TITLE - Early temporal prediction of Type 2 Diabetes Risk Condition from a General Practitioner Electronic Health Record: A Multiple Instance Boosting Approach

AUTHORS - Bernardini Michele, Morettini Micaela, Romeo Luca, Frontoni Emanuele, Burattini Laura.

PUBLISHED: 6 May 2020

YEAR - 2020

DOI - 10.1016/j.artmed.2020.101847

JOURNAL - Artificial Intelligence in Medicine

VOLUME 105

PAGES 1-11

PUBLISHER- Elsevier

PMID - 32505428

Full text link - https://www.sciencedirect.com/science/article/pii/S0933365719306384?via%3Dihub

Early temporal prediction of Type 2 Diabetes Risk Condition from a General Practitioner Electronic Health Record: A Multiple Instance Boosting Approach

Michele Bernardini[a], Micaela Morettini[a], Luca Romeo[a,b], Emanuele Frontoni[a,*], Laura Burattini[a]

[a]*Department of Information Engineering (DII), Università Politecnica delle Marche, Ancona, Italy*
[b]*Cognition, Motion and Neuroscience and Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genova, Italy*

Abstract

Early prediction of target patients at high risk of developing Type 2 diabetes (T2D) plays a significant role in preventing the onset of overt disease and its associated co-morbidities. Although fundamental in early phases of T2D natural history, insulin resistance is not usually quantified by General Practitioners (GPs). Triglyceride-glucose (TyG) index has been proven useful in clinical studies for quantifying insulin resistance and for the early identification of individuals at T2D risk but still not applied by GPs for diagnostic purposes. The aim of this study is to propose a multiple instance learning boosting algorithm (MIL-Boost) for creating a predictive model capable of early prediction of worsening insulin resistance (low vs high T2D risk) in terms of TyG index. The MIL-Boost is applied to past electronic health record patients' information stored by a single GP. The proposed MIL-Boost algorithm proved to be effective in dealing with this task, by overcoming the other state-of-the-art ML competitors (*Recall* from 0.70 and up to 0.83). The proposed MIL-based approach is able to extract hidden patterns from past EHR temporal data, even not directly exploiting triglycerides and glucose measurements. The major advantages of our method can be found in its ability to model the temporal evolution of longitudinal EHR data while dealing with small

[*]Corresponding author

Email addresses: `m.bernardini@pm.univpm.it` (Michele Bernardini), `m.morettini@univpm.it` (Micaela Morettini), `l.romeo@univpm.it` (Luca Romeo), `e.frontoni@univpm.it` (Emanuele Frontoni), `l.burattini@univpm.it` (Laura Burattini)

sample size and sparse observations (e.g., a small variable number of prescriptions for non-hospitalized patients). The proposed algorithm may represent the main core of a clinical decision support system.

Keywords: Type 2 Diabetes; Machine Learning; Predictive Medicine; Temporal Analysis; Electronic Health Record; Clinical Decision Support System.

1. Introduction

Type 2 Diabetes (T2D) is a chronic metabolic disorder characterized by high blood glucose concentration (i.e., hyperglycemia). T2D affects millions of people worldwide and predisposes to the development of severe cardiovascular and renal complications [1]. Early prediction of target patients at high risk of developing T2D plays a significant role in preventing the onset of overt disease and its associated comorbidities. Unfortunately, it is estimated that the first 10 years of T2D natural history - when the disorder is easiest to treat - are wasted [2].

The most powerful predictor of future development of T2D is represented by "insulin resistance", a reduced sensitivity of tissues to insulin action in lowering blood glucose concentration [3]. As insulin resistance worsens, more global defects in insulin secretion occur and, at the end, hyperglycemia arises [4]. Although fundamental in early phases of T2D natural history, insulin resistance is not usually quantified by General Practitioners (GPs) since specific blood tests - which are not included in those usually performed in routine examinations - as well as mathematical computations, are required [5].

A simple surrogate assessment of insulin resistance can be obtained through the triglyceride-glucose (TyG) index, based on routine triglyceride and glucose measurements [6, 7]. TyG index has been proven useful in clinical studies for the early identification of individuals at T2D risk and its predictive value was shown to be stronger than the one observed for triglyceride and glucose measurements taken singularly [8]. These findings highlight the usefulness of this index for the identification of individuals with early risk of developing T2D. However TyG index is still not applied by GPs for diagnostic purposes. In fact, this methodology may be ideally straightforward on an individual basis; however, scheduling an appointment for laboratory screening across a patient panel of thousands becomes challenging.

In this context, a Clinical Decision Support System (CDSS) predicting TyG changes over time may allow for better predictions of target groups with high risk of T2D. Such a CDSS may provide to GPs reminders for routine lab testing, recommendations for specific medication choices, and prescription of specialist examinations for a more accurate assessment of the metabolic status. Design of a CDSS is usually based on Electronic Health Record (EHR) systems, which are important tools in the daily GPs activities to store a considerable volume of data [9]. Moreover, Machine Learning (ML) techniques have been widely used for extracting information from such large amount of data. In particular, ML have been proven useful in set-up powerful predictive models for T2D [10], but still never focused on early temporal prediction of T2D risk (i.e. insulin resistance worsening prediction). One of the main challenge in this context is the modelling of the temporal evolution of EHR data. The Multiple instance learning (MIL) is one of the ML techniques that has been proven useful to accomplish this challenge, even though in a different domain [11, 12].

The aim of this study was to propose the core of a new CDSS based on a MIL boosting (i.e., MIL-Boost) algorithm. The proposed algorithm was applied to past EHR patient information stored by a single GP in order to create a predictive model capable of early prediction of worsening insulin resistance (low vs high T2D risk) in terms of TyG index.

2. Related work

In recent literature several approaches have been proposed to predict chronic pathologies onset from heterogeneous and longitudinal EHR data [13–19].

Usually, the most important requirement to perform this predictive task is the availability of a large amount of transversal (i.e., number of patients) and longitudinal (i.e., number of temporal observations of the same patient) data, which commonly come from hospitals or clinical research structures but are not always easily accessible or publicly available in the general practice scenario. The authors in [15] predicted multiple chronic diseases from longitudinal EHR data through an unsupervised Deep Learning (DL) model (e.g., deep neural network of stack of denoising autoencoders). However, this approach may suffer from a lack of interpretability because is not able to explicitly provide a top feature rank importance. On the contrary, other work proposed supervised techniques to predict chronic cardiovascular [16] and kidney diseases [17–19] by providing also model interpretability. The authors in [16] employed Logistic Regression (LR), Random Forest (RF), Gradient Boosting Trees (Boosting), Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models to predict 10-year cardiovascular disease events. In addition, the authors in [17] used also a temporal multi-task procedure to predict the short-term progression (i.e., 1 year) of estimated glomerular filtration rate (eGFR). They proposed a L2-regularized LR model to rank the predictors importance within each fixed past time-window (i.e., 6 months). Similarly, the authors in [18] determined the progression of kidney disease through the prediction of the future eGFR from 1 to 3 years by applying a RF regression model. The authors in [19] aimed to predict levels of albuminuria to evaluate renal function changes across a 5-year time window. Time-interval relations patterns were employed to discover the most relevant laboratory exams as predictive risk factors.

Focusing on T2D, in literature lots of work have already been proposed for classification [20–25] and/or prediction [10, 26, 27] tasks. Studies related to the classification task did not focus on predicting the temporal evolution of T2D condition across EHR longitudinal data. Differently, studies performing a prediction task employed standard ML models to predict the T2D diagnosis using past EHR observations divided in a fixed number of time windows. Moreover, although the authors in [10] used EHR data of GPs, the considered features space contains also glycaemic information. In order to handle limited longitudinal EHR data, the authors in [28] proposed a semi-supervised learning solution, that consists of a generative adversarial network coupled with a CNN to augment the training set data and improve the risk prediction performance, respectively. Their proposed model, also compared with LR, RF, LSTM, and Support Vector Machine (SVM) obtained the best predictive performance, but was not able to quantify the importance of the best predictors. Differently from all the above cited work [10, 26–28], our task aims to predict insulin-resistance as an early factor of T2D risk condition.

The limited amount and sparsity of longitudinal observations for each patient reflect the main challenges of our task. Because of these differences in the task definition, we have decided to perform the experimental comparison with respect to other state-of-the-art ML models (i.e., Decision Tree (DT) [10, 25–27]; RF [18, 26, 27]; KNN [26, 27]; Boosting [21]; SVM with linear kernel (SVM Lin) and SVM with Gaussian kernel (SVM Gauss) [10, 26, 27]; and SVM with Lasso regularizer (SVM Lasso) [23]), employed in literature to solve tasks closer to our setting. Similarly to [17], we compared these state-of-the art models according to time-invariant and temporal majority vote procedures.

3. Materials

3.1. Clinical Data: inclusion and exclusion criteria

The FIMMG dataset[1] has been collected from a single General Practitioner's Electronic Health Record which consists of 2433 patients. Our clinical data represent a subset of the FIMMG dataset with a longitudinal observational time-period up to 9 years according to the following criteria (see Fig. 1): i) exclusion of all diagnosed diabetic patients according to the International Classification of Disease 9th Revision (ICD-9) (since they can be farmacologically treated) ii) inclusion of only demographic, monitoring and laboratory exam fields (since continuous EHR features are collected more frequently over time); and iii) inclusion of patients with at least a single measurement of triglycerides (TG; mg/dl) and fasting glycemia (Gb; mg/dl) collected simultaneously.

For each *i-th* patient, a different number (t_i) of (TG$_j$, Gb$_j$) pairs measurements were collected, where j identified the temporal instance with $\{1, \ldots, j, \ldots, t_i\}$. Accordingly, the TyG$_j$ index was computed according to [8]:

$$\text{TyG}_j = \frac{\ln(TG_j \cdot Gb_j)}{2} \tag{1}$$

On the basis of the insulin resistance threshold of TyG (TyG$_{th}$ $= 8.65$) reported

[1] http://vrai.dii.univpm.it/content/fimmg-dataset

in [8], each observation can be classified as low ($\text{TyG}_j < \text{TyG}_{th}$) or high ($\text{TyG}_j \geq \text{TyG}_{th}$) risk. We let seq_{ij} be the d-dimensional EHR features vector of the j-th instance for the i-th patient. If a single EHR feature has multiple records between two TyG measurements its median value was taken into account. Missing values of monitoring and laboratory exams features were indicated as *NaN*.

3.2. Problem formulation

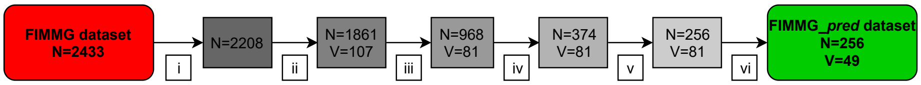


Figure 1: Inclusion and exclusion criteria (N identifies the number of EHR patients, and V the number of EHR features)

In order to better evaluate the temporal evolution of the patient's T2D risk condition, more strict inclusion criteria (see Fig. 1) were added to i), ii) and iii) as follows: iv) patients with at least 3 instances (for ensuring sufficient medical history to be investigated); v) patients with a temporal distance $\Delta_{(t_i-1)t_i}$ between the two last instances equal or greater than 12 months (to guarantee, also in agreement with GPs, a consistent and robust predictive temporal window [17]); and vi) EHR features that contain an overall amount of *NaN* less than a threshold of 90% ($th_{nan}= 90\%$). The rationale choice behind this threshold is the need of a predictive model in the clinical scenario that is consistent even with large proportions of missing data (up to 90%), as previously did in other studies [29, 30].

The proposed approach predicts the future TyG_{it_i} ($\hat{\text{TyG}}_i$) considering only the past instances (i.e., $\{seq_{i1}, \ldots, seq_{i(t_i-1)}\}$) (see Fig. 2).

Table 1 shows the final configuration of our clinical data, named FIMMG_*pred* dataset[2], after the application of all six inclusion/exclusion criteria to the original FIMMG dataset.

[2] http://vrai.dii.univpm.it/content/fimmgpred-dataset

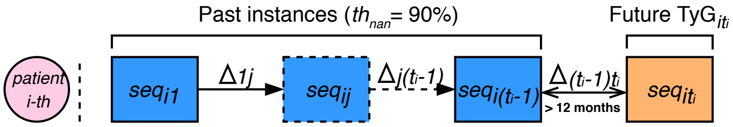


Figure 2: For each *i-th* patient, the temporal distance between past instances (i.e., Δ_{1j}, $\Delta_{j(t_i-1)}$) is variable, while between the last 2 instances (i.e., $\Delta_{(t_i-1)t_i}$) is at least equal or greater than 12 $months$.

Table 1: FIMMG_*pred* dataset comprised of a total of 256 patients and 681 past instances related to 49 EHR features. The detailed list of the 45 laboratory exams features is reported in Appendix A (see Tab. A.1). Both age and blood pressure measurements were computed at $seq_{i(t_i-1)}$ time point. In the case that blood pressure measurements were missing at $seq_{i(t_i-1)}$, the closest past time-window blood pressure was taken into account (i.e. zero-order hold interpolation).

Dataset description	**Count**	**Mean (std)**
Total patients	256	-
Controls (TyG<8.65)	179 (70%)	-
High-risk controls (TyG≥8.65)	77 (30%)	-
Observation period (*years*)	9	-
Past instances $\{seq_{i1}, \ldots, seq_{i(t_i-1)}\}$	681	-
Fields (EHR features)	**Count**	**Mean (std)**
Demographic	2	
Gender:		
Male	126 (49%)	-
Female	130 (51%)	-
Age (*years*)	-	68($\pm$14)
Monitoring	2	
Blood pressure ($mmHg$)		
Systolic	-	136($\pm$14)
Diastolic	-	82($\pm$7)
Laboratory exams	45	

4. Methods

4.1. Preprocessing

In order to retrieve information from *NaN* stored in FIMMG_*pred* dataset we exploited the *K-Nearest Neighbor (KNN) imputation*, consisting in replacing the *NaN* according to the KNN strategy [31]. The hyper-parameter K was set to 1 in order to preserve the initial data structure [31]. As already done in a similar context [32], the

K-Nearest Neighbor (KNN) imputation was selected as the best strategy after exploring other data imputation techniques (*extra values imputation*, *median imputation*).

4.2. Multiple instance learning boosting algorithm

MIL paradigm has attracted much attention in the last several years, and has been proven useful in various domains, including bioinformatics [33], text processing [34] and computer vision [35] and biomedical image analysis [36].

In the MIL paradigm the data is assumed to have some ambiguity in how the labels are associated. Differently from traditional supervised learning, labels are assigned to a set of inputs (bags) rather than providing input/label pairs. Thus, during the learning process, the classifier receives a set of *bags* along with the corresponding ground-truth (i.e., label). Each bag contains multiple instances. In this framework, the data is assumed to have some ambiguity in how the labels are associated: a bag is labeled positive if there is at least one positive instance [37]. Hence, the MIL task can be addressed to both estimate the instance and bag labels.

The MIL-Boost algorithm is originated from the work presented in [38] by starting with the standard multiple instance assumption [37] and the boosting algorithm [39]. The main idea behind the boosting algorithm is to sequentially train several weak classifiers $h_k \in H$ and combine them into a strong classifier $\mathbf{h}$ [37]. The combination is performed in an additive way by weighting each weak classifier h_k:

$$\mathbf{h} = \sum_{k=1}^{K} \alpha_k h_k(x) \tag{2}$$

where α_k are positive weights, K refers to the number of weak classifiers and x is the feature vector. The employed weak classifier is the logistic regression. The gradient boosting framework evolves the standard boosting formulation by considering each classifier h_k the best sequential approximation in the classifiers space *H* of the relative loss function based on a previous estimation [40, 41].

The general idea behind the application of MIL-Boost is to consider as instances the set of past observations (seq_{ij}) related to different patients (i.e., bags). In the MIL paradigm, the instance probabilities of the MIL-Boost algorithm are derived as follows:

$$p_{ij} = \sigma(\mathbf{h}(seq_{ij})) \tag{3}$$

where $\sigma(\cdot)$ is the logistic function $\frac{1}{1+exp(-(\cdot))}$. The instance probability is related to the bag probability as follows:

$$p_i = g_j(p_{ij}) \tag{4}$$

where $g(\cdot)$ is a function that approximates the max operator (i.e., noisy OR function). The loss function is the negative binomial log-likelihood. For each patient, the last TyG$_{it_i}$ measurement was assumed as the bag label (0 [negative bag] if the TyG$_{it_i} <$ 8.65, 1 [positive bag] if the TyG$_{it_i} \geq 8.65$) of the proposed MIL-Boost algorithm where the past instances $\{seq_{i1}, \ldots, seq_{i(t_i-1)}\}$ are the instance predictors (see Fig. 2 and Fig. 3a). The MIL-Boost algorithm (see Fig. 3a) groups the past instances into bags of instances. Thus, our task is to predict the bag label according to the estimated bag probability (p_i).

In the proposed MIL-based approach each bag is allowed to have different size (i.e. different number of instances $t_i - 1$), by taking into account the sparse sample size of longitudinal data (i.e. the laboratory exams for non-hospitalized patients are not prescribed on a regular basis over time).

Although we modeled the single bag as a set of multiple instances, we did not assume explicitly an ordinal and defined structure of the instance (e.g. by including the instance ordering number [ion] in the feature set).

4.3. Experimental procedure

We evaluated the performance of the MIL-Boost using a Tenfold Cross-Validation over subjects (CVOS-10) procedure[3] to measure the prediction of early T2D risk condition. All subjects were divided in ten folds and selecting alternately nine folds for

[3] The code to reproduce the experimental results is available at the following link: `https://github.com/michelebernardini/Early-temporal-prediction-of-type-2-diabetes-risk`

training and one fold for testing in order to generalize across unseen patients. This setup is closer to clinical diagnosis purposes, since the ML algorithm needs to generalize the decision rules, learnt from subjects who already have a diagnosis, across new unseen subjects.

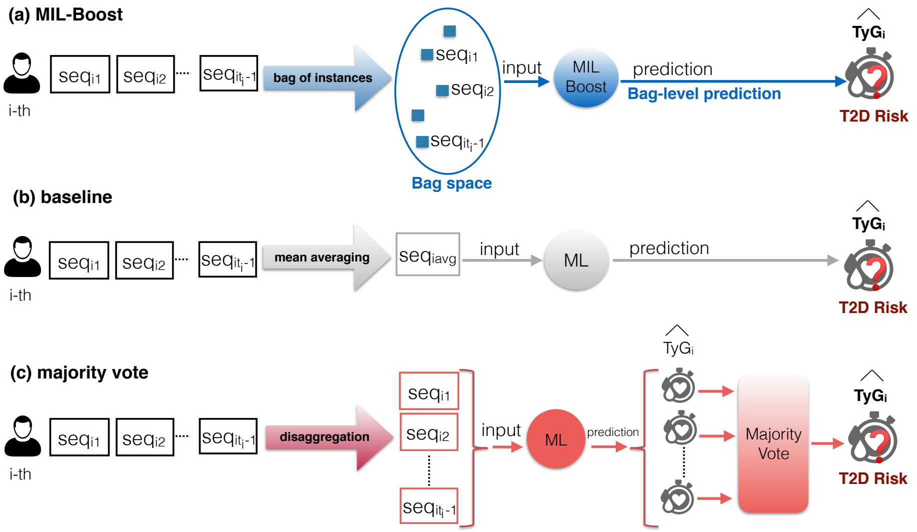


Figure 3: Overview of the experimental procedures: a) MIL-Boost, b) time-invariant baseline, and c) temporal majority vote.

The experimental procedure was evaluated by considering two different configurations: i) "yesTyG" where triglycerides and glycaemia were included as separate EHR predictors; ii) "noTyG" where triglycerides and glycaemia were not included.

4.3.1. Measures

The predictive performance was evaluated according to the following measures:

– *Accuracy*: the percentage of correct predictions;

– *Macro-precision*: the *Precision* is calculated for each class and then take the unweighted mean. The *Precision* reflects the percentage of true positive over the predicted condition positive;

– *Macro-recall*: the *Recall* is calculated for each class and then take the unweighted mean. The *Recall* reflects the percentage of true positive over the condition positive (sensitivity);

- *Macro-F1*: the harmonic mean of *precision* and *recall* averaged over all classes;
- *Area Under Receiver Operating Characteristic curve (AUC)*: represents the probability that the classifier will rank a randomly chosen positive sample higher than a randomly chosen negative one.

From now on we refer to the *Macro-precision*, *Macro-recall* and *Macro-F1* as *Precision*, *Recall* and *F1* respectively.

4.4. Experimental Comparisons

We decided to compare the MIL-Boost algorithm with respect to other ML algorithms employed in literature closer to our setting (see Sec. 2), such as: DT [10, 25–27]; RF [18, 26, 27]; KNN [26, 27]; Boosting [21]; SVM Lin and SVM Gauss [10, 26, 27]; and SVM Lasso [23]. These state-of-the-art approaches were also combined with the KNN imputation technique described in Sec. 4.1 to provide a fair comparison with the proposed MIL-Boost. Moreover, we compared these state-of-the art methods according to the approach proposed by [17] where a time-invariant and a temporal majority vote procedures were used. Further comparisons were performed with respect to other standard MIL-algorithms: MIL-DT [ID3-MI] [42], MIL-RF [MIForests] [43] and MIL-SVM [34] with linear and Gaussian kernel.

4.4.1. Time-invariant baseline

In the time-invariant baseline experimental procedure (see Fig. 3b) a single instance was computed for each bag/patient as the average of the past EHR features (seq_{iavg}). The Ty$\hat{\mathrm{G}}_{it_i}$ was predicted without taking into account the temporal evolution of the past clinical history.

4.4.2. Temporal majority vote

The temporal majority vote experimental procedure (see Fig. 3c) is able to combine the temporal information in the longitudinal data. A single instance learning ML model was trained by all the past instances seq_{ij} of the trained subjects for predicting the $\hat{\text{TyG}}_i$. Each past instance ($\{seq_{i1}, \ldots, seq_{i(t_i-1)}\}$) of the $i-th$ patient provides a total of $t_i - 1$ predictions of $\hat{\text{TyG}}_i$. The final output $\hat{\text{TyG}}_i$ was computed by computing the majority vote of each single prediction for each patient.

4.5. Validation procedure

Table 2 summarizes the range of the hyperparameters optimized for each ML model during the CVOS-10. The chosen hyperparameters were summarized according to how many times the value was chosen in the CVOS-10 models (count) for the noTyG procedure. In particular, the hyperparameters tuning was performed implementing a grid-search and optimizing the *Recall* in a nested CVOS-5. *Recall* was preferred over other optimization objectives, because minimising the false negative rate has more clinical relevance for a screening purpose. Hence, each split of the outer loop was trained with the optimal hyperparameters tuned in the inner loop. Although this procedure was computationally expensive, it allowed to obtain an unbiased and robust performance evaluation [44]. For all models the *Accuracy*, *F1*, *Precision* and *Recall* were computed by selecting the best threshold in the nested CVOS-5. The predicted bag label was assigned according to the best threshold and the model scores. This procedure aims to deal with the natural unbalanced setting of this task.

5. Results

Figure 4 shows the overall temporal distance between consecutive instances ($\Delta_{j(j+1)}$) per patient. It turns out that in our dataset in average laboratory exams are repeated for each patient at regular time intervals of almost 400 days.

Figure 5 shows the TyG_{it_i} index distribution for the final configuration of the FIMMG_*pred* dataset (see Tab. 1). The data follow a Normal distribution (according to a Kolmogorov Smirnov Test, $D = 0.041$, $p = 0.753$) with mean $\mu = 8.41$ and standard deviation $\sigma = 0.52$.

Figure 6 quantifies the *NaN* occurrences for each EHR feature (i.e., demographic, monitoring and laboratory exams) stored in the FIMMG_*pred* dataset.

Table 2: Range of Hyperparameters (Hyp) for each model: Decision Tree (DT), Regression Forest (RF), K-Nearest Neighbor (KNN), Gradient Boosting Trees (Boosting), Support Vector Machine with linear kernel (SVM Lin), Support Vector Machine with Gaussian kernel (SVM Gauss), Support Vector Machine with Lasso regularizer (SVM Lasso), and Multiple Instance Learning Boosting (MIL-Boost). The chosen hyperparameters were summarized according to how many times the value was chosen in the CVOS-10 models (count) for the noTyG procedure.

Model	**Hyp**	**Range(count)**
DT [10, 25–27]	max # of splits	$\{\mathbf{5}(3), 10(2), 15(2), 20(2), 25(0), 50(1)\}$
RF [18, 26, 27]	# of DT # of predictors to select	$\{\mathbf{5}(4), 10(2), 20(1), 30(2), 40(0), 50(1)\}$ $\{\frac{all}{4}(0), \frac{all}{3}(0), \frac{all}{2}(0), \mathbf{all}(10)\}$
KNN [26, 27]	max # of neighbors	$\{1(1), 3(2), 5(1), 7(1), 9(0), 11(0), \mathbf{13}(3), 15(2)\}$
Boosting [21]	max # of splits max # of weak classifiers	$\{1(0), 5(2), 10(2), \mathbf{20}(3), 30(1), 40(1), 50(1)\}$ $\{1(1), 5(0), 10(1), 20(0), \mathbf{30}(3), 40(2), \mathbf{50}(3)\}$
SVM Lin [10, 26, 27]	Box Constraint	$\{10^{-2}(0), 0.1(1), 1(1), 10(2), \mathbf{10^2}(5), 10^3(1)\}$
SVM Gauss [10, 26, 27]	Box Constraint Kernel Scale	$\{10^{-4}(0), 10^{-3}(1), \mathbf{10^{-2}}(9), 0.1, 1, 10, 10^2, 10^3\}$ $\{10^{-4}(0), 10^{-3}(2), \mathbf{10^{-2}}(8), 0.1, 1, 10, 10^2, 10^3\}$
SVM Lasso [45]	Lambda	$\{10^{-5}(0), \mathbf{10^{-4}}(5)10^{-3}(3), 10^{-2}(2), 0.1(0), 1(0), 10(0)\}$
MIL-DT	max # of splits	$\{5(2), 10(1), 15(2), 20(1), 25(1), \mathbf{50}(3)\}$
MIL-RF	# of DT # of predictors to select	$\{5, 10, 20(1), \mathbf{30}(5), 40(2), 50(2)\}$ $\{\frac{all}{4}(0), \frac{all}{3}(0), \frac{all}{2}(0), \mathbf{all}(10)\}$
MIL-SVM Lin	Box Constraint	$\{10^{-4}(0), 10^{-3}(0), 10^{-2}(0), \mathbf{0.1}(7), 1(3), 10(0)\}$
MIL-SVM Gauss	Box Constraint Kernel Scale	$\{10^{-5}(0), \mathbf{10^{-4}}(6), 10^{-3}(0), 10^{-2}(4), 0.1(0), 1(0)\}$ $\{10^{-5}(0), \mathbf{10^{-4}}(7), 10^{-3}(3), 10^{-2}(0), 0.1(0), 1(0)\}$
MIL-Boost	learning rate # of weak classifiers	$\{10^{-5}(0), \mathbf{10^{-4}}(3), 10^{-3}(1), 10^{-2}(0), 0.1, \mathbf{1}(3), 10(2), 10^2(1)\}$ $\{1(0), \mathbf{5}(10), 10(0), 15(0)\}$

5.1. *Predictive performance*

Table 3 shows the predictive performance of the MIL-Boost and the performed comparison. The comparison was carried out both with the different experimental procedures (i.e., time-invariant baseline, temporal majority vote) and with the different configurations (i.e., yesTyG, noTyG).

Figure 7 shows the performance comparison in terms of averaged *Recall* and standard deviation of Majority vote and MIL-algorithms over all CVOS-10 folds

The *Recall* obtained for both MIL configurations follows a Normal distribution according to the one-sample Kolmogorov-Smirnov test (yesTyG: D = 0.198, $p = 0.76$; noTyG: D = 0.161, $p = 0.92$).

Accordingly, the statistical comparisons in terms of *Recall* between the proposed approach and the other ML models for each configuration were performed by a two-sample t-test (significance level = 0.05). Results evidenced that MIL-Boost is statistically superior ($p < 0.05$) than baseline yesTyG: KNN, SVM Gauss; baseline noTyG:

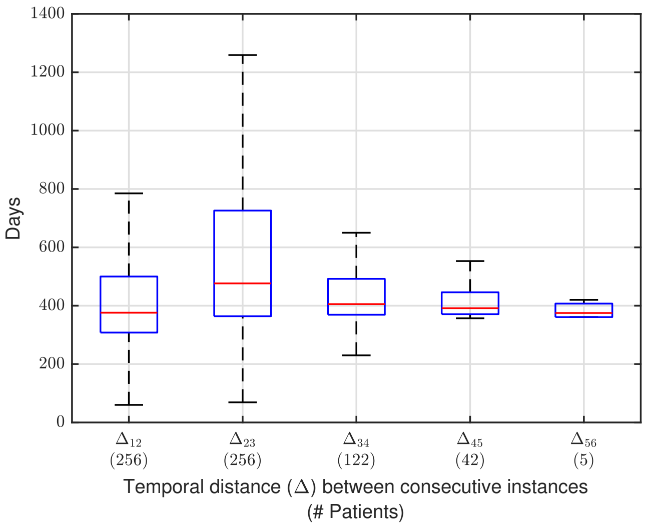


Figure 4: Overall temporal distance distribution between consecutive instances ($\Delta_{j(j+1)}$) per patient. The amount of patients is indicated below in round brackets.

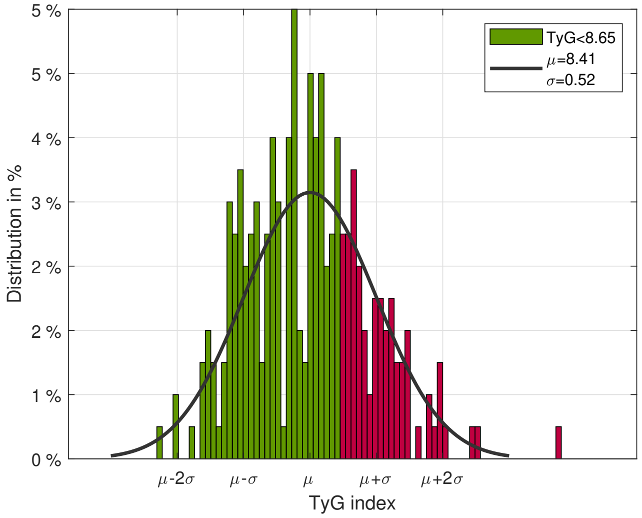


Figure 5: TyG$_{it_i}$ index distribution with mean $\mu = 8.41$ and standard deviation $\sigma = 0.52$. TyG index threshold (TyG$_{th}$ = 8.65) separates the green side (179 patients) from the red side (77 patients) of the graph.

DT, RF, Boosting, KNN, SVM lasso, SVM Gauss; majority vote yesTyG: KNN, SVM lin, SVM lasso, SVM Gauss; and majority vote noTyG: RF, KNN, SVM Lasso, SVM

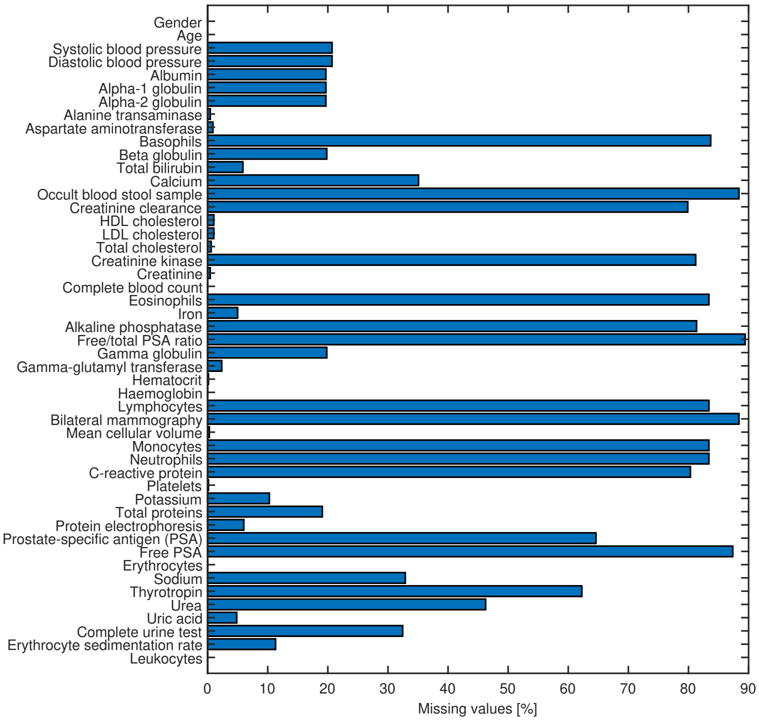


Figure 6: Percentage (%) of missing values (*NaN*) for each of the 49 EHR features: demographic, monitoring, and laboratory exams.

Gauss. Moreover MIL-Boost is statistically superior ($p < 0.05$) than noTyG: MIL-DT, MIL-RF and MIL-SVM Gauss and yesTyG: MIL-SVM lin and MIL-SVM Gauss.

5.2. Model interpretability

The top-10 rank features were listed in descending order of percentage importance for the temporal MIL-Boost experimental procedure in yesTyG configuration (see Fig. 8) and in noTyG configuration (see Fig. 9). The most discriminative predictors were extracted in according to the weights ω_K of the last updated weak logistic regressor h_K averaged over the 10 folds, where K is the maximum $\#$ of classifiers tuned during the validation stage. The percentage of the top-10 rank features was about 46.30 $\%$ and 40.91$\%$, respectively.

Table 3: Results of baseline, majority vote and MIL-Boost experimental procedures by considering (i.e., yesTyG) or not considering (i.e., noTyG) triglycerides and glucose information. Best results are evidenced in bold for both (i.e., yesTyG, noTyG) configurations. *Recall* is underlined because it is chosen as the hyperparameters optimization metric during the validation stage.

Baseline	*Accuracy*		*F1*		*Precision*		*Recall*		*AUC*	
	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG
DT	0.77	0.67	0.72	0.60	0.75	0.61	0.71	0.61	0.79	0.64
RF	0.77	0.68	0.72	0.57	0.74	0.61	0.72	0.58	0.84	0.66
Boosting	0.76	**0.70**	0.71	0.59	0.73	0.62	0.72	0.59	0.82	0.58
KNN	0.69	0.63	0.57	0.49	0.62	0.50	0.58	0.51	0.64	0.56
SVM lin	0.73	0.67	0.68	0.62	0.70	0.63	0.68	0.62	0.75	0.66
SVM lasso	0.77	0.65	0.70	0.57	0.76	0.60	0.70	0.57	0.80	0.63
SVM Gauss	0.70	**0.70**	0.41	0.41	0.35	0.35	0.50	0.50	0.50	0.50
Majority vote	*Accuracy*		*F1*		*Precision*		*Recall*		*AUC*	
	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG
DT	0.78	0.68	0.74	0.62	0.74	0.65	0.76	0.66	0.84	**0.74**
RF	0.77	0.65	0.73	0.57	0.73	0.60	0.75	0.59	0.83	0.69
Boosting	0.79	**0.70**	0.74	0.61	0.75	0.63	0.75	0.62	0.87	0.68
KNN	0.63	0.60	0.50	0.42	0.51	0.41	0.52	0.46	0.64	0.54
SVM lin	0.75	0.64	0.69	0.57	0.70	0.59	0.71	0.60	0.81	0.65
SVM lasso	0.77	0.66	0.69	0.57	0.71	0.59	0.70	0.59	0.81	0.66
SVM Gauss	0.63	0.66	0.38	0.39	0.31	0.33	0.50	0.50	0.46	0.50
MIL-algorithm	*Accuracy*		*F1*		*Precision*		*Recall*		*AUC*	
	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG	yesTyG	noTyG
MIL-Boost	0.83	**0.70**	0.81	**0.68**	0.82	**0.69**	0.83	**0.70**	0.89	**0.71**
MIL-DT	0.84	0.59	0.84	0.56	0.84	0.57	0.87	0.58	0.91	0.59
MIL-RF	**0.87**	0.63	**0.86**	0.60	**0.86**	0.60	**0.89**	0.61	**0.94**	0.64
MIL-SVM lin	0.67	0.72	0.40	0.67	0.34	**0.69**	0.50	0.68	0.47	0.52
MIL-SVM Gauss	0.67	0.67	0.40	0.40	0.34	0.34	0.50	0.50	0.51	0.49

Decision Tree (DT), Regression Forest (RF), K-Nearest Neighbor (KNN), Gradient Boosting Trees (Boosting), Support Vector Machine with linear kernel (SVM Lin), Support Vector Machine with Gaussian kernel (SVM Gauss), Support Vector Machine with Lasso regularizer (SVM Lasso), and Multiple Instance Learning Boosting (MIL-Boost)

5.3. Sensitivity to missing values

Figure 10 shows the trend of the MIL-Boost *Recall* as a function of the missing values threshold th_{nan} for both feature space configurations. For yesTyG configuration, the lower the th_{nan}, the more the *Recall* increases (up to almost 0.90), while for noTyG configuration, the maximum *Recall* (th_{nan}= 90%) does not increase by decreasing the th_{nan} and thus, it appears that *Recall* is not affected by the EHR features elimination. Standard deviation in noTyG configuration is globally greater than yesTyG one. A multiple comparison t-test confirms how there are not any significative differences ($p < .05$) across *NaN* thresholds.

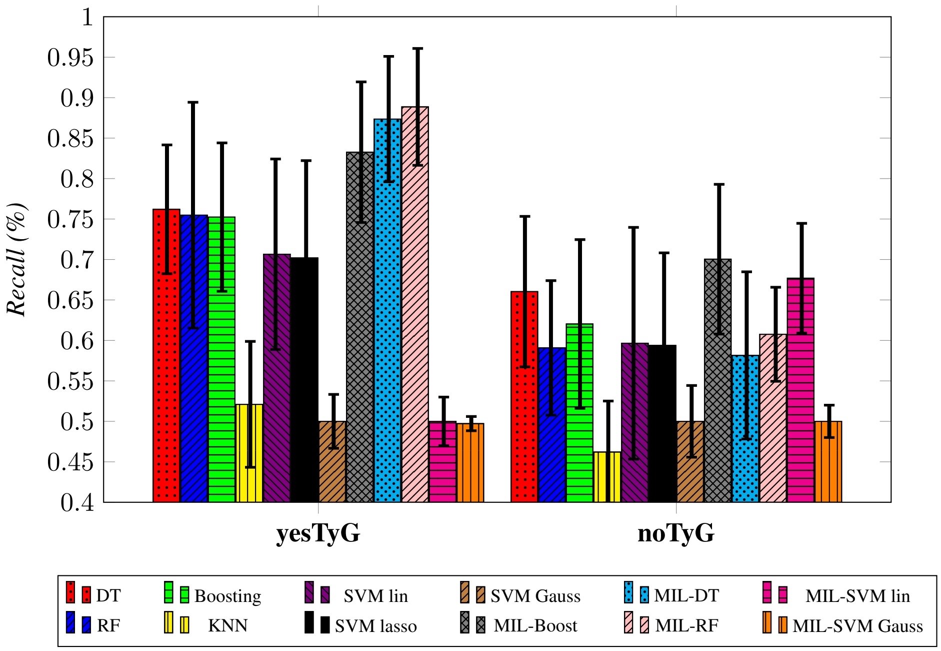


Figure 7: Averaged Recall and standard deviation of Majority vote and MIL-algorithms over all CVOS-10 procedure.

5.4. *Sensitivity to the sparsity of the data*

We have computed the Recall of MIL-Boost vs a measure of the sparsity of the data (see Figure 11). Since the sparsity can be due to a small variable number of exam prescriptions, the number of past instances $(t_i - 1)$ was selected as a measure of the sparsity in the data. The lower the number of past instances and the higher the sparsity in the data.

Although the performance decrease as the sparsity in the data increases, the *Recall* remains always over chance level (0.5).

6. Discussion

6.1. *Predictive performance*

This study proposed a model that captures temporal information for the early prediction of worsening insulin resistance (low vs high T2D risk) in terms of TyG index. The model learns from past routine measurements either including or excluding

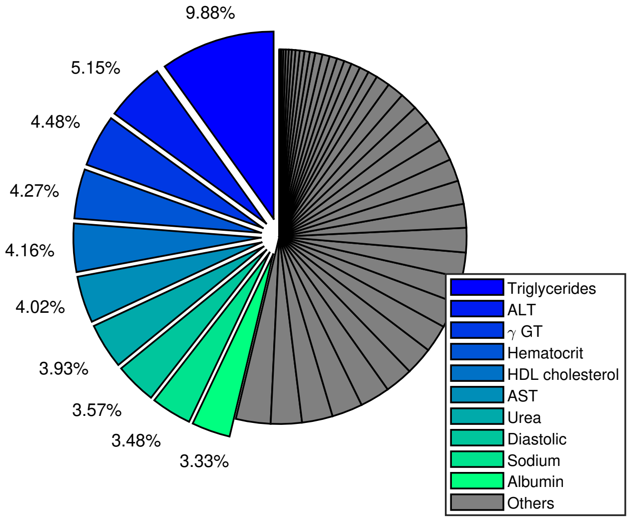


Figure 8: Top-10 rank features for MIL-Boost experimental procedure (yesTyG configuration). The percentage importance of the Others features was about 54%. Glycaemia ranks the 12^{nd} position with 2.68 %.

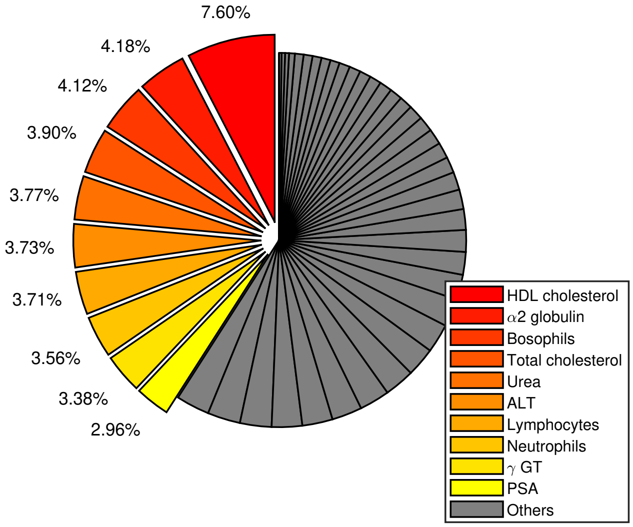


Figure 9: Top-10 rank features for MIL-Boost experimental procedure (noTyG configuration). The percentage importance of the Others features was about 59%.

triglycerides and glucose measurements, which are the ones used to compute TyG index. The proposed MIL-Boost algorithm proved to be effective in dealing with this

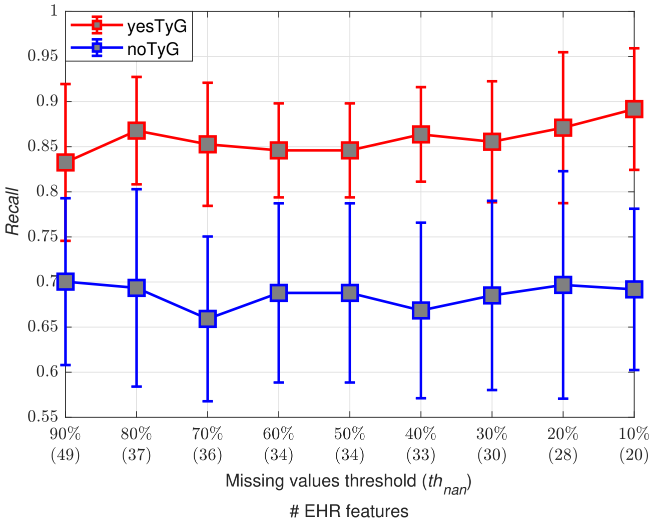


Figure 10: Trend of the MIL-Boost *Recall* and its standard deviation as a function of the missing values threshold th_{nan}. The amount of EHR features is indicated in round brackets for each th_{nan}.

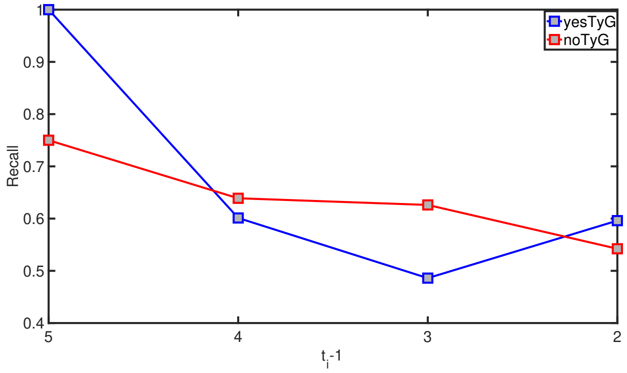


Figure 11: MIL-Boost Recall vs sparsity of the dataset in terms of the number of past instances $(t_i - 1)$.

task, by overcoming the other state-of-the-art ML models for both configurations (*Recall*, yesTyG: 0.83, noTyG: 0.70) and overcoming the other MIL-based approaches for noTyG configuration. In particular, the higher performance of MIL-Boost with respect to other MIL-based approaches in the more challenging clinical scenario (i.e. noTyG) highlights how the proposed approach is able to extract hidden patterns from past EHR temporal data, even not directly exploiting triglycerides and glucose measurements.

The TyG index, core element of this study, has been exploited by the same authors also in a previous work [32]. However, the present study is basically different with the previous one in terms of tasks (identification vs. forecast) and dataset considered (present observations/future observations) and thus, also for the adopted methodologies. The TyG-er approach [32] deals with the identification of the TyG index from routine data, i.e. the extraction of the most relevant non-glycemic (routinary) clinical factors strictly associated with the insulin-resistance condition; associations have been investigated by looking at clinical factors and TyG observed at the same time point. Results of [32] highlighted clinical factors having a well-known (as for example cholesterol), but also a non-trivial (as for example leukocytes and protein profile) association with insulin resistance. Knowledge of non-trivial associations provides hints for further investigation in clinical studies. On the other side, this work deals with the prediction of future (i.e. forecast) TyG worsening, starting from the knowledge of past values of routine clinical factors. Thus, the proposed MIL algortihm explored the relation among the sparse observations of the clinical factors in order to improve the prediction of TyG worsening and thus of the T2D risk. As desirable, clinical factors highlighted by the TyG-er approach [32] are also the more relevant ones for TyG worsening prediction.

The recent advances of DL and the huge amount of the data have laid the foundations to apply DL methodologies to EHR data for predictive tasks [46]. However, in most cases EHR data pertain to hospitalized subjects [46–48], thus being characterized by a huge set of longitudinal and more specific measurements. The major advantages of our method with respect to other approaches reported in literature [15–19] can be found in its ability to deal with a lower and a more sparse sample size of transversal and longitudinal data (e.g., a lower number of prescriptions for non-hospitalized patients). The proposed MIL-Boost algorithm may deal with the sparse nature of this setting, where different subjects (i.e., bags) may have a different number of observation (i.e., instances) over time (see Section 5.4). Our MIL-based approach relaxes the constraint imposed by some other work [10, 17, 26, 27] by modeling a variable number of observations for each patient (i.e. in the proposed MIL-based approach each bag is allowed to have different size). Additionally, we did not employ any preprocessing step (e.g., resampling strategies) to deal with the natural unbalance of this task.

Moreover, we did not find any statistical changes related to the inclusion of the temporal information (i.e. instance ordering number [ion]) in the model for the yesTyG ($p = 0.793$) and noTyG ($p = 0.375$) configuration. Results evidenced that MIL-Boost *Recall* of the noTyG configuration is slightly higher (0.70 vs 0.68) if the ion is not included in the feature set. On the other hand the MIL-Boost *Recall* of the yesTyG configuration is slightly lower if the ion is not included in the feature set. Although the experimental results might suggest that the temporal ordering of the exams is not relevant for predicting the early T2D risk condition, future work might be addressed to model the temporal evolution of the instance inside the bag by imposing a sequential constraint (e.g. by applying a laplacian regularizer which encourages the temporal smoothness between two exams).

Additionally, concerning the sensitivity to missing values, we found that the proposed model is affected only by the yesTyG configuration, because the progressive EHR feature elimination gives more importance to triglycerides and glycaemia as discriminant predictors; while for noTyG configuration the *Reecall* trend appears more stable. However, t-test confirms how there are not any significative differences ($p < .05$) across *NaN* thresholds. This fact implies that features with many missing values are not discriminative, and thus, suggests how the distribution of missing values is the same for all the two classes (i.e. the missing values mechanism is not informative about the classification target [49]).

6.2. Clinical significance

MIL-Boost predicts the deterioration of TyG index, whose efficacy in discriminating subjects at low and high T2D risk has been recently recognized in clinical settings [8, 50]. Our approach could lay the foundations for a CDSS having an important impact from the therapeutic point of view. Besides planning targeted screening, such a CDSS may allow pharmacological and non-pharmacological interventions administration by GPs in an early pathophysiological T2D state, thus when they are more effective. Non-pharmacological interventions may include timely promotion by GPs of a healthy diet and/or regular physical activity, which have been shown to modify early T2D mechanisms correlated to insulin resistance [51, 52].

The model interpretability results of our study provided novel insight into the best combination of conventionally used (HDL, ALT, γGT) and non-conventionally used (urea, $\alpha 2$ globulin, bosophils, lymphocytes, neutrophils) biomarkers for diagnosing early T2D risk condition. Evidence in recent literature can be found to support our model interpretability results [32, 53]. Glycaemia appears redundant (12^{nd} rank) in case of presence of triglycerides and other clinical factors in the yesTyG configuration. Thus, it turned out that triglycerides are more relevant than glycaemia in order to predict the future TyG status of the patient. Additionally, regarding the complementary set of features, ALT, gamma-GT, HDL cholesterol, and urea keep remaining within the top-10 rank features. Notice that glycated haemoglobin (HbA1c), an important clinical factor used for T2D diagnosis and monitoring, was not included in our analysis. However, HbA1c is not included in routinely examinations since GPs usually prescribe HbA1c assessment when T2D is strongly suspected or already diagnosed. In our dataset (which does not consider already-diagnosed T2D patients), HbA1c was measured in less that 10% of the cases and it has been discarded according to the exclusion criteria vi) (i.e., EHR features that contain an overall amount of *NaN* less than a threshold of 90%)

6.3. Future work

Starting from the knowledge of the best features, the higher interpretability of our approach may favor the acceptance of the experimental findings by the medical community and allow an easier implementation of a CDSS. The proposed MIL-Boost approach performed on the FIMMG_*pred* dataset, collected by the same GP, could be also extended and applied to other EHRs stored by multiple GPs. In fact, the computational-time efficiency of our algorithm allows to easily re-train the model over new EHR data (see Fig. B.1a). Such competitiveness in terms of computational efficiency (see Fig. B.1) allows the proposed algorithm to be embedded also in a cross-platform framework for low-cost mobile devices. Since missing values represent one of the main problems of this kind of data, future work should also try to investigate the effect of more advanced strategies (e.g., collaborative filtering, matrix factorization, etc.).

Of note, the methodology proposed in the present study is not meant to replace current diagnostic T2D methodology, which will be applied in the case of TyG classified as "high". Our aim was to provide a support to screen patients at risk for T2D at the very beginning and a classification setup may be effective. Of course, a continuous prediction of TyG changes over time resulting from a regression setup is desirable and will be explored in future studies.

The final application of the proposed approach will be the integration of the MIL-Boost on the FIMMG Nu.Sa. cloud platform [54] to achieve a real-world application of a data driven CDSS. The actual FIMMG Nu.Sa. suite has more than 20 statistical or ML based applications to support GPs in their daily activities and, the proposed new approach will be the first based on a predictive and high-interpretable ML model able to capture EHR temporal data.

7. Conclusions

As demonstrated by the high predictive performances, the model interpretability, and the capability to deal with missing values and sparsity of data, the proposed MIL-Boost is a reliable approach for the early prediction of T2D risk condition (low vs high T2D risk) using past EHR temporal information collected from a single GP. The proposed algorithm may represent the main core of a CDSS.

8. Acknowledgments

This work was supported by the strategic project of the Università Politecnica delle Marche, entitled “MIRAGE - Multidisciplinary Innovative Research actions on AGE” (Project n. I32F17000080005). The founding source has no role in the conduct of the research and in the preparation of the article. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflict of interest statement

The authors declare that they have no competing interests.

References

References

[1] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes, A. Ohlrogge, B. Malanda, IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045, Diabetes Research and Clinical Practice 138 (2018) 271 – 281.

[2] M. Y. Bertram, T. Vos, Quantifying the duration of pre-diabetes, Australian and New Zealand journal of public health 34 (3) (2010) 311–314.

[3] R. Taylor, Insulin resistance and type 2 diabetes, Diabetes 61 (4) (2012) 778–779.

[4] M. E. Cerf, Beta cell dysfunction and insulin resistance, Frontiers in endocrinology 4 (2013) 37.

[5] B. Antuna-Puente, E. Disse, R. Rabasa-Lhoret, M. Laville, J. Capeau, J.-P. Bastard, How can we measure insulin sensitivity/resistance?, Diabetes & Metabolism 1849 (3) (2011) 169–264.

[6] L. E. Simental-Mendía, M. Rodríguez-Morán, F. Guerrero-Romero, The Product of Fasting Glucose and Triglycerides As Surrogate for Identifying Insulin Resistance in Apparently Healthy Subjects, Metabolic Syndrome and Related Disorders 6 (4) (2008) 299–304.

[7] F. Guerrero-Romero, L. E. Simental-Mendía, M. González-Ortiz, E. Martínez-Abundis, M. G. Ramos-Zavala, S. O. Hernández-González, O. Jacques-Camarena, M. Rodríguez-Morán, The Product of Triglycerides and Glucose, a Simple Measure of Insulin Sensitivity. Comparison with the Euglycemic-Hyperinsulinemic Clamp, The Journal of Clinical Endocrinology & Metabolism 95 (7) (2010) 3347–3351.

[8] D. Navarro-González, L. Sánchez-Íñigo, J. Pastrana-Delgado, A. Fernández-Montero, J. A. Martinez, Triglyceride–glucose index (TyG index) in comparison with fasting plasma glucose improved diabetes prediction in patients with normal fasting glucose: The vascular-metabolic CUN cohort, Preventive Medicine 86 (2016) 99 – 105.

[9] A. E. Anderson, W. T. Kerr, A. Thames, T. Li, J. Xiao, M. S. Cohen, Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general united states population: A cross-sectional, unselected, retrospective study, Journal of Biomedical Informatics 60 (2016) 162 – 168.

[10] A. Talaei-Khoei, J. M. Wilson, Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables, International Journal of Medical Informatics 119 (2018) 22 – 38.

[11] K. Sikka, A. Dhall, M. Bartlett, Weakly supervised pain localization using multiple instance learning, in: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–8.

[12] K. Sikka, A. Dhall, M. S. Bartlett, Classification and weakly supervised pain localization using multiple segment representation, Image and Vision Computing 32 (10) (2014) 659–670.

[13] A. N. Richter, T. M. Khoshgoftaar, A review of statistical and machine learning methods for modeling cancer risk using structured clinical data, Artificial Intelligence in Medicine 90 (2018) 1 – 14.

[14] J. Guo, X. Yuan, X. Zheng, P. Xu, Y. Xiao, B. Liu, Diagnosis labeling with disease-specific characteristics mining, Artificial Intelligence in Medicine 90 (2018) 25 – 33.

[15] R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Scientific Reports 6 (2016) 26094.

[16] J. Zhao, Q. Feng, P. Wu, R. A. Lupu, R. A. Wilke, Q. S. Wells, J. C. Denny, W.-Q. Wei, Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction, Scientific Reports 9 (1) (2019) 717.

[17] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, J. V. Guttag, Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration, Journal of Biomedical Informatics 53 (2015) 220 – 228.

[18] J. Zhao, S. Gu, A. McDermaid, Predicting outcomes of chronic kidney disease from emr data based on random forest regression, Mathematical Biosciences 310 (2019) 24–30.

[19] A. Shknevsky, Y. Shahar, R. Moskovitch, Consistent discovery of frequent interval-based temporal patterns in chronic patients' data, Journal of Biomedical Informatics 75 (2017) 83–95.

[20] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, Y. Chen, A machine learning-based framework to identify type 2 diabetes through electronic health records, International Journal of Medical Informatics 97 (Supplement C) (2017) 120–127.

[21] S. Perveen, M. Shahbaz, A. Guergachi, K. Keshavjee, Performance analysis of data mining classification techniques to predict diabetes, Procedia Computer Science 82 (2016) 115–121.

[22] W. Yu, T. Liu, R. Valdez, M. Gwinn, M. J. Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes, Medical Informatics and Decision Making 10 (1) (2010) 16.

[23] M. Bernardini, L. Romeo, P. Misericordia, E. Frontoni, Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine, IEEE Journal of Biomedical and Health Informatics (2019) 1–1doi:10.1109/JBHI.2019.2899218.

[24] M. F. Faruque, I. H. Sarker, et al., Performance analysis of machine learning techniques to predict diabetes mellitus, in: International Conference on Electrical, Computer and Communication Engineering, IEEE, 2019, pp. 1–4.

[25] A. J. Hall, A. Hussain, M. G. Shaikh, Predicting insulin resistance in children using a machine-learning-based clinical decision support system, in: C.-L. Liu, A. Hussain, B. Luo, K. C. Tan, Y. Zeng, Z. Zhang (Eds.), Advances in Brain Inspired Cognitive Systems, Springer International Publishing, Cham, 2016, pp. 274–283.

[26] S. Mani, Y. Chen, T. Elasy, W. Clayton, J. Denny, Type 2 diabetes risk forecasting from emr data using machine learning, in: AMIA annual symposium proceedings, Vol. 2012, American Medical Informatics Association, 2012, p. 606.

[27] A. Pimentel, A. V. Carreiro, R. T. Ribeiro, H. Gamboa, Screening diabetes mellitus 2 based on electronic health records using temporal features, Health Informatics Journal 24 (2) (2018) 194–205.

[28] Z. Che, Y. Cheng, S. Zhai, Z. Sun, Y. Liu, Boosting deep learning risk prediction with generative adversarial networks for electronic health records, in: International Conference on Data Mining, IEEE, 2017, pp. 787–792.

[29] P. Madley-Dowd, R. Hughes, K. Tilling, J. Heron, The proportion of missing data should not be used to guide decisions on multiple imputation, Journal of Clinical Epidemiology 110 (2019) 63 – 73.

[30] E. W. Steyerberg, Missing values, in: Clinical Prediction Models, Springer, 2019, pp. 127–155.

[31] L. Beretta, A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, Medical Informatics and Decision Making 16 (3) (2016) 74.

[32] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, L. Burattini, Tyg-er: An ensemble regression forest approach for identification of clinical factors related to insulin resistance condition using electronic health records, Computers in Biology and Medicine 112 (2019) 103358.

[33] T. G. Dietterich, R. H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial intelligence 89 (1) (1997) 31–71.

[34] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in neural information processing systems, 2003, pp. 577–584.

[35] B. Babenko, M.-H. Yang, S. Belongie, Robust object tracking with online multiple instance learning, IEEE transactions on pattern analysis and machine intelligence 33 (8) (2011) 1619–1632.

[36] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, D. Rueckert, A. D. N. Initiative, et al., Multiple instance learning for classification of dementia in brain mri, Medical image analysis 18 (5) (2014) 808–818.

[37] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, S. Vluymans, Multiple instance learning: foundations and algorithms, Springer, 2016.

[38] C. Zhang, J. C. Platt, P. A. Viola, Multiple instance boosting for object detection, in: Advances in neural information processing systems, 2006, pp. 1417–1424.

[39] R. E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Machine learning 37 (3) (1999) 297–336.

[40] J. Friedman, T. Hastie, R. Tibshirani, et al., Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), The annals of statistics 28 (2) (2000) 337–407.

[41] L. Mason, J. Baxter, P. L. Bartlett, M. R. Frean, Boosting algorithms as gradient descent, in: Advances in neural information processing systems, 2000, pp. 512–518.

[42] Y. Chevaleyre, J.-D. Zucker, Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. application to the mutagenesis

problem, in: Conference of the Canadian Society for Computational Studies of Intelligence, Springer, 2001, pp. 204–214.

[43] C. Leistner, A. Saffari, H. Bischof, Miforests: Multiple-instance learning with randomized trees, in: European Conference on Computer Vision, Springer, 2010, pp. 29–42.

[44] G. C. Cawley, N. L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research 11 (Jul) (2010) 2079–2107.

[45] J. Zhu, S. Rosset, R. Tibshirani, T. J. Hastie, 1-norm support vector machines, in: Advances in neural information processing systems, 2004, pp. 49–56.

[46] B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis, Journal of Biomedical and Health Informatics 22 (5) (2018) 1589–1604.

[47] L. Rasmy, Y. Wu, N. Wang, X. Geng, W. J. Zheng, F. Wang, H. Wu, H. Xu, D. Zhi, A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set, Journal of Biomedical Informatics 84 (2018) 11 – 16.

[48] T. Pham, T. Tran, D. Phung, S. Venkatesh, Predicting healthcare trajectories from medical records: A deep learning approach, Journal of Biomedical Informatics 69 (2017) 218 – 229.

[49] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, Pattern classification with missing data: a review, Neural Computing and Applications 19 (2) (2010) 263–282.

[50] S. Low, K. C. J. Khoo, B. Irwan, C. F. Sum, T. Subramaniam, S. C. Lim, T. K. M. Wong, The role of triglyceride glucose index in development of type 2 diabetes mellitus, Diabetes Research and Clinical Practice 143 (2018) 43 – 49.

[51] V. H. Telle-Hansen, K. B. Holven, S. M. Ulven, Impact of a healthy dietary pattern on gut microbiota and systemic inflammation in humans, Nutrients 10 (11).

[52] M. Morettini, F. Storm, M. Sacchetti, A. Cappozzo, C. Mazzà, Effects of walking on low-grade inflammation and their implications for type 2 diabetes, Preventive Medicine Reports 2 (2015) 538 – 547.

[53] A. Abbasi, A.-S. Sahlqvist, L. Lotta, J. M. Brosnan, P. Vollenweider, P. Giabbanelli, D. J. Nunez, D. Waterworth, R. A. Scott, C. Langenberg, et al., A systematic review of biomarkers and risk of incident type 2 diabetes: an overview of epidemiological, prediction and aetiological research literature, PloS one 11 (10) (2016) e0163721.

[54] E. Frontoni, A. Mancini, M. Baldi, M. Paolanti, S. Moccia, P. Zingaretti, V. Landro, P. Misericordia, Sharing health data among general practitioners: The nu.sa. project, International Journal of Medical Informatics 129 (2019) 267 – 274.

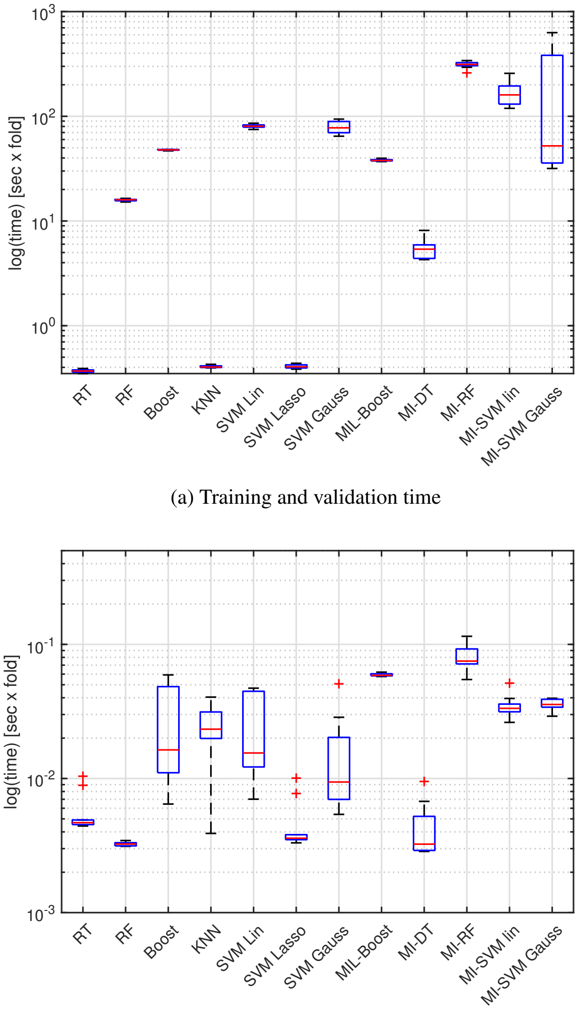
Appendix A. Full list of the laboratory exams

Table A.1: Detailed list of the 45 laboratory exams evaluated for this study.

#	Laboratory exams	#	Laboratory exams
1	Albumin	24	Hematocrit (HCT)
2	Alpha-1 globulin (α1 globulin)	25	Haemoglobin (HGB)
3	Alpha-2 globulin (α2 globulin)	26	Lymphocytes
4	Alanine transaminase (ALT)	27	Bilateral mammography
5	Aspartate aminotransferase (AST)	28	Mean cellular volume (MCV)
6	Basophils	29	Monocytes
7	Beta globulin (β globulin)	30	Neutrophils
8	Total bilirubin	31	C-reactive protein (CRP)
9	Calcium (Ca)	32	Platelets (PLT)
10	Occult blood stool sample	33	Potassium (K)
11	Creatinine clearance (Cockroft)	34	Total proteins
12	HDL Cholesterol	35	Protein electrophoresis
13	LDL Cholesterol	36	Prostate-specific antigen (PSA)
14	Total Cholesterol	37	Free prostate-specific antigen (free PSA)
15	Creatinine kinase (CK)	38	Erythrocytes (RBC)
16	Creatinine	39	Sodium (Na)
17	Complete blood count (CBC)	40	Thyrotropin (TSH)
18	Eosinophils	41	Urea
19	Iron (Fe)	42	Uric acid
20	Alkaline phosphatase (ALP)	43	Complete urine test
21	Free/total prostate-specific antigen ratio (free/total PSA ratio)	44	Erythrocyte sedimentation rate (ESR)
22	Gamma globulin (γ globulin)	45	Leukocytes (WBC)
23	Gamma-glutamyl transferase (γGT)		

Appendix B. Computation-time analysis

The computation time analysis for the training and validation stage is shown in Figure B.1a, while for the testing stage in Figure B.1b.



(a) Training and validation time

(b) Testing time

Figure B.1: Comparison in terms of computation time.