Received: 2 August 2016 | Revised: 17 December 2016 | Accepted: 17 March 2017

DOI: 10.1111/exsy.12214

WILEY Expert Systems

ARTICLE

Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications

Kalliopi Dalakleidi[1] | Konstantia Zarkogianni[1] | Anastasia Thanopoulou[2] | Konstantina Nikita[1]

[1] Faculty of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece

[2] Diabetes Center, 2nd Department of Internal Medicine, Medical School, University of Athens, Athens, Greece

Correspondence
Kalliopi Dalakleidi, Candidate, Faculty of Electrical and Computer Engineering, National Technical University of Athens, 9, Iroon Polytechniou Str., 15780 Zografos, Athens, Greece.
Email: kdalakleidi@biosim.ntua.gr

Funding Information
Hellenic State Scholarships Foundation

Abstract

The aim of the present study is to comparatively assess the performance of different machine learning and statistical techniques with regard to their ability to estimate the risk of developing type 2 diabetes mellitus (Case 1) and cardiovascular disease complications (Case 2). This is the first work investigating the application of ensembles of artificial neural networks (EANN) towards producing the 5-year risk of developing type 2 diabetes mellitus and cardiovascular disease as a long-term diabetes complication. The performance of the proposed models has been comparatively assessed with the performance obtained by applying logistic regression, Bayesian-based approaches, and decision trees. The models' discrimination and calibration have been evaluated using the classification accuracy (ACC), the area under the curve (AUC) criterion, and the Hosmer–Lemeshow goodness of fit test. The obtained results demonstrate the superiority of the proposed models (EANN) over the other models. In Case 1, EANN with different topologies has achieved high discrimination and good calibration performance (ACC = 80.20%, AUC = 0.849, *p* value = .886). In Case 2, EANN based on bagging has resulted in good discrimination and calibration performance (ACC = 92.86%, AUC = 0.739, *p* value = .755).

KEYWORDS

bagging ensemble, Bayesian models, cardiovascular disease, decision trees, diabetes, logistic regression, neural networks ensemble, risk

1 | INTRODUCTION

Diabetes mellitus (DM) is a chronic metabolic disease characterized by elevated blood glucose levels, due to insufficient insulin secretion (type 1 diabetes mellitus [T1DM]) and/or insulin resistance (type 2 diabetes mellitus [T2DM]). According to the International Diabetes Federation (IDF; IDF Diabetes Atlas, 2015), it is estimated that 415 million of people have DM and that this number is expected to rise beyond 642 million until 2040. T2DM is the most prevalent form of diabetes accounting for 91% cases in the adult population in high-income countries (IDF Diabetes Atlas, 2015). Due to its asymptomatic nature at the early stages of the disease, it is estimated that 193 million of cases are currently undiagnosed, which denotes that a vast amount of patients with T2DM are progressing towards complications unawares (IDF Diabetes Atlas, 2015).

Prolonged high blood glucose levels can lead to serious diseases affecting the heart and blood vessels, eyes, kidneys, and nerves. T2DM is strongly associated with disabling and mortality related long-term macrovascular and microvascular complications. Cardiovascular disease (CVD) is the most common cause of death and disability among patients with DM (IDF Diabetes Atlas, 2015).

The onset and the progress of T2DM may be delayed or even prevented by initiating appropriate intervention (Zarkogianni et al., 2015a). In particular, a number of prevention programs have demonstrated that effective lifestyle behavioral changes (e.g., diet and physical activity) can greatly reduce the risk of developing T2DM (IDF Diabetes Atlas, 2015). On the other hand, optimal diabetes treatment plan plays a crucial role in controlling the disease progression. Within this context, computational risk prediction models for the onset and the evolution of T2DM can greatly support clinical decision making and facilitate self-disease management (Verdú et al., 2016; Zarkogianni et al., 2015a).

There are several studies in the existing literature focusing on the development of T2DM risk engines (Zarkogianni et al., 2015a). Logistic

Copyright © 2017 John Wiley & Sons, Ltd

regression (Tabaei & Herman, 2002), Cox proportional hazards model (Tuomilehto et al., 2010), recursive partitioning (Xie et al., 2010), and Weibull parametric survival model (Kahn et al., 2009) are the most commonly used methodologies for building these models. The prediction horizon varies from 5 to 15 years and the models' discriminative ability, as measured by means of c-statistic, ranges from approximately 71% to 86% with the latter being achieved by applying the full Framingham 7-year risk calculator (Wilson et al., 2007). The most commonly identified T2DM risk predictors are age, family history of diabetes, body mass index, hypertension, waist circumference, sex, ethnicity, fasting glucose level, glycosylated hemoglobin, lipids, uric acid, or γ-glutamyltransferases, smoking status, and physical activity (Abbasi et al., 2012; Collins, 2011). Taking into account that T2DM has genetic predisposition, several genotype risk scores have been developed (Bao et al., 2013). Most of them receive as input the genetic variants that have been identified and/or confirmed within the frame of the genome-wide association studies for T2DM (Manolio, 2010) and the genome-wide association studies metadata analysis studies. Although putative causal genes corresponding up to 15–20% increase in the T2DM risk have been detected, the obtained area under the curves (AUCs) have been reported within the range 55% to 68%. On top of this, the addition of these genetic markers into the input space of conventional risk models has not improved significantly their predictive power (Wang et al., 2016). More genetic variants strongly correlated with T2DM need to be identified in order to provide valuable information towards predicting the risk.

Risk prediction models for long-term diabetes complications have been mainly focused on CVD and diabetic retinopathy (Brown et al., 2000; Skevofilakas, Zarkogianni, Karamanos, & Nikita, 2010; Stevens et al., 2001; Zarkogianni et al., 2015a). Referring to the former case, which constitutes the most important diabetes-related complication, the European Association for the study of diabetes recommends using Framingham and Diabetes Epidemiology: Collaborative analysis of Diagnostic criteria in Europe as preferred prediction models for calculating the CVD risk. However, these models are applicable to the general population and underestimate the risk in the population of diabetes. On the other hand, IDF guidelines recommend using the UK Prospective Diabetes Study (UKPDS) risk engine, which is dedicated to the T2DM population but results in varying discriminative performance (c-statistic: 65–86%) and poor calibration. Age, sex, systolic blood pressure, smoking status, atrial fibrillation, ethnicity, glycosylated hemoglobin, total cholesterol, HDL cholesterol, along with fasting, and 2-hr glucose constitute the most commonly used risk factors.

Having recognized the need of developing more efficient risk prediction models for the incidence of T2DM and its complications, the present study focuses on the comparative assessment of different statistical and machine learning methods towards producing reliable risk scores within a 5-year time frame. In this context, models based on artificial neural networks (ANNs) and ensemble learning are investigated for the first time with respect to their ability to capture the onset and the evolution of T2DM.

ANNs have been used in a great number of diagnostic decision support systems for medical applications, and they have demonstrated good predictive power (Buller, Buller, Innocent, & Pawlak, 1996; Verma & Zakos, 2001; Zarkogianni, Vazeou, Mougiakakou, Prountzou, & Nikita, 2011; Zarkogianni et al., 2015b). Ensembles of ANNs (EANN) can improve both the generalization abilities and the performance of an individual ANN, by compensating with each other the errors produced by each ANN (Sharkey, 1996). EANNs can be constructed by applying variations on each ANN member in terms of initial randomized weights, topology, learning algorithm, training data, and input space. There are several different ways of combining the outputs of each member of the EANN, such as averaging, weighted averaging, nonlinear combining, Bayesian, probabilistic, and stacked generalization methods (Mougiakakou, Valavanis, Nikita, & Nikita, 2007; Sharkey, 1996; Yang, Yang, Zhou, & Zomaya, 2010).

In this study, two different types of EANN have been investigated: (a) bagging ensemble of ANNs and (b) ensembles of ANNs with different topologies. Moreover, the majority voting scheme and averaging have been adopted. For the development and the evaluation of the models, two datasets have been used: (a) The Pima Indian Diabetes (PID) dataset (Blake & Merz, 1998) and (b) the Hippokration dataset, which has been granted from the General Hippokrateion Hospital of Athens (Dagliati et al., 2014).

Several classification techniques have been applied on the PID dataset such as ANNs, Bayesian-based approaches, fuzzy logic, decision trees, K-means, Support Vector Machines, random forests (RF), Genetic algorithms, and K-Nearest Neighbors, for producing the risk of developing T2DM (Anirudha, Kannan, & Patil, 2014; Belciug & Gorunescu, 2014; Bioch, Meer, & Potharst, 1996; Carpenter & Markuzon, 1998; Gürbüz & Kiliç, 2014; Ilango & Ramaraj, 2010; Kahramanli & Allahverdi, 2008; Michie, Spiegelhalter, & Taylor, 1994; Nanni, Fantozzi, & Lazzarini, 2015; Patil, Joshi, & Toshniwal, 2010; Perez, Yanez-Marquez, Camacho-Nieto, Lopez-Yanez, & Arguelles-Cruz, 2015; Purwar & Singh, 2015; Seera & Lim, 2014; Sutanto & Ghani, 2015; Yilmaz, Inan, & Uzer, 2014; Zhu, Xie, & Zheng, 2015). In order to provide evidence of advancing the current state of the art, the performance of the proposed models has been comparatively assessed with those obtained by applying logistic regression, Bayesian-based approaches, and decision trees.

2 | MATERIAL AND METHODS

The proposed work is focused on two cases:

- **Case 1.** *Prediction of the risk of developing T2DM within the 5-year time frame.*
- **Case 2.** *Prediction of the 5-year risk of experiencing the first fatal or nonfatal CVD incidence as a long-term T2DM complication.*

2.1 | Datasets

2.1.1 | Case 1: PID dataset

The PID dataset contains data from the 5-year follow-up of 768 Pima Indian women at least 21 years old living near Phoenix, Arizona, USA (Blake & Merz, 1998). In this population, 268 women developed T2DM within the 5-year time frame. As it is presented in Table 1,

TABLE 1 Description of the Pima Indian Diabetes dataset

Variables	Mean value ± standard deviation
Number of times pregnant	3.85 ± 3.37
Plasma glucose concentration at 2 hr in an oral glucose tolerance test	120.89 ± 31.97
Diastolic blood pressure	69.11 ± 19.36
Triceps skin fold thickness	20.54 ± 15.95
2-hour serum insulin	79.80 ± 115.24
Body mass index	31.99 ± 7.88
Diabetes pedigree function	0.47 ± 0.33
Age	33.24 ± 11.76

several clinical, physical, and epidemiological risk factors measured at the baseline visit have been taken into consideration in order to produce the risk of developing T2DM.

2.1.2 | Case 2: Hippokrateion dataset

The Hippokrateion dataset has been granted from the General Hippokrateion Hospital of Athens. It contains data from the 5-year follow-up of 560 T2DM patients without experiencing CVD incidence before the first visit. In this dataset, 40 out of the 560 T2DM patients (7.14%) experienced fatal or nonfatal CVD within the 5-year follow-up period. It comprises 27 features, which are summarized in Table 2, providing information related to demographics, lifestyle, laboratory examinations, complications or comorbidities, and treatment.

2.2 | Methods

2.2.1 | Ensembles of ANNs

In general, ANNs are inspired by the way a biological brain solves problems using large clusters of neurons connected by axons. ANNs are typically organized in weighted interconnected layers containing a number of nodes each of which applies an activation function. The input layer is responsible for feeding the patterns to the hidden layers where the main processing is performed. The hidden layers communicate with the output layer in order for the latter to produce the final decision. During the learning stage, the weights of the ANNs are adjusted according to the input training patterns.

The multilayer feedforward neural network constitutes the base classifier of the EANNs. The back-propagation learning algorithm has been used for training the ANNs, and the initial weights have been calculated based on the Nguyen–Widrow method (Nguyen & Widrow, 1990). In order to construct the EANNs, two different approaches have been followed resulting in two model versions:

- Model version 1: following the Bagging Ensemble approach, the bootstrap sampling with replacement procedure has been applied to the initial training dataset in order to produce different training datasets for each member of the EANN.
- Model version 2: the ensemble includes ANNs with different number of hidden layers and neurons in the hidden layers.

In both model versions, the outputs of each member of the ensemble have been combined on the basis of the majority voting scheme and averaging for classification and regression purposes, respectively.

2.2.2 | Logistic regression

The binary logistic model (BLM) describes the relationship between the independent predictors and the outcome variables, by generating the coefficients of a linear formula to predict the logit transformation of the probability (Tay, 2016).

Logistic model tree (LMT) is based on the combined use of logistic regression and decision tree learning (Landwehr, Hall, & Frank, 2005).

TABLE 2 Description of the Hippokrateion dataset

Continuous variables	Mean value ± standard deviation	Categorical variables	Number (percentage)
Age	58.56 ± 10.70	Hypertension	No: 300 (53.57%), yes: 260 (46.43%)
Diabetes duration	7.68 ± 7.38	Angiotensin-converting enzyme inhibitor	No: 445 (79.46%), yes: 115 (20.54%)
Body mass index	29.50 ± 5.54	Sex	Male: 263 (46.96%), female: 297 (53.04%)
Systolic blood pressure	139.47 ± 20.55	Diabetic parents	No: 304 (54.29%), yes: 256 (45.71%)
Diastolic blood pressure	82.71 ± 10.74	Retinopathy	No: 485 (86.61%), yes: 75 (13.39%)
Glycosylated Hemoglobin	7.44 ± 1.82	Calcium antagonists	No: 463 (82.68%), yes: 97 (17.32%)
Blood glucose	164.95 ± 56.20	Diuretics	No: 481 (85.89%), yes: 79 (14.11%)
Total cholesterol	226.43 ± 49.92	B-blockers	No: 507 (90.54%), yes: 53 (9.46%)
Triglycerides	167.08 ± 110.68	Smoker	No: 289 (51.61%), yes: 146 (26.07%), only in the past: 125 (22.32%)
High-density lipoprotein cholesterol	48.27 ± 16.41	Proteinuria	No: 513 (91.61%), microalbuminuria: 28 (5.00%), albuminuria: 19 (3.39%)
Low-density lipoprotein cholesterol	147.34 ± 42.34	Hypolipid diet	No: 469 (83.75%), Statines: 74 (13.21%), fibrates: 17 (3.04%)
		Aspirin	No: 509 (90.89%), 100 mg: 44 (7.86%), 325 mg: 7 (1.25%)
		Diet	No: 412 (73.57%), yes: 148 (26.43%)
		Sulfonylurea	No: 411 (73.39%), yes: 149 (26.61%)
		Diguanides	No: 513 (91.61%), yes: 47 (8.39%)
		Insulin	No: 504 (90%), yes: 56 (10%)

In particular, an LMT is a decision tree with linear regression models at its leaves, which are produced according to the LogitBoost algorithm (Friedman, Hastie, & Tibshirani, 2000).

2.2.3 | Bayesian model-based approaches

In general, a Bayesian network consists of a directed acyclic graph and a set of probability distributions for each node (Pearl, 1988). Nodes and arcs in the directed acyclic graph represent random variables and direct correlations between variables, respectively.

A specialized form of the Bayesian network is the probabilistic Naïve Bayes classifier, which relies on two simplifying assumptions. Firstly, the predictive attributes are conditionally independent given the class, and, secondly no hidden or latent attributes influence the prediction process (John & Langley, 1995; Witten & Frank, 2005).

2.2.4 | Decision tree-based models

A decision tree has been built based on the C4.5 algorithm, which uses the concept of information entropy (Quinlan, 1993). At each node of the tree, the attribute that most effectively splits its set of samples into one class or the other is selected according to the normalized information gain.

A rule-based classifier that infers rules by partial C4.5 decision trees (PART) has been applied (Frank & Witten, 1998). Each rule is constructed using the best leaf of each partial C4.5 decision tree.

An ensemble of C4.5 decision trees has been created following the diverse ensemble creation by oppositional relabeling of artificial training examples (DECORATE) approach (Melville & Mooney, 2003). DECORATE uses specially constructed artificial training examples, which are given category labels that disagree with the current decision of the ensemble.

The use of RF has also been investigated in this study (Breiman, 2001). RF combines a multitude of decision trees, which are trained with subsets randomly drawn from the initial training dataset.

2.2.5 | Evaluation criteria

The PID dataset has been split into 50% for training and 50% for testing. Referring to the Hippokrateion dataset, 70% of the whole dataset has been used for training and the remaining 30% for testing the models' performance.

The models' predictive performance has been measured for both discrimination and calibration. In order to evaluate the models' discriminative ability, the classification accuracy (ACC) and the AUC has been calculated (Swets, 1988). The ACC represents the percentage of the correct predicted outcomes. The AUC constitutes the most reliable measure of the ability of model to separate the individuals who developed the disease from those who did not, by providing higher risk scores to the former case. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. An AUC of 100% indicates perfect discrimination ability, and an AUC of 50% proves worthless performance.

Calibration measures how close the predictions are to the actual probability. The most commonly used measure of calibration is the Hosmer–Lemeshow goodness of fit test (Lloyd-Jones, 2010), which forms subgroups, typically using the deciles of the estimated risk. Within each subgroup, the actual against the predicted number of the disease incidents is compared by applying the χ^2 test.

3 | RESULTS AND DISCUSSION

3.1 | Parameters tuning

The ensemble sizes in model versions 1 and 2 have been chosen equal to 50 and five ANNs, respectively. Because the input space dimension is different for each case, the chosen number of neurons in the hidden layers is also different per case (Table 3).

In the LMT, the minimum number of instances at which a node is considered for splitting is 15, and a logistic model is built at a node only if it contains at least five instances. The structure learning algorithm of the Bayesian network is the hill climbing method K2 (Cooper & Herskovits, 1992). The naive Bayes network structure is used initially. The Bayesian score metric is employed during the learning of the network structure. The number of parents of each node is one. Direct estimates of the conditional probability distributions are calculated.

Regarding the decision tree-based models, the confidence factor used for pruning has been set to 0.25. The minimum number of instances per leaf (Decision Tree, DECORATE) and rule (PART) has been equal to two. The number of folds used for reduced-error pruning in Decision Tree and PART has been chosen to three. The desired number of member classifiers and the maximum number of iterations in DECORATE have been determined to 15 and 50, respectively. The number of the base classifiers in RF has been set to 100.

3.2 | Evaluation of the models' performance

The models' discrimination and calibration performance are presented in Tables 4 and 5, respectively. In Table 4, the highest obtained ACC and AUC for each case is highlighted in bold. In Table 5, all the bold *p* values are higher than .05 and indicate acceptable calibration performance. It can be shown that the model version 2 has achieved superior performance over all the other models in Case 1 (ACC = 80.20%, AUC = 0.849, *p* value = .886). On the other hand, model version 1 has demonstrated the best performance in Case 2 (ACC = 92.86%, AUC = 0.739, *p* value = .755). Although BLM has achieved the highest ACC (93.45%) in Case 2, it has resulted in low discrimination ability (AUC = 0.612).

The obtained AUC values are higher in Case 1 from those in Case 2 for all the models, as opposed to the ACC values. This is due to the unbalanced nature of the Hippokrateion dataset. From Table 5, it can be inferred that the model version 1, along with the Bayesian-based models and decision tree-based models apart from the RF have bad calibration performance in Case 1. Moreover, model version 2, along with BLM, Bayesian-based models and decision tree based models except from Bayes Net and RF have bad calibration performance in Case 2.

In order to justify the effectiveness of the ensemble approach, the performance of the EANNs has been compared with the one obtained by

TABLE 3 Number of neurons in the hidden layers

	Case 1	Case 2
Model version 1	{3}	{27}
Model version 2	{4}, {5}, {6}, {4}-{3}, {4}-{2}	{27}, {28}, {29}, {30}, {31}

Note. In model version 1, the number of neurons is identical for all the members of the ensemble. In model version 2, each member of the ensemble has different number of hidden layers and neurons.

TABLE 4 Models' discrimination performance as measured by the ACC and the AUC

	Case 1		Case 2	
Algorithm	ACC	AUC	ACC	AUC
Model version 1	**80.50**	**0.867**	92.86	**0.739**
Model version 2	80.20	0.849	92.86	0.731
ANN	79.43	0.848	92.86	0.535
BLM	80.47	0.858	**93.45**	0.612
LMT	77.60	0.840	92.86	0.487
Bayes net	71.35	0.803	92.86	0.500
Naïve Bayes	75.52	0.824	92.26	0.572
Decision tree	74.22	0.698	91.07	0.360
PART	73.70	0.753	87.50	0.333
DECORATE	75.26	0.817	89.29	0.439
RF	75.00	0.828	92.86	0.688

Note. ACC = classification accuracy; ANN = artificial neural network; AUC = area under the curve; BLM = binary logistic model; DECORATE = diverse ensemble creation by oppositional relabeling of artificial training examples; LMT = logistic model tree; RF = random forests.

The highest obtained ACC and AUC for each case is highlighted in bold.

TABLE 5 Models' calibration performance as measured by applying the Hosmer–Lemeshow goodness of fit test

	Case 1 *p* value	Case 2 *p* value
Model version 1	.031	**.755**
Model version 2	**.886**	.000
ANN	**.235**	**.065**
BLM	**.122**	.000
LMT	**.146**	**.105**
Bayes net	.000	**.353**
Naïve Bayes	.000	.000
Decision tree	.000	.000
PART	.000	.000
DECORATE	.000	.000
RF	**.522**	**.400**

Note. ANN = artificial neural network; BLM = binary logistic model; DECORATE = diverse ensemble creation by oppositional relabeling of artificial training examples; LMT = logistic model tree; RF = random forests.

All the bold p values are higher than .05 and indicate acceptable calibration performance.

applying a single ANN. Although ANN has achieved good performance in Case 1 (ACC = 79.43%, AUC = 84.80%, *p* value = .235), it has resulted in low discrimination ability (AUC = 53.50%) in Case 2. In both Cases, the superiority of the EANN over the ANN has been demonstrated.

Overall, the EANNs' ability to produce more reliable risk scores compared to the other models is attributed to the diversity obtained among the members of each ensemble. Due to the unbalanced nature of the Hippokrateion dataset, all the models, except from the proposed ones, have been over fitted to the majority class during the training stage, which is proven by the low values of AUC. Another important finding is that model version 1 is more capable of handling this unbalanced dataset than model version 2 in terms of calibration (*p* value = .755 vs. .000).

A comparison between the obtained results and those reported in the literature is carried out. Although a direct and fair comparison is not feasible due to different datasets, input spaces, and evaluation frameworks, substantial inferences can be obtained. Referring to Case 1, Detect-2 (Alssema et al., 2011), Australian type 2 diabetes risk assessment tool (AUSDRISK; Chen et al., 2010), European prospective investigation into cancer study-Norfolk (EPIC-Norfolk; Simmons et al., 2007), and German Diabetes Risk Score (GDRS) (Schulze et al., 2007) are the most well-studied 5-year risk prediction models. These models have been based on logistic and Cox proportional hazard regression. The reported AUCs are 76.40%, 78%, 76.20%, and 84%, respectively. Detect-2 and AUSDRISK have had good calibration performance, and the calibration *p* values for the EPIC-Norfolk and GDRS have not been reported. The model version 2 has resulted in slightly greater AUC (84.90%).

Regarding Case 2, taking into account that the UKPDS risk engine is one of the most widespread CVD risk prediction models dedicated to the population of T2DM, its performance has been evaluated within the framework of the present study on the Hippokrateion dataset and has been compared with the one obtained by applying model version 1. The UKPDS risk engine resulted in lower AUC (58.74%) than model version 1 (73.90%) and not acceptable calibration performance (*p* value = .00).

It should be pointed out that the proposed models have been developed and evaluated using data corresponding to homogeneous populations in terms of race and ethnicity, and in Case 1, only the female population has been considered. This benefits the proposed models compared to the aforementioned state of the art models especially when the reported results have been produced by validating the models using independent cohorts of individuals (AUSDRISK, EPIC-Norfolk, UKPDS). Moreover, EANNs have higher complexity than logistic and Cox hazard regression, making thus the interpretation of the predicted outputs more difficult. On the other hand, this sophisticated technique has the capacity to handle unbalanced datasets and to result in greater accuracy than simpler models (e.g., BLM and LMT).

Based on the outcomes of the proposed work, it can be inferred that ANNs combined with ensemble learning have great potential to support medical decision-making for the management of diabetes. It should be stressed that feature selection is out of the scope of the present study, because all the risk factors, which have been taken into consideration, have been well established in the literature of strongly influencing the onset of T2DM and the incidence of CVD as long-term T2DM complication. Future work concerns the validation of the proposed models using other cohorts of individuals or patients. On top of this, the latest achievements towards identifying molecular biomarkers associated with the onset and progress of T2DM using high-throughput-omic technologies such as microarrays, next generation sequencing, and mass spectrometry (Floegel et al., 2013) pave the way for enriching the models' input space by taking, also, into account the biological profile. The enhanced integration of heterogeneous datasets, from behavioral down to molecular (genomic, transcriptomic, epigenomic, proteomic, and metabolomic) level, constitutes important challenge and has great potential to early detect indicative abnormalities relevant to the onset and the progress of the disease and to increase the accuracy of the risk scores.

4 | CONCLUSIONS

A comparative assessment of different machine learning and statistical methodologies towards the development of risk prediction models for the incidence and the evolution of T2DM has been conducted. The obtained results justify the need to apply more sophisticated techniques in order to achieve accuracy and reliability. EANNs can significantly contribute in this direction by having the capacity to handle the unbalanced nature, which usually occurs in medical datasets, and furthermore to capture an individual's health evolution.

ACKNOWLEDGEMENT

Kalliopi Dalakleidi was supported by a scholarship for PhD studies from the Hellenic State Scholarships Foundation "IKY fellowships of excellence for postgraduate studies in Greece-Siemens Program."

REFERENCES

Abbasi, A., Peelen, L. M., Corpeleijn, E., Schouw, Y., Stolk, R. P., Spijkerman, A., … Beulens, J. (2012). Prediction models for risk of developing type 2 diabetes: Systematic literature search and independent external validation study. *BMJ*, 345. https://doi.org/10.1136/bmj.e5900.

Alssema, M., Vistisen, D., Heymans, M. W., Nijpels, G., Gluemer, C., Zimmet, P. Z., … Dekker, J. M. (2011). The evaluation of screening and early detection strategies for type 2 diabetes and impaired glucose tolerance (DETECT-2) update of the Finnish diabetes risk score for prediction of incident type 2 diabetes. *Diabetologia*, *54*, 1004–1012.

Anirudha, R. C., Kannan, R., & Patil N. (2014) Genetic algorithm based wrapper feature selection on hybrid prediction model for analysis of high dimensional data, presented at 9th international conference on industrial and information systems, Gwalior, India, 2014, IEEE.

Bao, W., Hu, F. B., Rong, Y., Bowers, K., Schisterman, E. F., Liu, L., & Zhang, C. (2013). Predicting risk of type 2 diabetes mellitus with genetic risk models on the bases of established genome-wide association markers: A systematic review. *American Journal of Epidemiology*, *178*, 1197–1207.

Belciug, S., & Gorunescu, F. (2014). Error-correction learning for artificial neural networks using the Bayesian paradigm. Application to automated medical diagnosis. *Biomedical Informatics*, *52*, 329–337.

Bioch, J. C., Meer, O., & Potharst R. (1996) Classification using Bayesian neural nets, presented at International conference on neural networks, Washington DC, 1996, IEEE, 3, pp. 1488–1493.

Blake, C. L., & Merz C. J. (1998) UCI repository of machine learning databases. http://www.ics.uci.edu/mlearn/MLRepository.html, Department of Information and Computer Science, University of California, Irvine, CA.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Brown, J. B., Russell, A., Chan, W., Pedula, K., & Aickin, M. (2000). The global diabetes model user friendly version 3.0. *Diabetes Research and Clinical Practice*, *50*, 15–46.

Buller, D., Buller, A., Innocent, P. R., & Pawlak, W. (1996). Determining and classifying the region of interest in ultrasonic images of the breast using neural networks. *Artificial Intelligence in Medicine*, *8*, 53–66.

Carpenter, G. A., & Markuzon, N. (1998). ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases. *Neural Networks*, *11*, 323–336.

Chen, L., Magliano, D. J., Balkau, B., Colagiuri, S., Zimmet, P. Z., Tonkin, A. M., … Shaw, J. E. (2010). AUSDRISK: An Australian type 2 diabetes risk assessment tool based on demographic, lifestyle and simple anthropometric measures. *The Medical Journal of Australia*, *192*, 197–202.

Collins, G. (2011). Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. *BMC Medicine*, *9*, 1–14.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*, 309–347.

Dagliati, A., Sacchi, L., Bucalo, M., Segagni, D., Zarkogianni, K., Millana, A. M., … Bellazzi, R. (2014). A data gathering framework to collect type 2 diabetes patients data, presented at IEEE-EMBS international conference on biomedical and health informatics (BHI), Valencia, 2014, IEEE, pp. 244–247.

Floegel, A., Stefan, N., Yu, Z., Muehlenbruch, K., Drogan, D., Joost, H. G., … Pischon, T. (2013). Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes*, *62*, 639–648.

Frank, E., & Witten I. H. (1998) *Generating accurate rule sets without global optimization, presented at Fifteenth International Conference on Machine Learning, Madison, WI. San Francesco, 1998*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 144–151.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, *28*, 337–407.

Gürbüz, E., & Kiliç, E. (2014). A new adaptive support vector machine for diagnosis of diseases. *Expert Systems*, *31*, 389–397.

IDF Diabetes Atlas, Seventh edition, 2015.

Ilango, S. B., & Ramaraj N. (2010) A hybrid prediction model with F-score feature selection for type II diabetes databases, presented at 1st amrita ACM-W celebration on women in computing in India, Coimbatore, India, 2010, ACM, 1–4.

John, G. H., & Langley P. (1995) *Estimating continuous distributions in Bayesian classifiers, presented at Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, Montreal, Quebec*, 1995, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 338–345.

Kahn, H. S., Cheng, Y. J., Thompson, T. J., Imperatore, G., & Gregg, E. W. (2009). Two risk-scoring systems for predicting incident diabetes mellitus in U.S. adults age 45 to 64 years. *Annals of Internal Medicine*, *150*, 741–751.

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*, *35*, 82–89.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, *95*, 161–205.

Lloyd-Jones, D. M. (2010). Cardiovascular risk prediction: Basic concepts, current status, and future directions. *Circulation*, *121*, 1768–1777.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *The New England Journal of Medicine*, *363*, 166–176.

Melville, P., & Mooney, R. J. (2003). *Constructing diverse classifier ensembles using artificial training examples, Presented at Eighteenth International Joint Conference on Artificial Intelligence, 2003*. (pp. 505–510). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine learning, neural and statistical classification*. NJ, USA: Ellis Horwood Upper Saddle River.

Mougiakakou, S., Valavanis, I., Nikita, A., & Nikita, K. (2007). Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers. *Artificial Intelligence in Medicine*, *41*, 25–37.

Nanni, L., Fantozzi, C., & Lazzarini, N. (2015). Coupling different methods for overcoming the class imbalance problem. *Neurocomputing*, *158*, 48–61.

Nguyen, D., & Widrow B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In Proceedings of the International Joint Conference on Neural Networks, San Diego, CA.

Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for type-2 diabetic patients. *Expert Systems with Applications*, *37*, 8102–8108.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Perez, M. A., Yanez-Marquez, C., Camacho-Nieto, O., Lopez-Yanez, I., & Arguelles-Cruz, A.-J. (2015). Collaborative learning based on associative models: Application to pattern classification in medical datasets. *Computers in Human Behaviour*, *51*, 771–779.

Purwar, A., & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data. *Expert Systems with Applications*, *42*, 5621–5631.

Quinlan, R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Schulze, M. B., Hoffmann, K., Boeing, H., Linseisen, J., Rohrmann, S., Moehlig, M., … Joost, H. G. (2007). An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care*, *30*, 510–515.

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert Systems with Applications*, *41*, 2239–2249.

Sharkey, A. (1996). On combining artificial neural nets. *Connection Science*, *8*, 299–314.

Simmons, R. K., Harding, A. H., Wareham, N. J., Griffin, S. J., & EPIC-Norfolk Project Team (2007). Do simple questions about diet and physical activity help to identify those at risk of type 2 diabetes? *Diabetic Medicine*, *24*, 830–835.

Skevofilakas, M., Zarkogianni, K., Karamanos, B.G., & Nikita K. S. (2010). A hybrid decision support system for the risk assessment of retinopathy development as a long term complication of type 1 diabetes mellitus, presented at 32nd IEEE EMBS conference, Buenos Aires, Argentina, 2010, IEEE, pp. 6713–6716.

Stevens, R. J., Kothari, V., Adler, A. I., Stratton, I. M., & United Kingdom Prospective Diabetes Study (UKPDS) Group. (2001). The UKPDS risk engine: A model for the risk of coronary heart disease in type 2 diabetes. *Clinical Science*, *101*, 671–679.

Sutanto, D. H., & Ghani, M. K. A. (2015). Improving classification performance of k-nearest neighbor by hybrid clustering and feature selection for non-communicable disease prediction. *Engineering and Applied Sciences*, *10*, 6817–6825.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.

Tabaei, B., & Herman, W. (2002). A multivariate logistic regression equation to screen for diabetes: Development and validation. *Diabetes Care*, *25*, 1999–2003.

Tay, R. (2016). Comparison of the binary logistic and skewed logistic (Scobit) models of injury severity in motor vehicle collisions. *Accident Analysis & Prevention*, *88*, 52–55.

Tuomilehto, J., Lindstroem, J., Hellmich, M., Lehmacher, W., Westermeier, T., Evers, T., … Chiasson, J. L. (2010). Development and validation of a risk-score model for subjects with impaired glucose tolerance for the assessment of the risk of type 2 diabetes mellitus: The STOP-NIDDM risk-score. *Diabetes Research and Clinical Practice*, *87*, 267–274.

Verdú, J., Sambo, F., Di Camillo, B., Cobelli, C., Facchinetti, A., Fico, G., … Zarkogianni, K. (2016). Predictive, preventive and personalized medicine in diabetes onset and complication (MOSAIC project). *The EPMA Journal*, *7*, 42–43.

Verma, B., & Zakos, J. (2001). A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. *IEEE Transactions on Information Technology in Biomedicine*, *5*, 46–54.

Wang, X., Strizich, G., Hu, Y., Wang, T., Kaplan, R. C., & Qi, Q. (2016). Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction. *Diabetes*, *8*, 24–35.

Wilson, P. W., Meigs, J. B., Sullivan, L., Fox, C. S., Nathan, D. M., & D'Agostino, R. B. Sr. (2007). Prediction of incident diabetes mellitus in middle aged adults: The Framingham offspring study. *Archives of Internal Medicine*, *167*, 1068–1074.

Witten, I. H., & Frank, E. (2005). *Data mining. Practical machine learning tools and techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Xie, J., Hu, D., Yu, D., Chen, C. S., He, J., & Gu, D. (2010). A quick self-assessment tool to identify individuals at high risk of type 2 diabetes in the Chinese general population. *Journal of Epidemiology and Community Health*, *64*, 236–242.

Yang, P., Yang, Y., Zhou, B., & Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, *5*, 296–308.

Yilmaz, N., Inan, O., & Uzer, M. S. (2014). A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases. *Medical Systems*, *38*, 48.

Zarkogianni, K., Vazeou, A., Mougiakakou, S. G., Prountzou, A., & Nikita, K. S. (2011). An insulin infusion advisory system based on autotuning nonlinear model-predictive control. *IEEE Transactions on Biomedical Engineering*, *58*, 2467–2477.

Zarkogianni, K., Litsa, E., Mitsis, K., Wu, P. Y., Kaddi, C. D., Cheng, C. W., … Nikita, K. S.. (2015a). A review of emerging technologies for the management of diabetes mellitus. *IEEE Transactions on Biomedical Engineering*, *62*, 2735–2749.

Zarkogianni, K., Mitsis, K., Litsa, E., Arredondo, M. T., Fico, G., Fioravanti, A., & Nikita, K. S. (2015b). Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. *Medical & Biological Engineering & Computing*, *53*, 1333–1343.

Zhu, J., Xie, Q., & Zheng, K. (2015). An improved early detection method of type-2 diabetes mellitus using multiple classifier system. *Information Sciences*, *292*, 1–14.

AUTHOR BIOGRAPHIES

Kalliopi Dalakleidi received her diploma from the School of Electrical and Computer Engineering at the National Technical University of Athens (NTUA) in 2005. Then, she worked as an ontology engineer on the automatic extraction of semantics from multimedia content in the framework of the European Project BOEMIE. Since 2012, she has focused on the development of mathematical models and algorithms on the glucose metabolism for the diagnosis and treatment of diabetes and its complications and is working towards her PhD. She is a teaching and laboratory assistant in the Simulation of Physiological Systems, Biomedical Engineering Laboratories and Medical Imaging and Image Processing undergraduate courses of the School of Electrical and Computer Engineering.

Konstantia Zarkogianni received the diploma in Electrical and Computer Engineering (2003) from the Aristotle University of Thessaloniki, Greece, the MSc Degree in Electronic and Computer Engineering (2005) from the Technical University of Crete, Greece, and the PhD degree (2011) from the NTUA, Greece. Since 2005, she is a member of the BIOmedical Simulations and IMaging Laboratory of NTUA. Her current research interests include clinical decision support systems, control systems, physiological systems modeling, diabetes management, multiscale modeling. She has authored or coauthored 10 papers in refereed international journals, one chapter in book, and 18 papers in international conference proceedings. She has participated as research associate and principle investigator in national and EU funded projects. She has been a guest editor of the special issue on Emerging Technologies for the Management of Diabetes Mellitus (Springer Journal of Medical and Biological Engineering and Computing [MBEC], 2015). She is a member of the Editorial Board of the SpringerPlus journal and reviewer for international scientific journals (IEEE Transactions on Biomedical Engineering, IEEE Journal of Biomedical and Health Informatics, Springer Medical & Biological Engineering & Computing, Elsevier Journal of Biomedical Informatics, and JSM Diabetology and Management). She is a member of the Institute of Electrical and Electronics Engineers and the Technical Chamber of Greece.

Anastasia Thanopoulou was born in 1965 in Indianapolis, Indiana, USA. She received her Medical Doctor's Degree in 1989 from the Aristotelian University of Salonica with grade Excellent, having a National Scholarship for her studies, and her Doctorate Degree from the same University in 1998 with Grade A. In 1999 she received a Certified Diploma—Specialisation in Internal Medicine—and in 2002 the Sub-Specialization in Diabetes Mellitus. From 1999 till 2005, she served as senior instructor of Medicine in the Technological Educational Foundation of Athens and as a Research Associate in the Diabetes Center of the 2nd Department of Internal Medicine, Medical School, National and Kapodistrian University of Athens, Hippokration Hospital, Athens, Greece. From 2005 till 2011, she served as a lecturer in Internal Medicine and Diabetes Mellitus in Medical School, National and Kapodistrian University of Athens, Greece, and from 2011 till present as assistant professor (from 2015 in Tenure Rank). From 2012, she is the scientific director of the Diabetes Center of the 2nd Department of Internal Medicine, Medical School, National and Kapodistrian University of Athens, Hippokration Hospital, Athens, Greece. She is the advisor for Academic Matters of the medical students of her clinic and responsible for the field of Internal Medicine for the School of Dentistry, National and Kapodistrian University of Athens (2011–present). At present, she is the supervisor in five doctoral theses of medical doctors. She has given more than 100 invited lectures in International and Greek Symposia and Conferences. She is a member of seven International and Greek professional societies and has served as the secretary of the Board of the Diabetes and Nutrition Study Group of the European Association for the Study of Diabetes (2010–2013). She has received various honors and awards. She is or has been a program director–principal investigator or coinvestigator of multiple scientific projects either funded by the Greek state or by other bodies. Some of the projects are part of international collaborations. She has been a coorganizer in 18 Greek and two International Conferences and Chair of the Local Organizing Committee of the 30th International Symposium on Diabetes and Nutrition of the Diabetes and Nutrition Study Group of the European Association for the Study of Diabetes, held in Athens, Greece, in 2012. She has authored or coauthored 61 abstracts of oral presentations in international congresses and 55 in Greek congresses. She has authored or coauthored 22 full articles in international SCI journals with personal mean impact factor 10.341. She has served as reviewer in 20 international journals and one Greek, reviewing of 63 papers. She is the coeditor of a Greek book and has written chapters in 28 educational and other Greek books. Her current research interests include diabetes comorbidities, nutrition and diabetes, drugs adverse effects, insulin resistance, liver, and diabetes.

Konstantina Nikita received the diploma in Electrical Engineering (1986) and the PhD degree (1990) from the NTUA, Greece. She then received the MD degree (1993) from the Medical School, University of Athens, Greece. Since 1990, she has been working as a researcher at the Institute of Communication and Computer Systems. In 1996, she joined the School of Electrical and Computer Engineering, NTUA, as an assistant professor, and since 2005, she serves as a professor at the same school. Her current research interests include biomedical signal and image processing and analysis, biomedical informatics, simulation of physiological systems, medical imaging, biological effects, and medical applications of radiofrequency electromagnetic fields. Dr. Nikita has authored or coauthored 154 papers in refereed international journals, 38 chapters in books, and over 300 papers in international conference proceedings. She has authored or edited two books (Simulation of Physiological Systems and Medical Imaging Systems) in Greek and five books in English published by Springer and Wiley. She holds two patents. She has been the technical manager of several European and National R&D projects. She is an associate editor of the IEEE Transactions on Biomedical Engineering, the IEEE Journal of Biomedical and Health Informatics, Wiley Bioelectromagnetics, and the Journal of Medical and Biological Engineering and Computing and a guest editor of several international journals. Dr. Nikita has received various honors or awards, among which, the Bodossakis Foundation Academic Prize for exceptional achievements in "Theory and Applications of Information Technology in Medicine" (2003).

How to cite this article: Dalakleidi K, Zarkogianni K, Thanopoulou A, Nikita K. Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications. *Expert Systems*. 2017;e12214. https://doi.org/10.1111/exsy.12214