# A Hybrid Genetic Algorithm for the Selection of the Critical Features for Risk Prediction of Cardiovascular Complications in Type 2 Diabetes Patients

Kalliopi V. Dalakleidi, Konstantia Zarkogianni, Vassilios G. Karamanos, Anastasia C. Thanopoulou and Konstantina S. Nikita

***Abstract*—The purpose of this study is to present a hybrid approach based on the combined use of a genetic algorithm (GA) and a nearest neighbours classifier for the selection of the critical clinical features which are strongly related with the incidence of fatal and non fatal Cardiovascular Disease (CVD) in patients with Type 2 Diabetes Mellitus (T2DM). For the development and the evaluation of the proposed algorithm, data from the medical records of 560 patients with T2DM are used. The best subsets of features proposed by the implemented algorithm include the most common risk factors, such as age at diagnosis, duration of diagnosed diabetes, glycosylated haemoglobin (HbA1c), cholesterol concentration, and smoking habit, but also factors related to the presence of other diabetes complications and the use of antihypertensive and diabetes treatment drugs (i.e. proteinuria, calcium antagonists, b-blockers, diguanides and insulin). The obtained results demonstrate that the best performance was achieved when the weighted k-nearest neighbours classifier was applied to the CVD dataset with the best subset of features selected by the GA, which resulted in high levels of accuracy (0.96), sensitivity (0.80) and specificity (0.98).**

## I. Introduction

PATIENTS with Type 2 Diabetes Mellitus (T2DM) are at increased risk of developing Cardiovascular Disease (CVD). T2DM is a metabolic disorder that is characterised by high blood glucose due to insulin resistance of the cells or reduced insulin production by the islet cells of the pancreas. CVD is the most serious long-term diabetes complication being the major cause of death in people with diabetes, accounting for 50% or more of all diabetes fatalities and severe disabilities. CVD affects the heart and blood vessels and may lead to fatal complications such as Coronary Heart Disease (CHD) and stroke. Guidelines for the management of T2DM advocate calculating CVD risk to guide the initiation of appropriate treatment. It is widely recognized that risk prediction models should provide accurate and validated estimates of the risk, in order to be useful for clinical decision making. Few studies have been devoted to the development of risk prediction models specific for diabetes population [1]. Several studies present CVD risk prediction models which include diabetes as a factor in the model. The majority of the risk prediction models, refer to the calculation of the 5 year CVD risk for patients with T2DM [2], [3], [4], [5], using the most common risk predictors such as the age, the sex, the duration of diagnosed diabetes, the glycosylated haemoglobin (HbA1c) concentration and the smoking habit.

An important issue towards the development of risk prediction models is the selection of clinically relevant predictors. The computational complexity of the problem of finding the optimal feature subset has been proved to be NP-hard [6]. Some of the proposed approaches in the existing literature are the branch and bound approaches, the sequential forward/backward search algorithms [7], stochastic algorithms, including simulated annealing, ant colony optimization, neural networks [8], [16], and genetic algorithms [9], [10], [11]. In particular, for the selection of critical features for diabetes complications in the current literature, the ReliefF model has been used, which is implemented by a K-Nearest Neighbours (KNN) classifier [12]. Neural networks combined with a Genetic Algorithm (GA) [13] have also been used for the identification of critical subsets of risk factors for CVD complications.

Often the medical datasets used are imbalanced, which means that some classes are highly underrepresented compared to other classes. This imbalance can make many conventional machine learning algorithms less effective, especially in predicting minority class examples. To overcome the problem of classification of imbalanced datasets, a number of solutions have been proposed at the data and algorithm level [14], while others are based on the use of ensemble techniques, such as BEV, SMOTEBoost and AdaBoost [15]. The present study aims at the design, development and evaluation of a feature selection algorithm able to identify the critical risk factors that are strongly related with the incidence of fatal and nonfatal CVD in patients with T2DM, and can be used as part of a CVD risk prediction decision making tool in the context of diabetes

Manuscript received August 5, 2013.

Kalliopi V. Dalakleidi is with the National Technical University of Athens, Greece (phone: 0030-210-7722968 e-mail: kdalakleidi@ biosim.ntua.gr).

Konstantia Zarkogianni is with the National Technical University of Athens, Greece (phone: 0030-210-7722968 e-mail: kzarkog@ biosim.ntua.gr).

Vasilios G. Karamanos is with the Diabetes Center, 2nd Department of Internal Medicine and Research Laboratory, National University of Athens, Hippokration General Hospital, Vas Sofias 114, 11527 Ambelokipoi, Athens, Greece (phone: 0030-210-7769000 e-mail: diavitologiko1@hippocratio.gr).

Anastasia C. Thanopoulou is with the Diabetes Center, 2nd Department of Internal Medicine and Research Laboratory, National University of Athens, Hippokration General Hospital, Vas Sofias 114, 11527 Ambelokipoi, Athens, Greece (phone: 0030-210-7769000 e-mail: a_thanopoulou@hotmail.com).

Konstantina S. Nikita is with the National Technical University of Athens, Greece (phone: 0030-210-7722285 e-mail: knikita@ece.ntua.gr).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

TABLE I: DESCRIPTION OF THE CVD DATASET

| Continuous Variables | Mean Value± Standard Deviation | Categorical Variables | Number (Percentage) |
|---|---|---|---|
| Age | 58.56±10.70 | Hypertension | No: 300(54%), Yes:260(46%) |
| Age when firstly diagnosed with T2DM | 50.88±10.64 | Angiotensin Converting Enzyme Inhibitor | No: 445 (80%), Yes: 115 (20%) |
| Diabetes Duration | 7.68±7.38 | Sex | Male: 263(47%), Female: 297(53%) |
| Body Mass Index | 29.50±5.54 | Diabetic Parents | No: 304(54%), Yes: 256(46%) |
| Systolic Blood Pressure | 139.47±20.55 | Retinopathy | No: 485(87%), Yes: 75(13%) |
| Diastolic Blood Pressure | 82.71±10.74 | Calcium Antagonists | No: 463(83%), Yes: 97(17%) |
| Mean Arterial Pressure | 101.63±12.73 | Diuretics | No: 481(86%), Yes 79(14%) |
| Pulse of blood pressure | 56.76±15.81 | B-blockers | No: 507(91%), Yes: 53(9%) |
| Glycosylated Haemoglobin | 7.44±1.82 | Smoker | No: 289(52%), Yes: 146(26%), Only in the Past: 125(22%) |
| Blood Glucose | 164.95±56.20 | Proteinuria | No: 513(92%), Microalbuminuria: 28(5%),Albuminuria: 19(3%) |
| Total Cholesterol | 226.43±49.92 | Hypolipid diet | No:469(84%), Statines:74(13%), Fibrates:17(3%) |
| Triglycerides | 167.08±110.68 | Aspirin | No: 509(91%), 100mg: 44(8%), 325mg: 7(1%) |
| High-density lipoprotein cholesterol | 48.27±16.41 | Diet | No: 412(74%), Yes: 148(26%) |
| Low-density lipoprotein cholesterol | 147.34±42.34 | Sulfonylurea | No: 411(73%), Yes: 149(27%) |
| Non-HDL Cholesterol | 178.16±46.23 | Diguanides | No: 513(92%), Yes: 47(8%) |
| Lipid Ratio | 5.10±2.06 | Insulin | No: 504(90%), Yes: 56(10%) |

management.

The proposed method is based on the combined use of a GA and a KNN classifier. The novelty and effectiveness of the presented approach in comparison with existing approaches lies in the use of a Dual Weighted K-Nearest Neighbours (DWKNN) classifier in the fitness function (FF) of a GA with two sequential levels of evaluation of the candidate subsets. The DWKNN classifier outperforms the other implementations of this classifier in handling the problem of imbalanced datasets. Moreover, the presence of two sequential levels of evaluation for the ranking of the candidate chromosomes in the FF of the GA ensures better discrimination ability.

## II. MATERIAL AND METHODS

### A. Dataset

The proposed feature selection algorithm was developed and evaluated using data from the medical records of 560 T2DM patients, the CVD dataset, collected from a 5-year follow up at the Hippokration General Hospital of Athens. In this dataset, 40 out of the 560 T2DM patients (7.14%) developed fatal or non-fatal CVD during their 5 year follow up period. The dataset comprises 32 features, which are summarized in Table I, providing information related to demographics, lifestyle, laboratory examinations, complications/comorbidities and treatment.

### B. Methods

In the present study, the feature selection task is performed by a hybrid GA, where the proposed subsets of features are ranked using different versions of a KNN classifier.

#### *B1. The KNN algorithm and its variations*

Due to its robustness, elegance, and feasibility of realization, the KNN algorithm continues to be one of the most widely used classification techniques. According to the KNN algorithm, a query instance is assigned to the class represented by a majority of its k-nearest neighbours in the training set. No matter what the distance measures are employed in pattern classification, it has been proven [14] that the asymptotic error rate of the KNN algorithm approaches the optimal Bayesian error rate, when both the number $n$ of the samples and the number $k$ of neighbours tend to infinity.

*The WKNN Algorithm*

To weigh the closer neighbours more heavily than the farther ones, a Weighted K-Nearest Neighbours (WKNN) algorithm [10] is applied, in which the distance-weighted function $w_i$ to the $i$-th nearest neighbor is defined as,

$$w_i = \frac{k+1-i}{\sum_{m=1}^{k} m} \quad (1)$$

where $m$ is an integer in the interval $(1,k)$ and $k$ is the total number of the neighbours. Thus, all weights are in the interval $\left( \frac{1}{\sum_{m=1}^{k} m}, \frac{k}{\sum_{m=1}^{k} m} \right)$, and a neighbour with smaller distance is weighted more heavily than one with greater distance.

*The DWKNN Algorithm*

In order to address the effect of the number of neighbours on the classification performance, a DWKNN algorithm has been proposed [14]. The DWKNN algorithm gives different weights to the k nearest neighbours depending on distances between them and their ranking according to their distance from the query object. The distance-weighted function $w_i$ of the $i$-th nearest neighbor is computed as,

$$w_i = \begin{cases} \frac{d_k^{NN} - d_i^{NN}}{d_k^{NN} - d_1^{NN}} \times \frac{1}{i}, d_k^{NN} \neq d_1^{NN} \\ 1, d_k^{NN} = d_1^{NN} \end{cases} \quad (2)$$

where $d_i^{NN}$ *is* the distance of the $i$-th nearest neighbour from the query object, $d_1^{NN}$ is the distance of the nearest

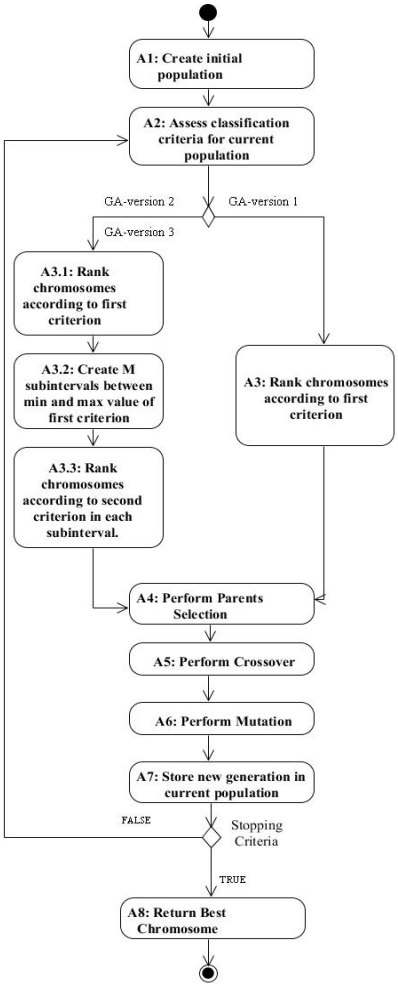


Fig. 1. Flow diagram of the GA and its variations

neighbour, and $d_k^{NN}$ is the distance of the k-furthest neighbour. Thus, the weight of the nearest neighbor is 1, and the weight of the furthest *k*-th neighbor is 0, whereas other weights are distributed between 0 and 1.

*B2. The GA and its variations*

The flow diagram of the GA and its variations is shown in Figure 1. The different versions of the GA vary, firstly, on the use of the WKNN classifier or the use of the DWKNN classifier for the ranking of the candidate chromosomes in the FF in Step A2. Secondly, the different versions of the GA vary on the number of criteria used for the ranking of the chromosomes in the FF in Step A3. The 10-fold cross validation method is used to assess the classification performance of the candidate subsets of features. After several experiments implemented aiming to reduce the computation time and increase the feature selection performance of the GA the values of the parameters were chosen appropriately.

In Step A2, the value of the FF for each candidate chromosome is calculated according to,

$$FF = b * (Se) + (1 - b) * (\frac{NIF}{L}) \quad (3)$$

where $b$ is one parameter of the FF that favors the importance of the classification performance of the candidate chromosome versus the importance of the number of not important features in the candidate chromosome. The variable *Se* is the value of the sensitivity of the chosen classifier (WKNN or DWKNN) for the CVD dataset with the use of the subset of features represented by the candidate binary chromosome, *NIF* is the number of not important features for the candidate chromosome and *L* is the total number of features.

*GA-version 1*

In Step A2, the value of the FF is calculated with the use of the sensitivity of the WKNN classifier. In Step A3, the candidate chromosomes are ordered in ascending order with respect to their FF value.

*GA-version 2*

In Step A2, the value of the FF is calculated with the use of the sensitivity of the WKNN classifier. In Step A3.1, the candidate chromosomes are ordered in ascending order with respect to their FF value. Then, in Step A3.2, the minimum and the maximum value of the FF are calculated and the interval between these two values is divided in *M* subintervals. In Step A3.3, in each subinterval, the candidate chromosomes are ordered in ascending order with respect to the accuracy of the classification of the WKNN classifier.

*GA-version 3*

In Step A2, the value of the FF is calculated with the use of the sensitivity of the DWKNN classifier. In Step A3.3, in the second level of ranking, the candidate chromosomes are ordered in ascending order with respect to the accuracy of the classification of the DWKNN classifier.

## III. Results and Discussion

### A. Experimental settings and performance analysis

In this section, a detailed presentation and discussion of the results acquired from the proposed methodology in the CVD dataset are given. The different versions of the GA presented in the previous section are used in several experiments in order to identify which one produces the best subsets of critical features related with the incidence of fatal and non fatal CVD events in patients with T2DM. In Figure 2, the sensitivity and accuracy achieved by the best subsets of features selected by GA-version 1, GA-version 2 and GA-version 3 for different number of the *k* nearest neighbours of the WKNN (GA-version 1 and GA-version 2) or DWKNN (GA-version 3) classifier, are shown. It can be easily observed that the addition of the second level of ranking of the candidate chromosomes in GA-version 2 and GA-version 3 improves the classification results of the selected subsets for most of the different numbers of *k* nearest neighbours in comparison with GA-version 1, where only one level of ranking is applied. Moreover, the use of the

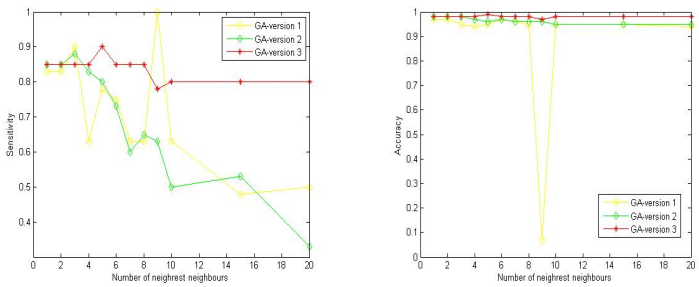


Fig. 2. Sensitivity (left panel) and Accuracy (right panel) versus the number of nearest neighbors, achieved by applying the different versions of the GA.

DWKNN classifier in GA-version 3 provides superior performance over WKNN in GA-version 1 and GA-version 2 for most of the different numbers of the *k* nearest neighbours. Next, the WKNN classifier was combined with all features and with the best subsets of features selected by the different versions of the GA. The accuracy, sensitivity and specificity of the classifier are presented in Table II, as well as the number of features that each algorithm selects. The mean number of features that were selected in the three best subsets is 13,33. As it can be observed, the performance measures are improved with the use of the best subset selected by the GA-version 3 for *k*=5.

The best subset selected by the GA-version 3 for *k*=5 consists of the following features: the age when the patient was firstly diagnosed with T2DM, the diastolic blood pressure, the mean arterial pressure, the HbA1c concentration, the blood glucose, the cholesterol, the HDL cholesterol, the LDL cholesterol, the non HDL cholesterol, the gender, the incidence of DM in the parents of the patient, the use of calcium antagonists, the use of b-blockers, the proteinuria, the diguanides treatment and the insulin treatment. It can be observed that the above subset of features includes the most common risk predictors reported in the literature, and highlights the significance of the inclusion of more risk factors related with the presence of diabetes complications and the use of antihypertensive and diabetes treatment drugs at the beginning of the follow up period. It is also important to notice, that among the features of the CVD dataset, there are sets of variables that are similar or related to each other, such as the current age, the age at diagnosis and diabetes duration. In the existing literature, different studies follow different approaches regarding these variables. For example, in the UKPDS study [1] only age at diagnosis and diabetes duration are included, whereas, in the Fremantle study [5] only current age is used. In the case of the best subset selected by GA-version 3 the use of the age when the patient was firstly diagnosed with T2DM is proposed, however, this is highly dependent on the classifier used.

TABLE II: CLASSIFICATION PERFORMANCE ACHIEVED BY THE WKNN CLASSIFIER (K=3) WITH THE BEST SUBSETS OF FEATURES SELECTED BY THE GA

| # of features | ACC | Se | Sp |
|---|---|---|---|
| *All* | 0.94 | 0.75 | 0.96 |
| *11(GA-version 1)* | 0.93 | 0.73 | 0.94 |
| *13(GA-version 2)* | 0.93 | 0.78 | 0.95 |
| *16 (GA-version 3)* | 0.96 | 0.80 | 0.98 |

## IV. CONCLUSION

The purpose of this study was to implement a hybrid approach based on a GA and a KNN classifier aiming at the selection of critical features for the 5-year risk prediction of CVD in patients with T2DM. In order to overcome the problem of unbalanced data the dual weighted k-nearest neighbours classifier was used. Moreover, two sequential levels of evaluation were applied for the ranking of the candidate chromosomes in order to enhance the discrimination performance. Future work will address handling similar features with respect to the simplification and improvement of the CVD risk calculation for a specific population. Moreover, the use of more risk factors related to diabetes complications and treatment will be investigated, in order to improve the CVD risk assessment and, thus, enable appropriate modifications of the treatment. The enhancement of the proposed algorithms with the ability not only to classify but also stratify the patients according to their risk (<5%, <10%, <20%) of developing a CVD event during the follow up period will also be addressed by means of ensemble classifiers.

## REFERENCES


[1] S. Dieren et al., "Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review," *Heart*, March 2012.

[2] A. P. Kengne et al., "Contemporary model for cardiovascular risk prediction in people with type 2 diabetes," *European Journal of Cardiovascular Prevention and Rehabilitation*, vol. 0, no. 00, Feb. 2011.

[3] J. Cederholm, K. Eeg-Olofsson, B. Eliasson, B. Zethelius, P. M. Nilsson, S. Gudbjornsdottir, "Risk prediction of cardiovascular disease in Type 2 Diabetes," *Diabetes Care*, vol. 31, no. 10, Oct. 2008.

[4] C. R. Elley, E. Robinson, T. Kenealy, D. Bramley, P. L. Drury, "Derivation and validation of a new cardiovascular risk score for people with type 2diabetes," *Diabetes Care*, no. 33, pp. 1347-1352, June 2010.

[5] W. A. Davis, M. W. Knuiman, T. M. E. Davis, "An Australian cardiovascular risk equation for type 2 diabetes: the Fremantle Diabetes Study," *Internal Medicine Journal*, vol. 40, pp. 286-292, May 2010.

[6] E. Amaldi and V. Kann, "On the approximation of minimizing non zero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, pp. 237-260, Dec. 1998.

[7] J. Xie, J. Lei, W. Xie, X. Gao,Y. Shi, X. Liu, "Novel Hybrid Feature Selection Algorithms for Diagnosing Erythemato-Squamous Diseases," *Lecture Notes and Computer Science*, vol. 7231, Springer, pp. 173-185, April 2012.

[8] M. Skevofilakas, K. Zarkogianni, B. G. Karamanos, K. S. Nikita, "A hybrid Decision Support System for the Risk Assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus," *32nd International Conference of the IEEE Engineering in Medicine and Biology Society*, Sep. 2010.

[9] D.E. Goldberg, "*Genetic Algorithms in Search, Optimization, and Machine Learning,*" Addison-Wesley Publishing Company Inc., Massachusetts, 1989.

[10] Y. Marinakis, G. Ntounias, J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification", *Computers in Biology and Medicine*, vol. 39, pp. 69-78, Jan. 2009.

[11] J. D. Kelly, L. Davis, "A hybrid genetic algorithm for classification," *Learning and Knowledge Acquisition*, pp. 645-650, 1991.

[12] B. H. Cho, H. Yu, K. W. Kim, T. H. Kim, I. Y. Kim, S. I. Kim, "Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods," *Artificial Intelligence in Medicine*, vol. 42, pp. 37-53, Jan. 2008.

[13] I. K. Valavanis, S. G. Mougiakakou, K. A. Grimaldi, K. S. Nikita, "A multifactorial analysis of obesity as CVD risk factor: Use of neural network based methods in a nutrigenetics context," *BMC Bioinformatics*, vol. 11, pp. 453, Sep. 2010.

[14] J. Gou, T. Xiong, Y. Kuang, "A novel weighted voting for K-Nearest neighbour rule," *Computers*, vol. 6, no. 5, pp. 833-840, July 2011.

[15] C. Li, "Classifying imbalanced data using a bagging ensemble variation," in *Proc. ACM-SE 45, 45th Annual Southeast Regional Conf.*, pp. 203–208, 2007.

[16] A. Verikas, M. Bacauskiene, "Feature selection with neural networks," *Pattern Recognition Letters 23*, pp. 1323–1335, 2002.