Gene – Nutrition Interactions in the Onset of Obesity as Cardiovascular Disease Risk Factor based on a Computational Intelligence Method

Ioannis K. Valavanis, Stavroula G. Mougiakakou, Member, IEEE, Stathis Marinos, George Karkalis, Keith A. Grimaldi, Rosalynn Gill, Konstantina S. Nikita, Senior Member, IEEE

Abstract—Identification of gene-gene and gene-environment interactions that contribute in the onset of a multi-factorial disease supports the prevention of diseases like the Cardiovascular Disease (CVD). Body Mass Index (BMI), a measure of human obesity, is an independent risk factor of CVD. Furthermore, it is known that a subject’s BMI is affected both by his/her lifestyle, e.g. nutrition, and genetic profile. Aim of the paper is to predict a subject’s onset of obesity using lifestyle and genetic information. The prediction is performed by a computational intelligence based system using a Parameter Decreasing Method (PDM) combined with an Artificial Neural Network (ANN). The system uses an initial set of 63 input variables corresponding to sex, average nutrition intake measurements, and genetic variations to identify the 32 most important ones that affect BMI. The selected variables are the ones to interact with each other towards the complex trait of BMI, which is used as a 2-class output variable (BMI ≤ 25 vs. BMI > 25) in the ANN. The system achieved a mean accuracy of the system evaluated by a 3-cross validation resampling technique equal to 77.89%.

I. INTRODUCTION

OBESITY is one of the independent risk factors of Cardiovascular Disease (CVD) [1], the class of diseases that involve the heart and/or blood vessels (arteries and veins) and can affect the cardiovascular system. CVD is a major death leading cause: heart disease and stroke, the most common CVD, are the first and third leading causes of death for both men and women in the United States, accounting for nearly 40% of all annual deaths [2]. In Europe, CVD causes over 4.35 million deaths which corresponds almost to half (49%) of deaths and is responsible for the one third of years lost due to early death [3]. CVD prevalence has increased in the developing countries, as well, due to changes of nutrition towards westernized diet high in sugar and fats and a sedentary lifestyle [4]-[6].

Many large scale studies have shown a positive relationship between CVD mortality and Body Mass Index (BMI), a measure of human obesity [7]-[10]. More specifically, the relative risk of myocardial fracture and Coronary Heart Disease (CHD) death for men in the overweight category (BMI > 25) was 1.21 when compared to the reference group (BMI < 25), as reported in the Framingham Heart Study [8] which followed up 2455 United States citizens for 44 years. Such studies show that intervention in one’s obesity is advisable towards the prevention of CVD. However, one certain diet cannot have the same effect on individuals’ weight loss since weight (and other CVD related risk factors) is influenced not only by individuals’ diet but also by his/her genetic profile [11]-[13]. This is demonstrated by the emergence of nutrigenetics, which studies the effects of fixed genetic variations, e.g. of a Single Nucleotide Polymorphism (SNP), on responsiveness to diet and introduces a new era in health management [13], [14]. In particular, Arkadianos et al [15] concluded after following-up individuals for 300 days that nutrigenetically tailored diets resulted in long-term BMI reduction and improvements in blood glucose level.

BMI is a complex trait influenced by genes, nutrition and generally one’s lifestyle. In order to predict a subject’s onset of obesity, in terms of BMI, using lifestyle and genetic information, computational methods can be applied. Recently, computational methods have been used to reveal gene-gene and/or gene-environment interactions included in the etiology of multi-factorial diseases, like CVD, Type 2 diabetes mellitus, and cancer. These methods adopt either “traditional approaches”, e.g. Logistic Regression (LR), Artificial Neural Networks (ANNs), Parameter Decreasing Methods (PDMs), Genetic Programming (GP), or non-parametric approaches, e.g. two-step approaches,
combinatorial methods, Multifactor Dimensionality Reduction (MDR) method, recursive partition methods etc [16]. They conclude to a set of informative predictors (gene and/or environmental factors) that affect a multi-factorial disease trait solely or by contributing to gene-gene and/or gene-environment interactions and sometimes include them in a corresponding predictive model. A LR model optimized by a Genetic Algorithm (GA) was proposed to predict a binary trait by using information of multiple SNPs, environmental effects and their interactions [17]. The model was successfully applied for the analysis of Type 2 diabetes data without and with angiopathy. A conditional LR model and the MDR method were comparatively assessed in the examination of complex genetic interactions in the case of myocardial infarction [18]. In [19], an ANN was used for the diagnostic prediction of allergic diseases such as atopic dermatitis, allergic conjunctivitis, allergic rhinitis and bronchial asthma, using patients’ data regarding SNPs that are thought to affect these diseases. A PDM combined with an ANN (PDM-ANN) was used in [20] to analyze 25 SNPs of 17 genes in order to predict development of childhood allergic asthma and select susceptible SNPs. Ten (10) SNPs were identified as the more informative ones and were the only ones to be used by the predictive model. During the last years, evolutionary strategies (e.g. GA and GP) were used to select the predictors fed to an ANN and/or to optimize their architecture [21]-[23]. The recently developed Genetic Programming Neural Network (GPNN) represents ANNs by tree-like structures and has been used in order to both identify factors related to Parkinson’s disease and optimize internal parameters and architecture (e.g. neuron activation weights) of the ANN-based corresponding predictive model [24]. A strong relation between DSTL gene and sex with Parkinson’s disease was found. Furthermore, an MDR method was developed for collapsing high-dimensional data into a single dimension thus permitting interactions to be detected in relatively small sample sizes [25]. The method is used for detecting and characterizing gene–gene and gene–environment interaction effects on risk of common complex multifactorial diseases. The MDR method was successfully applied to identify gene-gene interactions that significantly predict prostate cancer risk [25], genes and pathways that are most important in the etiology of hypertension [27], and genotype combinations contributing to Type 2 diabetes mellitus [27]. Finally, the random forest approach, a recursive partition tree-based method, was used to select predictors associated with a complex multifactorial disease like asthma [29].

Scope of the current work is to investigate the causality beneath BMI in the context of nutrigenetics. A large set of subjects for which sex, nutrient habits, and genetic profile is known is fed to a PDM-ANN model. The most informative factors from sex, nutrition and genes that affect BMI are identified and applied in a corresponding ANN model able to predict the BMI related class to which a subject belongs to (BMI ≤ 25 vs. BMI > 25). Accuracy, sensitivity and specificity measurements combined with the resampling technique of cross validation are utilized to evaluate the constructed predictive model.

The rest of the paper is organized as follows: In Section II the used data and the applied methodology for data analysis are presented followed (Section III) by the corresponding experimental results and the discussion. Finally, in Section IV the overall conclusions of the presented study are given.

II. MATERIALS AND METHODS

In the following the dataset used in this study and the methodology for design and development of the PDM-ANN system are described.

A. Dataset

A set of 2341 white people that underwent a nutrigenetic test was used in the current study. Genotype, nutrient intake data and BMI measurement were collected from all subjects. Genotype and nutrient data were acquired using the Sciona MyCell™ kit (Sciona Inc., Boulder, CO, USA). Within this test, subjects completed a comprehensive diet and lifestyle questionnaire, while cheek cell samples were taken from them for genetic testing purposes. Nutrient intake measurements were determined from the responses to the diet questionnaire and depicted their average diet habits. A total of 38 nutrition related measurements were calculated. These included the total calories intake per day, as well the daily intake through food or supplements for various substances, i.e. calcium, allium, caffeine, cruciferous, folic acid, cholesterol, omega 3, refined carbohydrate, saturated fat, and vitamins A, B6, B12, C, D, E. For the substances that were taken as supplements by at least one subject (e.g. calcium and vitamins), two other measurements were calculated apart from “intake in food”: “intake in supplement” and “total intake”. Nutrition intake measurements were categorized into four classes. To this end, for “intake in food” and “total intake” measurements the quartiles were found, while for the “intake in supplement” measurement the first class corresponded to a zero intake and the remaining classes corresponded to the bottom 33.3%, middle 33.3% and top 33.3% of non-zeros values in the measurement. The cheek cell samples were sent by courier to Sciona Inc. and genetic testing was carried out using a Sequenom Mass Array system for 24 genetic variations (SNPs or Insertion(I)/Deletion(D)) related to nutrition: ACE II/DD, APOC3 C3175G, CBS C699T, CETP G279A, COL1A1 G Sp1 T, GSTM1 deletion, GSTP1 A313G, GSTP1 C341T, GSTT1 deletion, IL 6 G634C, IL 6 G174C, LPL 1595G, MTHFR C677T, MTHFR A1298C, MTR A2756G, MS MTRR A66G, ENOS G894T, PPAR gamma 2 Pro12Ala, MnSOD C28T, SOD3 C760T, TNF alpha G308A, VDR Fok1, VDR Bsm1, VDR Taq1. Each genetic variation is featured one out of three (3) forms (e.g. AA, GG and AG for the CETP G279A SNP, and II, DD, ID

In conclusion, the dataset used in this study consisted of 2341 white subjects, each of whom underwent a nutrigenetic test including a comprehensive diet and lifestyle questionnaire, as well as the collection of cheek cell samples for genetic testing. The dataset included a set of 38 nutrition related measurements and 24 genetic variations (SNPs or Insertion(I)/Deletion(D)) related to nutrition. The genetic variations were used to identify factors related to Parkinson’s disease and optimize internal parameters and architecture of the corresponding predictive model.
for the ACE II/DD variation), resulting to three-class variable. For each subject, the sex is also known and is used as a two-class variable (male vs. female). Finally, BMI was calculated for all subjects: 877 out of 2341 were characterized as normal (BMI ≤ 25), while the rest 1474 subjects as overweight (BMI > 25). A total of 63 variables, composed by the subsets of nutrient intake measurements, genetic variations and sex, compose the initial set of input variables fed to a PDM-ANN model.

B. Parameter Decreasing Method Artificial Neural Network (PDM-ANN) System

The PDM-ANN based system combines the PDM (a backwards features selection method) with an ANN in order to identify from the initial set of input variables those variables that are most related to the output and contain the causality to it [16], [20]. The ANN used here corresponds to a multi-layer feed-forward perceptron network of one input layer fed with the selected input variables, one hidden layer of variable number of hidden neurons and the output layer of one neuron [30]. The output neuron is used to assign subjects in one of the two BMI related classes: 0 corresponds to normal (C1), and 1 to overweight (C2). The hyperbolic tangent sigmoid and log sigmoid are used as activation functions in the hidden and output layer, respectively. Due to the use of hyperbolic tangent sigmoid function the labels of categorical input variables were changed to certain values in the range [-1.0, +1.0], e.g. the labels [1st, 2nd, 3rd and 4th class] of a 4-class variable were changed to [-1.000, -0.333, +0.333, +1.000], while the use of log-sigmoid function in the output neuron, results in output values in the range [0.0, +1.0]. A value of less than +0.5 corresponds to C1, while a value greater or equal to +0.5 corresponds to C2.

The ANN is fully connected and trained using the batched back-propagation algorithm with adaptive learning rate and momentum [30]. The optimal number of hidden neurons, as well as the appropriate values of learning rate (lr) and momentum (mc), were estimated using a trial-and-error process. Initial learning rate and momentum were set to lr=0.01 and mc=0.9, respectively, while several values for the number of hidden neurons (number of hidden neurons = 2, 4, 6, 8, 10, 12) were tested.

PDM-ANN starts by constructing an ANN fed by all 63 input variables. The ANN is trained and evaluated using the three-Cross Validation (3-CV) resampling technique. Thus, the procedure of training and testing the ANN is repeated three times. Each time, the ANN is trained using 2/3 (~67%) randomly chosen cases of the available 2341 cases and tested using the remaining cases (~33%). Each of the ANNs obtained by 3-CV (ANN1, ANN2, ANN3) is evaluated using the mean value \((\bar{A}_{(63)}, \bar{A}_{(63)}, \bar{A}_{(63)})\) of the classification accuracies obtained in the corresponding 3-CV training and testing sets. We note here that the accuracy achieved by a predictive model, e.g. an ANN, in a set is the fraction of cases in the set that are correctly classified by the model.

The 3-CV technique outputs a fitness value (%):

\[
F_{63} = (\bar{A}_{(63)} + \bar{A}_{(63)} + \bar{A}_{(63)}) / 3
\]  

(1)

for the initial ANNs that use all 63 input variables. This fitness value evaluates the set of all 63 input variables, as well, and it is a measure of its information content towards the BMI measurement. The procedure continuous by deleting one variable from the total number of variables and constructing the 3-CV ANNs that use the remaining variables as input. The 62-dimensional input variables set that yields the best average of mean accuracies \((\bar{A}_{(62)}, \bar{A}_{(62)}, \bar{A}_{(62)})\) in the 3-CV sets is the one chosen at this dimensionality threshold and is assigned the fitness value \(F_{62}\), similarly as when using all 63 input variables. As the procedure continuous, the input set of \(N\) variables is derived from the one that uses \(N+1\) inputs by subtracting the least informative variable by means of mean accuracy in the 3-CV training and testing sets and is assigned a fitness value \(F_N\). This process is repeated until one variable remains. The best set of 3-CV ANNs and the corresponding input variables set are selected based on the values of fitness function \(F\). The variables that are included in the selected set of variables are the most informative ones and contain the causality to the output. The ones left out are either redundant or contain no causality to the output.

III. RESULTS AND DISCUSSIONS

The PDM-ANN method was used to identify the most important predictors among the aforementioned genetic variations, nutrient intake measurements and sex that affect BMI when used as a two class output variable (C1 vs. C2), Fig. 1 shows the obtained \(F_N\) values per number of selected variables \(N\). It is observed that mean accuracy obtained in the 3-CV training and testing sets when using \(N\) input variables (corresponding to \(F_N\) value) is kept in the range 77.7%-79.6% up to \(N=32\), while it starts to decrease after the number of variables fed to the ANNs gets less than this certain dimensionality. It continuous to decrease up to \(N=1\), where a mean accuracy equal to 62.3% is obtained. The fact that the variables subtracted by the PDM up to the dimensionality threshold \(N=32\) don’t affect the \(F_N\) values indicates that these certain variables have not an important impact to the BMI measurement or are redundant with other variables not subtracted by PDM. On the other hand, the one by one subtraction of variables after dimensionality \(N=32\) causes a sequential reduce to \(F_N\) values and thus we can infer that the variables subtracted after this dimensionality threshold have a high information content towards BMI measurement. Consequently, the subset of 32 input variables provided by the backwards features selection of PDM can be considered as the optimal one for discriminating subjects according to BMI measurement in
the current study, in terms of obtained mean accuracy and number of variables included as well.

Table I presents the $F_N$ values and mean accuracies in the 3-CV training and testing sets per number of selected variables $N$ for $N=1, 2, 3, 4, 5$ and $30, 31, 32$. The in-between cases are not reported here both for room saving purposes and because they don’t offer much to the knowledge presented in Table I. $F_N$ values show the decrease of the mean accuracy obtained in 3-CV sets as the dimensionality reduces after $N=32$, while the same happens for accuracies obtained in the training sets. The selected 32-dimensional subset of variables is marked with bold in Table I and contains sex, 19 nutrition related variables e.g. calories and cholesterol with an obvious impact on one’s BMI, and twelve (12) genetic variations. Thus, the finally selected subset of variables contains factors reflecting both one’s sex, lifestyle, i.e. nutrition, and genetic profile. These interact with each other (gene-environment, gene-gene interactions) towards the complex trait of BMI. The corresponding predictive model based on ANNs achieved a mean accuracy equal to 77.89% in the 3-CV training and testing sets. Regarding its ability to generalize into totally unknown data, the corresponding mean accuracy achieved in the 3-CV testing sets is 60.22%.

Furthermore, the predictive model fed by the 32 inputs was evaluated in terms of sensitivity and specificity in the 2-class problem, as well. Taking the class C2 (BMI > 25) as positive and the class C1 (BMI ≤ 25) as negative, mean sensitivity and specificity in 3-CV sets were measured and are reported in Table II. Results show that the trained ANNs can discriminate better the positive cases (mean sensitivity in 3-CV testing sets: 69.15%) than the negative cases (mean specificity in 3-CV testing sets: 46.08%). This is, actually, a desirable result since the aim of the constructed ANN-based system is to predict one’s future status of BMI using subjects’ sex, nutrition habits and genetic profile. Thus, it is more important for the system to be able to predict a high BMI, characterized as an independent risk factor for CVD, than to predict a low BMI. It is important to note, here, that even when training the ANNs with one variable, i.e. the Cholesterol-Intake in Food, a quite high $F_{N-1}$ value (~62%) is obtained, as well. This can be explained by the effect that the intake of cholesterol on one’s BMI can have. However, the subset of 32 input variables seems to discriminate much better the subjects into the two BMI related classes ($F_{N=32}$ ~78%) and is the one to be considered optimal in the current study.

Future work includes the use of an ANN trained by a hybrid method based on GAs along with the back propagation algorithm with learning rate and momentum (GA-ANN) [31]. The hybrid method will select the most informative factors that affect BMI on the basis of the available dataset and adjust automatically the ANN architecture. This will permit the comparative assessment between the two methods (PDM-ANN and GA-ANN) and identify the most accurate and reliable. Furthermore, the inclusion of variables that reflect other aspects of one’s lifestyle, e.g. exercise, along with clinical measurements would improve the ability of an ANN-based model to predict the status of BMI.

The best prediction model will be included to a web-based integrated platform which will be able to i) assess the risk of person to develop CVD and to provide personalized advices regarding required lifestyle changes (e.g. nutrition habits) in order to reduce that risk. The general outline of the integrated system is presented in Fig. 2. After a person has completed the web-based lifestyle questionnaire and taken the genetic tests and clinical measurements, the computational intelligence module of the platform responds whether there is a risk that she/he will develop a CVD.
The result will be included in a report along with personalized information on the subject’s nutrition and genetic profile and personalized advices on appropriate changes in his/her lifestyle in order to reduce the CVD risk.

### TABLE I

<table>
<thead>
<tr>
<th>Selected Variables (N)</th>
<th>$F_N$ (%)</th>
<th>Mean Training Accuracy (%)</th>
<th>Mean Testing Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol-Intake in Food (1)</td>
<td>62.23</td>
<td>63.16</td>
<td>61.29</td>
</tr>
<tr>
<td>Sex, Cholesterol-Intake in Food (2)</td>
<td>64.24</td>
<td>64.17</td>
<td>64.32</td>
</tr>
<tr>
<td>Sex, Cholesterol-Intake in Food, Vitamin A-Total Intake (3)</td>
<td>65.52</td>
<td>65.71</td>
<td>65.34</td>
</tr>
<tr>
<td>Sex, Cholesterol-Intake in Food, Omega 3-Intake in Supplement, Vitamin A-Total Intake (4)</td>
<td>66.12</td>
<td>67.37</td>
<td>64.87</td>
</tr>
<tr>
<td>Sex, Cholesterol-Intake in Food, Omega 3-Intake in Supplement, Vitamin A-Total Intake, VDR Fok1 (5)</td>
<td>65.75</td>
<td>68.85</td>
<td>62.66</td>
</tr>
<tr>
<td>all variables in case N=32 except {Vitamin B12-Intake in Food, TNF alpha G308A} (30)</td>
<td>77.29</td>
<td>94.74</td>
<td>59.83</td>
</tr>
<tr>
<td>all variables in case N=32 except {TNF alpha G308A} (31)</td>
<td>77.30</td>
<td>94.21</td>
<td>60.39</td>
</tr>
<tr>
<td>Sex, Calories, Calcium-Intake in Food, Calcium-Supplement Only, Allium-Total Intake, Folic Acid-Supplement, Cholesterol-Intake in Food, Cholesterol-Intake in Supplement, Omega 3-Intake in Food, Omega 3-Intake in Supplement, Saturated Fat-Intake in Supplement, Vitamin A-Total Intake, Vitamin A-Intake in Food, Vitamin A-Intake in Supplement, Vitamin B6-Total Intake, Vitamin B6-Intake in Food, Vitamin B6-Intake in Supplement, Vitamin B12-Total Intake, Vitamin B12-Intake in Food, Vitamin C-Total Intake, CBS C699T, CETP G279A, COL1A1 G Sp1 T, GSTM1 deletion, GSTTI A313G, GSTTI deletion, IL 6 G634C, MTHFR C677T, MmSOD C28T, TNF alpha G308A, VDR Fok1, VDR BsmI (32)</td>
<td>77.89</td>
<td>95.56</td>
<td>60.22</td>
</tr>
</tbody>
</table>

### TABLE II

<table>
<thead>
<tr>
<th>Measurement</th>
<th>3-CV Training Sets</th>
<th>3-CV Testing Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Sensitivity (%)</td>
<td>98.14</td>
<td>69.15</td>
</tr>
<tr>
<td>Mean Specificity (%)</td>
<td>91.15</td>
<td>46.08</td>
</tr>
</tbody>
</table>

**IV. CONCLUSIONS**

In this paper a system based on the combined use of PDM and ANN has been presented aiming to identify factors related to the onset of obesity, as an example of CVD risk factor, using persons’ information regarding sex, nutrition habits and genetic variations. The system was able to select a set of 32 variables containing information about ones’ sex, lifestyle, i.e. nutrition, and genetic profile. The corresponding ANN achieved a mean accuracy equal to 77.89% in the 3-CV training and testing sets, while its mean ability to predict the risk of high BMI, characterized as an independent risk factor for CVD, was 69.15% in the 3-CV testing sets.

The above presented results are encouraging towards the design and development of a computational intelligent based system, which will be able to assess the risk for CVD based on a subject’s sex, genetic and lifestyle information.
REFERENCES


