IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 27, 2022, accepted May 16, 2022, date of publication May 27, 2022, date of current version June 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3178419

A Non-Invasive Approach for Total Cholesterol Level Prediction Using Machine Learning

NAHUEL GARCÍA-D'URSO[1], PAU CLIMENT-PÉREZ[1], MIRIAM SÁNCHEZ-SANSEGUNDO[2], ANA ZARAGOZA-MARTÍ[3], ANDRÉS FUSTER-GUILLÓ[1], AND JORGE AZORÍN-LÓPEZ[1]

[1]Department of Computer Technology, University of Alicante, 03690 Alicante, Spain
[2]Department of Health Psychology, Health Sciences Faculty, University of Alicante, 03690 Alicante, Spain
[3]Department of Nursing, Health Sciences Faculty, University of Alicante, 03690 Alicante, Spain

Corresponding author: Andrés Fuster-Guilló (fuster@ua.es)

This work was supported by the Spanish State Research Agency (AEI) funded by MCIN/AEI/10.13039/501100011033 under Grant PID2020-119144RB-I00.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee at the University of Alicante, Application No. UA-2021-11-18.

ABSTRACT Artificial intelligence techniques have been increasingly applied in healthcare to help in many areas, from assisting clinical diagnoses to preventing diseases. In this paper, a machine learning approach to predict cholesterol levels using non-invasive and easy-to-collect data is presented. Specifically, it uses clinical and anthropometric data gathered by nutritionists during weight loss intervention (dieting) periods. The prediction power analysis of different patient variables is aimed at improving both non-invasive diagnosis quality and screening of associated diseases. Moreover, a clustering analysis has been carried out to identify different groupings of patients that might share some characteristics that have so far remained inconspicuous but might contain a valuable diagnosis or prognosis information for clinical experts. The experiments show a mean absolute percentage error rate (MAPE) of 4.39% in cholesterol estimation via regression, as well as clustering of patients within four profiles in which variable values share commonalities among cluster members.

INDEX TERMS Digital health assessment, 3D body reconstruction, clinical data regression, patient data clustering, pattern recognition.

I. INTRODUCTION

Finding correlations between anthropometric measurements (AMs) and laboratory findings is of great interest in the medical field [1], as it would lead to less invasive means of patient exploration. Examples of this can be found in the literature: from atherogenic markers [2], or diabetes assessment [3], or cardiovascular risk [4]. In some occasions, AMs can correlate to other AMs of patients bodies, which is useful for weight and height estimation from other measures, and helps in dose assessment for ICU patients [5].

Using a novel dataset including patient data from multi-modal sources (anthropometric measurements, as well as body sampling, etc.) the aim of this paper is to estimate cholesterol levels accurately from these non-invasive means, as these can be more cost-effective (no laboratories, experts, or reactives required), and can be used for screening purposes, i.e. avoiding more costly tests when not required. In this paper, the cholesterol prediction level is related to the prediction of the total cholesterol level, indicated as milligrams per decilitre (mg/dl). That is, the aim is to minimize the prediction error of the total blood cholesterol as much as possible, using the available gathered patient data. The dataset collected contains data from initially overweight patients (regardless of cholesterol levels), taken at several points in time during a dietary intervention. This regression will be assessed quantitatively (i.e. by measuring the prediction error).

Furthermore, this paper also aims at clustering patients into different profiles, according to shared common characteristics. Groups of patients will emerge from the data, in an unsupervised manner, therefore this will need to be evaluated qualitatively (e.g. of the methodologies involved in this paper).

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaojie Ju.

The remainder of this paper is structured as follows: next, related work in the literature will be introduced in the motivation section (Sec. II); then, the materials and methods section (Sec. III) introduces how cholesterol level prediction works, as well as which data were collected from patients. First, a description of the dataset is provided (Sec. III-B); then, the regression-based prediction is introduced (Sec. III-C), as well as the clustering employed (Sec. III-D), and the experimental setup (Sec. III-E). After this, results obtained so far are presented (Sec. IV). Finally, some conclusions are drawn (Sec. V), and future work on fully-automated estimation is further explained.

II. MOTIVATION

In the medical field of obesity, the use of indices such as body mass index (BMI), Body Adiposity Index (BAI), waist circumference (WC) and waist-to-height ratio (WHtR), have been recognized as simple and effective measures to diagnose and account a wide range of pathologies including cardiovascular diseases, hypertension, metabolic syndrome or dyslipidemia [6]. The potential of these indices in clinical practice is based on the simplicity and effectiveness of their use that allows to estimate the risk of obesity and mechanism involved in diverse pathologies associated with chronic inflammatory response caused by obesity such as high levels of blood pressure, visceral fat and central obesity and cholesterol [7], [8]. In addition, as compared to laboratory findings, these indices are not only less invasive, but in some occasions can also be cheaper. Since medical devices used to process blood samples, experts' time, facilities, etc. have a higher impact in healthcare provision systems, than proposed AM-based alternatives.

Often times, these types of devices have limited use for medical purposes, however, Jiang *et al.* [9] propose to use RGB-D devices to calculate BMI from 3D captured data, by estimating the height of each individual along with an approximate weight extracted from the volume. As a result, in their study, they can estimate the BMI within an error of 2.54 kg/m^2. Furthermore, Lu *et al.* [10] propose to use Kinect devices for 3D body reconstruction, in order to perform body composition analysis (fat, muscle, bone). Their reconstruction error is of 2.048 mm (RMSE) according to their experiments, with 82% accuracy for body composition analysis results.

Recent advancements in neural networks have enabled the loosening of some prior constraints, such as the possibility to use RGB devices, without depth sensors. An example of this is presented by Smith *et al.* [11], which introduce a method to recover 3D body data from 2D silhouettes from a pair of images. Another example is that of Trujillo-Jiménez *et al.* [12], which propose to perform precise anthropometry from handheld devices, using *body2vec*: a specially trained neural network that performs body segmentation (and background removal) prior to point cloud estimation and reconstruction from video. They match the estimated reconstructed models against two standards: a ''silver'' one consisting of LIDAR data, and a ''gold'' standard consisting of expert-provided AMs. However, they can obtain ''useful'' AMs from the videos, but are unable to accurately reconstruct the body in full.

Accurately retrieving the 3D body reconstruction of patients could open possibilities beyond the ''classical'' AMs that are taken today, as there could be other more inconspicuous AMs (or combinations and ratios of AMs) that correlate better with certain health parameters that could otherwise only be obtained via invasive, more expensive means. Furthermore, the deployment of cameras has another benefit, which is that, given the initial acquisition and installation of the hardware, further services (i.e. algorithms, that is ''software'') can be developed and deployed with minimum or no change, making it possible for future broadening of the explorations or analyses that can be performed to patients from a single 3D capture session. Furthermore, a single capture of an individual allows the experts to take more measures of additional parts of the body that were not initially considered, which might be necessary for the refinement of algorithms at a later stage.

It is in this context, that the Tech4Diet project[1] aims to obtain 4D models of patients undergoing weight loss (WL) programmes; that is, by capturing the 3D body reconstruction of intervention participants along several sessions in time (fourth dimension). Once this digital model of the patient is obtained, other metrics of health assessment can be derived. Ideally, even some laboratory (blood) sampling findings can be highly correlated to digitally-estimated AMs, and other body composition variables.

Advanced digital 3D body reconstruction based anthropometry which is not limited to waist, wrist and hip measures, but many more that can be obtained from a 3D scan of the body, could also show better correlations with body fat and muscle composition (e.g. arm, forearm, thigh, calf, ankle, etc.), which in turn could be related to total cholesterol level. Initial work in this regard has already been fruitful [13] (see Fig. 2), and accurate wrist, waist and hip measures have been acquired, and compared to those manually measured by an expert. However, the current method does not include a means to automatically determine the exact area where each measurement is to be taken [14], as depicted in Fig. 1 (i.e. where in the forearm is the wrist located, that is, which diameter around the 3D reconstructed forearm structure should be considered to be the wrist, exactly). This is currently under development as one of the aims of the Tech4Diet project. When completed, full body composition, as well as several anthropometric measurements will be automatically estimated, which will enable fully non-invasive, fully vision-based health assessment.

As part of these ongoing efforts, in this paper the focus is brought to the possible correlations that exist, and the regressions that can be made from manually annotated AMs, as well as other non-invasive variables taken from patients

[1] Project site: http://tech4d.dtic.ua.es/.

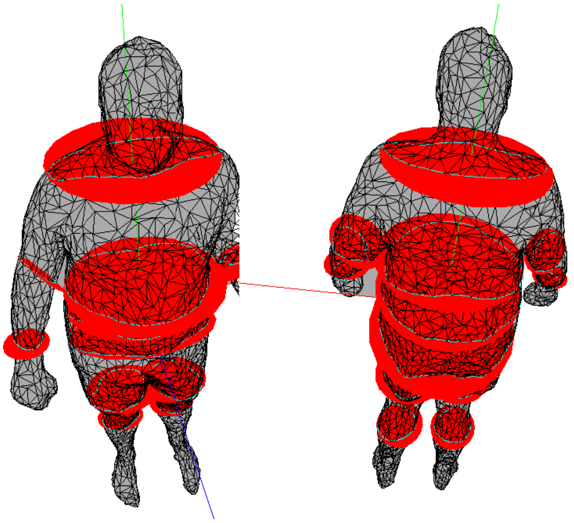

FIGURE 1. Example of digital anthropometry from 3D reconstructed bodies using computer vision with depth-sensing devices. Red discs represent slicing points for different body part measurements (waist, hip, neck, wrist, etc.).

using a full body analysis, weight scale, total cholesterol level and triglycerides. Cholesterol has been considered as the most important sterol synthesized by human cells [15]. High levels of cholesterol and triglycerides are usually associated with an increased intake of saturated fatty acids and BMI, which contributes to the risk of chronic and neurodegenerative diseases [16]. The mechanism involved in the increased risk of diseases is caused by deposition of fibrous tissues and fat in the arterial walls [17]. Traditional indices such BMI, WC and WHpR can be used to determine and estimate some indicators of obesity, but they do not account for differences between sexes [18]. New advances in research on human behaviour in obesity are being investigated to determine whether these indices can be used to diagnose different pathologies, and assess the risk of complications associated to obesity.

III. METHODOLOGY

The main objective of this research is to provide nutritionists with tools for better understanding of the relationship between patient variables and total cholesterol levels in patients.

One way this can be achieved is by helping predict the total blood cholesterol level using variables from patients in nutritional intervention. Another way, is being able to see the groups of patients that emerge from the data, to better determine commonalities of patients that have certain total cholesterol levels. That is, by clustering the patients into groups, some interesting common characteristics of a particular set of patients might be correlated to their total cholesterol outcomes, and these insights could be valuable to experts. This section introduces the materials and methods used for cholesterol prediction via regression, and cholesterol profile clustering.

A. DATA COLLECTION

In the first place, a data collection was carried to obtain nutritional and anthropometric information from 84 patients for 6 months. During this period, each patient was intervened on 4 times (4.38 median value). All study participants were informed (including their right to withdraw at any point, for any reason) and gave consent to take part. Furthermore, the study was conducted according to all ethics regulations regarding studies with human subjects from the University of Alicante. A total of 26 variables were taken on each session with a patient. These are summarized in Table 1, and are divided into five groups:

- Anthropometric measures,
- body composition analysis,
- lifestyle metrics,
- capillary blood sampling, and
- blood pressure.

More precisely, anthropometric measures, i.e. wrist, waist and hip measures were taken manually using a flexible measuring tape (0.1 cm precision, two measurements, then mean was used).

Body composition analysis was performed by a Tanita® MC 780-P MA smart scales (Tanita Corp., Arlington Heights, IL, USA), including weight (0.1 kg precision), fat and muscle

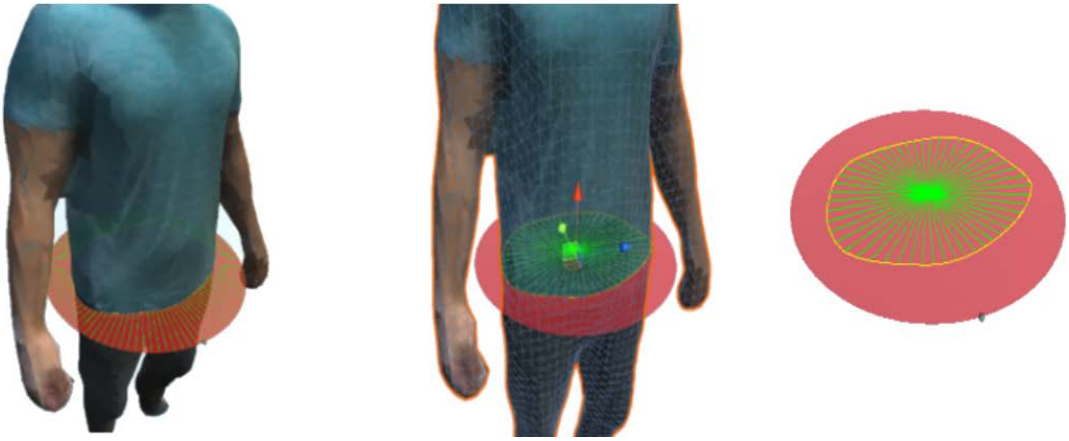

FIGURE 2. Visual example of the method used for digital anthropometry of the waist: a number of rays from the centre of the plane intersecting the body are used to estimate the contour size at the slicing point.

TABLE 1. Variables taken for each individual session of a participant (data point) in the dataset. A total of 26 variables are considered.

Type	Source	Measurement (unit)	N.
Anthropometric	flexible measuring tape	– Wrist (cm)	1
		– Waist (cm)	1
		– Hip (cm)	1
Body composition	Tanita® MC 780-P MA and Seca® 213 stadiometer	– Fat per limb/trunk (%)	5
		– Muscle per limb/trunk (%)	5
		– Fat & muscle, full body (%)	2
		– Visceral fat area (cm^2)	1
		– Weight (kg)	1
		– Height (m)	1
Other, Lifestyle	Interview	– Activity (score)	1
		– Gender	1
		– Age (y.o.)	1
Blood (capillary)	Accutrend® Plus	– Glucose (mg/dl)	1
		– Triglycerides (mg/dl)	1
		– Cholesterol (mg/dl)	1
Blood pressure	Omron® M3	– Systolic (mmHg)	1
		– Diastolic (mmHg)	1

percentages for the full body, as well as separately for each limb and the trunk, and percentage of visceral fat. Height (0.1 cm precision) was measured using a Seca® portable stadiometer 213 (Seca, Hamburg, Germany).

Lifestyle metrics, i.e. a physical activity score which was determined by using the International Physical Activity Questionnaire Short Version (IPAQ-SF). The IPAQ-SF [19] comprises 7 items assessing the frequency and duration of physical activity across three ranges of intensity—vigorous physical activity (VPA = 8.0 metabolic equivalents –or METs), moderate physical activity (MPA = 4.0 METs), and low physical activity (LPA = 3.3 METs)—undertaken across a set of domains including leisure time, domestic and gardening (yard) activities, and work related and transport-related activities during a period of one week. The final scores assigned in our study can take three possible values (1.2, 1.4 and 1.6, and are used as a multiplier for total energy expenditure –see below) and indicates the value of physical activity performed by the patient on a daily basis: value 1.2 refers to sedentary people, wheelchair users, etc. Value 1.4 refers to people who do little physical activity. Value 1.6 for people who practice sports daily on a semi-professional or professional basis.

For the blood sampling (capillary) measures, an Accutrend® Plus device (by Roche Diagnostics GmbH, Mannheim, Germany) was used to obtain glucose, triglycerides and cholesterol levels (from different fingers, using the Accuchek® Softclix® Pro lancing device (Roche Diagnostics GmbH, Mannheim, Germany).

Finally, blood pressure was obtained using an Omron® M3 blood pressure monitor (Omron Healthcare Europe, Hoofddorp, Netherlands) to get systolic and diastolic pressure values.

Regarding inclusion/exclusion criteria, the participants included 87 male and female Spanish volunteers with overweight and obesity recruited by advertisements on the website of the Tech4Diet project. The participants ranged in age from 22 to 63 years ($\bar{x} = 47.14$ years; $\sigma = 9.22$ years). The inclusion criteria were (i) having a BMI greater than 25 kg/m^2, (ii) being right-handed, (iii) being able to read and write fluently, and (iv) having Spanish as the mother tongue. The exclusion criteria were (i) currently being or having in the past year been on a dietary/nutritional treatment supervised by a nutritionist; (ii) the presence of endocrine-metabolic disorders including problems of the thyroid, pituitary gland, or adrenal gland and metabolic syndrome; (iii) having a prior history of neurological illness (e.g., stroke or Parkinson's disease); (iv) having a history of head injury (causing a loss of consciousness for more than 30 min); (v) having a history of severe psychopathology according to the DSM-IV-TR diagnostic criteria; and (vi) currently receiving psychiatric treatment. Initial participants were recruited from September to November 2020. From the 101 individuals approached, 14 (16.09%) were excluded due to meeting exclusion criteria: five (5.75%) had followed dietary treatments over the past year, six (6.9%) had histories of endocrine-metabolic disorders, two (2.3%) were taking psychopharmacological medications due to mental health disturbances, and one (1.15%) had a history of head injury. The final sample included 84 male and female participants with overweight and obesity, disregarding those for which not enough sessions had been captured.

The body mass index was calculated as weight over height squared (kg/m^2). BMI was interpreted according to the World Health Organization (WHO) classification. The BMI cut-off point for overweight was defined as 24 kg/m^2, while obesity was defined as a BMI of 30 kg/m^2.

B. DATASET PREPARATION

After the initial collection, data preprocessing was carried out in order to remove outliers, by applying z-score.

Following this, an imputation of data was carried out. The reason behind this is that some variables had not been collected for some participants in a given session, and were missing; or in cases in which a variable had an implausible value. To perform the imputation, missing values were filled in using data associated to another session from the same participant. For this, a multivariate iterative imputer was used, with k-nearest neighbours (with $k = 10$ neighbours) so that the data is completed using the most similar data points (likely from the same participant) around the data point with the missing value.

At the end of this process, the resulting dataset contains a total of 359 data points (feature vectors), with 26 variables (dimensions) each.

An extended version of the dataset can also include *derived* measures such as waist-to-hip ratio (WHR), for which all necessary measures are available:

$$\mathrm{WHR} = \frac{\mathrm{Waist}}{\mathrm{Hip}}. \tag{1}$$

The basal metabolic rate (BMR), or basal energy expenditure (BEE) that requires weight (w), height (h), gender,

TABLE 2. Basal metabolic rate (BMR) equations depending on different patient variables, i.e. gender, age (*a*), height (*h*) and weight (*w*). Harris–Benedict method.

Gender	Age (a)	
	$a \leq 65$	$a > 65$
Men	$66.773 + 13.751w + 5h - 6.775a$	$11.7w + 588$
Women	$655.095 + 9.563w + 1.849h - 4.675a$	$9.1w + 659$

TABLE 3. Regressors used during experimentation, along with most relevant parameters. Selected regressors are shown in bold at the top.

Acronym	Regressor name
GBR	Gradient Boosting regressor
RFR	Random Forest regressor
ETR	Extra Trees regressor
XGB	Extreme Gradient Boosting regressor
ABR	AdaBoost regressor
LSVR	Linear Support Vector regressor ($it_{\text{max}} = 10^5$)
DTC	Decision Tree regressor
Huber	Huber regressor ($\epsilon = 1.65$)
EN	ElasticNet ($\alpha = 0.009$; $\ell_1$ratio $= 0.9$; $it_{\text{max}} = 10^3$)
KNN	k-Nearest Neighbours regressor ($n = 2$, $p = 1$)

TABLE 4. Clustering methods tested, along with most relevant parameters used.

Method	Parameters
Agglomerative clustering	clusters $= 4$
HDBSCAN	$\text{min}_{\text{size}} = 25$, $\text{min}_{\text{samples}} = 10$
k-Means	clusters $= 4$ (tried 1–11) init $=$ 'kmeans++'
Spectral clustering	clusters $= 4$

and age (a) and depends on these variables for calculation is shown in Table 2. From the BMR, the total energy expenditure (TEE) can be calculated, using the activity score as a multiplier:

$$\text{TEE} = \text{BMR} \cdot \text{Activity} \tag{2}$$

When including these, each data point contains a total of 29 variables.

C. REGRESSION

After data preprocessing is done, the data is fed to 10 different regressors for training, these are show in Table 3, along with their most relevant parameters. In all cases, for reproducibility, the same random seed is used for all regressors. From this point onwards, regressors will be named using their acronyms, as per the table. The selection of the top four regressors on the table has been decided based on lowest error rates for prediction. Each regressor is provided with 324 samples for training and 35 more for testing.

D. CLUSTERING

As explained, another way in which cholesterol level prediction can be approached is by looking at cholesterol (and/or triglycerides) levels, and checking what other variables are relevant to the observed outcome, or in which way other variables correlate (or not) with the observed levels.

By looking at the patterns that emerge from grouping similar data together, it could be possible to gain valuable insight that can better assist nutritionist or other experts in estimating cholesterol levels by using indirect (non-invasive) means of analysis.

For this purpose, clustering techniques, in the unsupervised machine learning family of methods, is used. Clustering consists on finding a number of groups or clusters in which data can be grouped based on similarity, and split based on dissimilarity. This is also sometimes referred to as distance.

Several clustering methods have been investigated, which are summarized in Table 4. For k-Means, the initialization step is performed using 'kmeans++', this selects initial cluster centres for k-mean clustering in a smart way to speed up convergence. In HDBSCAN, min_{size} refers to the smallest size grouping that you wish to consider a cluster, and $\text{min}_{\text{samples}}$ refers to the number of samples in a neighbourhood for a point to be considered a core point.

E. EXPERIMENTAL SETUP

1) REGRESSION EXPERIMENTS

Two sets of experiments were conducted to validate the presented approach. First, from all the variables available per participant, subsets were taken, including or excluding entire groups of variables depending on their type (anthropometric, body composition, etc). The aim of this experiment was to determine which group or category of variables is most useful to determine total cholesterol levels (which could be interesting per se). On a second experiment, automated feature selection is applied, to determine the best subset of variables that yields the lowest error. In this case, the focus is on finding variables that are the most discriminant for the task.

Regarding the first set of experiments, Table 5 introduces the different variable sets used, namely: when using anthropometric variables only (A); or body composition only (BC); or a combination of anthropometric, and body composition (ABC); or both, including blood pressure (ABCP); or using only anthropometric and blood pressure (AP); or body composition and blood pressure only (BCP); or when all variables were used (All), except for total cholesterol for obvious reasons. That is, in all experiments the X set of variables excludes the total cholesterol, which is the y value to calculate from the regression function learned $f(X, \Phi) = y$ by means of adjusting the set of parameters Φ of the regressor.

The anthropometric variables set (A) includes, apart from the direct measures, the waist-to-hip ratio (WHR) from Eqn. 1. Furthermore, the body composition set (BC) includes the BMR (Table 2), the TEE (Eqn. 2), as well as the gender and age of the participant. The activity score is used to obtain the TEE, but is not fed directly to any model.

TABLE 5. Summary of different groups of variables taken into consideration for each experiment.

Variable set	Anthropo-metric	Body composition	Blood pressure	Blood sample
A	✓			
BC		✓		
ABC	✓	✓		
ABCP	✓	✓	✓	
AP	✓		✓	
BCP		✓	✓	
All	✓	✓	✓	✓

Regarding the second part of the experiments, with automated feature selection, three different feature selection schemes are applied. These schemes are based on the most relevant (discriminative) features from the random forest regressor using either recursive feature elimination (RFE), or the drop-column importance (DCI) of different variables, or the permutation feature importance (PFI).

The RFE method works by recursively removing attributes from the dataset and building a model (Random Forest Regressor) on the attributes that remain. On each iteration, the least important feature, i.e. the feature with the lowest weight assigned by the estimator, is removed. In our case, the weight assigned is computed as the Gini importance. Instead, PFI measures the importance of a feature by calculating the decrease in the model score after a single feature has been randomly shuffled (i.e. the values have been randomly re-assigned to other individuals). This procedure breaks the relationship between the feature and the target, therefore the drop in the model score is indicative of how much the model depends on the feature. Finally, the DCI method differs from PFI in that each feature is removed in each iteration instead of randomly shuffled among individuals.

In both cases, that is, in both sets of experiments, parameter values to the regressors were left unchanged (default values), or were optimized via Grid Search cross-validation and distributed asynchronous hyperparameter optimization, i.e. HyperOpt [20].

2) CLUSTERING EXPERIMENTS

Regarding clustering experiments, three different tests have been conducted to test which clustering method is best, as well as to check for data separability by performing a principal component analysis (PCA) beforehand.

One of the experiments entails applying each of the techniques in Table 4 to find possible clusters, and see how well each performs in terms of inter-cluster distances (cluster separation). For this, the two main principal components (from PCA) are used in 2D scatter plots. Furthermore, a t-test is provided to conclude whether the clusters are quantitatively different, i.e. with regards to the 10 most significant variables found during the regression experiments just introduced above.

TABLE 6. Mean absolute percentage error rates (MAPE) when using anthropometric variables only (A), body composition only (BC), both (ABC), both plus blood pressure (ABCP), anthropometric and pressure (AP), body composition and pressure (BCP), or all variables (All). Best value (lowest) is marked in bold, second best is underlined.

Variable set	Parameter values	Regressor error			
		RFR (%)	ETR (%)	GBR (%)	XGB (%)
A	Default	7.44	8.31	6.40	9.02
A	Optimized	4.93	5.03	4.99	6.88
BC	Default	6.14	6.53	6.05	6.97
BC	Optimized	**4.58**	4.99	5.30	5.70
ABC	Default	6.96	6.17	6.13	7.98
ABC	Optimized	5.22	5.08	5.29	6.34
ABCP	Default	6.22	6.68	6.45	6.68
ABCP	Optimized	5.32	4.87	5.39	6.08
AP	Default	6.06	6.59	6.49	7.41
AP	Optimized	4.65	4.88	5.15	6.70
BCP	Default	5.09	5.51	6.05	6.89
BCP	Optimized	4.60	4.90	5.04	4.99
All	Default	5.59	5.79	5.82	6.66
All	Optimized	4.69	4.98	4.72	5.44

Then, a series of variables are plotted against each other, to determine whether different cholesterol level grouping is present, showing a correlation with other metrics in the data.

IV. RESULTS

Following the same scheme presented in the experimental setup, this section is divided into regression and clustering, and each subsection showing the results for each of the experiments or tests performed.

A. *REGRESSION RESULTS*

Table 6 shows the mean absolute percentage error (MAPE) scores for the first part of the regression experiments described above, i.e. those using manually picked subsets of variables, as shown in Table 5. From the results shown, it can be observed that body composition (BC) data alone is highly correlated with cholesterol outcomes, this is likely to be derived from the fact that fat mass percentages for each limb, the trunk, or the whole body are relatively important to assess dyslipidemia. For this, the Random Forest regressor (RFR) is the best-performing, specially when parameters are optimized. This is even better than using all variables, which include triglycerides, which are correlated to cholesterol outcomes too. These results are highly relevant, as the BC set of variables does not contain any blood sampling, that is, a cholesterol level can be regressed from fully non-invasive means of participant exploration. In all cases, hyperparameter optimization seems to lead to better results, as should be expected.

Regarding the second set of regression experiments, Table 7 shows the results for the different feature selection schemes used, with the last row showing the results when

TABLE 7. Mean absolute percentage error rates (MAPE) for regression with different feature selection schemes, with the top 10, 15 or 20 most relevant features. Using relevant features from either recursive feature elimination (RFE), or permutation feature importance (PFI), or drop-column importance (DCI) methods. Best value (lowest error) appears in bold, second best is underlined.

Features	Regressor error (test)		
	RFR (%)	**ETR** (%)	**GBR** (%)
All features (baseline)	5.36	5.51	5.64
Top 10 RFE	5.65	<u>5.24</u>	5.43
Top 15 RFE	5.66	5.36	5.67
Top 20 RFE	5.53	5.55	5.32
Top 10 PFI	5.32	5.46	5.95
Top 15 PFI	5.80	5.42	5.39
Top 20 PFI	5.83	5.95	5.50
Top 10 DCI	6.50	6.86	6.25
Top 15 DCI	5.74	5.62	5.57
Top 20 DCI	6.10	6.33	6.11
Top 10 RFE *(optimized)*		**4.39**	

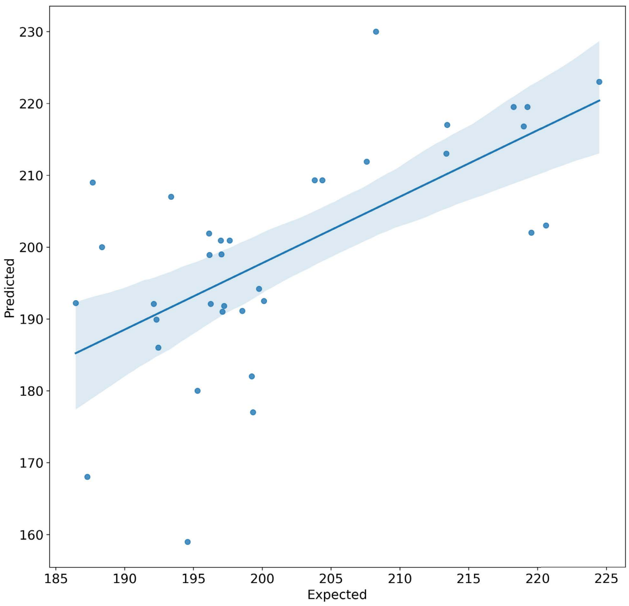

FIGURE 3. Total cholesterol prediction scatter (expected–predicted result).

parameters are optimized, and Table 8 shows the ranking for the top 20 features, according to these three schemes: three main columns show the feature selection schemes, and rows show the ranking (best to worse), or the contribution of each feature according to each scheme. The sets are divided into the top 10, 15, and 20 best variables to use. Differences in score values among different schemes are not comparable, since each works differently with regard to the calculation of the relevance of each variable.

Looking at the regression results in Table 7, it can be observed that the best result (lowest MAPE) without optimization (underlined) is obtained when using the top 10 features from random forest regressor (RFR) with the recursive feature elimination (RFE) using the extra trees regressor (ETR) to estimate the cholesterol levels. The bottom-most row presents the result when hyperparameters are optimized for the underlined result. As can be seen, in this case the error is reduced by 0.85% from 5.24% to 4.39%. This is an 16.2% reduction in error with respect to the non-optimized, and the overall best result so far. An expected–predicted scatter plot is shown in Fig. 3 for this best case.

From Table 8, some insights can be drawn from the set of selected features: BMR seems to be a good predictor

TABLE 8. Ranking of the 10, 15, and 20 most relevant features according to different feature selection schemes: recursive feature elimination (RFE), drop-column importance (DCI), and permutation feature importance (PFI).

Rank	recursive elimination (RFE)		drop-column importance (DCI)		permutation importance (PFI)	
	feature name	score	feature name	score	feature name	score
1	BMR	1.67	systolic_BP	0.0047	BMR	0.2924
2	glucose	1.54	BMR	0.0045	age	0.1480
3	fat_L_leg	1.41	WHR	0.0019	triglycerides	0.1289
4	triglycerides	1.28	weight	0.0017	glucose	0.1002
5	hip	1.16	global_fat	0.0007	WHR	0.0965
6	systolic_BP	1.03	height	0.0006	systolic_BP	0.0817
7	age	0.90	visceral_fat	0.0004	fat_trunk	0.0699
8	fat_trunk	0.77	hip	0.0001	TEE	0.0569
9	TEE	0.64	triglycerides	-0.0043	fat_L_leg	0.0492
10	WHR	0.51	diastolic_BP	-0.0038	weight	0.0464
11	global_musc	0.39	muscle_R_arm	-0.0033	fat_R_leg	0.0455
12	wrist	0.26	fat_trunk	-0.0031	hip	0.0453
13	weight	0.13	muscle_trunk	-0.0031	wrist	0.0437
14	gender	-1.67	fat_L_arm	-0.0026	global_fat	0.0428
15	muscle_L_arm	-1.54	waist	-0.0025	visceral_fat	0.0412
16	muscle_R_leg	-1.41	gender	-0.0024	global_musc	0.0398
17	muscle_R_arm	-1.28	age	-0.0022	diastolic_BP	0.0370
18	fat_L_arm	-1.16	muscle_R_leg	-0.0022	waist	0.0275
19	muscle_L_leg	-1.03	muscle_L_arm	-0.0019	height	0.0221
20	height	-0.90	global_musc	-0.0012	fat_L_arm	0.0217

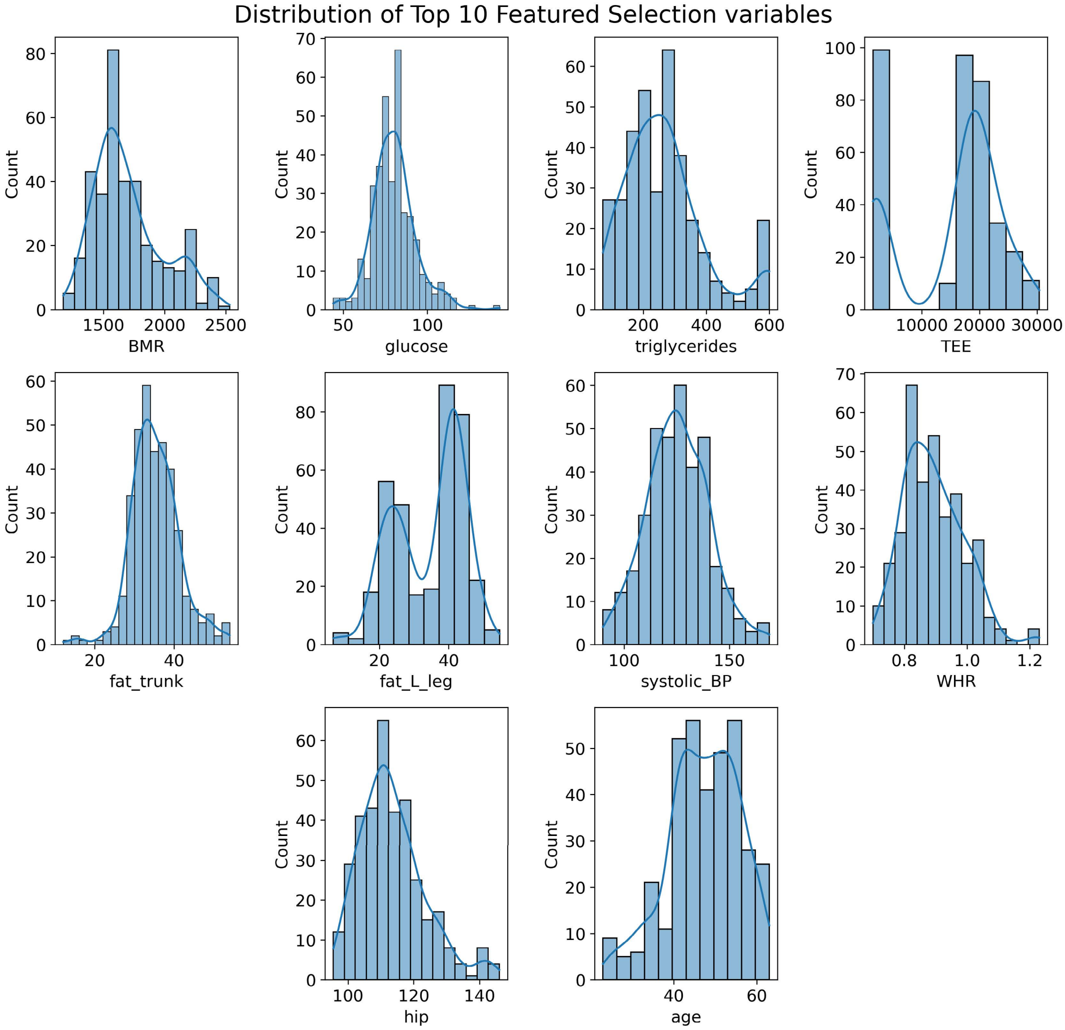

FIGURE 4. Histogram plots for the top 10 selected features (RFE Top 10 patient variables, from Table 8) along with a curve fitted to each distribution.

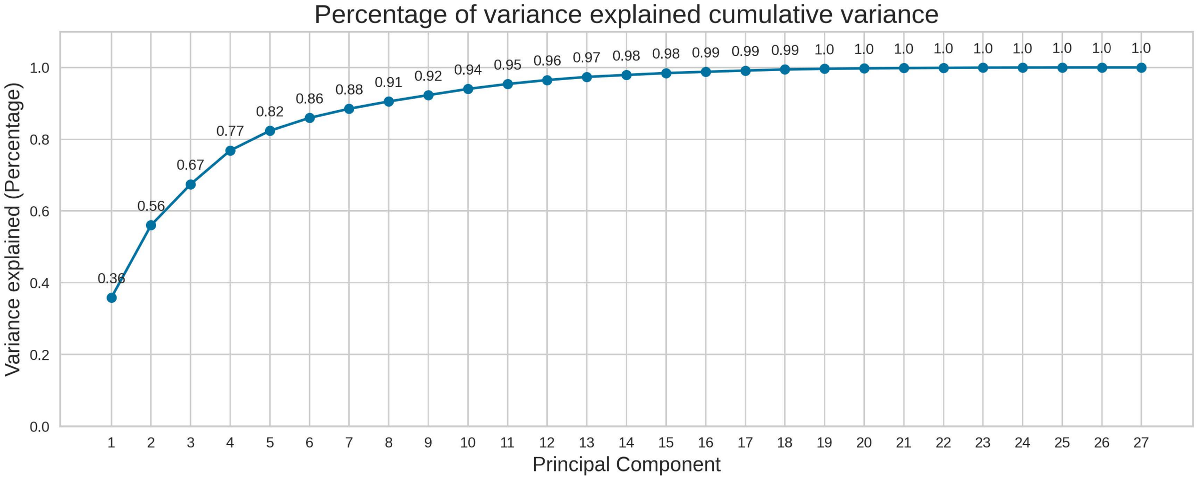

FIGURE 5. Variance explained by each new principal component identified by the analysis, shown as cumulative variance (in percentage).

for cholesterol values, also WHR tends to be in the top 10 features. TEE appears twice in the top 10, but not for the DCI method. Weight appears in the top 10, except for RFE. Systolic blood pressure (BP) appears in the top 10 in all cases, as do some fat measures (global, trunk, left leg), which makes sense, given the relationship between body fat make-up and

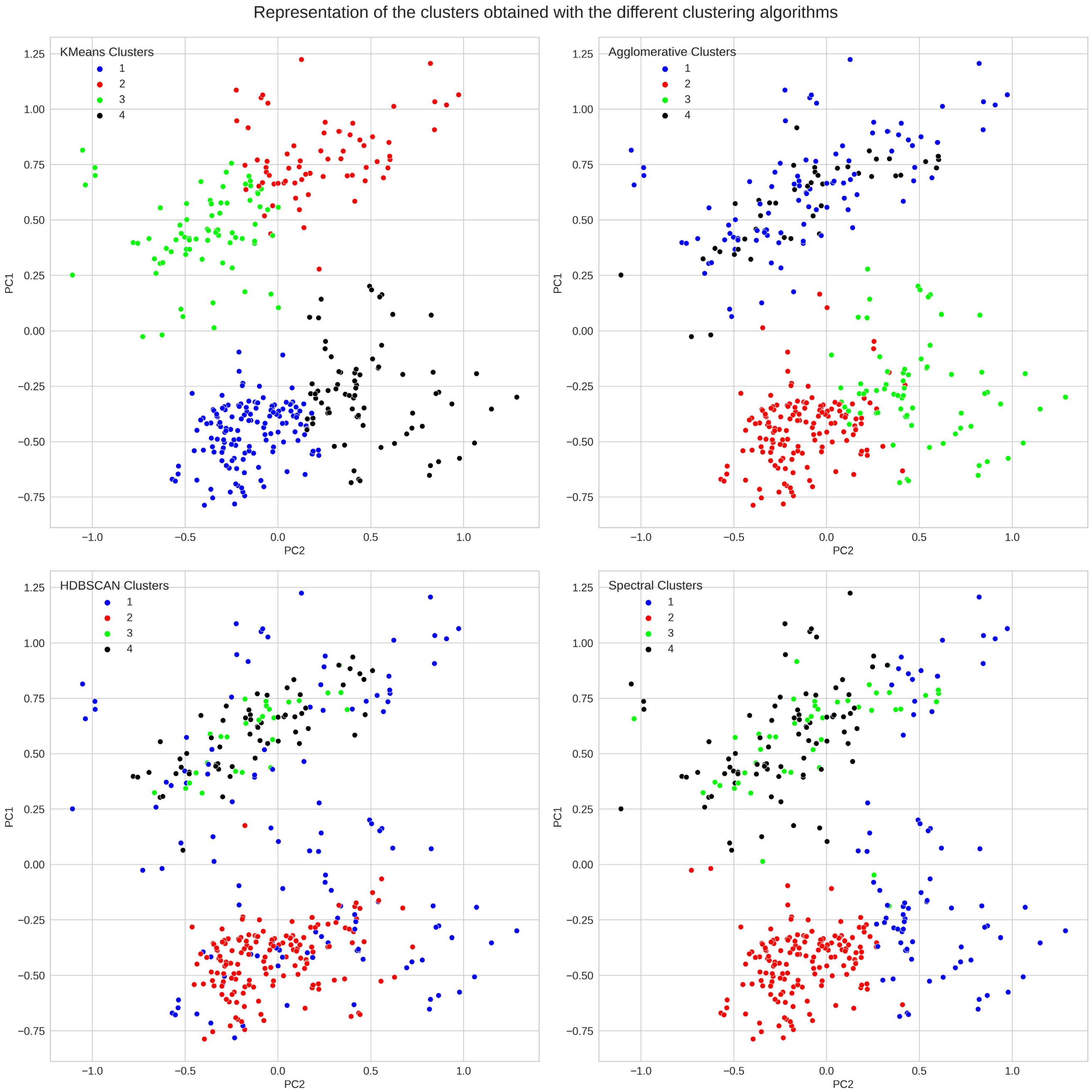

FIGURE 6. Clustering of patient data using the two first principal components with different clustering methods.

blood lipid levels, as well as the metabolic syndrome (MS) in which high blood pressure, high cholesterol and high blood glucose are present. As expected, blood sample related features (glucose, triglycerides) are related to cholesterol, and therefore make it to the top 10 features as well, in most cases.

To end the regression part, Fig. 4 shows histogram plots for the top 10 most discriminative features of the random forest regressor (RFR) when using recursive features elimination (RFE), which is the set of features resulting in the best result for feature selection, and the best overall presented in this paper.

B. CLUSTERING RESULTS

With regard to clustering, PCA is first applied to obtain the principal components. Figure 5 shows the cumulative variance explained by each new principal component extracted from the data.

Next, different clustering methods are tested, to determine how well they perform (in terms of cluster separability, inter-cluster distances). Figure 6 shows the results of clustering with each tested algorithm using the results from the two principal components. At the top-left is k-Means clustering, showing four well separate clusters; top-right is agglomerative clustering, in which clusters 1 and 4 fall in

TABLE 9. **Results for the t-test on pairs of *k*-Means clusters.**

(a) Clusters 1-2

measure	t-statistic	p-value
BMR	-39.7194	0.0000
fat_R_leg	36.4553	0.0000
fat_L_leg	36.4241	0.0000
musc_trunk	-33.7583	0.0000
weight	-31.8879	0.0000
glucose	-2.4438	0.0154
diastolic_BP	-2.2594	0.0249
triglycerides	1.9426	0.0534
cholesterol	1.2677	0.2063
age	-1.0224	0.3078

(b) Clusters 1-3

measure	t-statistic	p-value
fat_L_leg	38.8551	0.0000
fat_R_leg	37.0211	0.0000
fat_L_arm	34.4117	0.0000
fat_R_arm	33.6072	0.0000
musc_R_arm	-22.6513	0.0000
TEE	-2.8130	0.0054
glucose	-2.7433	0.0066
diastolic_BP	-1.8610	0.0641
triglycerides	0.6676	0.5051
age	0.3660	0.7147

(c) Clusters 1-4

measure	t-statistic	p-value
weight	-18.2533	0.0000
waist	-16.6569	0.0000
hip	-14.2395	0.0000
global_fat	-14.1982	0.0000
visceral_fat	-13.3829	0.0000
TEE	-2.8367	0.0050
glucose	-2.2528	0.0253
triglycerides	-1.8874	0.0605
height	-1.2385	0.2169
cholesterol	0.6963	0.4870

(d) Clusters 2-3

measure	t-statistic	p-value
weight	18.6431	0.0000
BMR	17.6106	0.0000
hip	13.5344	0.0000
waist	11.2463	0.0000
visceral_fat	9.4054	0.0000
age	1.2676	0.2070
triglycerides	-1.2187	0.2250
diastolic_BP	-1.0116	0.3134
TEE	0.4422	0.6590
glucose	-0.0658	0.9477

(e) Clusters 2-4

measure	t-statistic	p-value
BMR	24.8332	0.0000
fat_L_leg	-24.1473	0.0000
fat_R_leg	-22.0764	0.0000
musc_R_leg	19.1647	0.0000
height	19.0005	0.0000
age	-1.5945	0.1131
systolic_BP	0.7594	0.4489
TEE	0.5903	0.5559
cholesterol	-0.5639	0.5737
glucose	0.3220	0.7479

(f) Clusters 3-4

measure	t-statistic	p-value
fat_L_leg	-27.1934	0.0000
fat_R_leg	-24.7240	0.0000
global_fat	-23.2887	0.0000
global_musc	21.9274	0.0000
fat_L_arm	-20.7438	0.0000
visceral_fat	-1.6838	0.0943
diastolic_bp	0.7119	0.4777
systolic_bp	-0.4753	0.6352
glucose	0.4345	0.6646
TEE	0.1466	0.8837

the same area; bottom-left is HDBSCAN, which shows much mixture between labelled elements; finally spectral clustering shows four clusters, some have clear boundaries such as 1 and 2, but 3 and 4 show overlap between them, and to a minor extent with cluster 1.

As can be observed, the best result is obtained when using *k*-Means, as the boundaries of each cluster are more clear, and there seems to be very little or no overlap, which makes it possible to split the data into four different groups of patients. A deeper analysis of cluster composition is performed next. To do so, each of the sets of variables shown in Table 5 (A, BC, P) is taken to find differences in variable values that are relevant for each cluster.

The first cluster (cluster 1) is conformed by 141 data points, being the largest cluster. With respect to the variables of the A set, these individuals have the lowest waist-to-hip ratio (WHR). Regarding the BC set of variables, this group shows low values of visceral fat, being the cluster with the least amount of this type of fat, even below the global average (-3.38). However, these individuals are not the ones with the least amount of fat per body part (limbs/torso) or as a whole, but they are the ones with the least muscle mass. They are also the shortest (least height). All this makes this group also the group with the lowest mean weight of all clusters, and below the global average (-13.17). It also has the lowest BMR and TEE. Furthermore, regarding blood sample variables, this group has the lowest total cholesterol, and the same trend is observed with blood pressure.

In cluster 2, we find the smallest set of data points, with a total of 68. Albeit having a hip circumference similar to the average, the value of waist circumference is elevated, the highest of all clusters. This makes this group also have the highest WHR. If observing BC variables, we find that individuals do not have a high fat percentage, but they do have the highest amount of muscle mass. This, as opposed to cluster 1, makes them the heaviest. They also have a high amount of visceral fat, BMR and TEE. Observing blood sampling values, this cluster stands out by the fact that its triglyceride scores are very low. Yet, blood pressure is the highest.

In cluster 3, looking at the A set of variables, it can be highlighted that the average hip circumference is much lower than the global mean, and therefore, the lowest of all clusters too. Observing BC variables, this cluster is the one with the lowest global and limb/torso fat values. The values of cholesterol are the lowest, even below the global mean (-6.78), and the glucose values are the highest. In this cluster, the diastolic blood pressure values are the highest, and are 8.53 points above the mean.

From cluster 4, it can be observed that the mean hip circumference surpasses all other clusters, and is 20 centimetres higher than the next cluster in this regard. Individuals in this cluster have the largest amount of fat of all clusters. And, although it has muscle mass scores that are average, the global muscle mass is the lowest of all clusters, i.e. -8.99 points below the global mean. This cluster has glucose and cholesterol levels that are close to the mean, but triglycerides are the highest.

Finally, to obtain a quantitative measurement of cluster disjointness, a t-test has been applied on pairs of the clusters obtained by *k*-Means. The test was performed on the 10 most discriminative variables used for regression before. The results from this test are shown on Table 9. The top 5 rows of each table show the best, and the bottom five rows show the worst p-values for each cluster pair. We conclude that, at the 5% significance level, the two clusters are significantly different from each other in terms of all 10 variables.

V. CONCLUSION

In this paper, we have presented a machine learning approach for total cholesterol estimation using regression from non-invasive patient variables consisting of anthropometric, body composition, blood pressure, lifestyle, and, optionally, capillary blood sampling. This information is useful for initial patient screening, resulting in reduced costs of operation for healthcare providers, when patients do not require further, more expensive tests. Additionally, clustering has been used to characterize the different groups or clusters of individuals that emerge from the data, which can provide valuable insights to clinical experts.

In the first set of regression experiments, variables have been manually divided into groups, according to their nature. Several subsets of feature groups have been taken and results calculated. It has been shown that, by using solely body composition features, it has been possible to achieve an overall error of 4.58% in total cholesterol (TC) estimation, when applying hyperparameter optimization.

In the second set of regression experiments, automated feature selection has been performed. When using the top 10 best (most discriminative) features, the error can be further reduced to 4.39%, after applying hyperparameter optimization. However, it needs to be taken into account, that, due to the performed feature selection, the feature set contains some blood sampling features (glucose and triglycerides). Therefore, this is not fully non-invasive. Nonetheless, capillary blood sampling is less invasive than venous blood extraction, and uses fewer resources, as no laboratory equipment, experts, and expensive reagents are necessary.

Finally, four clustering methods have been compared, and using principal component analysis (PCA) it has been possible to split the data into four patient profiles, each showing characteristic ranges of values in their anthropometric (A), body composition (BC), blood pressure (P), or blood sampling variables.

As future research lines, our next steps in the short and medium term are aimed at reducing the error by including: a) digital anthropometry metrics, b) more participants in the study, and c) further variables, especially those that are part of an automated body composition analysis, given the good results of the first set of experiments conducted in the present study. More accurate cholesterol levels for the ground truth can also be obtained by employing a venous blood sample analysis, rather than a capillary blood sample. This could improve the results, given the lower error range of laboratory blood analysis versus the capillary blood sample device used. Finally, in the long term, we are aiming for complete digital anthropometry from 4D models (3D + time) of the patient body to extract further measurements that reveal more information or have a higher correlation with other health indicators.

REFERENCES

[1] K. Bulhoes and L. Araújo, "Metabolic syndrome in hypertensive patients: Correlation between anthropometric data and laboratory findings," *Diabetes care*, vol. 30, no. 6, pp. 1624–1626, 2007.

[2] I. Elekima and A. Inokon, "A study of correlation of anthropometric data with atherogenic indices of students of rivers state university, Port Harcourt, Nigeria," *Asian J. Res. Med. Pharmaceutical Sci.*, vol. 6, no. 1, pp. 1–12, Feb. 2019.

[3] Y. Khader, A. Batieha, H. Jaddou, M. El-Khateeb, and K. Ajlouni, "The performance of anthropometric measures to predict diabetes mellitus and hypertension among adults in Jordan," *BMC Public Health*, vol. 19, no. 1, pp. 1–9, Dec. 2019.

[4] L. Lampignano, R. Zupo, R. Donghia, V. Guerra, F. Castellana, I. Murro, C. Di Noia, R. Sardone, G. Giannelli, and G. De Pergola, "Cross-sectional relationship among different anthropometric parameters and cardio-metabolic risk factors in a cohort of patients with overweight or obesity," *PLoS ONE*, vol. 15, no. 11, Nov. 2020, Art. no. e0241841.

[5] D. Rativa, B. J. T. Fernandes, and A. Roque, "Height and weight estimation from anthropometric measurements using machine learning regressions," *IEEE J. Translational Eng. Health Med.*, vol. 6, pp. 1–9, 2018.

[6] B. C. C. Lam, G. C. H. Koh, C. Chen, M. T. K. Wong, and S. J. Fallows, "Comparison of body mass index (BMI), body adiposity index (BAI), waist circumference (WC), waist-to-hip ratio (WHR) and waist-to-height ratio (WHtR) as predictors of cardiovascular disease risk factors in an adult population in Singapore," *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0122985.

[7] M.-A. Cornier, J.-P. Després, N. Davis, D. A. Grossniklaus, S. Klein, B. Lamarche, F. Lopez-Jimenez, G. Rao, M.-P. St-Onge, A. Towfighi, and P. Poirier, "Assessing adiposity: A scientific statement from the American heart association," *Circulation*, vol. 124, no. 18, pp. 1996–2019, Nov. 2011.

[8] D. P. Guh, W. Zhang, N. Bansback, Z. Amarsi, C. L. Birmingham, and A. H. Anis, "The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis," *BMC Public Health*, vol. 9, no. 1, pp. 1–20, Dec. 2009.

[9] M. Jiang, Y. Shang, and G. Guo, "Computational approach to body mass index estimation from dressed people in 3D space," *IET Image Process.*, vol. 14, no. 7, pp. 1248–1256, May 2020.

[10] Y. Lu, S. Zhao, N. Younes, and J. K. Hahn, "Accurate nonrigid 3D human body surface reconstruction using commodity depth sensors," *Comput. Animation Virtual Worlds*, vol. 29, no. 5, p. e1807, Sep. 2018.

[11] B. M. Smith, V. Chari, A. Agrawal, J. M. Rehg, and R. Sever, "Towards accurate 3D human body reconstruction from silhouettes," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 279–288.

[12] M. A. Trujillo-Jiménez, P. Navarro, B. Pazos, L. Morales, V. Ramallo, C. Paschetta, S. De Azevedo, A. Ruderman, O. Pérez, C. Delrieux, and R. Gonzalez-José, "Body2vec: 3D point cloud reconstruction for precise anthropometry with handheld devices," *J. Imag.*, vol. 6, no. 9, p. 94, Sep. 2020.

[13] A. Fuster-Guilló, J. Azorín-López, M. Saval-Calvo, J. M. Castillo-Zaragoza, N. Garcia-D'Urso, and R. B. Fisher, "RGB-D-based framework to acquire, visualize and measure the human body for dietetic treatments," *Sensors*, vol. 20, no. 13, p. 3690, Jul. 2020.

[14] A. Fuster-Guilló, J. Azorín-López, J. MiguelCastillo-Zaragoza, C. Manchón-Pernis, L. FernandoPérez-Pérez, and A. Zaragoza-Martí, "Multidimensional measurement of virtual human bodies acquired with depth sensors," in *Proc. Int. Workshop Soft Comput. Models Ind. Environ. Appl.* Cham, Switzerland: Springer, 2020, pp. 721–730.

[15] H. Ma and K.-J. Shieh, "Cholesterol and human health," *The J. Amer. Sci.*, vol. 2, no. 1, pp. 46–50, 2006.

[16] F. B. Hu, J. E. Manson, and W. C. Willett, "Types of dietary fat and risk of coronary heart disease: A critical review," *J. Amer. College Nutrition*, vol. 20, no. 1, pp. 5–19, Feb. 2001.

[17] G. Soliman, "Dietary cholesterol and the lack of evidence in cardiovascular disease," *Nutrients*, vol. 10, no. 6, p. 780, Jun. 2018.

[18] C.-H. Hsu, J.-D. Lin, C.-H. Hsieh, S. C. Lau, W.-Y. Chiang, Y.-L. Chen, D. Pei, and J.-B. Chang, "Adiposity measurements in association with metabolic syndrome in older men have different clinical implications," *Nutrition Res.*, vol. 34, no. 3, pp. 219–225, Mar. 2014.

[19] C. L. Craig, A. L. Marshall, M. SjöströM, A. E. Bauman, M. L. Booth, B. E. Ainsworth, M. Pratt, U. Ekelund, A. Yngve, J. F. Sallis, and P. Oja, "International physical activity questionnaire: 12-country reliability and validity," *Med. Sci. Sports Exercise*, vol. 35, no. 8, pp. 1381–1395, Aug. 2003.

[20] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 115–123.

NAHUEL GARCÍA-D'URSO received the bachelor's and master's degrees in computer science from the University of Alicante, Spain, where he is currently pursuing the Ph.D. degree with the Department of Computer Technology, and collaborates in research regarding the Tech4Diet project in a multidisciplinary environment along with nutrition experts of the Faculty of Health Sciences. His research interests include machine learning, data science, and computer vision.

PAU CLIMENT-PÉREZ received the B.Sc. degree in computer engineering and the master's degree in computer technologies from the University of Alicante, in September 2009, and October 2010, respectively, and the Ph.D. degree in computer vision from Kingston University London, in October 2016. He has industrial experience in a smart video surveillance company (2016–2018). He has been working as a Postdoctoral Researcher for several EU (PAAL) and national (Gloria2) research projects, since 2018. His research interests include active assisted living, privacy preservation, and more recently fisheries management, all this via computer vision and deep learning for different recognition tasks.

MIRIAM SÁNCHEZ-SANSEGUNDO received the Ph.D. degree in health sciences from the University of Alicante, Spain. She is currently working as a Research Professor with the University of Alicante. Her research interests include the diagnosis and advances in research on human behavior, the design of intervention programs to improve the prevention of obesity, neuropsychology, and mental disorders.

ANA ZARAGOZA-MARTÍ received the Diploma degree in human nutrition and dietetics and the M.Sc. degree in clinical and community nutrition from the University of Alicante, the M.Sc. degree in public health from the Miguel Hernández University of Elche, the master's degree in nutrition, dietetics, and diet therapy from the University of Navarre, and the Ph.D. degree in health sciences from the University of Alicante.

She is currently an Assistant Professor and a Researcher with the University of Alicante. Her main research activity has been aimed at recognizing the relationships between the Mediterranean diet and the health outcomes of different population groups, resulting in numerous scientific publications in high-impact scientific journals. Her lecturing activity is aimed at Clinical Nutrition and research methodology as part of the degree in human nutrition and dietetics, and the master's degree in clinical and community nutrition at the University of Alicante.

ANDRÉS FUSTER-GUILLÓ received the B.S. degree in computer science engineering from the Polytechnic University of Valencia, Spain, in 1995, and the Ph.D. degree in computer science from the University of Alicante, Spain, in 2003. Since 1997, he has been a member of the faculty of the Department of Computer Science and Technology, University of Alicante, where he is currently an Associate Professor. He was a Deputy Coordinator of the Polytechnic School and the Director of the Secretariat for Information Technology, University of Alicante. During this period, he has coordinated and participated in several strategic technological projects, including Open University (transparency portal and open data), UACloud, and Smart University, among others. He has published over 80 articles in different areas of research, including computer vision, 3D vision, machine learning, artificial neural networks, and open data.

JORGE AZORÍN-LÓPEZ received the degree in computer engineering, in 2001, and the Ph.D. degree in computer science from the University of Alicante, Spain, in 2007. Since 2001, he has been a Faculty Member of the Department of Computer Technology, University of Alicante, where he is currently an Associate Professor and the Academic Secretary. His current research interests include 3D computer vision, computational intelligence, machine learning, deep learning, ambient intelligence, human activity analysis, and visual inspection. In these lines of research, he has worked in 20 research projects (five of them as a coordinator) funded by national, regional, and local public and private entities. He has authored more than 100 contributions in several journals, conferences, and book chapters.