2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Machine Learning Model for Predicting CVD Risk on NHANES Data*

G. A. Klados**, K. Politof**, E. S. Bei, K. Moirogiorgou, N. Anousakis-Vlachochristou, G. K. Matsopoulos, *Member, IEEE* and M. Zervakis, Senior *Member, IEEE*

Abstract— **Cardiovascular disease (CVD) is a major health problem throughout the world. It is the leading cause of morbidity and mortality and also causes considerable economic burden to society. The early symptoms related to previous observations and abnormal events, which can be subjectively acquired by self-assessment of individuals, bear significant clinical relevance and are regularly preserved in the patient's health record. The aim of our study is to develop a machine learning model based on selected CVD-related information encompassed in NHANES data in order to assess CVD risk. This model can be used as a screening tool, as well as a retrospective reference in association with current clinical data in order to improve CVD assessment. In this form it is planned to be used for mass screening and evaluation of young adults entering their army service. The experimental results are promising in that the proposed model can effectively complement and support the CVD prediction for the timely alertness and control of cardiovascular problems aiming to prevent the occurrence of serious cardiac events.**

I. Introduction

As reported by the World Health Organization (WHO), CVD is the number one cause of death globally, accounting for about 17.9 million deaths per year. CVDs form a source of great morbidity and mortality, directly affecting the economy of the societies [1]. CVD is a group of diseases that includes coronary heart disease (CHD), cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis, and pulmonary embolism [1]. Reducing the incidence by identifying those individuals at highest risk of CVDs from earlier symptoms of markers can play an important role as part of an overall solution along with the appropriate clinical evaluation and treatment.

Cardiac risk stratification can be merely viewed as an assessment scheme used to evaluate a patient's risk of developing cardiovascular disease (CVD) [2].

*Research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EΔK-03505, e-MASS).

G. Klados, K. Politof, E. S. Bei, K. Moirogiorgou and M. Zervakis are with the School of Electrical and Computer Engineering, Technical University of Crete, Chania, GR-73100, Hellas (phone: +30-28210-37003; email: gklados@isc.tuc.gr, kpolitof@isc.tuc.gr, abei@isc.tuc.gr, dina@display.tuc.gr, michalis@display.tuc.gr). **Equally contributed.

N. Anousakis-Vlachochristou is with the Naval Hospital of Athens, Athens, GR-11521, Hellas (email: anousakisvn@gmail.com).

G. K. Matsopoulos is with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens, GR-15780, Hellas (email: gmatso@esd.ece.ntua.gr).

Previous work on historic, non-laboratory data provide evidence that several risk scores including age, sex, smoking, diabetes, systolic blood pressure, treatment of hypertension and body-mass index, might bear valuable information for CVD risk assessment [3]. Most of these factors may be seen as prodromal symptoms and warnings to the individual, which can supplement the risk score obtained from classical biochemical measurements (such as cholesterol values). For this reason, the self-assessment questionnaires are used to supplement most clinical procedures.

Based on NHANES 2003-2004 data and other NHANES datasets, previous work has focused on issues such as exploring the distribution of cardiorespiratory fitness and its association with obesity and leisure-time physical activity, analyzing the relationship of plasma fatty acids and cardiovascular fitness, examining the prevalence of coronary heart disease, investigating the associations between cardiovascular health metrics and family history of premature heart disease, or unveiling associations of multi-morbidity with functional limitations [4-9]. Furthermore, patient health questionnaires have been successfully and systematically used in assessing and screening for health conditions, including heart diseases [10].

This study focuses mainly on subjective self-evaluation aiming to derive such historic markers (or signs) that are most significant in assessing the development of CVD progression. The utilization of these markers can be seen in population screening for risk assessment, as well as in supplementing the regular examination means of current health status (physiological and biochemical examination, electrocardiogram, etc.) for CVD evaluation, through a machine learning model. The purpose of the current study was dual: 1) to collate and extract the most important physical and biological variables correlated to CVD, by taking advantage of the questionnaire- and examination-based CVD information held in NHANES 2003-2004 data; and 2) to generate a machine learning model capable to exploit this knowledge – selected patient's history events and physical examination values - in order to propose a CVD risk evaluation tool [4]. Our study aims primarily to assess a CVD-significant patient health questionnaire extracted out of the multiple health aspects in the NHANES dataset. The paper proceeds as follows. Section II presents the study framework describing the data categories, subject and study settings. Section III presents the proposed evaluation process with the machine learning scheme. The results are presented in Section IV concluding in section V.

II. METHODOLOGICAL FRAMEWORK

A. Dataset Selection

Data was obtained from the database of the National Health and Nutrition Examination Health Survey (NHANES) 2003–2004, conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention [4]. Participants in the NHANES surveys, who are selected by a multistage stratified probability sampling technique, undergo a variety of interviews, physical examinations and laboratory tests. The findings of the NHANES surveys are used to determine the prevalence of major diseases and risk factors for diseases. Aiming at the development of a prediction model for CVD risk assessment, we focused on four different data components, namely 'Physical Functioning', 'Medical Conditions', 'Cardiovascular Health', and 'Cardiovascular Fitness', from which we selected a number of variables that related to cardiovascular health, as illustrated in Fig. 1. Details on the questionnaires and examination components can be found in the NHANES 2003-2004 reference manuals [4].

Selection Framework: All risk factors are considered categorical variables encoded in an integer scale, e.g., as yes, no or missing. We mention here that category D induces certain dependence on the variables through the sequence of questions asked in the questionnaire. For example, the positive answer to CDQ001 leads to CDQ002, whereas the negative answer leads to CDQ010 and the rest variables are ignored, but are actually denoted as 'missing' in the database. 'Missing' encodes the response of non-interest and it is denoted as "false-missing".

B. Subjects and Setting

Considering all 33 selected variables included in the above four categories, all subjects were classified under three groups: 1) 'healthy', 2) 'non-healthy' and 3) 'needs further tests'. Inclusion criteria for the 'non-healthy' group were as follows: (i) heart problem for category A, (ii) a personal history of congestive heart failure, CHD, angina/angina pectoris, or heart attack prior to entry into the NHANES study for category B, (iii) presence of cardiovascular conditions (variable CVDEXCL2, Fig. 1) for category C, and (iv) presence of angina I/angina II or angina, according to Rose questionnaire criteria or related criteria respectively, for category D [4, 10-11]. An individual is classified as 'needs further tests' when: (i) reported daily difficulties caused by diabetes, or hypertension/high blood pressure, or stroke problem, or weight problem for category A, (ii) recorded a low fitness level, or excluded per medications, or experienced a priority 2 stop for category C (Fig. 1), and (iii) did not fulfill the criteria for angina and angina I/angina II, or experienced pain or discomfort in chest and had occasionally experienced pain in one or more areas, and/or experienced severe pain in chest for more than half hour, and/or felt shortness of breath on stairs/inclines (Fig 1). All other subjects were assigned to the 'healthy', in terms of cardiac status, group.

Due to intercoupling of the 33 features, we selected the most descriptive (by using all 33 features), i.e., we adopted a simplification of the questions that bear similar connotation. We create a highly sparse space which requires a large number of participants in order to create conclusive results. Accordingly, the questions regarding physiological functioning PFQ063A, PFQ063B, PFQ063C, PFQ063D and PFQ063E were encoded into one variable designated as PFQ063AE. In addition, MCQ160B, MCQ160C, MCQ160D, and MCQ160E were examined as one variable and designated as MCQ160BE. Similarly, CVQ220A, CVQ220B, CVQ220C and CVQ220E were regarded as one variable designated as CVQ220. The above simplifications reduced the number of features from 33 to 22 variables reducing also the complexity of classifier training.

C. Study Limitations

As demonstrated in Fig. 2, a major challenge of the NHANES 2003-2004 dataset is the age variability of participants to all reported variable categories. Each category is addressed to a specific age group, differently positioned on the age range of 18 to 84 years. In order to jointly use those categories, the age group must be carefully restricted in order to remain with a sufficient number of individuals. Combining the categories of data, we resorted to the four cases presented in Fig. 2, with different age groups per case and varying numbers of participants for cardiac risk stratification. The first case involves variables related to all categories related to physical and medical condition, as well as information on cardiovascular fitness and health. The second and third cases consider the physical and medical status, along with one measure of the cardiovascular status. This selection increases the numbers of participants from 3 to 4 times. Further focusing only on the physical and health status in the fourth case, the number of participants increases and the age range covers the entire domain.

Category	SAS Variable Name and Description (Label)
A[a] Physical Functioning	PFQ063A-PFQ063B-PFQ063C- PFQ063D-PFQ063E: Health problems causing difficulty including: Heart problem, Hypertension/high blood pressure, Diabetes, Stroke problem, Weight problem
B[a] Medical Conditions	MCQ160B: Ever told had congestive heart failure
	MCQ160C: Ever told you had coronary heart disease
	MCQ160D: Ever told you had angina/angina pectoris
	MCQ160E: Ever told you had heart attack
C[b] Cardiovascular Fitness	CVDFITLV: Cardiovascular fitness level
	CVDEXCL2: Excluded per cardiovascular conditions
	CVDEXCL5: Excluded per (due to) medications
	CVQ220A: Priority 2 Stop, excessive HR in stage 2
	CVQ220B: Priority 2 Stop, excessive HR in stage 1
	CVQ220C: Priority 2 Stop, excessive BP
	CVQ220E: Priority 2 Stop, significant drop in SBP
	CVQ220G: Priority 2 Stop, variability in HR
D[a] Cardiovascular Health	CDQ001: SP ever had pain or discomfort in chest
	CDQ002: SP get it walking uphill or in a hurry
	CDQ003: During an ordinary pace on level ground
	CDQ004: If so does SP continue or slow down
	CDQ005: Does standing relieve pain/discomfort
	CDQ006: How soon is the pain relieved
	CDQ009A: Pain in right arm
	CDQ009B: Pain in right chest
	CDQ009C: Pain in neck
	CDQ009D: Pain in upper sternum
	CDQ009E: Pain in lower sternum
	CDQ009F: Pain in left chest
	CDQ009G: Pain in left arm
	CDQ009H: Pain in epigastric area
	CDQ008: Severe pain in chest more than half hour
	CDQ010: Shortness of breath on stairs/inclines

a. Questionnaire Data, b. Examination Data.

Figure 1. Major Categories under Study.

Case	Total number of subjects in 4 cases and distribution of the total population						
	Involved Categories	*Number of Variables*	*Age group*	*Total*	*Healthy*	*Non-Healthy*	*Needs further tests*
1st	A, B, C, D	22	40 - 49	759	439	138	182
2nd	A, B, C	6	18 - 45	2670	1860	211	599
3rd	A, B, D	18	40 - 84	3076	1970	642	464
4th	A, B	2	18 - 84	5363	3624	782	957

Figure 2. Detailed data for the population under study in four different cases.

III. Proposed Machine learning Model

A. Machine Learning Algorithm

The main goal, after completing the data pre-processing (i.e., category simplification and age group selection), was to find an efficient decision function able to separate the training data with the known class labels.

We used a Support Vector Machine (SVM) classifier as in [12-14], with a Radial Basis Function (RBF) kernel, variable γ depending on the dispersion of each attribute and $C = 1.0$. As shown in Fig. 2, the case datasets are not balanced for the three classes, with prevalence on the healthy condition. This unbalanced distribution of classes must be carefully addressed as the SVM model is expected to suffer from a certain bias in decision making.

B. Addressing Class Bias

In order to improve the generalization performance for the imbalanced biomedical data, we employed a validation approach for SVM based on data bootstrapping, as shown in Fig. 3. Initially, the data set is divided into training set (80%) and test set (20%) by stratified random data selection. The external training set formed by the 80% of original data is then divided into 'healthy', 'non-healthy' and 'needs further tests' sets. However, these sets do not contain the same size of samples. This problem is addressed by creating many balanced sets by stratified random selection from the original sets. The newly formed classes are then joined and mixed to form the bootstrap set. In every repetition, the data were shuffled in order to acquire randomness into the sets.

Then, the performance evaluation of SVM was achieved through 5-fold cross validation, after performing 100 repetitions of the 5-fold stratified data groups (grey solid border). The SVM is initialized before each fold's creation and trained on the four training folds. Then it is internally tested on the remaining fifth fold, but also externally tested on the left-out testing set formed by the 20% of original data. This is the first form of classifier evaluation demonstrated in Fig. 3, the so called CLF1, for validating the predictive ability of the categorical CVD health variables. The internal evaluation addresses the performance under a balanced training and control dataset.

In addition, the external evaluation is related to the evaluation under a balanced training set and an unbalanced testing set. The overall procedure is performed two times (grey dash-dotted border in Fig. 3) while the creation of new equal sets 10 times (grey dashed border in Fig. 3). Finally, we have also considered the case of unbalanced training–testing sets, in which 80% of the data is used for the training of SVM classifier while 20% for its evaluation (CLF2 evaluation). This process is repeated 100 times, with the classifier initialized on every iteration.

C. Feature Elimination

In order to further support the evaluation of the predictive power of features, we also considered the problem of feature selection through recursive feature elimination SVM-RFE procedure [13-14], applied on the CLF2 scheme. The RFE procedure returns a weight for each feature, which is associated with its effect on the SVM classification. In the present work, feature weights were calculated with an average of 1000 iterations using the linear kernel SVM-RFE.

IV. Results

The results obtained by the SVM-RFE algorithm are shown in Fig. 4. The reverse weight of each characteristic reflects its average importance in the SVM model. Characteristics marked with a value greater than 5 were considered the least significant and were eliminated (PFQ063AE, CVQFITLV, CDQ001, CDQ002, CDQ005, CDQ009A, CDQ009B) (Fig. 4, 5th-6th column). When interpreting the data, one must take into account the following aspects: (i) the possibility of inaccurately estimating the self-perceptions of difficulties (PFQ063AE) regarding their level of health [15], and (ii) the fact that the reference population does not include the subpopulation represented by individuals who were not eligible for the NHANES fitness test (CVQFITLV) [5]. The observed ranking of the self-reported history data of CVD health (CDQ005, CDQ001, CDQ009B, CDQ002, CDQ009A), it may be partially explained by the questionnaire structure and the sequence of questions (see selection framework). As a result, we created case 5 which consists of all attributes except those that were removed (Fig. 4).

In this section, we present for each classifier (CLF1 inner, CLF1 outer and CLF2) and in each case, the average performance metrics (Accuracy, Precision, Recalls and F1 score). The results for cases 1-5 are shown in Table I. In general, we observe that except case 4, all other combinations of health-related characteristics give quite satisfactory results, especially in identifying the truly cases. Even with reduced set of parameters, the case 5, where the characteristics have been selected by cross-validation on the common dataset (as in case 1) through the SVM-RFE methodology, provides satisfactory results.

Comparing the results for the various cases tested, we observe that lowest scores occurred from the case 4 which had only the physical and health condition measures. Additionally, the case 1 had slightly better results than the case 3, with the latter had the same characteristics except for cardiovascular fitness measures. Thereby, the cardiovascular fitness condition could increase the classification performance. Furthermore, the case 2 with the cardiovascular fitness supporting the medical and physical condition provided the best results. Hence, we can conclude that if we include all cardiovascular health characteristics, it deteriorates the efficiency. Finally, the case 5 engaging only the RFE selected characteristics by excluding some of the characteristics of cardiovascular fitness and health conditions and the physical measure, as mentioned above. Although the case 5 did not exceed the performance of the case 2, it attained higher scores than the case 1 that has the full feature set. In comparison with the other cases, the case 2 ensures that the mix-up between the certain cases is limited to minimum, i.e., from control (healthy) to disease/unsure (non-healthy/need further test) or *vice-versa*. However, the sample considered in either case 1 or 5 is small and should be further examined with additional data. Furthermore, class balancing in CLF selects only one partition of small (balanced) size for training. If we increase the size of partitions or if we test bootstrap partitions with appropriate data shuffling, we expect the performance of balanced CLF to be further improved.

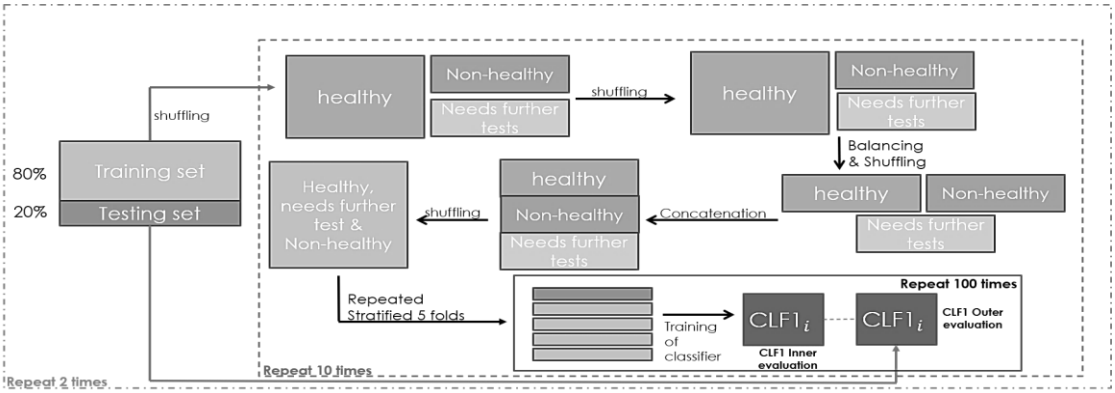

Figure 3. The implementation diagram of machine learning for balanced training−control set (CLF1 internal evaluation) and for balanced training set−unbalanced control set (CLF1 external evaluation).

V. CONCLUSION

This study focused on the potential of patient questionnaires with historic subjective and examination objective health data to identify possible risks for heart diseases. Besides screening, such information may also support the diagnostic power of physiological-biochemical exams clinically performed in CVD. The evaluation scheme considered SVMs with rigorous feature selection. After several tests, the categories related to medical condition and cardiovascular health and fitness show promising potential in assessing the CVD risk, with the category of fitness showing particular efficiency. Based on the results, we infer that the 6 variables used in case 2 have efficient performance, but the set could be enhanced with variables from the category of CVD health, as indicated by SVM-RFE. Our approach achieved comparable results, in respect of cardiovascular diagnosis, with the study [16] and slightly improved especially with the cases 2 and 5.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. No conflict of Interest reported.

SVM					
Characteristics	*Values*	*Characteristics*	*Values*	*Characteristics*[a]	*Values*
1. CDQ004	1	9. CDQ009D	2.018	16. CDQ005	5.702
2. CDQ006	1	10. CDQ009E	2.303	17. CDQ001	5.743
3. MCQ160B E	1.009	11. CDQ009F	2.478	18. CDQ009B	5.823
4. CVDEXCL5	1.154	12. CDQ003	3.134	19. CDQ002	7.496
5. CDQ008	1.225	13. CDQ009C	4.239	20. CDQ009A	8.438
6. CDQ009H	1.381	14. CDQ009G	4.439	21.PFQ063AE	9.714
7. CVQ220	1.528	15. CVDEXCL2	4.61	22. CVDFITLV	11.952
8. CDQ010	1.614				
a. These characteristics were removed.					

Figure 4. Results from SVM-RFE.

TABLE I. METRICS AND CLF1 INNER, OUTER AND CLF2

SVM Metrics (CLF1 Inner – Outer | CLF2)				
Cases	***Accuracy Score***	***Precision Score***	***Recalls Score***	***F1s***
1st	0.868 – 0.887 | 0.943	0.873 – 0.889 | 0.935	0.900 – 0.876 | 0.956	0.696 – 0.688 | 0.846
2nd	0.936 – 0.979 | 0.984	0.864 – 0.956 | 0.964	0.964 – 0.984 | 0.990	0.759 – 0.911 | 0.933
3rd	0.817 – 0.923 | 0.931	0.770 – 0.879 | 0.886	0.887 – 0.934 | 0.945	0.563 – 0.743 | 0.767
4th	0.635 – 0.787 | 0.841	0.617 – 0.711 | 0.778	0.761 – 0.772 | 0.897	0.320 – 0.406 | 0.581
5th	0.856 – 0.929 | 0.935	0.872 – 0.924 | 0.929	0.910 – 0.939 | 0.958	0.707 – 0.808 | 0.839

REFERENCES

[1] WHO, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] S. Singh, and R. Zeltser, "*Cardiac Risk Stratification,"* in: *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2020.

[3] A. Pandya, et al., "A comparative assessment of non-laboratory-based versus commonly used laboratory-based cardiovascular disease risk scores in the NHANES III population," *PLoS One*, vol. 6, no. 5, pp. e20416, May 2011.

[4] NHANES: https://www.cdc.gov/nchs/nhanes/index.htm

[5] C.Y. Wang, et al., "Cardiorespiratory fitness levels among US adults 20-49 years of age: findings from the 1999-2004 National Health and Nutrition Examination Survey," *Am J Epidemiol*., vol. 171, no. 4, pp. 426-435, Feb. 2010.

[6] P.L. Tsou, and C.J. Wu, "Sex-Dimorphic Association of Plasma Fatty Acids with Cardiovascular Fitness in Young and Middle-Aged General Adults: Subsamples from NHANES 2003-2004," *Nutrients*, vol. 10, no. 10, 1558, Oct. 2018.

[7] S.S. Yoon, et al., "Trends in the Prevalence of Coronary Heart Disease in the U.S.: National Health and Nutrition Examination Survey, 2001-2012," *Am. J. Prev. Med*., vol. 51, no. 4, pp. 437-445, Oct. 2016.

[8] R. Moonesinghe, et al., "Prevalence and Cardiovascular Health Impact of Family History of Premature Heart Disease in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007-2014," *J. Am. Heart Assoc*., vol. 8, no. 14, e012364, July 2019.

[9] K. Jindai, et al., "Multimorbidity and Functional Limitations Among Adults 65 or Older, NHANES 2005–2012," *Prev. Chronic Dis*., vol. 13, 160174, Nov. 2016.

[10] S.Heyden, et al., "Angina Pectoris and the Rose Questionnaire," *Arch. Intern. Med*., vol. 128, no. 6, pp. 961–964, 1971.

[11] A. Koyanagi, et al., "Correlates of physical activity among community-dwelling adults aged 50 or over in six low- and middle-income countries," *PLoS ONE*, vol. 12, no. 10, e0186992, Oct. 2017.

[12] W.-H. Weng, "Machine Learning for Clinical Predictive Analytics," in: *Leveraging Data Science for Global Health*. L. A. Celi et al. (eds.), 2020, ch. 12.

[13] I. Guyon, et al., "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, pp. 389–422, Jan. 2002.

[14] H. Sanz, et al., "SVM-RFE: selection and visualization of the most relevant features through non-linear kernels," *BMC Bioinformatics*, vol. 19, 432, Nov. 2018.

[15] S. Komanduri, et al., "Prevalence and risk factors of heart failure in the USA: NHANES 2013 - 2014 epidemiological follow-up study," *J. Community Hosp. Intern. Med. Perspect*., vol. 7, no. 1, pp. 15-20, Mar. 2017.

[16] A. Dinh, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 19, 211, 2019. https://doi.org/10.1186/s12911-019-0918-5