussed in Section 2. In Section 3 the main parts of the methods for the long-term risk prediction of cholesterol are displayed. In particular, the design of a training and testing dataset, feature selection and the experiments setup are presented. Section 4 concludes the paper.

2 RELATED WORKS

Many studies have been conducted, in relation to high cholesterol risk assessment, but in most of the cases these are directly linked to risk of cardiovascular diseases.

The Framingham Heart Study (Mahmood et al., 2014) led to the development of LDL Cholesterol

Goal Level Calculator¹, which is mainly based on demographic and lifestyle variables.

Jian Liu et al (Liu et al., 2005), assess coronary heart disease (CHD) risk within levels of the joint distribution of non-HDL and LDL cholesterol among individuals with and without diabetes. The results of their study indicated that diabetes condition is a strong risk factor for coronary heart disease (CHD) death. Cengiz Colak et al, (Colak et al., 2016) study was carried out to predict the cholesterol level in patients with Myocardial Infarction (MI) using Artificial Neural Networks (ANN) and Support Vector Machines (SVM) models. The results pointed out that ANN has higher predictive performance in comparison with SVM in predicting cholesterol level. Moreover, the prediction of blood cholesterol levels from genotype data was studied by Francesco Reggiani et. al, (Reggiani et al., 2020). In addition, Jin Sol Lee et al, (Lee et al., 2018) considered the prediction of cholesterol ratios within a Korean population.

Finally, the most significant risk factors contributing to the development of high cholesterol, as suggested by the different research works, are summarized in Table 2.

3 LONG-TERM CHOLESTEROL RISK ASSESSMENT

Although high cholesterol levels are usually asymptomatic, their assessment is important as a risk screening tool for cardiovascular diseases. In particular, the ratios $\frac{totalcholesterol}{HDL}$ and $\frac{LDL}{HDL}$ provide a good indication of risk of cardiovascular diseases, as shown in Table 1.

Table 1: Cholesterol Levels and Risk of Cardiovascular Diseases.

$\frac{totalcholesterol}{HDL}$		$\frac{LDL}{HDL}$		Risk
Male	Female	Male	Female	
< 3.4	< 3.3	1	1.5	Low
5.0	4.5	3.6	3.2	Average
9.5	7.0	6.3	5.0	Moderate
> 23	> 11	8	6.1	High

3.1 Training

The training and test dataset for the high cholesterol risk prediction is based on the ELSA (Marmot et al.,

¹<https://www.mcw.edu/calculators/ldl-cholesterol-goal-level>

Table 2: Cholesterol risk scoring systems.

Author	Model	Tool Risk factors included
Framingham Heart 10-year risk LDL cholesterol Level	LDL cholesterol Level calculator	Gender, Age, Smoke status, High Blood Pressure, Medication for High Blood Pressure
Jian Liu et al (Liu et al., 2005)	Joint Distribution of Non-HDL and LDL Cholesterol and Coronary Heart Disease Risk Prediction Among Individuals with and Without Diabetes	Ethnicity, Gender, Age, Smoke status, Drinking Alcohol, High Blood Pressure
Cengiz Colak et al (Colak et al., 2016)	Prediction of Cholesterol Level in Patient with Myocardial Infarction Based on Data Mining Methods	Age, Gender, Smoke status, High Blood Pressure, Physical Activity, Sedentary lifestyle
Francesco Reggiani et al (Reggiani et al., 2020)	Prediction of blood cholesterol levels from genotype data	BMI, Age, Ethnicity, Gender, High Blood Pressure, Physical Activity
Jin Sol Lee et al (Lee et al., 2018)	Prediction of cholesterol ratios within a Korean population	Triglyceride, Age, Gender, BMI, High Blood Pressure

Table 3: Distribution per gender of newly diagnosed with High Cholesterol at 2-years follow-up in the original dataset.

Wave	High Cholest.	Male	Female	Total
Ref 2	No	2,339	2,829	5,168
F-up 3	Yes	400	552	952
Ref 4	No	2,531	3,161	5,692
F-up 5	Yes	292	293	585
Ref 6	No	2,196	2,797	4,993
F-up 7	Yes	167	173	340
All	No	7,066	8,787	15,853
	Yes	859	1,018	1,877

2018) database, which consists of reference waves 2, 4 and 6 as baseline and the respective waves 3, 5, and 7 for the 2-years follow-up assessment. Although the number of participants in ELSA waves selected as reference one (namely waves 2, 4, and 6) is very large, initially we drop out participants that already have high cholesterol levels at reference waves and participants that did not take the interview at both, the reference and the corresponding follow-up wave. In Tables 3 and 5, the distributions of selected participants that satisfied the above criteria, per age group and per gender are presented.

These distributions however correspond to a dataset that is not representative for the population, as they do not relate well to the prevalence of high cholesterol for these age groups as they have been re-

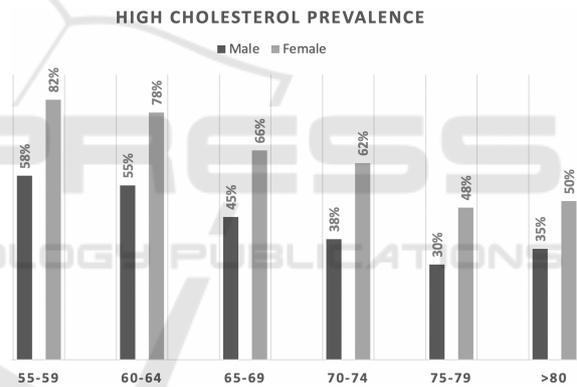


Figure 1: Age Group and Gender distributions in England.

ported at country level². The reported distributions per age group and per gender in England are shown in Figure 1.

Similarly, the prevalence of high cholesterol among people of older age at European level has been reported as being at over 80% for people aged 55-64, and at 79% for people aged 65-74 years old. It is worth noticing, that in Portugal (one of the pilot sites of SmartWork system), the prevalence of high cholesterol among adult population (25-74 years old) is at 63.3% (Rodrigues et al., 2016).

Taking into account these findings, we balanced the dataset using random undersampling, in order to reach distributions similar to those reported for the reference population. The distributions of the partici-

²<https://www.statista.com/statistics/983478/england-high-cholesterol-among-older-people/>

Table 4: Distribution per gender of newly diagnosed with High Cholesterol at 2-years follow-up in balanced dataset.

Wave	High Cholest.	Male	Female	Total
Ref 2	No	756	978	1,734
F-up 3	Yes	400	552	952
Ref 4	No	509	541	1,050
F-up 5	Yes	292	293	585
Ref 6	No	287	320	607
F-up 7	Yes	167	173	340
All	No	1552	1839	3,391
	Yes	859	1,018	1,877

pants per age group and gender in the balanced dataset are shown in Tables 4 and 6.

3.2 Features Selection

The initial features set considered for the training of the ML-based models included 106 variables, with 61 being categorical and 45 numeric attributes, among those collected at the reference waves of ELSA dataset. However, their importance was established using utilizing a feature selection method based on a variation of Random Forests (Genuer et al., 2010). According to this method, the attributes are ranked using the Gini importance score of the model's trees. The Gini index (Sundhari, 2011) is calculated as follows

$$Gini = 1 - \sum_{i=1}^c p_i^2, \quad (1)$$

where c is the number of classes and p_i the relative frequency of class i in the dataset.

To further reduce the selected number of attributes without harming the information content of the constructed dataset a stepwise backward elimination process was also employed using the Logistic model. The final list of features was reduced to 22 attributes, among which the most important are: total, HDL and LDL cholesterol levels at the reference wave, age group, gender, weight and BMI, drinking and smoking habits, physical activity, education level, diagnosis of other chronic conditions (e.g. diabetes, stroke, high blood pressure), and self-assessed health status.

3.3 Performance Evaluation

In order to assess the overall performance of the engineered dataset, several ML algorithms were employed, covering a wide range of classifier families. The models Naïve Bayes (NB) (Quinlan, 2014), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Artificial Neural Network (ANN) using

2 hidden layers, 5 Nearest Neighbors (5-NN) (Aha et al., 1991), Rotation Forest (RotF) (Rodríguez et al., 2006), Decision Trees (DT) (Quinlan, 1986), Logistic Model Trees (LM Trees) (Landwehr et al., 2005) and Random Forest (RF) (Breiman, 2001) were applied on the constructed dataset using a 10-cross validation (Stone, 1978) experimentation setup.

The model performance results can be seen in Table 7, with metrics of accuracy and recall being recorded for each prediction model. As can be observed, although all trained models present similar moderate accuracies, with the highest being 62.99% for the LM Trees classifier, the recall demonstrated by the models is relatively high, meaning that the prevalence of high cholesterol can be indicated with a satisfying probability in populations similar with the engineered dataset, and thus this fact is considered advantageous for the prognosis of the disease in the context of the SmartWork system. Finally, the best overall performance was obtained with the LM Trees model, which performed best both with respect to accuracy and recall metrics.

4 CONCLUSIONS

In this research work, a dataset was constructed based on the ELSA database, aiming at the prognosis of high cholesterol, targeting the elder office workers. Several ML methods were examined and the LM Trees classifier was found to wield the best prediction performance against the different families of classifiers. Although, the recorded accuracies were moderate, the results presented consistently high recalls, a fact that seems promising for the discrimination ability of the models regarding possible positive subjects.

Moreover, a possible path for increasing the achieved accuracy is the employment of deep learning (Liu et al., 2017) models and techniques, as they can provide complex decision boundaries and thus better fit the training data.

A limitation of the current study is that the engineered dataset does not consider in the features set the family history in relation to high cholesterol levels as such information was not available in ELSA.

Future directions of this research work could include, the further research on feature ranking and selection techniques such as Least Absolute Shrinkage and Selection Operator (LASSO) (Muthukrishnan and Rohini, 2016), and the imputation of the missing values of the dataset in order to improve the contained information of the dataset. Moreover, learning paradigms such as self-labeling (Triguero et al., 2015) can be proved very useful in exploiting the

Table 5: Distribution per age group of newly diagnosed with High Cholesterol at 2-years follow-up in the original dataset.

Wave	High Cholesterol	50-54	55-59	60-64	65-69	70-74	75+	Total
Ref wave 2	No	652	1,142	827	754	629	1,164	5,168
F-up wave 3	Yes	92	221	183	158	129	169	952
Ref wave 4	No	852	1,245	1,165	759	711	960	5,692
F-up wave 5	Yes	65	119	122	113	89	77	585
Ref wave 6	No	561	994	1,027	857	566	988	4,993
F-up wave 7	Yes	46	74	74	54	47	45	340
All waves	No	2,065	3,381	3,019	2,370	1,906	3,112	15,853
	Yes	203	414	379	325	265	291	1,877

Table 6: Distribution per age group of newly diagnosed with High Cholesterol at 2-years follow-up in the balanced dataset.

Wave	High Cholesterol	50-54	55-59	60-64	65-69	70-74	75+	Total
Ref wave 2	No	167	316	275	285	258	433	1,734
F-up wave 3	Yes	92	221	183	158	129	169	952
Ref wave 4	No	118	170	183	204	178	197	1,050
F-up wave 5	Yes	65	119	122	113	89	77	585
Ref wave 6	No	84	106	111	97	94	115	607
F-up wave 7	Yes	46	74	74	54	47	45	340
All waves	No	369	592	569	586	530	745	3,391
	Yes	203	414	379	325	265	291	1,877

Table 7: Performance Evaluation of ML models.

	NB	SVM	ANN	5-NN	RotF	DT	LM Trees	RF
Accuracy	62.69%	59.51%	61.42%	56.56%	61.86%	61.39%	62.99%	61.36%
Recall	68.90%	72.70%	66.70%	67.70%	69.60%	72.20%	73.50%	68.80%

vast amounts of unlabeled data presented in ELSA database, thus significantly increasing the constructed dataset size. Finally, the planned future updates on the ELSA database waves can be utilized to extend the engineered dataset, providing better modeling capabilities.

ACKNOWLEDGEMENTS

This work has been partially supported by the Smart-Work project (GA826343), EU H2020 and SC1-DTH-03-2018 - Adaptive smart working and living environments supporting active and healthy ageing.

REFERENCES

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, 6(1):37–66.

Anagnostis, P., Stevenson, J. C., Crook, D., Johnston, D. G., and Godsland, I. F. (2015). Effects of menopause, gender and age on lipids and high-density lipoprotein cholesterol subfractions. *Maturitas*, 81(1):62–68.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chen, S., Bergman, D., Miller, K., Kavanagh, A., Frownfelter, J., and Showalter, J. (2020). Using applied machine learning to predict healthcare utilization based on socioeconomic determinants of care. *Am J Manag Care*, 26(01):26–31.

Colak, C., Çolak, M. C., Ermiş, N., Erdil, N., and Özdemir, R. (2016). Prediction of cholesterol level in patients with myocardial infarction based on medical data mining methods.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Crouse, S. F., O’Brien, B. C., Rohack, J. J., Lowe, R. C., Green, J. S., Tolson, H., and Reed, J. L. (1995). Changes in serum lipids and apolipoproteins after exercise in men with high cholesterol: influence of intensity. *Journal of Applied Physiology*, 79(1):279–286.

- Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., and Moustakas, K. (2021). Machine learning tools for long-term type 2 diabetes risk prediction. *IEEE Access*, 9:103737–103757.
- Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, 31(14):2225–2236.
- Group, B. S. et al. (2000). Secondary prevention by raising hdl cholesterol and reducing triglycerides in patients with coronary artery disease. the bezafibrate infarction prevention (bip) study. *Circulation*, 102:21–27.
- Kocsis, O., Moustakas, K., Fakotakis, N., Vassiliou, C., Toska, A., Vanderheiden, G. C., Stergiou, A., Amaxilatis, D., Pardal, A., Quintas, J., et al. (2019). Smart-work: designing a smart age-friendly living and working environment for office workers. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 435–441.
- Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. *Machine learning*, 59(1-2):161–205.
- Lee, J. S., Cheong, H. S., and Shin, H. D. (2018). Prediction of cholesterol ratios within a korean population. *Royal Society open science*, 5(1):171204.
- Liu, J., Sempos, C., Donahue, R. P., Dorn, J., Trevisan, M., and Grundy, S. M. (2005). Joint distribution of non-hdl and ldl cholesterol and coronary heart disease risk prediction among individuals with and without diabetes. *Diabetes care*, 28(8):1916–1921.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26.
- Mahmood, S. S., Levy, D., Vasan, R. S., and Wang, T. J. (2014). The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, 383(9921):999–1008.
- Marmot, M., Oldfield, Z., Clemens, S., Blake, M., Phelps, A., Nazroo, J., et al. (2018). English longitudinal study of ageing: Waves 0–8, 1998–2017.
- Muthukrishnan, R. and Rohini, R. (2016). Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE international conference on advances in computer applications (ICACA)*, pages 18–20. IEEE.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.
- Quinlan, J. R. (2014). *C4.5: programs for machine learning*. Elsevier.
- Reggiani, F., Carraro, M., Belligoli, A., Sanna, M., Dal Prà, C., Favaretto, F., Ferrari, C., Vettor, R., and Tosatto, S. C. (2020). In silico prediction of blood cholesterol levels from genotype data. *PLoS one*, 15(2):e0227191.
- Rodríguez, J., Kuncheva, L., and Alonso, C. (2006). Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28:1619–30.
- Rodrigues, A. P., Kislaya, I., Antunes, L., Gaio, V., Barreto, M., Santos, A., Gil, A., Namorado, S., Lyshol, H., Nunes, B., et al. (2016). Prevalence of elevated cholesterol in portugal: National health examination survey results (2015) ana paula rodrigues. *The European Journal of Public Health*, 26(suppl_1):ckw174–112.
- Stone, M. (1978). Cross-validation: A review. *Statistics: A Journal of Theoretical and Applied Statistics*, 9(1):127–139.
- Sundhari, S. S. (2011). A knowledge discovery using decision tree by gini coefficient. In *2011 International Conference on Business, Engineering and Industrial Applications*, pages 232–235. IEEE.
- Triguero, I., García, S., and Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284.
- Wakabayashi, I. and Groschner, K. (2009). Modification of the association between alcohol drinking and non-hdl cholesterol by gender. *Clinica Chimica Acta*, 404(2):154–159.