Machine Learning approaches in Predictive Medicine using Electronic Health Records data

Michele Bernardini

UNIVERSITÀ POLITECNICA DELLE MARCHE
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA
CURRICULUM IN INGEGNERIA BIOMEDICA, ELETTRONICA E DELLE
TELECOMUNICAZIONI

Machine Learning approaches in Predictive Medicine using Electronic Health Records data

Ph.D. Dissertation of:
Michele Bernardini

Advisor:
Prof. Emanuele Frontoni

Coadvisor:
Dr. Luca Romeo

Curriculum Supervisor:
Prof. Franco Chiaraluce

XIX edition - new series

UNIVERSITÀ POLITECNICA DELLE MARCHE
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA
CURRICULUM IN INGEGNERIA BIOMEDICA, ELETTRONICA E DELLE
TELECOMUNICAZIONI

Machine Learning approaches in Predictive Medicine using Electronic Health Records data

Ph.D. Dissertation of:
Michele Bernardini

Advisor:
Prof. Emanuele Frontoni

Coadvisor:
Dr. Luca Romeo

Curriculum Supervisor:
Prof. Franco Chiaraluce

XIX edition - new series

UNIVERSITÀ POLITECNICA DELLE MARCHE
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA
FACOLTÀ DI INGEGNERIA
Via Brecce Bianche – 60131 Ancona (AN), Italy

Abstract

Traditional approaches in medicine to manage diseases can be briefly reduced to the "one-size-fits all" concept (i.e., the effect of treatment reflects the whole sample). On the contrary, precision medicine may represent the extension and the evolution of traditional medicine because is mainly preventive and proactive rather than reactive. This evolution may lead to a predictive, personalized, preventive, participatory, and psycho-cognitive healthcare. Among all these characteristics, the predictive medicine (PM), used to forecast disease onset, diagnosis, and prognosis, is the one this thesis emphasizes. Thus, it is possible to introduce a new emerging healthcare area, named predictive precision medicine (PPM), which may benefit from a huge amount of medical information stored in Electronic Health Records (EHRs) and Machine Learning (ML) techniques. The thesis ecosystem, which consists of the previous 3 inter-connected key points (i.e., PPM, EHR, ML), contributes to the biomedical and health informatics by proposing meaningful ML methodologies to face and overcome the state-of-the-art challenges, that emerge from real-world EHR datasets, such as high-dimensional & heterogeneous data; unbalanced setting; sparse labeling; temporal ambiguity; interpretability/explainability; and generalization capability. The following ML methodologies designed from specific clinical objectives in PM scenario are suitable to constitute the main core of any novel clinical Decision Support Systems [1] usable by physicians for prevention, screening, diagnosis, and treatment purposes: i) a sparse-balanced Support Vector Machine (SB-SVM) approach [2] aimed to discover type 2 diabetes (T2D) using features extracted from a novel EHR dataset of a general practitioner (GP); ii) a high-interpretable ensemble Regression Forest (TyG-er) approach [3] aimed to identify non-trivial clinical factors in EHR data to determine where the insulin-resistance condition is encoded; iii) a Multiple Instance Learning boosting (MIL-Boost) approach [4] applied to EHR data aimed to early predict an insulin resistance worsening (low vs high T2D risk) in terms of TyG index; iv) a novel Semi-Supervised Multi-task Learning (SS-MTL) approach [5] aimed to predict short-term kidney disease evolution (i.e., patient's risk profile) on multiple GPs' EHR data; v) A XGBoosting (XGBoost) approach [6] aimed to predict the sequential organ failure assessment score (SOFA) score at day 5, by utilising only EHR data at the admission day in the Intensive Care Unit (ICU). The SOFA score describes the COVID-19 patient's complications in ICU and helps clinicians to create COVID-19 patients' risk profiles. The thesis also contributed to the publication of novel publicly available EHR datasets (i.e., FIMMG dataset, FIMMG_*obs* dataset, FIMMG_*pred* dataset, *m*FIMMG dataset).

Sommario

L' approccio tradizionale in medicina per gestire le patologie può essere ridotto al concetto di "one-size-fits all", in cui l'effetto di una cura rispecchia l'intero campione. Però, la medicina di precisione può rappresentare l'estensione e l'evoluzione della medicina tradizionale perché risulta principalmente preventiva e proattiva piuttosto che prettamente reattiva. Questa evoluzione può portare a una Sanità predittiva, personalizzata, preventiva, partecipativa e psicocognitiva. Tra tutte queste caratteristiche, la tesi si focalizza sulla medicina predittiva. Quindi, si può introdurre un nuovo emergente paradigma di Sanità, chiamato medicina di precisione predittiva (PPM), che può beneficiare da tecniche di Machine Learning (ML) e da una enorme quantità di informazioni racchiuse nelle cartelle cliniche elettroniche (EHRs). L'ecosistema sanitario della tesi, costituito dai 3 punti chiave interconnessi (PPM, EHR, ML), offre un contributo al campo dell'informatica medica proponendo metodologie di ML con lo scopo di affrontare e superare le sfide dello stato dell'arte che emergono dagli EHR dataset, come: dati eterogenei e molto numerosi, sbilanciamento tra classi, labeling sparso, ambiguità temporale, interpretabilità, capacità di generalizzazione. Le seguenti metodologie di ML sviluppate per specifici task clinici nello scenario della PM sono adatte a costituire il nucleo principale di nuovi sistemi clinici di supporto alle decisioni, utilizzabili dai medici per scopi di prevenzione, screening, diagnosi e follow-up: i) un approccio sparse-balanced Support Vector Machine con lo scopo di predire il diabete di tipo 2 (T2D), utilizzando le informazioni estratte da un nuovo EHR dataset di un medico di medicina generale; ii) un approccio Regression Forest ensemble ad alta interpretabilità con lo scopo di identificare fattori clinici non di routine nei dati EHR per determinare dove sia racchiusa la condizione di insulino-resistenza; iii) un approccio di Multiple Instance Learning boosting applicato ai dati EHR volto a predire precocemente un peggioramento dell'insulino-resistenza (basso vs alto rischio di T2D) in termini di TyG index; iv) un nuovo approccio multitasking semi-supervisionato con lo scopo di predire l'evoluzione a breve termine della patologie renale (cioè il profilo di rischio del paziente) sui dati EHR di un cluster di medici di medicina generale; v) un approccio XGBoosting con lo scopo di predire il SOFA score al quinto giorno, utilizzando solo i dati EHR del giorno di ammissione in unità di terapia intensiva (ICU). Il SOFA score descrive le complicazioni del paziente COVID-19 in ICU e aiuta i medici a creare profili di rischio dei pazienti COVID-19. La tesi ha anche contribuito alla pubblicazione di nuovi EHR datasets open access (FIMMG dataset, FIMMG_*obs* dataset, FIMMG_*pred* dataset, *m*FIMMG dataset).

Acronyms

AI Artificial Intelligence

ML Machine Learning

DL Deep Learning

HIT Health Information Technology

EHR Electronic Health Record

PM Predictive Medicine

PPM Predictive Precision Medicine

CDSS Clinical Decision Support System

WHO World Health Organization

IDF International Diabetes Federation

NHS National Health System

ICD-9 International Classification of Diseases

GP General Practitioner

FIMMG Italian Federation of General Practitioners

NMI Netmedica Italia

T2D Type 2 Diabetes

IR Insulin Resistance

KD Kidney Disease

CKD Chronic Kidney Disease

eGFR estimated Glomerular Filtration Rate

SOFA Sequential Organ Failure Assessment

BMI Body Mass Index

AP Arterial Pressure

COVID-19 Coronavirus Disease 2019

ICU Intensive Care Unit

CVD Cardiovascular Disease

KNN K-Nearest Neighbors

NB Naive Bayes

LR Logistic Regression

DT Decision Tree

RF Random Forest

Boost Boosting

XGBoost eXtreme Gradient Boosting

SVM Support Vector Machine

MLP Multi Layer Perceptron

CNN Convolutional Neural Network

LSTM Long-Short Term Memory

GAN Generative Adversarial Network

MIL Multiple Instance Learning

MTL Multi-Task Learning

SSL Semi-Supervised Learning

SLA Self-Learning Algorithm

SMOTE Synthetic Minority Over-sampling Technique

LASSO Least Absolute Shrinkage and Selection Operator

CV Cross Validation

MAE Mean Absolute Error

MSE Mean Squared Error

ROC Receiver Operating Characteristic

AUC Area Under Curve

Contents

1. Background and motivation **1**
1.1. Predictive medicine and precision medicine . . . 3
1.2. Electronic health records . . . 4
1.3. Machine Learning for healthcare . . . 5
1.4. Thesis: Problems statement . . . 6
1.4.1. Thesis contribution . . . 8
1.5. Thesis overview . . . 9
1.6. Thesis outcomes: Publications . . . 10

2. Type 2 diabetes discovering **13**
2.1. Introduction . . . 13
2.2. Related work . . . 15
2.2.1. Sparse SVM for unbalanced dataset . . . 16
2.3. Clinical data: FIMMG dataset . . . 17
2.3.1. Data analysis . . . 18
2.4. Methods . . . 20
2.4.1. Background: 2-norm SVM . . . 20
2.4.2. Sparse 1-norm SVM . . . 21
2.4.3. Sparse Balanced SVM . . . 22
2.4.4. Experimental procedure . . . 23
2.5. Experimental results . . . 26
2.5.1. Case I . . . 26
2.5.2. Case II . . . 28
2.5.3. Case III . . . 29
2.5.4. Computation time analysis . . . 30
2.6. Discussion . . . 32
2.6.1. Clinical perspective . . . 32
2.6.2. SB-SVM effectiveness and outcomes . . . 32
2.6.3. Pattern discrimination and localization . . . 33
2.6.4. Clinical impact . . . 34

3. Insulin resistance: Clinical factors **35**
3.1. Introduction . . . 35
3.2. Clinical data: FIMMG_*obs* dataset . . . 36

3.3. Methods . . . 38
3.3.1. Preprocessing . . . 38
3.3.2. Regression Forest . . . 38
3.3.3. Features importance . . . 38
3.3.4. Experimental procedure . . . 39
3.4. Experimental results . . . 42
3.4.1. Preprocessing . . . 42
3.4.2. Pattern localisation . . . 44
3.4.3. Predictive performance . . . 44
3.5. Discussion . . . 47
3.5.1. Limitations and future work . . . 48

4. Insulin resistance: Type 2 diabetes early-stage risk condition 49
4.1. Introduction . . . 49
4.2. Related work . . . 50
4.3. Clinical data: FIMMG_*pred* dataset . . . 52
4.4. Methods . . . 54
4.4.1. Multiple Instance Learning boosting algorithm . . . 54
4.4.2. Experimental procedure . . . 55
4.4.3. Experimental comparisons . . . 56
4.4.4. Validation procedure . . . 57
4.5. Experimental results . . . 58
4.5.1. Predictive performance . . . 58
4.5.2. Model interpretability . . . 60
4.5.3. Sensitivity to missing values . . . 62
4.5.4. Sensitivity to the sparsity of the data . . . 63
4.5.5. Computation-time analysis . . . 63
4.6. Discussion . . . 63
4.6.1. Predictive performance . . . 63
4.6.2. Clinical significance . . . 66
4.6.3. Future work . . . 66

5. Clinical Decision Support System for type 2 diabetes quality care evaluation 69
5.1. Introduction . . . 70
5.1.1. Contributions . . . 71
5.2. Related work . . . 72
5.2.1. EHR use and sharing . . . 72
5.2.2. EHR analysis for CDDS . . . 72
5.3. Methods . . . 73
5.3.1. Data collection of T2D patients from EHRs and data sharing in a cloud infrastructure . . . 75

5.3.2. T2D patients enrollment . . . 77
5.3.3. Data indicators and features . . . 78
5.3.4. Enrolled patient management (screening and follow-up) . . . 81
5.3.5. Self-Audit & Data Quality . . . 85
5.4. CDSS analysis on a real-use case for quality care evaluation . . . 87
5.4.1. Dataset annotation . . . 87
5.4.2. Machine learning approach . . . 88
5.4.3. Machine learning results . . . 88
5.4.4. Impact of the proposed CDDS: Quality care evaluation for economic incentives . . . 88
5.5. Discussion . . . 89

6. Short-term kidney disease evolution 91
6.1. Introduction . . . 91
6.2. Related work . . . 93
6.3. Clinical data: mFIMMG dataset . . . 94
6.3.1. Preprocessing . . . 94
6.4. Methods . . . 96
6.4.1. Notations . . . 97
6.4.2. Baseline approaches . . . 97
6.4.3. Semi-Supervised Multi-Task Learning (SS-MTL) . . . 98
6.4.4. Experimental comparisons . . . 102
6.5. Experimental results . . . 103
6.5.1. State-of-the-art comparison: Semi-Supervised Learning (SSL) 103
6.5.2. State-of-the-art comparison: No-temporal . . . 104
6.5.3. State-of-the-art comparison: Stacked-temporal . . . 106
6.5.4. Multitask-temporal comparison . . . 107
6.5.5. Pattern localisation . . . 109
6.6. Discussion . . . 109
6.6.1. Predictive performance . . . 110
6.6.2. Clinical significance . . . 112
6.6.3. Limitations . . . 112
6.6.4. Future work . . . 113

7. COVID-19 Complications 115
7.1. Introduction . . . 116
7.1.1. Problem formulation . . . 116
7.2. Clinical data: RISC-19-ICU registry . . . 117
7.3. Methods . . . 118
7.3.1. Experimental procedure . . . 118
7.4. Experimental results . . . 119
7.5. Discussion . . . 120

8. Conclusions **123**
8.1. Final considerations 125
8.1.1. Open challenges 126

List of Figures

1.1. Thesis healthcare ecosystem: predictive precision medicine (PPM), Electronic Health Record (EHR), and Machine Learning (ML). 2
1.2. Example of possible Clinical Decision Support System (CDSS) designed inside the thesis healthcare ecosystem. 2

2.1. Overview of the Clinical Decision Support System (CDSS) architecture emerging from the SB-SVM approach. The Gerneral Practitioner (GP) stores the EHR data in Netmedica Italia (NMI) Cloud platform. The FIMMG dataset is composed of three different fields: demographic, monitoring and clinical. The related features were used for training the SB-SVM model and providing a T2D prediction. 19
2.2. Overview of the SB-SVM model architecture. A Tenfold Cross-Validation procedure was executed. The optimization of the SB-SVM hyperparameters was performed implementing a grid-search and optimizing the macro-recall score in a nested stratified Fivefold Cross-Validation. Hence, each split of the outer loop was trained with the optimal hyperparameters tuned in the inner loop. 24
2.3. SB-SVM performance in terms of ROC curves. 26
2.4. Performance comparison in terms of baseline models ROC curves. . . 28
2.5. Magnitude of the SB-SVM coefficients and l^0 measure values. Top 10-rank features are pointed out with red spots. 29
2.6. Comparison in term of computation time. 31

3.1. Overview of the TyG-er approach for the Tenfold Cross Validation Over Subjects (CVOS-10) procedure. The *id* number represents the patient; given N, the total number of the subjects, $M = \{1,...,m\}$, the training patients, and $I = \{1,...,i\}$, the testing patients, since $N = M + I$, it follows that $M \cap I = \emptyset$. The *seq* number identifies the temporal sequences of the TyG measurement for each *id*, where t is the last *seq* number for each *id*. The *inp* values represent the 80 EHR features (i.e., demographic, monitoring and laboratory exams). TyG_i represents the label of the Regression Forest (RF), while $\hat{\text{TyG}}_i$ is the prediction of the RF. 39

3.2. Overview of the TyG-er approach for the Leave Last Records Out (LLRO) procedure. The *id* number represents the patient; given $N = \{1,...,n\}$, the total number of the patients, it follows that all patients were included for training and testing (i.e., $i \in I$, $I \subset N$). The *seq* number identifies the temporal sequences of the TyG measurement for each *id*, where, if t is defined as the last *seq* number for each *id*, it follows that $1 < j < t$. The *inp* values represent the 80 EHR features (i.e., demographic, monitoring and laboratory exams). TyG_i represents the label of the Regression Forest (RF), while $\hat{\text{TyG}}_i$ is the prediction of the RF. 40
3.3. Preprocessing results. 43
3.4. Top 10 features listed in descending order of percentage importance according to the permutation approach (see Sec. 3.3.3) for each experimental procedure. 45

4.1. Inclusion and exclusion criteria (N identifies the number of EHR patients, and V the number of EHR features) 52
4.2. 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*. 53
4.3. Overview of the experimental procedures: a) MIL-Boost, b) time-invariant baseline, and c) temporal majority vote. 56
4.4. Overall temporal distance distribution between consecutive instances ($\Delta_{j(j+1)}$) per patient. The amount of patients is indicated below in round brackets. 59
4.5. TyG_{it_i} index distribution with mean $\mu = 8.41$ and standard deviation $\sigma = 0.52$. TyG index threshold ($\text{TyG}_{th} = 8.65$) separates the green side (179 patients) from the red side (77 patients) of the graph. 59
4.6. Percentage (%) of missing values (*NaN*) for each of the 49 EHR features: demographic, monitoring, and laboratory exams. 60
4.7. Average Recall and standard deviation of Majority vote and MIL-algorithms over all CVOS-10 procedure. 61
4.8. 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 %. 61
4.9. Top-10 rank features for MIL-Boost experimental procedure (noTyG configuration). The percentage importance of the Others features was about 59%. 62
4.10. 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}. 62

4.11. MIL-Boost Recall vs sparsity of the dataset in terms of the number of past instances ($t_i - 1$). 63
4.12. Comparison in term of computation time. 64

5.1. GP's workflow in T2D Integrated Management Care. 74
5.2. CDSS for T2D patients integrated management care managing patient enrolment and treatment in the same conceptual flow and using the same data features 75
5.3. The available EHR fields in the database 77

6.1. mFIMMG dataset preprocessing: labeled and unlabeled samples. . . . 95
6.2. Three different approaches. a) No-temporal: the temporal information is averaged across all time-windows; b) Stacked-temporal: the temporal information is preserved by concatenating longitudinally all the time-windows; and c) Multitask-temporal: each time-window is treated as a separate task. 97
6.3. SS-MTL: training experimental procedure. 101
6.4. MTL and SS-MTL approaches: Recall trend over fraction of labeled training samples $x, y \in Z_l$. In the legend, stars indicate that gender and age were included as predictors (Overall*), filled circles were not (Overall). 108
6.5. Pseudolabel samples $\tilde{x}', \tilde{y}' \in \widetilde{Z}_u$ selected from SLA procedure (after random downsampling) over fraction of labeled training samples $x \in Z_l$. In the legend, stars indicate that gender and age were included as predictors (Overall*), filled circles were not (Overall). 108

7.1. Experimental setup of the proposed method. 117
7.2. Flowchart of the proposed method. 118
7.3. Receiver operating characteristic (ROC) analysis over each fold. . . . 120
7.4. Visualisation of the top 10 most discriminative features for predicting the SOFA score according to the XGBoost algorithm. 120

List of Tables

1.1. EHR datasets qualitative comparison: data documentation (dd), data accessibility (da), data heterogeneity (dh), longitudinal observations (lo), number of patients (#rp), data type (dt), low (L), medium (M), and, high (H). 6

1.2. Machine Learning (ML) challenges (see Section 1.4) faced by the ML methodologies proposed in each chapter. 8

2.1. FIMMG dataset description: Data analysis of each EHR field considered for Case II and Case III. 18

2.2. Discarded features for Case II and Case III. 20

2.3. Range of Hyperparameters (Hyp) for the proposed Sparse Balanced-Support Vector Machine (SB-SVM) model and other tested approaches: Linear SVM (SVM Lin), Gaussian SVM (SVM Gauss), K-nearest neighbor (KNN), decision tree (DT), random forest (RF), logistic regression (LR) ridge, smoothly clipped absolute deviation (SCAD) SVM, 1-norm SVM, multi layer perceptron (MLP) and deep belief network (DBN). The threshold hyperparameter *th* is optimized for each model except for SMOTE-based approaches and 1-norm SVM (i.e., the approaches which have been already formulated to be consistent with the unbalanced setting). 25

2.4. SB-SVM: Comparison with other state-of-the-art approaches. 27

2.5. Top 10-rank features according to the SB-SVM magnitude coefficients: Blood Pressure (BP), Drugs (D), Exam prescriptions (EP), Exemptions (E), Pathologies (P). 30

3.1. FIMMG_*obs* dataset overview. 37

3.2. Range of Hyperparameters (Hyp) for each model: Regression Forest (RF), Regression Tree (RT), Boosting, Linear Support Vector Machine (SVM Lin), Gaussian SVM (SVM Gauss), and SVM Lasso. 42

3.3. Predictive performance of TyG-er and comparison with respect to the state-of-the-art (i.e., Regression Tree (RT), Boosting, Linear Support Vector Machine (SVM Lin), Gaussian SVM (SVM Gauss), and SVM Lasso). For each experimental procedure (i.e., Tenfold Cross-Validation (CV-10), Tenfold Cross-Validation Over Subjects (CVOS-10), and Leave Last Records Out (LLRO)), the best model in terms of mean squared error (*MSE*) as well as the best competitor were highlighted in bold. *MAE* represents the mean absolute error, while r and r_s represent the Pearson correlation and the Spearman's rank correlation. Standard deviation is indicated in round brackets. 46

4.1. Detailed list of the 45 laboratory exams evaluated for this study. . . . 53

4.2. 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. 57

4.3. 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. 58

5.1. T2D patients' enrolment CDSS output. Examples with real (anonymized) data. 78

5.2. Care quality Indicators under evaluation for improving the clinical performance. Data from a single GP. 81

5.3. Antidiabetics drugs subministration. Data from a single GP. 81

5.4. Complications in act. Data from a single GP. 81

5.5. Self Audit & Data Quality aggregated visualization. Sample data for a GP. Red show warning data for a particular feature of an enrolled patient. 86

5.6. Data delivered consultation 89

5.7. LAP score incentives 89

6.1. Distribution of eGFR for the labeled samples (#2176) in according with the CKD stages. 96

6.2. Final configuration of the mFIMMG dataset after the preprocessing stage. 97

6.3. Notations. 98

6.4. Range of hyperparameters (hyp) for each model: Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), Support Vector Machine (SVM) with Lasso regularizer, and Convex Fused Group Lasso (CFG) with Logistic regression model. 103

6.5. Experimental results comparison of the Self-Learning Algorithm (SLA) procedure with other Semi-Supervised Learning (SSL) techniques (i.e., Positive and Unlabeled learning [PU], Label Propagation [LP]). The comparison was performed only for the Overall* field of the baseline (i.e., no-temporal) approach. The best result in terms of *Recall* was highlighted in bold. 104

6.6. **No-temporal:** Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), and Support Vector Machine (SVM) with Lasso regularizer. In SLA procedure the same classifier adopted externally in 10-CV was used. Overall* indicates that also gender and age were included as predictors. Best result in terms of *Recall* was highlighted in bold for each field. 105

6.7. **Stacked-temporal:** Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), and Support Vector Machine (SVM) with Lasso regularizer. In SLA procedure the same classifier adopted externally in 10-CV was used. Overall* indicates that also gender and age were included as predictors. Best result in terms of *Recall* was highlighted in bold for each field. 106

6.8. **Multitask-temporal:** Decision Tree (DT) classifier was used to select pseudolabels in SLA procedure, except for Overall* where Support Vector Machine (SVM) with Lasso regularizer was used. Overall* indicates that also gender and age were included as predictors. Fraction (f) represents the amount of labeled samples used in the training stage. The table depicts the SS-MTL majvot configuration. Best result in terms of *Recall* was highlighted in bold for each field. 107

6.9. Top-10 predictors for SS-MTL majvot approach with f =30 %. Overall* indicates that also gender and age were included as predictors. D=Drugs; E=Exam; M=Monitoring. 109

7.1. Descriptive statistics. 119

7.2. Confusion matrices (rows are the true classes) of the XGBoost algorithm for solving the classification task. 119

8.1. Machine Learning (ML) challenges (see Section 1.4) faced by the ML methodologies proposed in each chapter. 123

A.1. Type 2 diabetes patient's features set. Example from a real anonymous patient. 129
A.1. Type 2 diabetes patient's features set. Example from a real anonymous patient. 130
A.2. Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%. . . 131
A.2. Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%. . . 132
A.2. Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%. . . 133
A.2. Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%. . . 134
A.2. Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%. . . 135
A.2. Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%. . . 136

Chapter 1.

Background and motivation

Medicine is an evolving field that updates its applications thanks to recent advances from a broad spectrum of sciences such as biology, chemistry, statistics, mathematics, engineering, and life and social sciences. Generally, discoveries in such sciences are applied to medicine with the aim of preventing, diagnosing, and treating a wide range of medical conditions.

The current traditional approach to manage diseases can be briefly reduced to the "one-size-fits all" concept (i.e., the effect of treatment reflects the whole sample). Although this view of medicine has been consolidated in the past for several decades, applications of effective treatment, for example, can lack efficacy and may have adverse or unpredictable reactions in individual patients [7].

Precision medicine is the extension and the evolution of the current traditional approach to patient management [8]. Unlike "one-size-fits all" approach, precision medicine is mainly preventive and proactive rather than reactive [9]. Barak Obama, who claimed the importance of "delivering the right treatments, at the right time, every time to the right person", has highlighted the critical impact of this emerging initiative in healthcare practice. The personalized approach has been therefore emerged as a promising enhanced substitute for oversimplified and reductive medicine to disease categorization and treatment. Precision medicine uses a wide spectrum of data, ranging from biological to social information, tailoring diagnosis, prognosis, and therapy on patient's needs and characteristics, converging to a patient-centered medicine [10]. This evolution may lead to predictive, personalized, preventive, participatory, and psychocognitive healthcare: among all, the predictive medicine (PM), which is used to forecast disease onset, prognosis, and therapy outcome, is the one this thesis emphasizes. Thus, It is possible to introduce a new emerging healthcare area, named predictive precision medicine (PPM) [11], defined as the merging of these two new fields (i.e. precision medicine, predictive medicine) of medical sciences by utilising biomarkers to forecast disease onset, progression and its treatment tailored on individual features, like omic, environmental and lifestyle elements. PPM represents a medical model that individualises the care of patients according to their risk of disease or their predicted response to intervention and thus has the potential to ensure the best response and highest safety margin for patient care. PPM could lead to significant

improvement from patients' life to the global population and healthcare systems.

PPM may soon fully benefit from a huge amount of medical information and Artificial Intelligence (AI) techniques [12, 13]. This new concept of medicine involves the medical institutions that collect every day a huge amount of healthcare information in Electronic Health Records (EHRs). Machine Learning (ML) and Deep Learning (DL) techniques perfectly fit with these structured big data and, consequently, permit to reach the PPM objectives. This new promising healthcare ecosystem (see figure 1.1), consisting of PPM, EHR, and ML, will better identify and treat chronic pathologies, reduce financial and time efforts, and improve patients' quality of life.

This thesis will focus on the potential of PPM as a new approach to health sciences and clinical practice, by proposing novel applicative ML methodologies as the main core of a Clinical Decision Support System (CDSS) as shown in Figure 1.2. The proposed ML methodologies, designed from specific clinical objectives, have been applied in several pathological scenarios, such as insulin resistance (IR), type 2 diabetes (T2D), kidney disease (KD), and COVID-19.

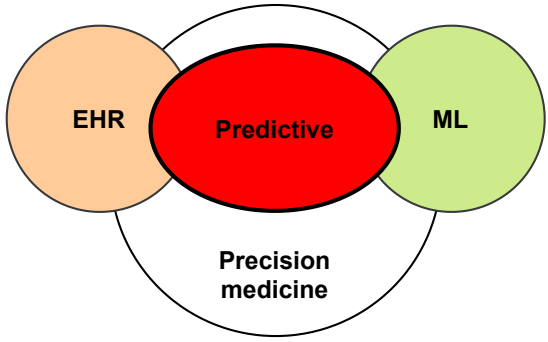


Figure 1.1.: Thesis healthcare ecosystem: predictive precision medicine (PPM), Electronic Health Record (EHR), and Machine Learning (ML).

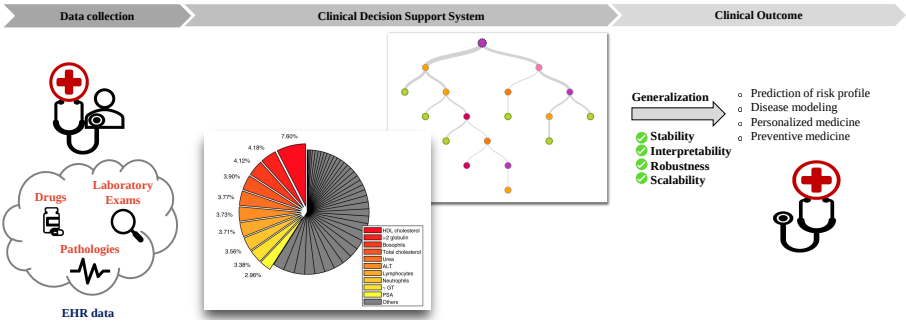


Figure 1.2.: Example of possible Clinical Decision Support System (CDSS) designed inside the thesis healthcare ecosystem.

1.1. Predictive medicine and precision medicine

PM is a relatively new area of Medicine. Broadly speaking, PM may be defined as the use of laboratory and genetic tests to predict either the onset of a disease in an individual or the trend of the current disease. PM may vary from estimating the risk for a certain outcome to predicting which treatment will be the most effective on the individual [12]. In this sense, lots of biomarkers could be used to forecast disease onset, prognosis, and therapy outcome. Any biomarker indicates a medical sign that can be measured objectively, accurately, and reproducibly; in this direction, the World Health Organization defined biomarker as "almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological". Thus, the concept of biomarker could be translated as any clinical factor that gives us information about the health of the individual. If the current traditional approach to clinical trials is "one-size-fits-all" (i.e., the effect of treatment reflects the whole sample), the future of medicine is to provide the "the right treatment for the right patient at the right time", identifying different subgroups depending on certain biomarkers that respond to an optimal treatment [14].

PM institutes preventive measures to either prevent the disease altogether or significantly decrease its impact upon the patient, so that healthcare professionals and the patient themselves can be proactive in instituting lifestyle modifications and increased physician surveillance. Thus, PM changes the paradigm of medicine from being reactive to being proactive and has the potential to significantly extend the duration of health and to decrease the incidence, prevalence, and cost of diseases. PM may be oriented for both healthy individuals and those with diseases, and its purpose concerns predicting susceptibility to a particular disease or predicting progression and treatment response for a given disease, respectively.

Unlike many preventive interventions that are applied to groups (e.g., immunization programs), PM is conducted on an individualized basis. PM is expected to be most effective when applied to multi-factorial diseases such as diabetes mellitus and chronic cardiovascular diseases that are prevalent in industrialized countries. However, complex diseases in the wider population are not only merely affected by heredity and patient's clinical condition, but also by external causes such as lifestyle and environment. Therefore, multiple environmental factors, particularly smoking, diet and exercise, infection, and pollution play important roles and may be as important as the patient's clinical condition.

Thus, diseases are influenced by various factors, some of which are generally well-known factors while others may be specific individual factors. While the former has been studied in great detail, the latter has not yet. Understanding individual factors may permit to prevent or manage the disease more effectively. This method of tailoring treatment, practices, and medical decisions to a patient based on specific personalized factors (i.e., patient's clinical history, biomarkers, environment, lifestyle) has

been defined as precision medicine [15].

The concept of precision medicine is relatively new and some of the potential advantages are as follows:

- Efficiency of care: precision medicine makes decisions based on individual-specific factors. Clinicians may be co-assisted and supported by proposed customized treatments for each of their patients, improving the range and the responsiveness of all possible interventions.
- Preventive care: the screening process can be used to diagnose diseases and even prevent such diseases by understanding the risk of an individual rather than simply reacting to an illness.
- Limit cost: targeted treatment based on patient's features mapping can reduce the cost of care with more informed treatment decisions and a greater chance of being effective. The cost will be potentially lower with the focus on preventive care rather than treatment of disease.
- Population health: studying patterns in a population as a whole, and as sections can help in identifying causes for particular diseases and design the treatment. Epidemiological studies can predict the likelihood of diseases and early detection.

While the drawbacks are as follows:

- Variability; excluding the patient's clinical condition, multiple external factors contribute to the predictive results. This aspect implies that outcomes obtained by precision medicine are more difficult to be quantified and objectively accepted by the scientific community. Furthermore, the potential false positives or false negatives that may arise from a screening program can cause substantial unnecessary stress on the individual.
- Infrastructure requirements: precision medicine might deeply impact healthcare, but it requires massive infrastructure investments to implement and update infrastructures and mechanisms of data collection, storage, and sharing. Additionally, the privacy and security of digital data have to be improved and assured.

1.2. Electronic health records

Health Information Technology (HIT) has become an essential part of the daily work of clinicians. The reason for a strong investment in HIT is the wider adoption of EHR systems. EHRs are expected to improve the national healthcare quality and efficiency [16, 17], for example, decreasing unnecessary services, such as repeated laboratory tests every time the patient changes hospital and office visits [18, 19]. Moreover, the

adoption of EHR reduces the workload of the clinician, healthcare costs, medical errors [18, 20], patient complications, and mortality [21, 22]. In addition, EHRs can document diagnostic investigations and medical treatment, provide CDSS and facilitate communication between health care personnel. EHRs represent important tools in daily activities to store a huge amount of data [23]. The users can range from General Practitioners (GPs) to physicians in hospitals and Intensive Care Units (ICUs). The most discriminative difference lies in time, particularly in record sampling frequency. In fact, in general practice, EHRs data cover a longer patient time history, but record sampling frequency is more dilated in time, while in the hospital or ICU the patient length of stay is shorter as well the record sampling frequency. Perhaps, EHR data usually may present a multitude of technical and operative issues, mainly due to a lack of digitalization and standardization, respectively. Commonly, EHR data present a non-uniform record sampling frequency, which leads to several missing values or missing views. Moreover, too sparse physician's annotations may render EHR data unusable. EHR data is often voluminous, multi-source and the enclosed clinical information may be hidden or not fully interpretable a priori [24].

Cybersecurity and privacy-preserving are not irrelevant aspects due to the sensitivity of the data.

Nowadays, there is a growing demand to extract large datasets from the EHRs for administrative reporting, clinical audits, and, mostly, for research purposes. In the research scenario, there is a lack of publicly EHRs availability, thus this aspect is assuming an increasing relevance. In the following table, several popular EHR datasets used in clinical practice were compared in terms of specific features, such as data documentation (dd), data accessibility (da), data heterogeneity (dh), longitudinal observations (lo), number of records/patients (#rp), and data type (dt). These features are evaluated in terms of different levels such as low (L), medium (M), and high (H).

1.3. Machine Learning for healthcare

Driven by an increase in computational power, storage, memory, and the generation of staggering volumes of data, AI techniques allow computers to perform a wide range of complex tasks with impressive accuracy. On the one hand, Informatics is crucial for precision medicine since it manages big data, creates learning systems, gives access for individual involvement, and supports precision intervention from translational research [35]; on the other hand, HIT is crucial for PM providing clinicians tools that give information about an individual at risk, disease onset and how to intervene [12]. The importance of AI in the field of medicine is confirmed by the fact that, for instance, in the United States the use of EHRs grew from 11.8% to 39.6% among physicians from 2007 to 2012 [36].

In this context, one of the aims of HIT is to convert EHR into knowledge that could be exploited for designing a CDSS. ML models can manage this enormous amount

Table 1.1.: EHR datasets qualitative comparison: data documentation (dd), data accessibility (da), data heterogeneity (dh), longitudinal observations (lo), number of patients (#rp), data type (dt), low (L), medium (M), and, high (H).

	dd	da	dh	lo	#rp	type
Nhanes [25]	H	H	H	L	10^3	Quest/Lab
MIMIC-III [26]	H	H	L	M	10^3	Hospital
HCup [27]	H	L	L	M	10^6	Hospital
ORBDA [28]	L	L	M	M	10^6	Hospital
CPRD [29]	H	M	H	H	10^6	GPs
THIN [30]	L	M	M	H	10^6	GPs
QResearch [31]	M	M	M	H	10^6	GPs
ResearchOne [32]	M	M	M	H	10^6	GPs
DNHSPD [33]	M	L	M	H	10^6	Pharmacy
Netmedica [1]	H	M	H	H	10^4	GPs
FIMMG [2]	H	H	H	M	10^3	GPs
mFIMMG [4]	H	H	H	M	10^4	GPs
RISC-19 [34]	H	M	H	H	10^3	Hospital

of data by predicting clinical outcomes and interpreting particular patterns sometimes unsighted by physicians [37]. ML techniques have been widely used for extracting information from such a large amount of data and have proven useful in improving diagnosis, outcome prediction, and management of chronic diseases [38, 39, 40]. This includes a possibility for the identification of high risk for medical emergencies such as transition into another disease state, for example, the progression from pre-diabetes to T2D using routinely-collected EHR data [4]. ML will also play a fundamental role in the development of learning healthcare systems, which describe environments that align science, informatics, economical incentives, and lifestyle education for continuous improvement and innovation. These systems, ranging from small group practices to large national providers, may combine multi-source data with ML techniques. The result consists of a continuous source of data-driven insights to optimise biomedical research, public health, and healthcare quality [41]. Accordingly, the widespread advances in the field of Deep Learning (DL) have also encouraged the application of DL approaches to clinical tasks (e.g., including outcome prediction) based on EHR data [42]. Therefore one of the limitations of DL research involves topics such as data heterogeneity and model interpretability/explainability [42, 43].

1.4. Thesis: Problems statement

In a PM scenario, ML methodologies were proposed to solve specific tasks for several pathological conditions, such as IR, T2D, KD, and COVID-19.

Managing and modeling real-world datasets extracted from EHRs usually lead to several ML challenges in terms of:

- High-dimensional & heterogeneous data: The number of patients may largely overcome the number of features (i.e. patient's information) or vice versa. Moreover, EHR data, which may originate from heterogeneous sources, usually own noisy and/or redundant features. Thus, the ML model should guarantee robustness against high-dimensional and heterogeneous data while, at the same time, ensuring a reasonable computational cost effort.

- Unbalanced setting: In a classification task usually the target class (i.e., the pathological condition to be predicted) is largely less representative than the other classes. The ML model should guarantee robustness against an unbalanced setting, ensuring also the correct representation of the minority classes more difficult to reach.

- Sparse labeling: Mostly in general practice, predictor and target values are sparse and not always available over time. Predictor values may contain a huge amount of missing values. The ML model should guarantee robustness against missing values, ensuring affordable data imputation techniques. Moreover, in supervised tasks, also the target values may not be constantly provided over time and lots of important clinical information cannot be used. Thus, dedicated ML paradigms are needed to exploit also the unlabeled information.

- Temporal ambiguity: Time-series data is of vital importance because it provides much more information than the "snapshots" presented by static data, and hence permits much greater insight. Time-series data is the key point of evidence-based CDSSs. With the increasing availability of EHRs, there is enormous unexpressed potential for providing accurate and actionable predictive models to time-series data for real-world concerns. The ML model should guarantee robustness against temporal ambiguity, being able to capture the patient's temporal information by adopting a suitable task-based temporal representation.

- Interpretability/explainability: Understanding which are the most discriminative predictors that contribute to the outcome of the model. It ensures that the outputs generated by machine learning models can be understood, rather than remaining "black boxes". This is particularly important in the healthcare domain where black box predictions are unlikely to be acceptable to patients, clinicians, or regulatory bodies.

- Generalization: The ML model's ability to well react on new unseen data rather than just the data that it was trained on. The approach of generalization must ensure that the data used to train the model is a good and reliable sample of the observation in the mapping that we want the model to learn. The ML model should guarantee robustness against the risk of overfitting, ensuring a high degree of generalization performance and scalability of the algorithm.

1.4.1. Thesis contribution

Table 1.2.: Machine Learning (ML) challenges (see Section 1.4) faced by the ML methodologies proposed in each chapter.

ML Methodologies	High-dimensional & heterogeneous data	Unbalance setting	Sparse labeling	Temporal ambiguity	Explainability/ Interpretability	Generalization
[2]T2D discovering (ch.2)	x	x	x		x	x
[3]IR: clinical factors (ch.3)			x	x	x	x
[4]IR: T2D early stage (ch.4)		x	x	x	x	x
[1]CDSS for T2D evaluation care (ch.5)	x		x		x	x
[5]KD early stage risk (ch.6)	x	x	x	x	x	x
[6]COVID-19 complications (ch.7)	x		x		x	x

The thesis healthcare ecosystem (i.e., PPM, EHR, ML), offers a contribution to the biomedical and health informatics field by proposing meaningful ML methodologies to face and overcome the ML challenge aspects previously listed (see Table 8.1):

- High-dimensional & heterogeneous data were managed during the preprocessing stage (i.e., features selection, standardization, outliers detection);
- Unbalanced setting was managed by adopting specific optimization metrics and/or optimal thresholds for the posterior probabilities of the decision function;
- Sparse labeling of the predictors was managed with standard static data imputation techniques (i.e, extra-values, mean, median, K-Nearest Neighbors (KNN)), while sparse labeling of the targets was managed by proposing semi-supervised learning (SSL) techniques;
- Temporal ambiguity was managed by proposing different experimental configurations (i.e., time-invariant, stacked-temporal, Multiple Instance Learning (MIL), Multi-Task Learning (MTL) with temporal relatedness/constraints);
- Interpretability/explainability was managed offering always a features importance ranking of the most discriminative predictors to clinically understand the outcome of the ML model;
- Generalization was managed by adopting regularization strategies.

The proposed novel ML methodologies in the PM scenario may constitute the main core of a CDSS usable by physicians for prevention, screening, diagnosis, and treatment purposes. Several pathological conditions, such as IR, T2D, KD, and COVID-19 were debated in this thesis, but nothing prevents from generalizing and scaling the proposed ML methodologies on other chronic pathological conditions as well.

This thesis also contributed to the publication of novel publicly available EHR datasets, such as:

- FIMMG dataset[1]
- FIMMG_*obs* dataset[2]
- FIMMG_*pred* dataset[3]
- *m*FIMMG dataset[4]

1.5. Thesis overview

An overview of the thesis structured into the following chapters is presented to facilitate the reading organization.

- **Chapter 1** has offered a qualitative and brief introduction of PPM, EHRs, and AI, the 3 main interconnected key points that constitute the thesis healthcare ecosystem. Then the problem statement and the contribution of the thesis have been provided.

The rest of the thesis proposes an in-depth study of four PM scenarios, such as IR, T2D, KD, and COVID-19. In particular, the following list shows the organization and an overview of the rest of the thesis:

- **Chapter 2** aims to exploit an ML methodology, named sparse-balanced SVM, for discovering T2D in general practice using features extracted from a novel EHR dataset, namely FIMMG dataset. Pathologies, exams, drugs, and exemption are used as predictors. Temporal information has not bee taken into account.
- **Chapter 3** aims to exploit a high-interpretable ML approach (i.e., ensemble Regression Forest combined with data imputation strategies), named TyG-er, to identify non-trivial clinical factors in EHR data to determine where the IR condition is encoded. A specific clinical biomarker, named TyG index is introduced. Temporal information is taken into account. Clinical data derive from a subset of the FIMMG dataset, but only laboratory test values are used in this study.
- **Chapter 4** aims to exploit a Multiple Instance Learning boosting (i.e., MILBoost) algorithm applied to past EHR patient information to create a predictive model capable of early prediction of worsening IR (low vs high T2D risk) in terms of TyG index. Temporal information is taken into account. Clinical data derive from a subset of the FIMMG dataset, but only laboratory test values are used in this study.

[1] http://vrai.dii.univpm.it/content/fimmg-dataset
[2] http://vrai.dii.univpm.it/content/fimmgobs-dataset
[3] http://vrai.dii.univpm.it/content/fimmgpred-dataset
[4] http://vrai.dii.univpm.it/content/mfimmg-dataset

- **Chapter 5** aims to develop a platform for GPs data sharing and standardized T2D patient management, guaranteeing the inter-operability of the platform with other healthcare databases. The proposed framework, equipped with a novel ML-based CDSS, processes and analyses the shared EHRs data for T2D screening purposes.
- **Chapter 6** aims to exploit a novel Semi-Supervised Multi-task Learning (SS-MTL) approach for predicting short-term KD evolution (i.e., patient's risk profile) on multiple GPs' EHR data, named mFIMMG dataset. The SS-MTL approach imposes a temporal relatedness between consecutive time-windows to predict the eGFR status over time and learns both from labeled and unlabeled samples in the learning procedure. Pathologies, exams, drugs, and laboratory test values are used as predictors.
- **Chapter 7** aims to propose the prediction of the SOFA score at day 5, by utilising only clinical data at the admission day in ICU. The temporal evolution of the SOFA score describes the COVID-19 patient's complications in ICU and its prediction helps to create patients' risk profiles. Approximately 100 ICUs participated at the RIsk Stratification in COVID-19 patients in the Intensive Care Unit (RISC-19-ICU) registry, but only a subsample of those patients has been utilised for this study. Temporal information has not been taken into account.

Each chapter (ch. 2 ÷ ch. 7), which differs in clinical tasks, pathological conditions and, ML methodologies i) reviews the state-of-the-art; ii) presents the adopted EHR dataset and preprocessing stage; iii) presents the proposed ML algorithm; iv) presents the experimental setup and measure; v) provides the experimental results for evaluating the performance of the proposed method; vi) discusses the obtained results and future work; and vii) presents the conclusions.

- **Chapter 8** offers the conclusions of each work presented in the previous chapters. Then, final considerations and open challenges of healthcare ecosystems (i.e, PPM, EHR, ML) are discussed.

1.6. Thesis outcomes: Publications

The thesis outcomes are available in the follow publications:

- M. Bernardini, L. Romeo, P. Misericordia, and 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.
- M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, and 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*, 2019.

- M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, and L. Burattini, "Early temporal prediction of Type 2 Diabetes Risk Condition from a General Practitioner Electronic Health Record: A Multiple Instance Boosting Approach", *Artificial Intelligence in Medicine*, 2020.

- E. Frontoni, L. Romeo, M. Bernardini, S. Moccia, L. Migliorelli, M. Paolanti, A. Ferri, P. Misericordia, A. Mancini, and P. Zingaretti, "A Decision Support System for Diabetes Chronic Care Models Based on General Practitioner Engagement and EHR Data Sharing", *IEEE Journal of Translational Engineering in Health and Medicine*, 2020.

- M. Bernardini, L. Romeo, E. Frontoni, and M.R. Amini, "A Semi-Supervised Multi-Task Learning Approach for Predicting Short-Term Kidney Disease Evolution", *IEEE Journal of Biomedical and Health Informatics*, 2020 [Accepted].

- J. Montomoli, L. Romeo, S. Moccia, M. Bernardini, L. Migliorelli, A. Donati, A. Carsetti, P. Garcia, T. Fumeaux, P. Guerci, R. Schuepbach, E. Frontoni, RISC-19-ICU Investigators, M. Hilty, "Predicting 5-day SOFA score at ICU admission in COVID-19 patients: a proof-of-concept study using prospectively collected data from 1613 patients in the RISC-19-ICU registry", *Journal of the American Medical Association*, 2020 [Submitted].

Chapter 2.

Type 2 diabetes discovering

The diagnosis of type 2 diabetes (T2D) at an early stage has a key role for an adequate T2D integrated management system and patient's follow-up. Recent years have witnessed an increasing amount of available Electronic Health Record (EHR) data and Machine Learning (ML) techniques have been considerably evolving. However, managing and modeling this amount of information may lead to several challenges such as overfitting, model interpretability and computational cost. Starting from these motivations, a ML method called Sparse Balanced Support Vector Machine (SB-SVM) for discovering T2D in a novel collected EHR dataset (named FIMMG dataset) was introduced. In particular, among all the EHR features related to exemptions, examination and drug prescriptions, only those collected before T2D diagnosis from a uniform age group of subjects were selected. The reliability of the introduced approach with respect to other ML and Deep Learning (DL) approaches widely employed in the state-of-the-art for solving this task was demonstrated. Results evidence that the SB-SVM overcomes the other state-of-the-art competitors providing the best compromise between predictive performance and computation time. Additionally, the induced sparsity allows to increase the model interpretability, while implicitly managing high dimensional data and the usual unbalanced class distribution.

2.1. Introduction

The World Health Organization (WHO) reported that the global prevalence of worldwide diabetes is around 9% (more than 400 million people). The 90% of people with diabetes suffers from T2D [44, 45]. T2D is on the rise worldwide and only in 2012 diabetes caused an estimated 1.5 million deaths. The WHO anticipates that worldwide deaths will double by 2030 [46]. In developing nations, more than the half of all diabetic cases goes undiagnosed. This can be attributed to the fact that T2D symptoms may be less marked than other types of diabetes (e.g., Type 1). Nonetheless, the International Diabetes Federation (IDF) stated that early diagnosis and opportune treatments can save lives while preventing or significantly delaying devastating complications [45]. Moreover, diabetes is the major cost on the economic balances of

national health systems (IDF indicates for the year 2015 a level of expenditure for the treatment of diabetic patients equal to 11.6% of the total world health expenditure).

A more efficient integrated management system, including General Practitioners (GPs) and specialists with multidisciplinary skills, could be a valid solution to alleviate the healthcare costs while preventing diabetes-related diseases (e.g., diabetic retinopathy, renal diabetes). Almost all GP outpatient clinics are now equipped with EHRs storing the health history of the patients as well as several heterogeneous information (i.e. demographic, monitoring, lifestyle, clinical). The reason for a strong investment in Health Information Technology (HIT) is that wider adoption of EHRs will reduce healthcare costs, medical errors [18, 20], patient complications and mortality [21, 22]. Moreover, the HIT will decrease the use of healthcare services such as laboratory tests and outpatient visits, [18, 19], while it will improve the national healthcare quality and efficiency [16, 17].

The high number of patients' information recorded in EHRs results in a large amount of stored data. In this context, one of the aims of biomedical informatics is to convert EHR into knowledge that may be exploited for designing a Clinical Decision Support System (CDSS). In this scenario, ML models are able to manage this enormous amount of data by predicting clinical outcomes and interpreting particular patterns sometimes unsighted by physicians [37].

The aim of this work is to exploit a ML methodology, named SB-SVM, for discovering T2D using features extracted from a novel EHR dataset, namely the FIMMG dataset. The proposed SB-SVM is able to manage high dimensional data by increasing the model interpretability and finding the most relevant features while dealing with the usual unbalanced class distribution. In the data analysis, among all the EHR features related to exemptions, examination and drug prescriptions, only those collected before T2D diagnosis were considered, while excluding all features that have already revealed a T2D patient's follow-up. Additionally, a subset of subjects enclosed from $60-80$ years range was considered, where the chronological age is not statistically relevant in order to discriminate T2D condition. The employed FIMMG dataset is available at the following link[1].

Three different research questions were formulated in order to measure the reliability of our approach with respect to the state-of-the-art methodologies:

- **Case I**: Is the SB-SVM approach able to predict T2D using all set of EHR features? (Section 2.3.1.1).
- **Case II**: Is the SB-SVM approach able to predict T2D using only a subset of EHR features collected before T2D clinical diagnosis? (Section 2.3.1.2).
- **Case III**: Is the SB-SVM approach able to predict T2D using only a subset of EHR features collected before T2D clinical diagnosis from a uniform age group of subjects? (Section 2.3.1.3).

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

2.2. Related work

In the last decade, with the increasing amount of available data, ML techniques for discriminating T2D condition have been considerably evolving. Accordingly, the widespread advances in the field of DL have encouraged the application of DL approaches to clinical tasks (including outcome prediction) based on EHR data [42]. Therefore one of the limitation of DL research involves topic such as data heterogeneity and model interpretability [42, 43]. Real-world datasets extracted from EHRs usually own high dimensionality data and several noisy and/or redundant features. Managing and modeling this amount of information may lead to several challenges such as i) overfitting; ii) reduction of interpretability; iii) computational cost increment. In this context, researchers tried to overcome these issues by performing features selection [47, 48, 49, 24] or engineering feature techniques [50]. Zheng et al. introduced an engineering features based framework able to improve the performance of traditional ML models (e.g., SVM, logistic regression (LR), decision tree (DT), k-nearest neighbor (KNN), random forest (RF), Naive Bayes (NB)) for predicting T2D condition [50]. Deviating from our approach, the feature extraction stage implemented by [50] required a further computational effort as well as the supervision of the physician in order to define high-level features. Sheikhi et al. performed the analysis not considering some features directly related to T2D (i.e., glycated hemoglobin (HbA1c)) [47]. Then, a feature selection based on LASSO and ridge LR was executed before predicting the diabetic condition. Kamkar et al. exploited the typical EHR tree structure in order to perform a feature selection based on Tree-LASSO algorithm [48]. Cho et al. predicted the onset of diabetic nephropathy from an unbalanced and irregular dataset. A feature selection has been employed with baseline statistical methods, ReliefF [51], SVM sensitivity analysis and recursive feature elimination (RFE) [49]. In [24] a filter feature selection strategy based on ReliefF was adopted as well, to rank the important attributes for identifying the diabetic condition while evidencing that the main discriminative feature is the age marker. Concerning the classification stage, they employed standard supervised algorithms such as NB, DT, and Instance-Based learners [24]. The main difference with the respect to the above-mentioned literature [47, 48, 49, 24] can be resumed according to the different strategy used to manage high dimensional data while achieving reliable performance. Our embedded method interacts with the classifier structure and is able to manage the huge amount of features while discovering the most relevant ones without any supplementary feature selection stage. Differently the wrapper feature selection approaches [49] can result in an external module with an increment of computational effort (e.g., beam search, sequential forward selection, sequential backward elimination) for the training and validation stage, while the filter methodologies [47, 48, 49, 24] rely on the evaluation of the data statistics (e.g., t-test, chi-square, information gain, ReliefF) which is not always linked with the learning algorithm.

The papers [52, 53, 54] are closer to our work, where the aim is to develop an algorithm able to manage high dimensional and unbalanced data without performing a supplementary features selection strategy. However, apart from the proposed methodology, our approach is also different in terms of experimental procedure for testing the reliability of the ML model in T2D early prediction.

In particular, DT model was proposed in [52] for discovering T2D in a uniform age group, while the ensemble (i.e., bagging) and boosting (i.e., AdaBoost) methodologies were employed for decreasing the generalization error when dealing with high dimensional data. In Wang et al. as well an ensemble strategy based on RF was employed in order to suggest antihyperglycemic medications for T2D patients [54]. The synthetic minority oversampling technique (SMOTE) algorithm [55] was used to solve the unbalanced class problem oversampling the minority class. Deviating from their approach, our method induces sparsity which leads to an increase in the interpretability of our ML model. Additionally, our method deals the high unbalanced setting without drawing synthetic samples [54], not always consistent and close with the real class data. As it will be evidenced from the experimental results (Section 2.5), our methodology is more effective with respect to the DT and bagging classifiers (i.e., RF) employed by [52, 54] even in a uniform age group [52].

In Yu et al. a SVM model was proposed for two different classification problems related to diabetes condition [53]. Their proposed SVM and LR models were able to discriminate between T2D and control subjects in high dimensional data with a large population not suffering from diabetes. The main difference of our approach with respect to [53] lies in the application of 1-norm regularizer in the hinge loss function that induces sparsity in the model coefficients.

2.2.1. Sparse SVM for unbalanced dataset

There were existing work which proposed the solution of a Sparse SVM while addressing the imbalance dataset problem in different domain ranging from clinical data [56] to image categorization [57]. Differently from [56] our method induced sparsity by applying the least absolute shrinkage and selection operator (LASSO), while the auhors in [56] employed the smoothly clipped absolute deviation (SCAD) penalty. Both regularizers are members of the L_q penalty functions and they can be adopted to automatically and simultaneously select variables retaining the most relevant features. Although LASSO and SCAD disclose the sparsity and continuty properties, the SCAD results also in a unbiasedness estimator [58]. However, our choice is motivated by the peculiarity of LASSO to perform better when the noise level of the features space (i.e., EHR data) is very high [58]. Additionally, in [56] the authors deal with the unbalanced setting by introducing an adaptive proportional weight within the objective function. This setting gives more importance to the minority class and, as result, the well-classified group gets the less weight. On the contrary, our method proposed

to adjust the decision threshold without modifying the SVM objective function.

Deviating from the work proposed in [57], our approach deals with the unbalanced setting changing directly the decision threshold of the inferred posterior probability while controlling the true positive/negative rate (macro-recall). Since under Bayesian decision theory our approach would be the optimal strategy [59], its reliability depends on the estimated posterior probability. On the other hand, the approach proposed in [57] is similar to the cost-sensitive SVM proposed in [60]. Differently from the standard SVM, the authors in [60] handled the unbalanced setting by penalizing differently each class. Our approach leads to a most intepretable strategy to deal with unbalanced classes, while estimating the posterior probability of the predicted classes.

The reliability of our approach is also confirmed by the state-of-the-art comparison performed in Section 2.5, which unveils a greater predictive accuracy of the SB-SVM with a lower computation effort.

In summary, the main contributions are the following:

- The collection and employment of the novel FIMMG dataset.
- The design of an algorithm core for treating and following chronic T2D patients.
- The introduction of the SB-SVM model able to achieve better performance with respect to the other state-of-the-art approaches.
- The experimental test performed in a clinical use case scenario.

2.3. Clinical data: FIMMG dataset

The largest Italian federation of GPs (FIMMG) have instituted Netmedica Italia (NMI) in order to offer HIT services to GPs in the national territory. NMI platform manages a cloud computing project that through the integration of GPs' databases (i.e., EHRs) is able to realize network medicine, audit process, data reporting, integrated management systems between GPs and Specialists for treating chronic pathologies. All the process guarantees maximum data security encrypted both during transfer and storage, and access is strictly allowed only to ones with the permission. NMI aggregates EHRs available from GPs in a unique standardized language and share them on a cloud platform. The FIMMG dataset (Figure 2.1), extracted from NMI cloud platform, belongs exclusively to a unique GP's EHR. The EHR contains a total of 2433 patients, including both those who are no longer followed by the GP or died. The physician's and patients' identities are anonymous. Three main fields compose the FIMMG dataset: *Demographic* (gender and age), *Monitoring* (blood pressure) and *Clinical* (pathologies, exemptions, exam and drug prescriptions) collecting a total of 1862 features. For each patient the date of feature registration and the number of its occurrence are reported too. This aspect assumes a relevant significance because it allows to trace up the patient's clinical history even in the time domain.

2.3.1. Data analysis

Table 2.1 reports the description of the FIMMG dataset as well as the EHR fields. Concerning the proportions between control and T2D patients, the FIMMG dataset is characterized by a strong imbalanced configuration with a ratio less than $10:1$ in the advantage of control patients not suffering from T2D, while the whole population is balanced in terms of gender. three experimental tests were performed:

Table 2.1.: FIMMG dataset description: Data analysis of each EHR field considered for Case II and Case III.

Dataset description	**Count (%)**	**Mean (std)**
Total patients:	2433	-
Control patients	2208 (0.91)	-
Diabetic patients	225 (0.09)	-
Total features	1841	-
Fields	**Count (%)**	**Mean (std)**
Demographic		
Gender:	-	-
Male	1186 (0.49)	-
Female	1247 (0.51)	-
Age (years)	-	58.00($\pm$23.58)
¡60	1374 (0.56)	-
60-80	535 (0.22)	-
¿80	524 (0.22)	-
Monitoring		
Blood pressure (mmHg)		-
Systolic	3	135.52($\pm$17.21)
Diastolic	3	80.83($\pm$8.65)
Clinical		
Pathologies	877	-
Exemptions	70	-
Exams	396	-
Drugs	490	-

Figure 2.1 shows an overview of the FIMMG dataset and the CDSS architecture emerging from the SB-SVM approach. The proposed SB-SVM algorithm can be seen as the main core of the CDSS framework.

2.3.1.1. Case I

In the starting configuration, all the features ($n = 1862$) related to the EHR fields shown in Table 2.1 have been considered for predicting the T2D condition. Obviously, from the starting feature set, the T2D information (i.e., T2D exemption code, T2D pathologies ICD-9 codes) have been already discarded. All the features were binarized according to the following values:

- 0: if the subject has never been affected by this pathology or associated with this exemption, or if the specific drug and exam have never been prescribed.
- 1: if the subject has been at least once affected by this pathology or associated with this exemption, or if the specific drug and exam have been at least once

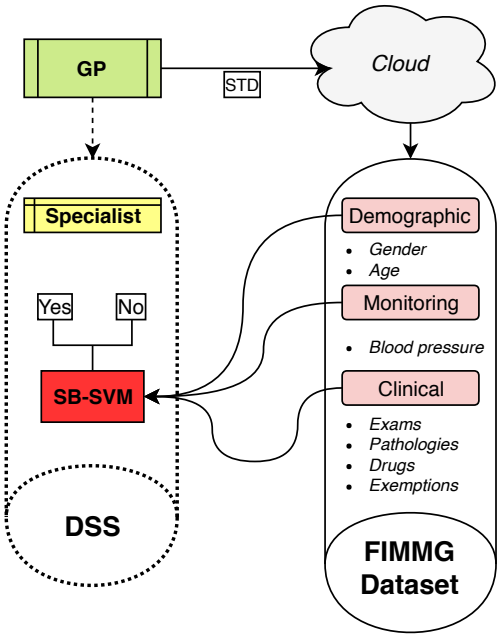


Figure 2.1.: Overview of the Clinical Decision Support System (CDSS) architecture emerging from the SB-SVM approach. The Gerneral Practitioner (GP) stores the EHR data in Netmedica Italia (NMI) Cloud platform. The FIMMG dataset is composed of three different fields: demographic, monitoring and clinical. The related features were used for training the SB-SVM model and providing a T2D prediction.

prescribed.

2.3.1.2. Case II

The following configuration setup (*Case II*) has been obtained discarding all exam and drug prescriptions collected after the T2D diagnosis. In particular, all ATC code A10 drugs used for diabetes have been excluded. Table 2.2 summarizes all discarded features for Case II and Case III.

For the purpose of predicting T2D early stage or an undiagnosed status, the previously removed features would excessively reveal the related pathology and affect the training of the predictive model. Table 2.1 shows the number of features considered for the Case II and Case III.

2.3.1.3. Case III

The chronological age is one of the most discriminative features of T2D according to previous findings [24] and as will be evidenced by the results in Figure 2.5b and Table

Table 2.2.: Discarded features for Case II and Case III.

Exam prescriptions
Gfr using MDRD formula, Glycaemia, HB1Ac, Microalbuminuria.
Drug prescriptions
Acarbose, Exenatide, Glibenclamide, Gliclazide, Glimepiride, Insulin aspart, Insulin glargine, Insulin glulisine, Insulin lispro, Linagliptin, Metformin, Metformin and linagliptin, Metformin and sitagliptin, Metformin and sulphonylurea, Metformin and empagliflozin, Pioglitazone, Repaglinide.

2.5. A uniform age distribution of subjects was analyzed, where it is not possible to reject the null hypothesis ($\alpha = 0.05$) that the age and the T2D condition comes from independent random samples from Gaussian distributions with equal means and equal unknown variances. Hence, the middle range interval $60-80$ years is selected, since it is not displayed statistically difference between age and T2D (i.e., $t_{533} = 1.267, p = 0.206$).

2.4. Methods

Given a training dataset $\sum_{i=1}^{m}(x_i, y_i)$ of m observations and n features, the input $x_i \in \Re^n$ is the feature vector and the output $y_i \in \{-1, 1\}$ is the class label. SVM is a non-probabilistic kernel-based decision machine which leads to a sparse solution. The estimation of model parameters corresponds to a convex optimization problem and the prediction of new inputs depends only on the kernel function evaluated on the subset of the training data points, named support vectors [61]. These properties allow to reduce the computational effort while improving the algorithm performance.

2.4.1. Background: 2-norm SVM

In the SVM problem the aim is to find a separating hyperplane: $\boldsymbol{w}^T \mathrm{x} + \boldsymbol{w}_0 = 0$, which maximizes the margin $\frac{1}{\|\boldsymbol{w}\|^2}$, that is defined to be the smallest distance between the decision boundary and any of the training points. Since the class-conditional distribution may not be linearly separable, the exact separation of the training data can lead to poor generalization. Thus, the general idea is to allow some of the training data points to be misclassified, allowing to overlap class distribution, by introducing the slack variables $\xi_i \geq 0$ where $i = 1, \ldots, M$. Hence, the SVM formulation can be seen

as the optimization of the "soft margin" loss function [61]:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{w}_0,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{M}\xi_i \\ \text{s.t.} \quad & \\ & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \boldsymbol{w}_0) \geq 1-\xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{2.1}$$

where $\boldsymbol{w}_0$ is the bias parameter and $C > 0$ is the box constraint and controls the overlapping between the two classes. The procedure for solving (2.1) is to compute a Lagrange function from the cost function and the corresponding constraints, by introducing a dual set of variables [61].

The SVM optimization problem of (2.1) can be written also in the form of error function:

$$\min_{\boldsymbol{w},\boldsymbol{w}_0} \quad \sum_{i=1}^{m}\Big[1-y_i\Big(\boldsymbol{w}_0+\sum_{j=1}^{q}w_j h_j(x_i)\Big)\Big]_+ + \lambda\|\boldsymbol{w}\|^2 \tag{2.2}$$

where λ is a tuning parameter that is inversely proportional to the box constraint C and $D = \{h_1(x),\ldots,h_q(x)\}$ is a set of basis functions that are usually chosen in the reproducing kernel Hilbert Space. Thus, the kernel trick allows the dimensions of the transformed features space to be very large, even infinite in some cases [62]. Note that the following optimization problem (2.2) has the form of *loss + penalty* and λ controls the trade-off between loss and penalty as well as between bias and variance. The loss function $(1-yf)_+$ is called the *hinge* loss while the 2-norm penalty is called the *ridge* penalty [62].

2.4.2. Sparse 1-norm SVM

In this work the *ridge* penalty was replaced with the 1-norm of $\boldsymbol{w}$, i.e., the *LASSO* penalty [63]. This penalty induced sparse solution in the model coefficients. The application of 1-norm SVM was widely used for solving high dimensional task while increasing sparsity as well as the interpretability of the model [64]. The considered sparse 1-norm SVM has the following optimization problem [62]:

$$\min_{\boldsymbol{w},\boldsymbol{w}_0} \quad \sum_{i=1}^{m}\Big[1-y_i\Big(\boldsymbol{w}_0+\sum_{j=1}^{q}w_j h_j(x_i)\Big)\Big]_+ + \lambda\|\boldsymbol{w}\| \tag{2.3}$$

This formulation combines the hinge loss with an l_1-constraint, bounding the sum of the absolute values of the coefficients. Note that for a high regularization term λ can encourage the sparsity of the model by driving some coefficients exactly to zero, so that, irrelevant features can be automatically removed from the model. This shrinkage has the effect of controlling the variances of the model coefficient, avoiding

overfitting, and improving the generalization performance especially when there are many highly correlated features. In this way it provides an automatic way for doing model selection in linear model [65]. Accordingly, the LASSO penalty corresponds to a double-exponential prior for the coefficients, while the ridge penalty corresponds to a Gaussian prior [62]. This reflects the greater tendency of the 1-norm SVM to produce some largely model coefficients in terms of magnitude and leave others at 0. In order to solve the optimization problem described in (2.3) the Sparse Reconstruction by Separable Approximation (SpaRSA) solver [66] was employed. The SpaRSA algorithmic framework [66] exploits the proximity operator used in [67] for solving large-scale optimization problems involving the sum of a smooth error term and a possibly nonsmooth regularizer. In this context, the proximal methods are specifically tailored to optimize an objective function of the form (2.3), gaining faster convergence rates and higher scalability to large nonsmooth convex problems [68]. In particular, the advantages of SpaRSA with respect to other competitors [69, 70] are resumed in [66].

It is worth noting here that when the predictor matrix X is not of full column rank, the LASSO solutions are not unique, because the criterion is not strictly convex [65]. The non-full-rank case can occur when $n > m$ or when $n \leq m$ due to collinearity of the EHR features [65]. The numerical algorithm (e.g., SpaRSA, iterative shrinkage thresholding and coordinate descent-based algorithms) can therefore compute valid solutions in the non-full rank case [65].

2.4.3. Sparse Balanced SVM

The margin of the predicted SB-SVM response was mapped into [0-1] interval by using a sigmoid function, without changing the SVM error function. The mapping was realized according to [71], adding a post-processing step where the sigmoid parameters were learned with regularized binomial maximum likelihood. The computed probabilistic outputs of SB-SVM reflects the predicted response (y_p) based on the threshold $th = 0.5$:

$$\begin{aligned} &P(y_p = T2D|f) = \frac{1}{1+exp(Af+B)} > th \quad y_p = T2D \\ &else \quad y_p = Control \end{aligned} \tag{2.4}$$

Several solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels [72, 73]. Differently, from the data-level solutions which include many different forms of re-sampling (e.g., random re-sample, directed re-sample, oversample with informed generation), our method works at the algorithm level. In particular, the decision threshold th was adjusted in the validation set in order to maximize the *macro-recall* metric while alleviating the effect of high unbalanced data. The prediction of the SB-SVM produces an uncalibrated value that is not a

probability. The post-processing step allows to transform the output of the SB-SVM classifier (i.e., distance from the margin) into posterior probability. Thus, the posterior probability represents a salient information which can be integrated in a CDSS for supporting the early-stage diagnosis by revealing the confidence level of the performed prediction. The idea behind the employed methodology [74] lies in the use of a parametric model (i.e., sigmoid model, see Eq. 2.4) to fit the posterior directly. Hence, the parameters A and B of the sigmoid function are adapted to give the best probability outputs [74]. Differently from [75] the employed sigmoid model has two parameters trained discriminatively, rather than one parameter.

The general idea behind the application of the proposed approach lies in the two main challenges that EHR data present: (i) high dimensional data with several irrelevant/noisy features and high degree of redundancy and (ii) natural unbalanced setting of this task. Starting from this motivations the introduction of 1-norm LASSO regularizer allow to induce sparsity while increasing the interpretability of the linear model. This is a salient information which may lead to understand not only the predicted outcome but also why and how the prediction was made. Additionally the LASSO regularizer may improve the accuracy prediction by reducing the variance of the predicted class by shrinking some coefficients to zero [65]. Accordingly, the shift of the decision threshold over the estimated posterior probability favours the minority class, increasing the recall rate over all the predicted testing set.

2.4.4. Experimental procedure

The proposed SB-SVM model was tested in three different scenarios according to the Data Analysis described in Section 2.3.1 in order to answer the hypothesis previously described.

Additionally, the proposed SB-SVM model was compared with respect to other ML approaches already used to solve this task (i.e., DT [24, 50, 52], RF [50, 52, 54], LR [49, 24], KNN [50], SVM Lin and Gauss [49, 50, 53]) and DL techniques (i.e., multi layer perceptron (MLP) and deep belief network (DBN)) already employed for the prediction of heart failure [76] and osteoporosis [77] respectively. The performed comparisons also include the data preprocessing comprised of features selection and data level solution for dealing with the nature unbalanced setting of the task (Table 2.4).

The assessment of the introduced ML model was performed according to the following measures:

- *Accuracy*: the percentage of correct predictions;
- *Macro-precision*: the percentage of true positive over the predicted condition positive (positive predicted value). The precision is calculated for each class and then take the unweighted mean.

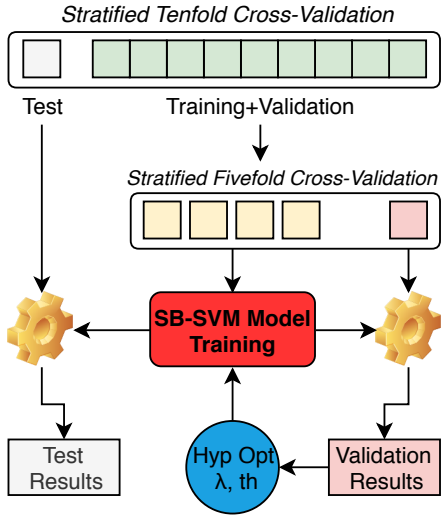


Figure 2.2.: Overview of the SB-SVM model architecture. A Tenfold Cross-Validation procedure was executed. The optimization of the SB-SVM hyperparameters was performed implementing a grid-search and optimizing the macro-recall score in a nested stratified Fivefold Cross-Validation. Hence, each split of the outer loop was trained with the optimal hyperparameters tuned in the inner loop.

- *Macro-recall*: the percentage of true positive over the condition positive (true positive rate or sensitivity). The recall is calculated for each class and then take the unweighted mean.
- *Macro-F1*: the harmonic mean of precision and recall averaged over all output categories.
- *Receiver Operating Characteristic (ROC)*: is designed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- *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.
- l^0 measure: it is employed as a measure of model sparsity [78]. It calculates the number of zero coefficients of the model: $\#\{j, coeff_j = 0\}$.

From now on the *Macro-precision*, *Macro-recall* and *Macro-F1* were referred as *Precision*, *Recall* and *F1* respectively.

The overview of the SB-SVM model architecture is shown in Figure 2.2. In both experiments a stratified Tenfold Cross-Validation (10-CV) was computed over subjects

procedure. The optimization of SB-SVM hyperparameters (i.e., λ, th) was performed implementing a grid-search and optimizing the *macro-recall* score in a nested stratified Fivefold Cross-Validation. Hence, each split of the outer loop was trained with the optimal hyperparameters tuned in the inner loop. Although this procedure is computationally expensive, it allows to obtain an unbiased and robust performance evaluation [79].

The regularization factor λ was picked inside the subset $\{0.001, 0.01, 0.1, 1, 10\}$, while the threshold was picked inside the subset $\{0, 0.01, 0.02, \ldots, 1\}$. Table 2.3 shows the different hyperparameters for all competitors' approaches, as well as the grid-search set.

Since the aim is to find the best threshold value of posterior probability/margin of the classifier, the threshold is a common hyperparameter for all tested methods, except for those using SMOTE or 1-norm SVM (i.e., the approaches which have been already formulated to be consistent with the unbalanced setting).[2].

The stability or reproducibility of the model is also an important aspect to be considered, especially in case of sparse solutions which can lead to unstable models [80]. This aspect was studied and reported by measuring the variance of the selected SB-SVM hyperparameters. Experimental results are reported in the next section.

Table 2.3.: Range of Hyperparameters (Hyp) for the proposed Sparse Balanced-Support Vector Machine (SB-SVM) model and other tested approaches: Linear SVM (SVM Lin), Gaussian SVM (SVM Gauss), K-nearest neighbor (KNN), decision tree (DT), random forest (RF), logistic regression (LR) ridge, smoothly clipped absolute deviation (SCAD) SVM, 1-norm SVM, multi layer perceptron (MLP) and deep belief network (DBN). The threshold hyperparameter *th* is optimized for each model except for SMOTE-based approaches and 1-norm SVM (i.e., the approaches which have been already formulated to be consistent with the unbalanced setting).

Work	Model	Features selection	Resampling	Hyp	Range
[49, 24, 50, 53]	SVM Lin	Ttest, ReliefF, RFE-SVM	none	Box Constraint	$\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$
[49, 24, 50, 53]	SVM Gauss	Ttest, ReliefF	none	Box Constraint Kernel Scale	$\{10^{-2}, 0.1, 1, 10, 10^{2}, 10^{3}, 10^{4}\}$ $\{10^{-2}, 0.1, 1, 10, 10^{2}, 10^{3}, 10^{4}\}$
[50]	KNN	none	none	n° of neighbors	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
[24, 50, 52, 55]	DT	none	SMOTE	max n° of splits	$\{10, 20, 30, 40, 50\}$
[50, 52, 54]	RF	none	SMOTE	n° of weak learners	$\{10, 20, 30, 40, 50\}$
[49, 24]	LR ridge	Ttest, ReliefF, RFE-SVM	none	λ	$\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$
	SB-SVM	none	none	λ	$\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$
[56]	SCAD SVM	none	none	λ	$\{0.10, 0.55, 1, 1.45, 1.90\}$
[57]	1-norm SVM	none	none	λ μ	$\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$ $\{0.1, 0.2, ..., 0.9\}$
[76, 77]	MLP, DBN	none	none	learning rate n° of hidden layers n° of units	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1\}$ $\{1, 2, 4, 8, 16\}$ $\{16, 32, 64, 128, 256\}$

[2]Matlab code to reproduce all results (and modify the setting) is available as supplementary material (Code Ocean platform)

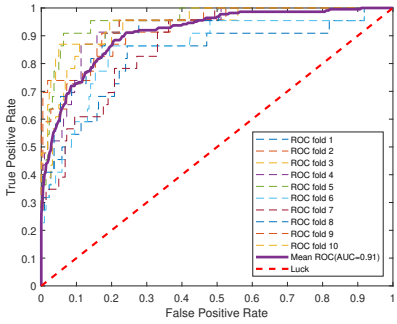
2.5. Experimental results

Our approach was tested using the FIMMG dataset described in Section 2.3 according to the Data analysis reported in Section 2.3.1 which aims to answer the three research questions previously described.

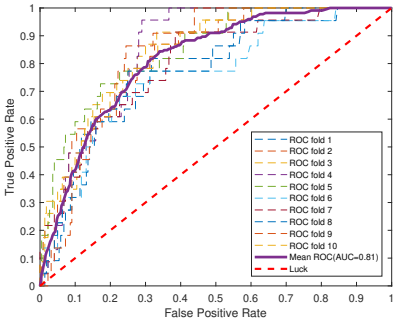
2.5.1. Case I

The the SB-SVM predictive performance for all cases is reported in Table 2.4. In particular, the higher results were obtained in Case I, where all features as well as all subjects were considered. Since *recall* was optimized in the validation set, it achieved best performance when compared to *precision*.

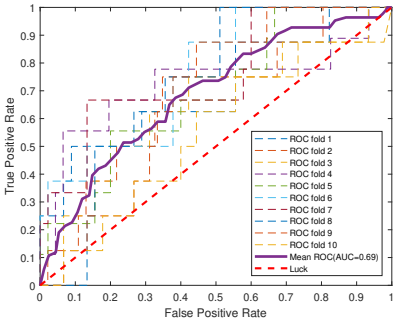
The ROC curves for each fold are shown in Figure 2.3a as well as the averaged curve. All points of ROC curves are above the chance level (i.e., red dotted line).



(a) Case I



(b) Case II



(c) Case III

Figure 2.3.: SB-SVM performance in terms of ROC curves.

The *recall* and *AUC* of SB-SVM are compared with respect to other state-of-the-art ML and DL approaches applied for T2D prediction (Table 2.4).

The *recall* and *AUC* for the SB-SVM are significantly higher ($p < .05$) than all baseline models, except than DT (*recall*: $t_{18} = 0.514$, $p = .61$; *AUC*: $t_{18} = 1.652$, $p = .12$) and RF (*recall*: $t_{18} = 0.514$, $p = .09$). Also the ROC curves confirm that the SB-SVM is above the other baseline models (Figure 2.4a). The SB-SVM considerably overcomes ($p < .05$) both SCAD SVM, MLP, resampling and RFE-SVM based models. On the contrary, the *recall* and *AUC* for the SB-SVM are statistically comparable with respect 1-norm SVM, DBN, Ttest and ReliefF based models.

Table 2.4.: SB-SVM: Comparison with other state-of-the-art approaches.

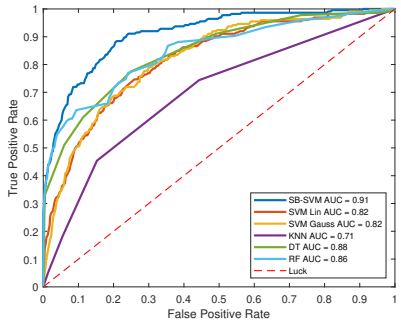
		CASE I		**CASE II**		**CASE III**	
Work	**Model**	**Recall%**	**AUC%**	**Recall%**	**AUC%**	**Recall%**	**AUC%**
	Baseline	mean std	mean std	mean std	mean std	mean std	mean std
[49, 50, 53]	SVM Lin	74.12 (4.02)	81.68 (5.60)	71.29 (3.65)	78.99 (4.30)	58.55 (5.80)	64.48 (5.39)
[49, 50, 53]	SVM Gauss	71.96 (4.22)	81.98 (4.84)	68.34 (4.41)	76.29 (4.40)	55.62 (8.08)	62.74 (9.22)
[50]	KNN	69.23 (4.97)	70.97 (5.06)	68.56 (5.57)	71.04 (6.09)	54.50 (7.16)	59.80 (8.10)
[24, 50, 52]	DT	80.99 (3.34)	87.79 (4.17)	72.98 (4.54)	77.56 (4.85)	58.98 (8.37)	61.87 (7.77)
[50, 52, 54]	RF	77.81 (5.66)	86.30 (4.24)	68.08 (6.36)	75.70 (4.61)	57.33 (5.74)	61.96 (9.47)
	Sparse SVM						
	SB-SVM	81.89 (4.03)	91.04 (4.16)	**74.64 (4.18)**	**81.43 (3.20)**	**65.33 (5.69)**	**68.90 (5.84)**
[56]	SCAD SVM	67.61 (4.41)	70.78 (4.20)	54.98 (4.09)	60.09 (4.13)	50.83 (9.97)	54.08 (10.41)
[57]	1-norm SVM	82.47 (3.47)	90.21 (3.65)	71.10 (4.27)	77.46 (5.76)	60.73 (7.15)	65.35 (8.38)
	Resampling						
[55]	DT + SMOTE	75.79 (4.72)	82.03 (2.73)	67.07 (3.06)	67.57 (4.26)	57.77 (8.67)	60.73 (10.35)
[54]	RF + SMOTE	71.63 (4.93)	86.34 (4.07)	58.15 (4.34)	77.10 (3.84)	57.66 (6.15)	68.57 (7.06)
	Features selection						
[49]	Ttest + LR ridge	80.91 (2.90)	89.81 (3.35)	73.14 (3.36)	78.89 (4.58)	61.35 (3.11)	67.47 (6.81)
[49]	Ttest + SVM Lin	76.81 (3.11)	88.99 (4.02)	72.42 (3.67)	79.00 (4.32)	54.07 (4.36)	60.56 (8.47)
[49]	Ttest + SVM Gauss	78.49 (3.07)	85.87 (4.34)	73.78 (2.62)	80.39 (4.02)	54.65 (7.23)	62.58 (5.15)
[49, 24]	ReliefF + LR ridge	83.02 (4.09)	91.39 (3.68)	74.03 (4.84)	80.34 (3.13)	57.54 (9.20)	66.66 (5.38)
[49, 24]	ReliefF + SVM Lin	**84.21 (3.24)**	91.24 (3.34)	74.36 (3.50)	81.01 (2.71)	58.23 (6.85)	66.16 (7.63)
[49, 24]	ReliefF + SVM Gauss	83.90 (3.15)	**91.85 (2.97)**	74.11 (2.38)	80.74 (1.84)	59.77 (5.27)	65.74 (6.53)
[49]	RFE-SVM + LR ridge	72.43 (5.27)	72.54 (5.31)	52.64 (1.51)	52.81 (1.67)	56.26 (4.12)	56.28 (4.24)
[49]	RFE-SVM + SVM Lin	71.87 (5.46)	72.26 (5.14)	52.27 (1.83)	52.31 (1.83)	55.23 (3.33)	55.50 (3.34)
	Deep Learning						
[76]	MLP	67.90 (3.55)	77.53 (4.31)	58.52 (5.43)	67.03 (6.31)	54.25 (5.37)	56.89 (7.72)
[77]	DBN	77.23 (4.23)	89.32 (3.47)	66.82 (5.91)	78.50 (6.97)	61.22 (10.26)	66.78 (14.68)

Once evaluated and compared the performance of the proposed SB-SVM method in order to discriminate between healthy and T2D patients, the aim is to identify which features contribute to the decision.

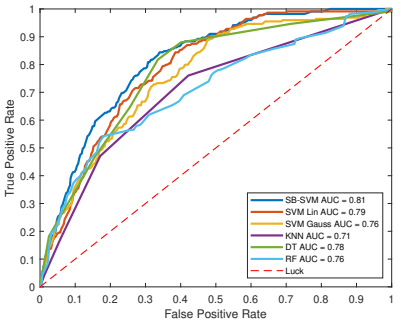
The SB-SVM offers a natural way for addressing this goal, in fact, compared to other approaches (i.e., SVM Gauss, MLP, DBN) the model is linear and easily interpretable. The computed l^0 measure is 0.39, while the magnitude of the SB-SVM coefficients is shown in Figure 2.5a.

The stability of the model was measured according to the variance of the SB-SVM hyperparameters (Var(λ)=1.44×10^{-5}, Var(th)=2.93×10^{-3}).

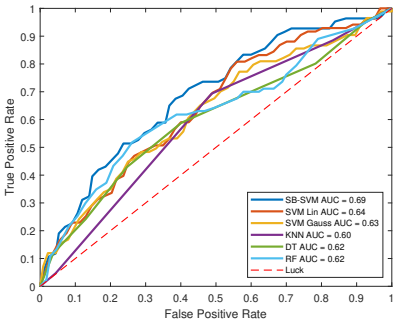
The top 10-rank features are summarized in Table 2.5 according to the magnitude of SB-SVM coefficients.



(a) Case I



(b) Case II



(c) Case III

Figure 2.4.: Performance comparison in terms of baseline models ROC curves.

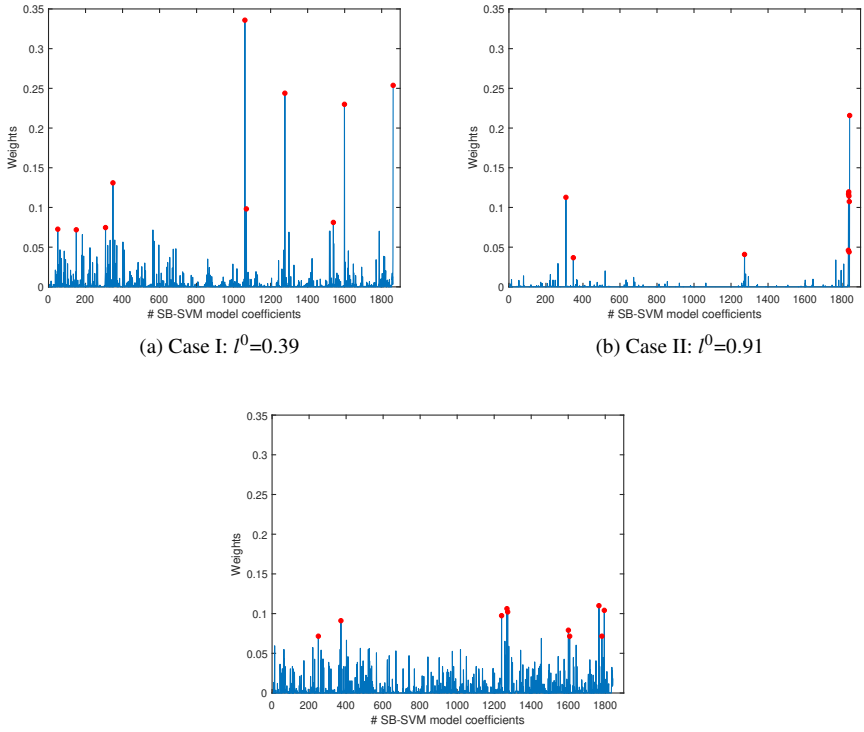
2.5.2. Case II

Due to the increasing difficulty of the classification task, *precision*, *recall*, and *AUC* decrease respectively of 5.48%, 7.25%, and 9.61%. However, all metrics are above ($p < .05$) chance level (0.5).

The ROC curves for each fold are shown in Figure 2.3b as well as the averaged curve. All points of ROC curves are above the chance level.

In the case II, once the task complexity has increased after discarding the features reported in Table 2.2, the *recall* and *AUC* of SB-SVM are greater than all the state-of-the-art ML and DL approaches (Table 2.4). Regarding the baseline models, the DT (*recall*: $t_{18} = 0.807$, $p = .43$; *AUC*: $t_{18} = 2.002$, $p = .06$) and SVM Lin (*recall*: $t_{18} = 1.811$, $p = .09$; *AUC*: $t_{18} = 1.369$, $p = .19$) are the closest models to SB-SVM. Also the ROC curves confirm that the SB-SVM is above the other baseline approaches (Figure 2.4b).

The performance of SB-SVM is superior but statistically comparable with respect to 1-norm SVM, DBN, ReliefF and Ttest features selection based models. On the



(a) Case I: l^0=0.39

(b) Case II: l^0=0.91

(c) Case III: l^0=0.57

Figure 2.5.: Magnitude of the SB-SVM coefficients and l^0 measure values. Top 10-rank features are pointed out with red spots.

contrary, SCAD SVM, MLP, resampling and RFE-SVM based models continue to be very far ($p < .05$) from the SB-SVM.

The interpretability of the SB-SVM model is increased according to the l^0 measure (0.91), while the magnitude of the SB-SVM coefficients is depicted in Figure 2.5b. The top 10-rank features are summarized in Table 2.5 according to the magnitude of the SB-SVM coefficients. The lower Var(λ)=3.34×10^{-36} and Var(th)=8.44×10^{-5} in the selection of the SB-SVM hyperparameters outlined the high stability of the model.

2.5.3. Case III

In the Case III, also a uniform distribution of subjects where age is no longer statistically significant for T2D prediction was considered. The *precision*, *recall* and *AUC* still remain above ($p < .05$) the chance level (0.5).

The ROC curves for each fold are shown in Figure 2.3c as well as the averaged

curve. All points of ROC curves are above the chance level.

The *recall* and *AUC* of SB-SVM continues to be greater than all the state-of-the-art ML and DL approaches (Table 2.4). The gain of the SB-SVM is increased especially in terms of *recall* with respect to other ML models (see Table 2.4). This aspect highlights the robustness of the proposed algorithm to maximize the *recall* while dealing with the natural unbalanced setting of the dataset. The performance of SB-SVM is significantly higher ($p < .05$) both in terms of *recall* and *AUC* compared with respect to the all baseline models. Instead, for what concern the other approaches, it is significantly greater ($p < .05$) than SCAD SVM, MLP, DT + SMOTE and RFE-SVM based models.

The magnitude of the SB-SVM coefficients is shown in detail in Figure 2.5c. The sparsity measure l^0 of the model is 0.57. The top 10-rank features of SB-SVM are summarized in Table 2.5. The variance in the selection of the SB-SVM hyperparameter is reasonably low (Var(λ)=1.89×10^{-5}, Var(th)=3.85×10^{-3}).

Table 2.5.: Top 10-rank features according to the SB-SVM magnitude coefficients: Blood Pressure (BP), Drugs (D), Exam prescriptions (EP), Exemptions (E), Pathologies (P).

Rank	Case I	Case II	Case III
1	HbA1c (*EP*)	Age	Arterial hypertension(stage II, III) (*E*)
2	Age	Mean diastolic (*BP*)	Weight (*EP*)
3	eGFR(MDRD formula) (*EP*)	Max diastolic (*BP*)	Arterial hypertension (*E*)
4	Metformin (*D*)	Mean systolic (*BP*)	Creatinine clearance (*EP*)
5	Heart failure (*P*)	Arterial hypertension (*P*)	Fundus oculi (*EP*)
6	Microalbuminuria (*EP*)	Max systolic (*BP*)	Aorta aneurysm (*P*)
7	Insulin glargine (*D*)	Min diastolic (*BP*)	Moxifloxacin (*D*)
8	Arterial hypertension (*P*)	Min systolic (*BP*)	Myasthenia gravis (*P*)
9	Hyper/Dyslipidaemia (*P*)	Creatinine clearance (*EP*)	Netilmicin (*D*)
10	Cancer pancreas (*P*)	Heart failure (*P*)	Myasthenia gravis (*E*)

In summary, the SB-SVM achieves always the greater *recall* and *AUC* over the other state-of-the-art models (Table 2.4) for the more challenging data analysis (i.e., Case II, Case III).

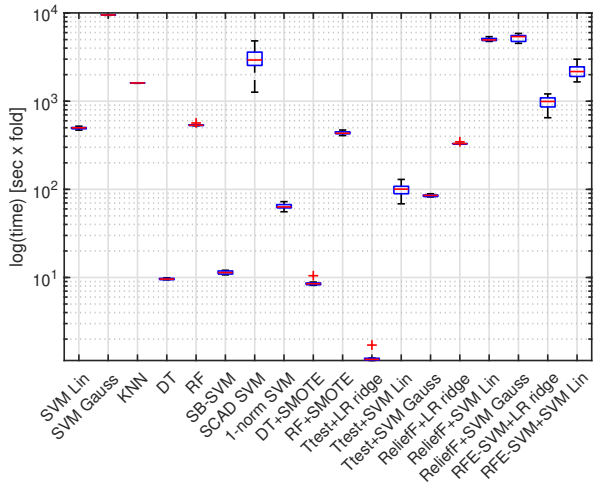
The motivation behind the selection of different top features lies in the substantial divergence between each case which represents a different performed task (Table 2.5).

2.5.4. Computation time analysis

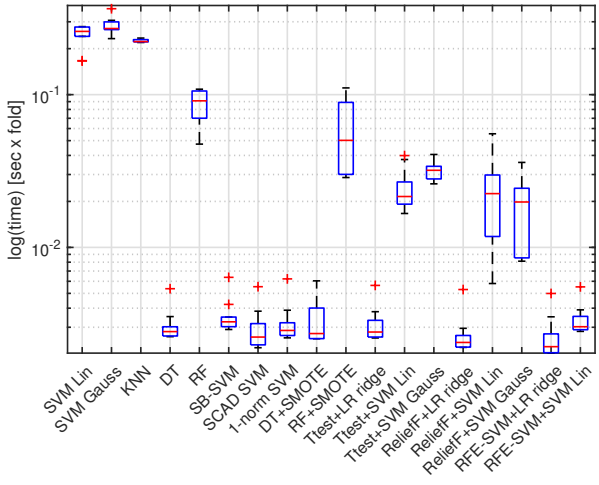
The computation time analysis for the training and validation stage is performed in Figure 2.6a, while the computation time for the testing stage is shown in Figure 2.6b.

The training stage of the features selection models which include a wrapper (i.e., RFE-SVM) or advanced filter approaches (i.e., ReliefF) is more time consuming than the SB-SVM. In particular the computation training effort of SB-SVM is lower than other linear (i.e., SVM Lin, Ttest + SVM Lin, ReliefF + SVM Lin, RFE-SVM + SVM Lin and 1-norm SVM), non-linear (i.e., SVM Gauss, Ttest + SVM Gauss and ReliefF

+ SVM Gauss) SVM-based approaches and ensemle-based classifiers (i.e., RF and RF + SMOTE). All the Sparse SVM approaches achieved a competitive computation time for the testing stage.



(a) Preprocessing, training and validation time



(b) Testing time

Figure 2.6.: Comparison in term of computation time.

2.6. Discussion

The proposed approach and the results shown before brings to the following main discussion points.

2.6.1. Clinical perspective

Faced with an often difficult and complicated management of the diabetic patient, the gold-standard diagnosis of T2D remains uncertain and challenging. In this condition, a CDSS solution which is able to provide a reliable prediction of T2D becomes essential in order to target timely and appropriately prevention strategies. This possibility becomes more interesting when the ML algorithm is able to identify a cluster of patients at high risk of developing the disease. Starting from these clinical motivations a ML approach, named SB-SVM, was proposed, able to predict T2D using the novel FIMMG dataset available at the following link[3].

2.6.2. SB-SVM effectiveness and outcomes

The experimental results demonstrate how the early stage T2D prediction is enclosed within EHR features. However, the discriminative power of each feature is not uniform, but sparsely distributed. In fact, the SB-SVM overcomes the other state-of-the-art approaches, particularly for the most challenging data analysis (i.e., Case II, Case III). The better recall and AUC suggest how the proposed SB-SVM methodology is a viable and natural solution to model the high dimensional EHR data while discarding noisy and/or redundant features. The LASSO regularizer is proven to be an effective strategy for dealing with high level of noise while significantly reduces both model error and model complexity [58, 65]. The FIMMG dataset comprises several noise components: i) missing values; ii) outliers; iii) unreadable encrypted information for privacy preserving; iv) GP's transcription typos (e.g., ICD-9 codes) and v) non-uniformity in the definition of certain medical examinations or pathologies.

The better effectiveness, as well as the higher interpretability of the SB-SVM model, are the key advantages of our methodology with respect to other ML and DL competitors. This is a salient information which leads to understand not only the predicted outcome, but also why and how the prediction has been made. On the other hand, DL approaches often lack this sort of algorithmic transparency. While the heuristic optimization procedures for neural networks are demonstrably powerful, it is not easy to understand how they work, and what inputs are most relevant for the prediction [43]. Accordingly, the experimental results outline some peculiarity of SB-SVM with respect to other state-of-the-art work: (i) the algorithm is robust to the natural high unbalanced setting (i.e., several healthy subjects and few T2D subjects) without drawing synthetic examples or employing additional resampling strategies, such as in [54];

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

(ii) the model is linear and easy interpretable due to the induced sparsity; (iii) it is able to manage high dimensional data implicitly selecting the most relevant features without requiring a further features selection [47, 48, 49, 24] or engineering features techniques [50]; (iv) the sparse solution is stable across the considered dataset and (v) the proposed classifier is trained only from a subset of EHR features recorded before the T2D diagnosis (i.e., Case II, Case III). In particular, all ATC code A10 drug prescriptions and exam prescriptions shown in Table 2.2 (i.e., Case II) were excluded. Simultaneously, a uniform age sample group was also selected, where the age does not enclose anymore a relevant information for discriminating the T2D condition (i.e., Case III).

The SB-SVM approach may be easily generalized for multi-class problem and for regression task. The multi-class problem can be set by combining multiple two-class SB-SVMs in order to build a multiclass classifier. This step can be performed by using the one-versus-the-rest or the one-versus-one approaches [81]. Currently, the performances of the SB-SVM are being tested in other EHR datasets available in the literature, with an increasing number of heterogeneous features and a higher number of subjects.

Another future direction may be the application of non-linear Kernel as Polynomial or Gaussian in order to map the features in non-linear space. This problem can be seen as constructing non-linear regression models with Gaussian basis functions instead of linear basis functions, using LASSO regularization [82]. Imposing LASSO in the non-linear model means to select not the original features but the most relevant non linear basis functions.

2.6.3. Pattern discrimination and localization

The main research questions previously answered are whether the SB-SVM is able to discover the T2D according to all set of EHR features (Case I), a subset of EHR features collected before the T2D diagnosis (Case II), even considering a uniform age group of subjects (Case III). This work focuses on the introduction of a ML model for pattern discrimination to answer these research questions. However, once the reliability and the robustness of our approach have been detected, it may possible to go ahead answering where the discriminatory information is encoded. The SB-SVM is able to implicitly localize the discriminative pattern while identifying the most relevant features for providing an accurate T2D prediction. These features were summarized in Table 2.5 and represent a salient information which can be exploited in order to support early stage diagnosis. It turns out that some of the top 10-rank features selected by the SB-SVM model for both experimental cases are consistent with those reported in the state-of-the-art regarding the T2D risk factors [45, 46]. In particular for the Case III, recent studies [83, 84] are also confirming the possibility that the antibiotics (e.g., Moxifloxacin, Netilmicin) exposure could increase the T2D risk.

2.6.4. Clinical impact

These experimental setups assume a considerable significance in the clinical use case, where a CDSS should provide a prediction in order to support the early-stage diagnosis while learning hidden patterns sometimes unsighted by physicians. Experimental results provide the evidence of the robustness of the SB-SVM methodology in the clinical scenario. Furthermore, the high stability of the model selection criterion outlines the great reproducibility as well as the high impact of the presented results for the considered dataset. The model is able to generalize across unseen subjects while selecting consistently the most relevant features over the whole dataset. However, the sparsity of the model coefficients is likely to be dataset/task dependent (it may change across different EHR dataset and different task). Moreover, the sample size and the disease heterogeneity of the employed EHR data may limit the generalization power of the study. Future work may be addressed to increase the number of subjects included in the FIMMG dataset while considering additional clinical features. From the clinical perspectives, the SB-SVM model may be useful also for the prediction of different pathological conditions (e.g., cardiovascular and neurological diseases). In addition to the resources offered by the so-called predictive medicine, in which the point analysis of genetic and biological components represent the constitutive elements of forthcoming widespread implementation, ML approaches that use EHR clinical data could offer and anticipate early care strategies.

Chapter 3.

Insulin resistance: Clinical factors

Insulin resistance (IR) is an early-stage deterioration of type 2 diabetes (T2D). Identification and quantification of IR requires specific blood tests; however, the triglyceride-glucose (TyG) index can provide a surrogate assessment from routine Electronic Health Record (EHR) data. Since IR is a multi-factorial condition, to improve its characterisation, this study aims to discover non-trivial clinical factors in EHR data to determine where the IR condition is encoded.

A high-interpretable Machine Learning approach (i.e., ensemble Regression Forest combined with data imputation strategies), named TyG-er was proposed. Three different experimental procedures were applied to test TyG-er reliability on the Italian Federation of General Practitioners dataset, named FIMMG_*obs* dataset, which is publicly available and reflects the clinical use-case (i.e., not all laboratory exams are prescribed on a regular basis over time).

Results detected non-conventional clinical factors (i.e., uricemia, leukocytes, gamma-glutamyltransferase and protein profile) and provided novel insight into the best combination of clinical factors for detecting early glucose tolerance deterioration. The robustness of these extracted clinical factors was confirmed by the high agreement (from 0.664 to 0.911 of Lin's correlation coefficient (r_c)) of the TyG-er approach among different experimental procedures. Moreover, the results of the three experimental procedures outlined the predictive power of the TyG-er approach (up to a mean absolute error of 5.68% and $r_c = 0.666, p < .05$).

The TyG-er approach is able to carry information about the identification of the TyG index, strictly correlated with the IR condition, while extracting the most relevant non-glycemic features from routine data.

3.1. Introduction

T2D is a widespread disease. It is estimated that it will affect approximately 693 million people worldwide by 2045 [85]. Due to the high prevalence and costs associated with management of T2D and its related complications, early identification of subjects at risk of developing T2D represents a key issue in public-health policy. Indeed, the later T2D onset is identified, the more aggressive and the more adapted to

the pathophysiological stage of T2D development the intervention should be [86]. In this context, General Practitioners (GPs) represent the first medical contact who may provide early identification of subjects at T2D risk.

T2D is characterised by a deterioration of glucose tolerance, thus resulting in an increased blood glucose concentration, termed hyperglycemia. Hyperglycemia arises in overt T2D and in intermediate pre-diabetic states; however, at an early stage of glucose tolerance deterioration, hyperglycemia has not yet occurred and the main alteration is represented by IR (i.e., a reduced sensitivity of tissues to insulin action in lowering blood glucose concentration), compensated by increased insulin secretion [87]. Differently from hyperglycemia, the correct identification and quantification of IR condition is not straightforward since it requires specific blood tests that are not included in those usually ordered by GPs in routine check-ups [88]. However, it is possible to provide a surrogate assessment of an IR condition by the TyG index [89, 90, 91], based on routine triglyceride and glucose measurements.

Since IR is a multi-factorial condition, identification of additional routine clinical factors (i.e., different from triglyceride and glucose measurements) could improve the characterisation of this condition and the effectiveness of early identification of subjects at T2D risk [92].

Machine Learning algorithms can be used to analyze EHR data to discover complex patterns and set up powerful models for GPs to screen the patient population and identify subjects at T2D risk [93]. Several studies, employing ML or model-based approaches, focused on management of T2D pathology [50, 52, 24, 93, 53, 2, 94, 95] but not on the identification of clinical factors, among routine measurements, of IR.

Starting from a clinical motivation and a gap in scientific literature, the present study aims to discover non-trivial clinical factors in EHR data to determine where the IR condition is encoded. To achieve this aim, a high-interpretable ML regression approach, named TyG-er approach, was proposed.

3.2. Clinical data: FIMMG_*obs* dataset

FIMMG_*obs*[1] is a publicly available collection of data stored by a single GP that includes demographic, monitoring and laboratory exam data. The FIMMG_*obs* dataset (see Tab. 3.1) includes 968 patients not affected by T2D. The longitudinal patient observational time-period was from 2010 to 2018. During this period, each patient (identified by an *id*) underwent multiple triglycerides (TG; mg/dl) and glycemia (Gb; mg/dl) measurements, acquired simultaneously and in fasting conditions, thus resulting in a sequence of TG_i, Gb_i pairs. For each *id*, a *seq* was defined as $\{1,\ldots,i,\ldots,t\}$ where t is the total number of pairs. For each i belonging to *seq*, the TyG_i was com-

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

Table 3.1.: FIMMG_*obs* dataset overview.

	Dataset description	**Count**	**Mean (Std)**
	Total patients	968	-
	Observation period (*years*)	9	-
	Total observations	2276	-
	Fields	**Count**	**Mean (Std)**
	Demographic	2	
	Gender:		
	Male	473	-
	Female	495	-
	Age (*years*)	-	61(±18)
	Monitoring	5	
	Blood pressure (*mmHg*):		
	Systolic	-	135(±16)
	Diastolic	-	82(±9)
	Height (*cm*)	-	160(±16)
	Weight (Kg)	-	83(±17)
	Body Mass Index (Kg/m^2)	-	32(±5)
#	Laboratory exams	#	Laboratory exams
1	Thyroglobulin antibodies (TgAb)	38	Gfr using MDRD formula
2	Thyroperoxidase antibodies (AbTPO)	39	Hematocrit (HCT)
3	Albumin	40	Haemoglobin (HGB)
4	Alpha-1-fetoprotein (α1 fetoprotein)	41	Immunoglobulin A (IgA)
5	Alpha-1 globulin (α1 globulin)	42	Immunoglobulin G (IgG)
6	Alpha-2 globulin (α2 globulin)	43	Immunoglobulin M (IgM)
7	Alanine transaminase (ALT)	44	Lactate dehydrogenase (LDH)
8	Amylase	45	Lymphocytes
9	Aspartate aminotransferase (AST)	46	Lipase
10	Basophils	47	Bilateral mammography
11	Beta globulin (β globulin)	48	Mean cellular volume (MCV)
12	Beta-2 globulin (β2 globulin)	49	Microalbuminuria
13	Total bilirubin	50	Monocytes
14	Carbohydrate antigen 19-9 (CA 19.9)	51	Neutrophils
15	Calcium (Ca)	52	C-reactive protein (CRP)
16	Occult blood stool sample	53	Brain natriuretic peptide (BNP)
17	Carcinoembryonic antigen (CEA)	54	Platelets (PLT)
18	Creatinine clearance (Cockroft)	55	Potassium (K)
19	Chlorine (Cl)	56	Total proteins
20	HDL Cholesterol	57	Protein electrophoresis
21	LDL Cholesterol	58	Prostate-specific antigen (PSA)
22	Total Cholesterol	59	Free prostate-specific antigen (free PSA)
23	Colonscopy	60	Prothrombin time (PT)
24	Creatinine kinase (CK)	61	Erythrocytes (RBC)
25	Creatinine	62	Reticulocytes
26	Complete blood count (CBC)	63	Sodium (Na)
27	Eosinophils	64	Free triiodothyronine (T_3)
28	Hepatitis B surface antigen (HBsAg)	65	Free thyroxine (T_4)
29	Hepatitis C antibodies (HCV)	66	Thyrotropin (TSH)
30	Rheumatoid factor (RF)	67	Urea
31	Ferritin	68	Uric acid
32	Iron	69	Complete urine test
33	Vitamin *B*9 (folate)	70	Urine culture
34	Alkaline phosphatase (ALP)	71	Erythrocyte sedimentation rate (ESR)
35	Free/total prostate-specific antigen ratio (free/total PSA)	72	Vitamin *B*12 (cobalamin)
36	Gamma globulin (γ globulin)	73	Leukocytes (WBC)
37	Gamma-glutamyl transferase (γGT)		

puted according to [91]:

$$\mathrm{TyG}_i = \frac{\ln(TG_i \cdot Gb_i)}{2} \tag{3.1}$$

Overall, the FIMMG_*obs* dataset comprises of 2276 TyG observations. On the basis of the TyG threshold (TyG_{th}=8.65) reported in [91], each observation can be classified as normal (TyG_i¡8.65) or at risk ($\mathrm{TyG}_i \geq$8.65).

FIMMG_*obs* includes also three different fields (i.e., demographic, monitoring and laboratory exams) resulting in a total of 80 EHR features (i.e., 2 demographic features, 5 monitoring features and 73 laboratory exams features). The list of the 73 laboratory-exams features is reported in Table 3.1; none of them provides glycemic information. For each *id* and for each *i* belonging to *seq*, a vector of 80 EHR features was considered. Missing values of monitoring and laboratory exams features were indicated as *NaN*.

3.3. Methods

The TyG-er approach combined data imputation strategies with a Regression Forest (RF) model to discover non-trivial clinical factors in EHR data for the identification of the TyG index ($\hat{\text{TyG}}_i$). After applying data imputation strategies, the 80 EHR features (*inp*) together with *id* and *seq* were considered as the features to train the RF model. The TyG_i computed according to Eq.3.1 represents the label of the RF model.

3.3.1. Preprocessing

In order to retrieve information from the missing values stored in FIMMG_*obs* were exploited the widely known [96]:

1. *Extra values imputation:* The *NaN* values were replaced with a numeric extra value (i.e., 999).

2. *Median imputation:* The *NaN* values were replaced with a median value computed in the training set [96].

3. *K-Nearest Neighbor (KNN) imputation:* The *NaN* values were replaced according to the KNN strategy [97]. The hyper-parameter K was set to 1 in order to preserve the initial data structure [97].

3.3.2. Regression Forest

Regression Forest (RF) is a Random Forest strategy for solving regression tasks. RF is a variant of bagging proposed by [98] and consists of an ensemble of regression trees (RTs) (i.e., n° of RT) generated by independent, identically distributed, random vectors. RF is designed by sampling from the observations, from the features (i.e., n° of features to be selected) and by varying two tree-parameters (i.e., max n° of splits and max n° of size) [99]. The best splitting features for each node was computed according to the sum of squared error.

3.3.3. Features importance

The influence of a feature in the RF model to identify $\hat{\text{TyG}}$ was measured according to a permutation of out-of-bag feature observation [100]. Hence, if a feature was relevant to identify the TyG index, then permuting its values should affect the model error. On the other hand, if a feature was not relevant, then permuting its values should not significantly affect the model error. Notice how the permutation approach offers an almost unbiased importance measure and is more consistent with respect to other approaches (e.g., the Gini index) [101].

3.3.4. Experimental procedure

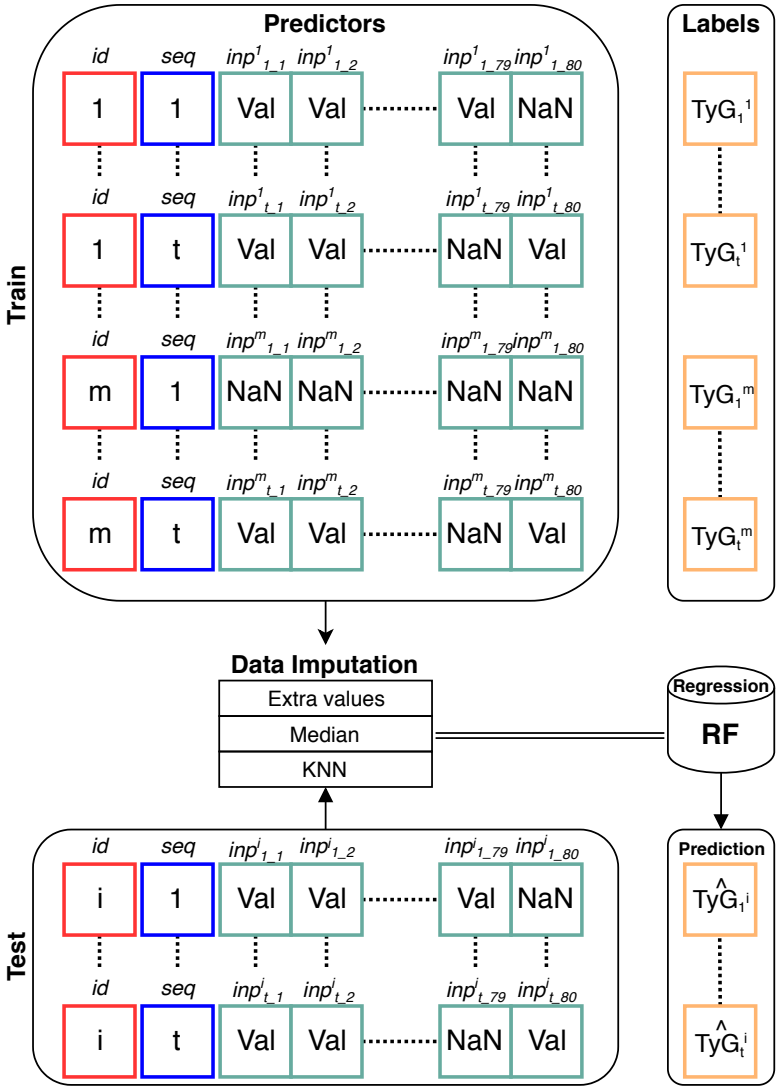


Figure 3.1.: Overview of the TyG-er approach for the Tenfold Cross Validation Over Subjects (CVOS-10) procedure. The *id* number represents the patient; given N, the total number of the subjects, $M = \{1, ..., m\}$, the training patients, and $I = \{1, ..., i\}$, the testing patients, since $N = M + I$, it follows that $M \cap I = \emptyset$. The *seq* number identifies the temporal sequences of the TyG measurement for each *id*, where t is the last *seq* number for each *id*. The *inp* values represent the 80 EHR features (i.e., demographic, monitoring and laboratory exams). TyG_i represents the label of the Regression Forest (RF), while $\hat{\text{TyG}}_i$ is the prediction of the RF.

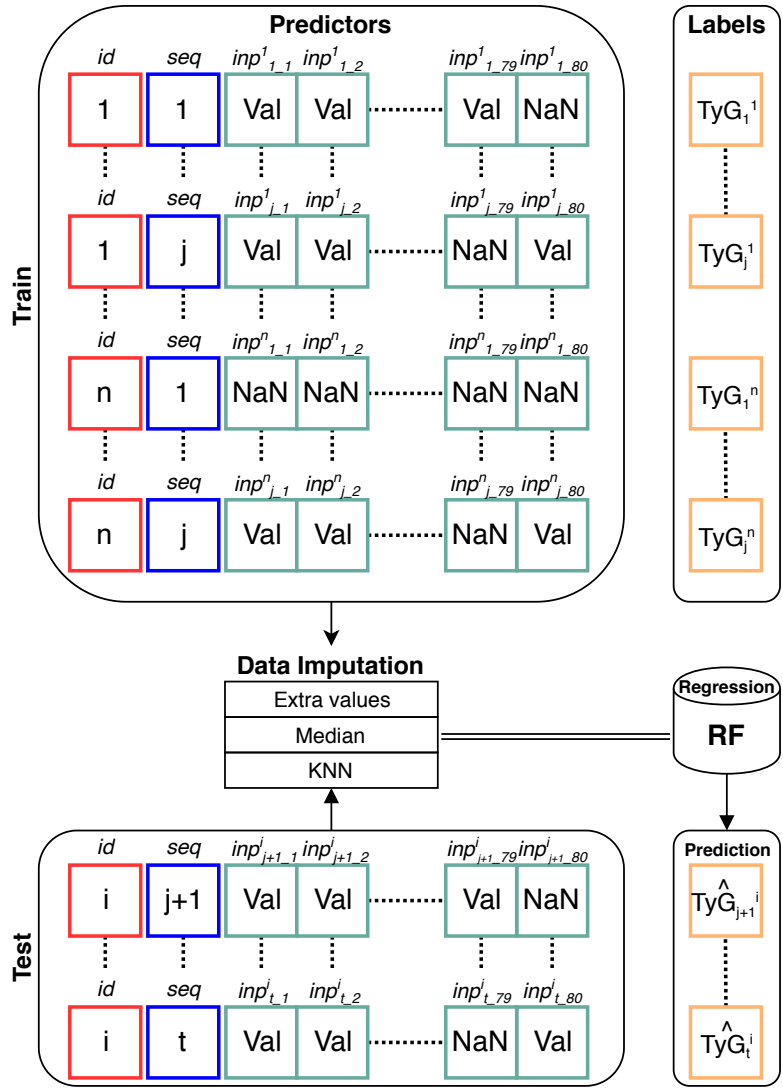


Figure 3.2.: Overview of the TyG-er approach for the Leave Last Records Out (LLRO) procedure. The *id* number represents the patient; given $N = \{1, ..., n\}$, the total number of the patients, it follows that all patients were included for training and testing (i.e., $i \in I$, $I \subset N$). The *seq* number identifies the temporal sequences of the TyG measurement for each *id*, where, if t is defined as the last *seq* number for each *id*, it follows that $1 < j < t$. The *inp* values represent the 80 EHR features (i.e., demographic, monitoring and laboratory exams). TyG_i represents the label of the Regression Forest (RF), while $\hat{\text{TyG}}_i$ is the prediction of the RF.

To measure the robustness of the discovered clinical factors, the TyG-er approach was tested for the TyG identification in three different experimental procedures [2].

The Tenfold Cross-Validation (CV-10) procedure represents the baseline experimental procedure, where an overall $\hat{\text{TyG}}$ was identified for each subject and compared with the average of all recorded TyG_i. At the same time, the features vector was represented by the average of the laboratory exams for each subject. CV-10 was implemented dividing all subjects into ten folds: selecting nine folds for training, and one fold for testing.

Two further experimental procedures were applied for TyG identification: (i) Tenfold Cross-Validation Over Subjects (CVOS-10) (see Fig. 3.1) and ii) Leave Last Records Out (LLRO) (see Fig. 3.2).

The generalisation across unseen patients of the TyG-er approach was implemented using a CVOS-10 procedure. On the other hand, the LLRO procedure allowed testing of how the TyG-er is able to generalise across unseen observations of the same patient. In the LLRO scenario, TyG-er was trained according to the first three observations (i.e., threshold $(th) = 3$) of all patients and tested with the remaining observations of the 50% of patients (i.e., the other observations of the remaining patients were used for the validation stage). The rational choice of the selected th affects the number of known observations for each patient.

3.3.4.1. Experimental Measurements

For each couple of experimental procedures the agreement of the different experimental procedures was measured according to Lin's concordance correlation coefficient (r_c) [102]. This index quantifies the agreement of the TyG-er approach among the three different experimental procedures for assigning a feature's importance. Lin's coefficient modifies the Pearson correlation coefficient by taking into account not only how close the feature's importance is to the line of best fit, but also how far that line is from the 45-degree mark (i.e., perfect agreement) [102].

The predictive performance of TyG-er in each experimental procedure was evaluated according to the following measures: (i) Pearson correlation (r), (ii) Spearman's rank correlation (r_s), (iii) mean absolute error (MAE), (iv) $MAE\%$ (computed with respect to the maximum range of TyG_i), v) mean squared error (MSE) and (vi) r_c. Statistical significance of the correlation tests was set at the 5% significance level.

3.3.4.2. Validation procedure

Concerning the CV-10 and CVOS-10 experimental procedures, the optimisation of the RF hyperparameters (i.e., n° of RT, max n° of splits, n° of features to select at random for each decision split) was performed implementing a grid-search and optimising the

[2]The code to reproduce all the results (testing different settings) is available at the following link: `https://github.com/michelebernardini/T2D-early-risk-identification`

MSE in a nested Fivefold Cross-Validation. *MSE* was preferred over other optimisation objectives, because the identification of the TyG exact numerical value has more clinical relevance than the identification of its trend over time. 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 the researchers to obtain an unbiased and robust performance evaluation [79]. The testing/validation split in the LLRO procedure was performed in order to uniformly stratify the distribution of the TyG index by minimising any bias between the validation and test sets. In particular, the overall TyG index was clusterised over patients performing a k-means strategy [103]. Thus, the optimal number of clusters (i.e., $k = 4$) was set according to a Calinski-Harabasz criterion [104]. Then, the test/validation subset was extracted in order to stratify the output of the k-means. Table 3.2 summarises the range of the hyperparameters optimised during the validation stage for each method.

Table 3.2.: Range of Hyperparameters (Hyp) for each model: Regression Forest (RF), Regression Tree (RT), Boosting, Linear Support Vector Machine (SVM Lin), Gaussian SVM (SVM Gauss), and SVM Lasso.

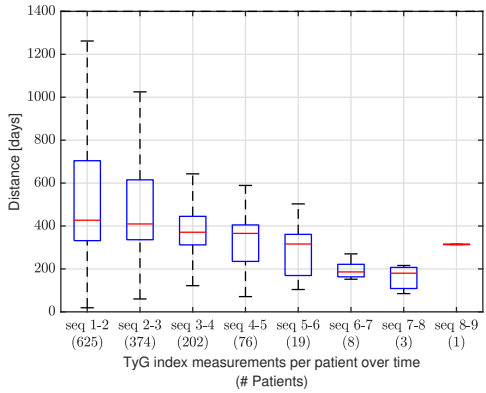
Model	**Hyp**	**Range**
RF	max n° of splits	$\{5, 10, 15, 20, 25\}$
	n° of RT	$\{50, 100, 150, 200, 250\}$
	n° of features to select	$\{\frac{all}{4}, \frac{all}{3}, \frac{all}{2}, all\}$
RT [24, 50, 52, 93]	max n° of splits	$\{5, 10, 15, 20, 25\}$
	min n° of leaf size	$\{50, 60, 70, 80, 90, 100\}$
Boosting [52]	max n° of splits	$\{100, 200, 300, 400, 500\}$
	max n° of cycles	$\{100, 200, 300, 400, 500\}$
SVM Lin [50, 53, 93]	Box Constraint	$\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$
SVM Gauss [50, 53, 93]	Box Constraint	$\{10^{-2}, 0.1, 1, 10, 10^2, 10^3, 10^4\}$
	Kernel Scale	$\{10^{-2}, 0.1, 1, 10, 10^2, 10^3, 10^4\}$
SVM Lasso [2]	Lambda	$\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$

3.4. Experimental results

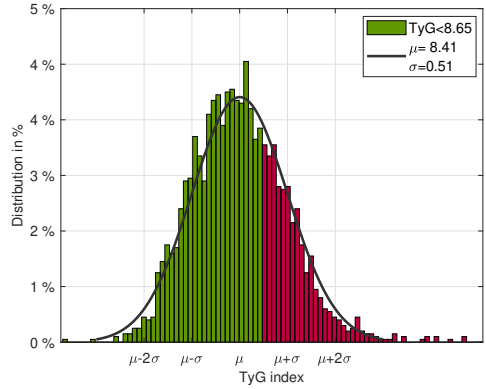
The results of our study are provided in terms of preprocessing (see Sec. 3.4.1 and Fig. 3.3), pattern localisation (see Sec. 3.4.2 and Fig. 3.4) and predictive performance (see Sec. 3.4.3 and Tab. 3.3).

3.4.1. Preprocessing

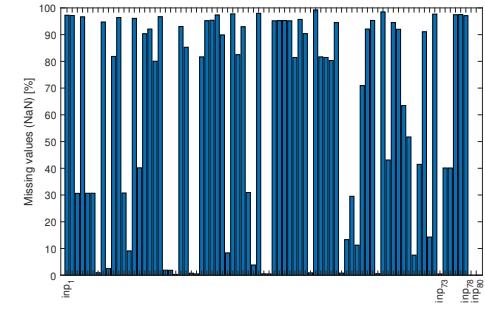
Figure 3.3a shows the overall temporal distance distribution between between consecutive TyG_i and TyG_{i+1} measurements per patient. It turns out that laboratory exams



(a) Overall temporal distance distribution between consecutive TyG_i and TyG_{i+1} measurements per patient. The amount of patients for each number of temporal sequences (*seq*) is indicated below in round brackets.



(b) TyG index normal distribution with mean $\mu = 8.41$ and standard deviation $\sigma = 0.51$. The TyG index threshold ($TyG_{th} = 8.65$) separates the green side (1593 observations) from the red side (683 observations) of the graph.



(c) Percentage (%) of missing values (*NaN*) for each of the 80 EHR features (*inp*), where $inp_1 \leq$ laboratory exams $\leq inp_{73}$, $inp_{74} \leq$ monitoring $\leq inp_{78}$ and demographic = $\{inp_{79}, inp_{80}\}$.

Figure 3.3.: Preprocessing results.

repeated more than once are progressively closer in time and executed by a decreasing number of patients.

Figure 3.3b shows the overall TyG index distribution for the FIMMG_obs dataset. The data follow a Normal distribution (according to a Kolmogorov Smirnov Test, $H = 0.025$, $p = 0.121$) with mean $\mu = 8.41$ and standard deviation $\sigma = 0.51$.

Details about missing value occurrences for each EHR feature (*inp*) (i.e., demographic, monitoring and laboratory exams) are reported in Figure 3.3c.

3.4.2. Pattern localisation

For each experimental procedure, TyG-er is represented by the best combination between data imputation and RF in terms of *MSE*. TyG-er is defined as RF combined with extra values for CV-10 ($MAE = 0.295$ corresponding to a $MAE\% = 5.98$) and CVOS-10 ($MAE = 0.310$ corresponding to a $MAE\% = 6.29$) procedures, while RF combined with KNN imputation represents the TyG-er approach for the LLRO ($MAE = 0.280$ corresponding to a $MAE\% = 5.68$) procedure.

The top 10 features were listed in descending order of percentage importance according to the permutation approach for each experimental procedure (see Fig. 3.4). Such further intra-model analysis is useful to extract and quantitatively compare the most discriminative features.

Figure 3.4a shows the most relevant TyG features for the CV-10 procedure; while Figure 3.4b and Figure 3.4c show respectively the most relevant TyG features of the TyG index over unseen patients (CVOS-10) and over unseen successive observations of the same patients (LLRO). The remaining features (i.e., Others) achieved individually a percentage importance of less than 2%.

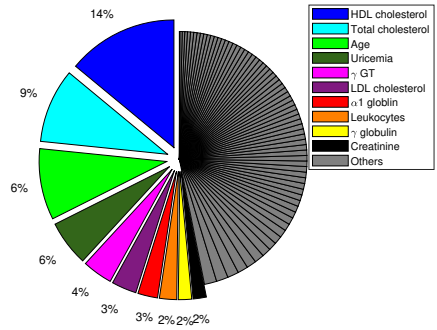
The agreement between the features' importance assigned by the TyG-er approach for each experimental procedure follows below:

- CV-10 vs CVOS-10: $r_c = 0.664$ $(p < .05)$, $CI = [0.608, 0.714]$
- CVOS-10 vs LLRO: $r_c = 0.720$ $(p < .05)$, $CI = [0.656, 0.775]$
- CV-10 vs LLRO: $r_c = 0.911$ $(p < .05)$, $CI = [0.872, 0.938]$

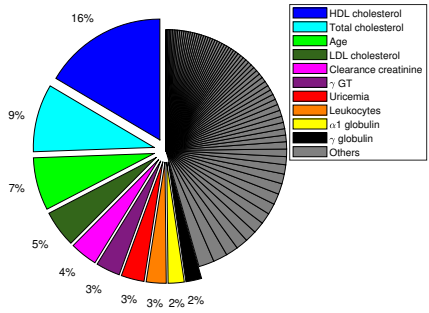
3.4.3. Predictive performance

The predictive performance of the three different procedures in terms of *MAE%* and r_c is reported below:

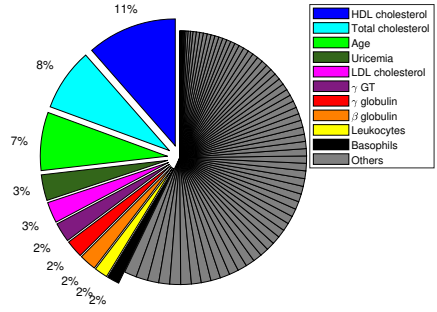
- CV-10: *MAE%*=5.98, r_c=0.545 $(p < .05)$
- CVOS-10: *MAE%*=6.29, r_c=0.543 $(p < .05)$
- LLRO: *MAE%*=5.68, r_c=0.666 $(p < .05)$



(a) Top 10 features for the Tenfold Cross-Validation experimental procedure. TyG-er consists of Regression Forest combined with extra values imputation.



(b) Top 10 features for the Tenfold Cross-Validation Over Subjects experimental procedure. TyG-er consists of Regression Forest combined with extra values imputation.



(c) Top 10 features for the Leave Last Records Out experimental procedure. TyG-er consists of Regression Forest combined with the K-Nearest Neighbour imputation.

Figure 3.4.: Top 10 features listed in descending order of percentage importance according to the permutation approach (see Sec. 3.3.3) for each experimental procedure.

Table 3.3 shows the predictive performance of TyG-er as well as the performed comparison with respect to other standard ML algorithms, widely used for the prediction of the T2D condition [52, 93, 50, 53, 2].

Table 3.3.: Predictive performance of TyG-er and comparison with respect to the state-of-the-art (i.e., Regression Tree (RT), Boosting, Linear Support Vector Machine (SVM Lin), Gaussian SVM (SVM Gauss), and SVM Lasso). For each experimental procedure (i.e., Tenfold Cross-Validation (CV-10), Tenfold Cross-Validation Over Subjects (CVOS-10), and Leave Last Records Out (LLRO)), the best model in terms of mean squared error (*MSE*) as well as the best competitor were highlighted in bold. *MAE* represents the mean absolute error, while r and r_s represent the Pearson correlation and the Spearman's rank correlation. Standard deviation is indicated in round brackets.

	CV-10				CVOS-10				LLRO			
Model	***r***	***rs***	***MAE***	***MSE***	***r***	***rs***	***MAE***	***MSE***	***r***	***rs***	***MAE***	***MSE***
Baseline												
RF	0.567	0.575	0.321 (0.026)	0.171 (0.024)	0.546	0.540	0.338 (0.048)	0.189 (0.055)	0.595	0.553	0.347	0.197
RT	0.436	0.443	0.342 (0.026)	0.196 (0.029)	0.389	0.388	0.368 (0.047)	0.222 (0.061)	0.505	0.416	0.366	0.210
Boosting	0.580	0.590	0.331 (0.031)	0.181 (0.026)	0.551	0.554	0.351 (0.052)	0.203 (0.061)	0.585	0.551	0.359	0.214
Extra values												
RF	0.637	0.633	0.295 (0.021)	**0.143 (0.023)**	0.633	0.611	0.310 (0.037)	**0.157 (0.036)**	0.706	0.651	0.305	0.147
RT	0.400	0.392	0.354 (0.018)	0.205 (0.024)	0.467	0.445	0.356 (0.037)	0.205 (0.046)	0.539	0.483	0.359	0.201
Boosting	0.636	0.633	0.298 (0.021)	**0.146 (0.023)**	0.623	0.607	0.319 (0.045)	**0.166 (0.047)**	0.698	0.647	0.311	0.157
SVM Lin	0.381	0.400	0.362 (0.023)	0.220 (0.024)	0.395	0.376	0.371 (0.038)	0.226 (0.042)	0.380	0.379	0.391	0.249
SVM Gauss	0.235	0.282	0.381 (0.039)	0.238 (0.037)	0.111	0.192	0.400 (0.053)	0.261 (0.071)	0.443	0.419	0.381	0.232
SVM Lasso	0.322	0.360	0.364 (0.028)	0.223 (0.024)	0.413	0.395	0.365 (0.043)	0.217 (0.047)	0.401	0.394	0.386	0.241
Median												
RF	0.638	0.637	0.297 (0.025)	0.144 (0.025)	0.620	0.608	0.312 (0.036)	0.160 (0.036)	0.705	0.642	0.307	0.147
RT	0.479	0.491	0.328 (0.034)	0.185 (0.036)	0.457	0.440	0.359 (0.040)	0.209 (0.050)	0.474	0.414	0.379	0.221
Boosting	0.639	0.635	0.298 (0.023)	0.146 (0.023)	0.614	0.603	0.319 (0.042)	0.168 (0.046)	0.717	0.660	0.302	0.151
SVM Lin	0.487	0.612	0.340 (0.031)	0.211 (0.042)	0.417	0.536	0.369 (0.056)	0.297 (0.262)	0.154	0.685	0.446	3.702
SVM Gauss	0.174	0.367	0.380 (0.037)	0.238 (0.037)	0.288	0.395	0.386 (0.058)	0.244 (0.075)	0.691	0.657	0.295	0.155
SVM Lasso	0.639	0.645	0.301 (0.022)	0.151 (0.018)	0.594	0.596	0.329 (0.042)	0.179 (0.047)	0.569	0.561	0.344	0.204
KNN												
RF	0.625	0.625	0.300 (0.025)	0.148 (0.023)	0.585	0.585	0.324 (0.038)	0.172 (0.041)	0.742	0.691	0.280	**0.129**
RT	0.464	0.476	0.334 (0.023)	0.189 (0.024)	0.456	0.456	0.357 (0.045)	0.208 (0.059)	0.499	0.470	0.362	0.215
Boosting	0.616	0.614	0.303 (0.026)	0.152 (0.024)	0.583	0.572	0.327 (0.045)	0.177 (0.051)	0.715	0.700	0.290	0.145
SVM Lin	0.556	0.603	0.315 (0.033)	0.222 (0.195)	0.459	0.536	0.372 (0.050)	0.515 (0.740)	0.150	0.652	0.468	4.495
SVM Gauss	0.238	0.344	0.372 (0.043)	0.228 (0.045)	0.164	0.302	0.384 (0.064)	0.245 (0.081)	0.783	0.764	0.254	**0.116**
SVM Lasso	0.603	0.639	0.306 (0.029)	0.180 (0.097)	0.593	0.593	0.328 (0.042)	0.179 (0.048)	0.198	0.640	0.420	2.103

The performed comparison highlighted how the TyG-er approach was the most reliable model for CV-10 ($r = 0.637, p < .05$; $r_s = 0.633, p < .05$; $MAE = 0.295$; $MSE = 0.143$) and CVOS-10 ($r = 0.633, p < .05$; $r_s = 0.611, p < .05$; $MAE = 0.310$; $MSE = 0.157$).

The SVM Gauss combined with the KNN data imputation strategy showed the best predictive power for LLRO. However, the TyG-er approach was found to be the best contender ($r = 0.742, p < .05$; $r_s = 0.691, p < .05$; $MAE = 0.280$; $MSE = 0.129$). Compared to the SVM Gauss, the TyG-er approach may offer a greater level of interpretability [105] while being less time consuming (200 seconds faster in the training stage).

3.5. Discussion

This study aimed to discover non-trivial clinical factors in EHR data to determine where the TyG information is encoded. To this end, a high-interpretable ML regression approach (the TyG-er approach) was proposed. Patients included in our analysis ranged from normal to high-risk condition (see Fig. 3.3b). T2D patients were excluded since they can be under anti-diabetic pharmacological treatment and TyG values may not be representative. The results of our study detected some non-conventional clinical factors and provided novel insight into the best combination of risk factors for detecting early glucose tolerance deterioration. The top three features of the TyG index selected by the proposed TyG-er approach, namely, high-density lipoprotein (HDL) cholesterol, total cholesterol and age, remained stable among the three experiments performed (CV-10, CVOS-10 and LLRO). The top three features provided high model interpretability since clinical studies in literature have shown their relationship with early glucose tolerance deterioration and IR [106]. Uricemia, leukocytes and gamma-glutamyltransferase (γGT) were selected by the proposed TyG-er approach as further important features. Altered values of uricemia and leukocytes are not usually conceived by GPs as primary features of early states of glucose tolerance deterioration; however, the results of our study confirmed what recent literature has suggested, i.e., that altered values of uricemia [107] and leukocyte activation [108] are correlated with an IR condition. Elevated serum γGT concentration is an accepted component of metabolic disturbance [109]; however, the results of our study confirmed its important contribution also in incipient disturbances in the glucose metabolism, such as IR [110]. Accordingly, clinical evidence showed that, in particular categories of individuals at high risk of developing T2D, fatty change of the liver linked to γGT is associated with IR [111]. Moreover, our results are also confirmed by recent evidence showing that γGT and uricemia can synergise in predicting the development of T2D [112].

Eventually, features related to the protein profile (α, β and γ globulins) were selected among the top 10 features. This result supports current trends aimed at searching for novel protein clinical factors of early glucose tolerance deterioration using tissues and/or biofluids (blood, serum, plasma, and urine) [113]. The robustness of those extracted clinical factors was confirmed by (i) the high agreement (according to the r_c value) of the TyG-er approach among the three different experimental procedures (see Sec. 3.4.2), and (ii) the high reliability in terms of predictive performance (see Tab. 3.3).

The results of the three experimental procedures outlined the predictive power of the TyG-er approach (see Sec. 3.4.3). In particular, the predictive power of the extracted clinical factors was demonstrated to generalise (i) across a different unseen subset of patients (i.e., CVOS-10) and (ii) across different trials of the same subset of patients (i.e., LLRO).

Although traditional methods [106, 107, 108, 112, 113, 110, 109] employed statistical analysis to answer the aim of our study, the TyG-er approach may improve the sensitivity of detection by combining multiple pieces of information across several clinical factors while dealing with the intrinsic presence of missing values.

3.5.1. Limitations and future work

To confirm generalisation capabilities of the TyG-er approach, future strategies will investigate the robustness of the results with different patient stratification (i.e., different TyG index values). Results of this study may differ when stratifying patients with respect to their metabolic risk (on the basis of TyG index values). Another future direction to prove the generalisation capabilities of the proposed method will be addressed to test the robustness of the extracted clinical factors into a multi-GP dataset.

Chapter 4.

Insulin resistance: Type 2 diabetes early-stage risk condition

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 comorbidities. Although fundamental in early phases of T2D natural history, insulin resistance (IR) is not usually quantified by General Practitioners (GPs). Triglyceride-glucose (TyG) index has been proven useful in clinical studies for quantifying IR 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 IR (low vs high T2D risk) in terms of TyG index. The MIL-Boost is applied to past electronic health record (EHR) patients' information stored by a single GP. The proposed MIL-Boost algorithm proved to be effective in dealing with this task, by performing better than 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 sample size and variability in the 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 (CDSS).

4.1. Introduction

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 [85]. 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 [114].

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 [115]. As IR worsens, more global defects in insulin secretion occur and, at the end, hyperglycemia arises [116]. Although fundamental in early phases of T2D natural history, IR is not usually quantified by GPs since specific blood tests - which are not included in those usually performed in routine examinations - as well as mathematical computations, are required [88].

A simple surrogate assessment of IR can be obtained through the triglyceride-glucose (TyG) index, based on routine triglyceride and glucose measurements [89, 90]. 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 [91]. 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 CDSS predicting TyG changes over time may allow for better predictions of target groups with high risk of T2D. Such a CDSS based on EHR data 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. Machine Learning model have already been utilised in developing successfully predictive models for T2D [93], but still never focused on early temporal prediction of T2D risk (i.e. IR 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 [117, 118].

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 IR (low vs high T2D risk) in terms of TyG index.

4.2. Related work

In recent literature several approaches have been proposed to predict chronic pathologies onset from heterogeneous and longitudinal EHR data [119, 120, 121, 122, 123, 124, 125].

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 [121] 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 [122] and kidney diseases [123, 124, 125] by providing also model interpretability. The authors in [122] 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 [123] 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 [124] 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 [125] 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 [50, 52, 53, 2, 126, 127] and/or prediction [128, 129, 93] 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 [93] used EHR data of GPs, the considered features space contains also glycaemic information. In order to handle limited longitudinal EHR data, the authors in [130] 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 [128, 129, 130, 93], 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, the experimental comparison was performed with respect to other state-of-the-art ML models (i.e., Decision Tree (DT) [128, 129, 93, 127]; RF [128, 129, 124]; KNN [128, 129]; Boosting [52]; SVM with linear kernel (SVM Lin) and SVM with Gaussian kernel (SVM Gauss) [128, 129, 93]; and SVM with Lasso regularizer (SVM Lasso)

[2]), employed in literature to solve tasks closer to our setting. Similarly to [123], these state-of-the art models were compared according to time-invariant and temporal majority vote procedures.

4.3. Clinical data: FIMMG_*pred* dataset

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. 4.1): i) exclusion of all diagnosed diabetic patients according to the International Classification of Disease 9th Revision (ICD-9) (since they can be pharmacologically 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 [91]:

$$TyG_j = \frac{\ln(TG_j \cdot Gb_j)}{2} \tag{4.1}$$

On the basis of the IR threshold of TyG ($TyG_{th} = 8.65$) reported in [91], each observation can be classified as low ($TyG_j < TyG_{th}$) or high ($TyG_j \geq TyG_{th}$) risk.

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*.

The following the full list of the laboratory exams:

4.3.0.1. Problem formulation

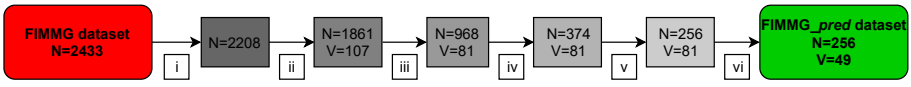


Figure 4.1.: Inclusion and exclusion criteria (N identifies the number of EHR patients, and V the number of EHR features)

In order to better evaluate the temporal evolution of the patient's T2D risk condition, more strict inclusion criteria (see Fig. 4.1) were added to i), ii) and iii) as follows: iv)

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

Table 4.1.: 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	44	Erythrocyte sedimentation rate (ESR)
22	Gamma globulin (γ globulin)	45	Leukocytes (WBC)
23	Gamma-glutamyl transferase (γGT)		

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 [123]); and vi) EHR features that contain an overall amount of *NaN* less than a threshold of 90% (th_{nan}= 90%). The rationale 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 done in other studies [131, 132].

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. 4.2).

Table 2.1 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.

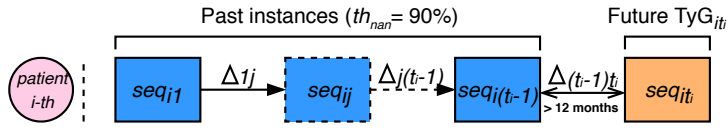


Figure 4.2.: For each *i-th* patient, the temporal distance between past instances (i.e., Δ_{1j}, $\Delta_{j(t_i-1)}$) is variable, while between the last 2 instances (i.e., $\Delta_{(t_i-1)t_i}$) is at least equal or greater than 12 *months*.

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

4.4. Methods

Preprocessing In order to handle missing data in the FIMMG_pred dataset, K-Nearest Neighbor (KNN) imputation was used, which replaces the NaN according to the KNN strategy [97]. The hyper-parameter K was set to 1 in order to preserve the initial data structure [97]. As already done in a similar context [3], the *K-Nearest Neighbor (KNN) imputation* was selected as the best strategy after exploring other data imputation techniques (*extra values imputation*, *median imputation*).

4.4.1. 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 [133], text processing [134], computer vision [135] and biomedical image analysis [136].

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 [137]. Hence, the MIL task can be addressed to both estimate the instance and bag labels.

The MIL-Boost algorithm originated from the work presented in [138] by starting with the standard multiple instance assumption [137] and the boosting algorithm [139]. 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 $\mathbf{h}$ [137]. The combination is performed in an additive way by weighting each weak classifier h_k:

$$\mathbf{h} = \sum_{k=1}^{K} \alpha_k h_k(x) \tag{4.2}$$

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 [140, 141].

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:

$$p_{ij} = \sigma(\mathbf{h}(seq_{ij})) \tag{4.3}$$

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

$$p_i = g_j(p_{ij}) \tag{4.4}$$

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 $\text{TyG}_{it_i} < 8.65$, 1 [positive bag] if the $\text{TyG}_{it_i} \geq 8.65$) of the proposed MIL-Boost algorithm where the past instances $\{seq_{i1}, \ldots, seq_{i(t_i-1)}\}$ are the instance predictors (see Fig. 4.2 and Fig. 4.3a). The MIL-Boost algorithm (see Fig. 4.3a) 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 the single bag was modeled as a set of multiple instances, an ordinal and defined structure of the instance was not explicitly assumed (e.g. by including the instance ordering number [ion] in the feature set).

4.4.2. Experimental procedure

The performance of the MIL-Boost was evaluated 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 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.

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. In both configurations (i.e. "yesTyG" and "noTyG") the past TyG index was never included among the EHR predictors.

The predictive performance was evaluated according to the measures defined in Section 2.4.4.

[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`

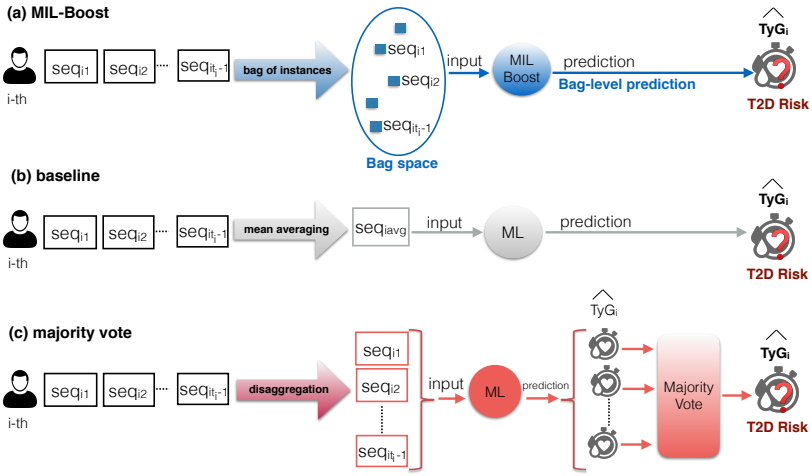


Figure 4.3.: Overview of the experimental procedures: a) MIL-Boost, b) time-invariant baseline, and c) temporal majority vote.

4.4.3. Experimental comparisons

The MIL-Boost algorithm was compared with respect to other ML algorithms employed in literature closer to our setting (see Sec. 4.2), such as: DT [93, 128, 129, 127]; RF [124, 128, 129]; KNN [128, 129]; Boosting [52]; SVM Lin and SVM Gauss [93, 128, 129]; and SVM Lasso [2]. These state-of-the-art approaches were also combined with the KNN imputation technique described in Sec. 4.4 to provide a fair comparison with the proposed MIL-Boost. Moreover, these state-of-the art methods were compared according to the approach proposed by [123] 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] [142], MIL-RF [MIForests] [143] and MIL-SVM [134] with linear and Gaussian kernel.

Time-invariant baseline In the time-invariant baseline experimental procedure (see Fig. 4.3b) a single instance was computed for each bag/patient as the average of the past EHR features (seq_{iavg}). The $\hat{\text{TyG}}_{it_i}$ was predicted without taking into account the temporal evolution of the past clinical history.

Temporal majority vote The temporal majority vote experimental procedure (see Fig. 4.3c) combines 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.

4.4.4. Validation procedure

Table 4.2.: 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 [93, 128, 129, 127]	max # of splits	$\{\mathbf{5}(3), 10(2), 15(2), 20(2), 25(0), 50(1)\}$
RF [124, 128, 129]	# 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 [128, 129]	max # of neighbors	$\{1(1), 3(2), 5(1), 7(1), 9(0), 11(0), \mathbf{13}(3), 15(2)\}$
Boosting [52]	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 [93, 128, 129]	Box Constraint	$\{10^{-2}(0), 0.1(1), 1(1), 10(2), \mathbf{10}^{2}(5), 10^{3}(1)\}$
SVM Gauss [93, 128, 129]	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 [62]	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	$\{10^{-5}(0), \mathbf{10}^{-4}(3), 10^{-3}(1), 10^{-2}(0), 0.1, \mathbf{1}(3), 10(2), 10^{2}(1)\}$ $\{1(0), \mathbf{5}(10), 10(0), 15(0)\}$

Table 4.2 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 us to obtain an unbiased and robust performance evaluation [79]. 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.

Table 4.3.: 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)

4.5. Experimental results

Figure 4.5 shows the TyG_{it_i} index distribution for the final configuration of the FIMMG_*pred* dataset (see Tab. 2.1). 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 4.4 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 4.6 quantifies the *NaN* occurrences for each EHR feature (i.e., demographic, monitoring and laboratory exams) stored in the FIMMG_*pred* dataset.

4.5.1. Predictive performance

Table 4.3 shows the predictive performance of MIL-Boost and comparisons with different experimental procedures (i.e., time-invariant baseline, temporal majority vote),

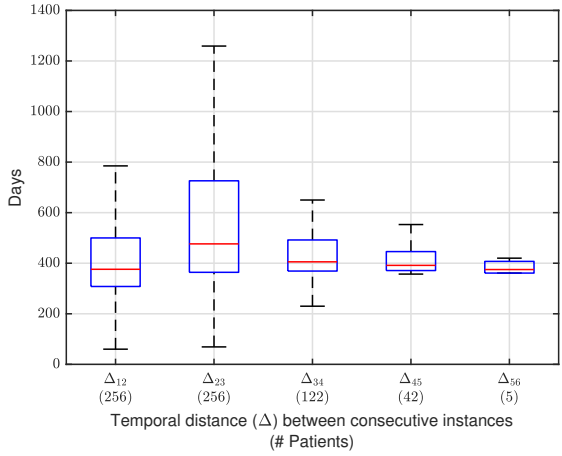


Figure 4.4.: Overall temporal distance distribution between consecutive instances ($\Delta_{j(j+1)}$) per patient. The amount of patients is indicated below in round brackets.

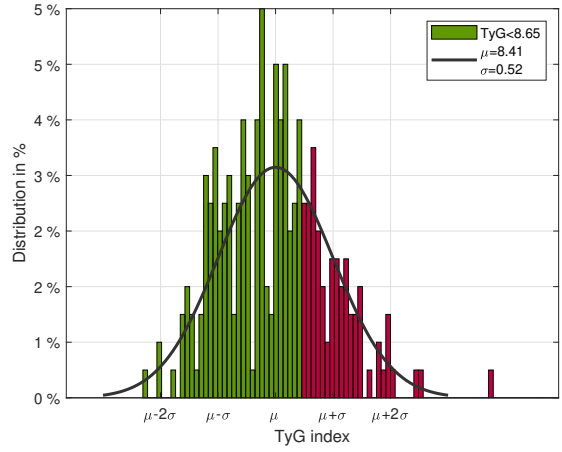


Figure 4.5.: TyG_{it_i} index distribution with mean $\mu = 8.41$ and standard deviation $\sigma = 0.52$. TyG index threshold ($TyG_{th} = 8.65$) separates the green side (179 patients) from the red side (77 patients) of the graph.

different configurations (i.e., yesTyG, noTyG), and different classification algorithms.

Figure 4.7 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: $D = 0.161$, $p = 0.92$).

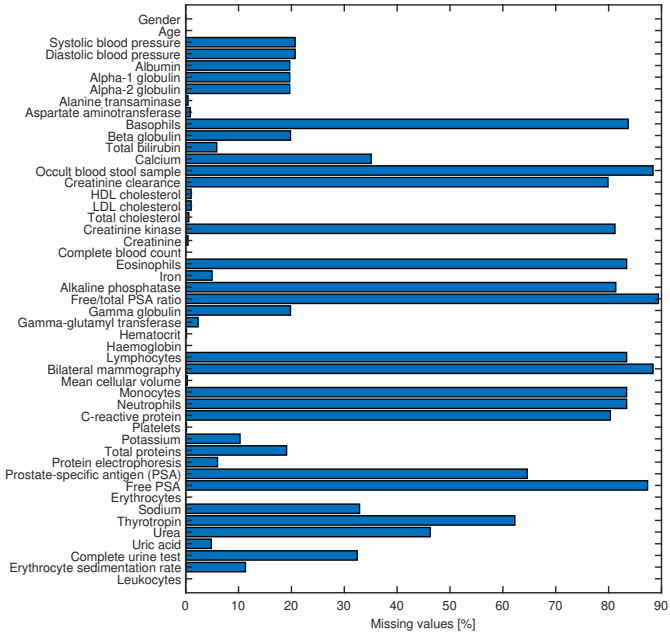


Figure 4.6.: Percentage (%) of missing values (*NaN*) for each of the 49 EHR features: demographic, monitoring, and laboratory exams.

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: 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 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.

4.5.2. Model interpretability

The top-10 rank features are listed in descending order of percentage importance for the temporal MIL-Boost experimental procedure in yesTyG configuration (see Fig. 4.8) and in noTyG configuration (see Fig. 4.9). The most discriminative predictors were extracted 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.

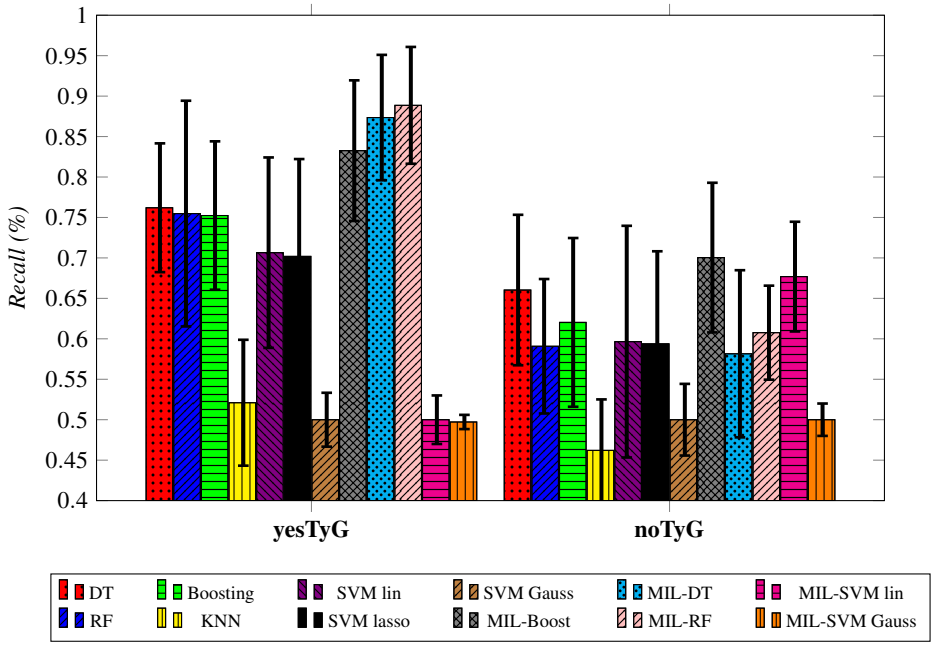


Figure 4.7.: Average Recall and standard deviation of Majority vote and MIL-algorithms over all CVOS-10 procedure.

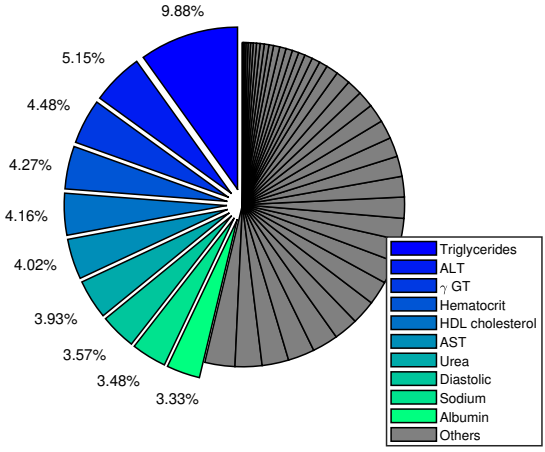


Figure 4.8.: Top-10 rank features for MIL-Boost experimental procedure (yesTyG configuration). The percentage importance of the Others features was about 54%. Glycaemia ranks the 12^{nd} position with 2.68 %.

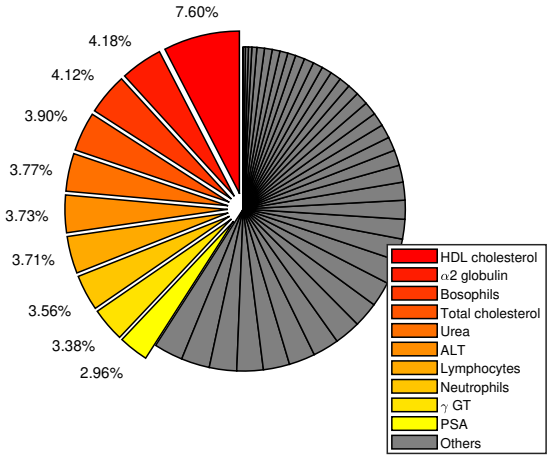


Figure 4.9.: Top-10 rank features for MIL-Boost experimental procedure (noTyG configuration). The percentage importance of the Others features was about 59%.

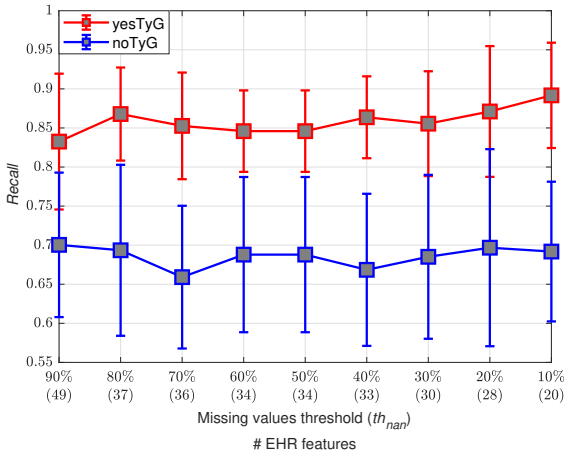


Figure 4.10.: Trend of the MIL-Boost *Recall* and its standard deviation as a function of the missing values threshold th_{nan}. The amount of EHR features is indicated in round brackets for each th_{nan}.

4.5.3. Sensitivity to missing values

Figure 4.10 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.

4.5.4. Sensitivity to the sparsity of the data

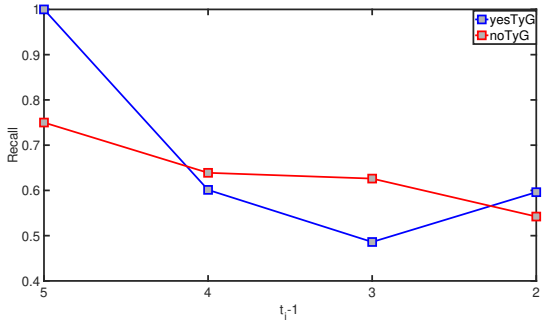


Figure 4.11.: MIL-Boost Recall vs sparsity of the dataset in terms of the number of past instances ($t_i - 1$).

The Recall of MIL-Boost was computed versus a measure of the sparsity of the data (see Figure 4.11). 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. An overall Recall was computed by aggregating the predictions of all the 10 folds for each value of "number of past instances" ($t_i - 1$).

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

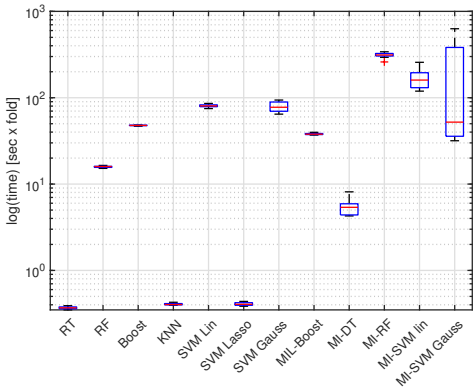
4.5.5. Computation-time analysis

The computation time analysis for the training and validation stage is shown in Figure 4.12a, while for the testing stage in Figure 4.12b.

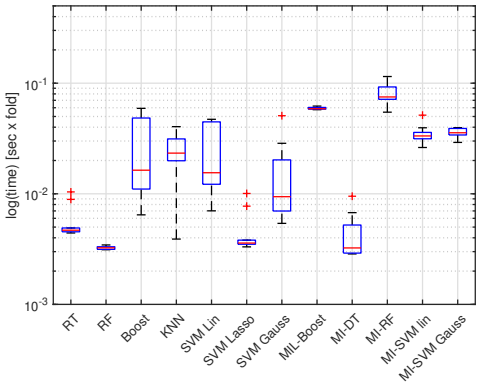
4.6. Discussion

4.6.1. Predictive performance

This study proposed a model that captures temporal information for the early prediction of worsening IR (low vs high T2D risk) in terms of TyG index. The model learns from past routine measurements either including or excluding triglycerides and glucose measurements, which are the ones used to compute the TyG index. The proposed



(a) Preprocessing, training and validation time



(b) Testing time

Figure 4.12.: Comparison in term of computation time.

MIL-Boost algorithm proved to be effective in dealing with this task, by performing better than the other state-of-the-art ML models for both configurations (*Recall*, yesTyG: 0.83, noTyG: 0.70) and other MIL-based approaches for the 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 [3]. However, the present study is basically different from the previous one in terms of tasks (identification vs. forecast), dataset considered (present observations/future observations) and for the adopted methodologies. The TyG-er approach [3] 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 [3] 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 IR. Knowledge of non-trivial associations provides hints for further investigation in clinical studies. Differently, 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 algorithm 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 [3] 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 [42]. However, in most cases EHR data pertain to hospitalized subjects [42, 144, 145], 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 [121, 122, 125, 124, 123] 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). Although the performance decrease as the sparsity in the data increases (see Section 4.5.4), the proposed MIL-Boost algorithm may deal with the variability of this setting, where different subjects (i.e., bags) may have a different number of observation (i.e., instances) over time. Our MIL-based approach relaxes the constraint imposed by some other work [128, 129, 93, 123] 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, no preprocessing step was employed (e.g., resampling strategies) to deal with the natural unbalance of this task.

Moreover, no statistical changes were found 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 could address 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, 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 predic-

tors; while for noTyG configuration the *Reecall* trend appears more stable. However, t-test confirms there are no significant 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 [96]).

4.6.2. 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 [146, 91]. 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 IR [147, 148].

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 [149, 3]. Glycaemia appears redundant (12^{nd} rank) in case of presence of triglycerides and other clinical factors in the yesTyG configuration. It appears 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. Note that glycated haemoglobin (HbA1c), an important clinical factor used for T2D diagnosis and monitoring, was not included in our analysis. HbA1c is not included in routine 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%).

4.6.3. 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. 2.6a). Such competitiveness in terms of computational efficiency (see Fig. 4.12) 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 [150] 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.

Chapter 5.

Clinical Decision Support System for type 2 diabetes quality care evaluation

Clinical Decision Support Systems (CDSS) have been developed and promoted for their potential to improve quality of health care. However, there is a lack of common clinical strategy and a poor management of clinical resources and erroneous implementation of preventive medicine.
To overcome this problem, this work proposed an integrated system that relies on the creation and sharing of a database extracted from GPs' Electronic Health Records (EHRs) within the Netmedica Italian (NMI) cloud infrastructure. Although the proposed system is a pilot application specifically tailored for improving the chronic type 2 diabetes (T2D) care it could be easily targeted to effectively manage different chronic-diseases. The proposed CDDS is based on EHR structure used by GPs in their daily activities following the most updated guidelines in data protection and sharing. The CDDS is equipped with a Machine Learning (ML) method for analyzing the shared EHRs and thus tackling the high variability of EHRs. A novel set of T2D care-quality indicators are used specifically to determine the economic incentives and the T2D features are presented as predictors of the proposed ML approach.
The EHRs from 41237 T2D patients were analyzed. No additional data collection, with respect to the standard clinical practice, was required. The CDDS exhibited competitive performance (up to an overall accuracy of $98\% \pm 2\%$ and macro-recall of $96\% \pm 1\%$) for classifying chronic care quality across the different follow-up phases. The chronic care quality model brought to a significant increase (up to 12%) of the T2D patients without complications. For GPs who agreed to use the proposed system, there was an economic incentive. A further bonus was assigned when performance targets are achieved.
The quality care evaluation in a clinical use-case scenario demonstrated how the empowerment of the GPs through the use of the platform (integrating the proposed CDDS), along with the economic incentives, may speed up the improvement of care.

5.1. Introduction

Type 2 diabetes (T2D) results from an ineffective use of insulin. The risk of developing T2D depends on an interplay of genetic and metabolic factors. For instance, high waist and body mass index (BMI) are associated with an increased risk, though the relationship may vary in different populations [151]. In 2015, the World Health Organization (WHO) estimated a global prevalence of diabetes around the 9%, with more than 90% of the patients being affected by T2D [44, 45]. Only in 2012, diabetes caused 1.5 million deaths, with more than 8 out of 10 deaths occurring in low and middle income countries. In developing countries, more than half of all diabetes cases goes undiagnosed due to the poor T2D symptoms, at least at the early T2D stage. The WHO anticipates that worldwide deaths from diabetes will double by 2030 [46]. As reported by the International Diabetes Federation (IDF), T2D early diagnosis and treatment can save lives and prevent, or significantly delay, complications [152].

T2D strongly impacts on the costs of national health systems (NHSs). According to the IDF [153], health expenditure for diabetes was estimated at US$ 105.5 billion in the European Region in 2010 (the 10% of the total health expenditure). This expenditure is expected to reach US$ 124.6 billion by 2030. The estimated costs for the European countries are around the 9%. In Italy, the total cost is about 15 billion €, with an increasing trend up to the 14.4% in 2040, slightly lower than the one expected at European level (18%) [45, 154]. T2D also causes a significant loss of productivity (work days lost, lower working efficiency, early retirement) and mortality and such social costs represent a heavy economic burden, not always easy to quantify, on society [155]. In Italy many legislative initiatives have taken place on the protection of the diabetic patients, which have merged over time into a National Diabetic Disease Plan (NDDP). Following legal considerations on T2D patient treatment that date back to 1987 [156], the NDPP from the Italian Ministry of Health has been released [157]. The NDPP has identified different areas of intervention to standardize treatments of prevention, diagnosis and monitoring of people with T2D living in Italy. The NDPP foresees a capillary network of GPs and other healthcare professionals (nurse, nutritionists, psychologist, podiatrist, cardiologist, nephrologist, neurologist, ophthalmologist, etc.) and provides regular consultation to approximately 50% of people suffering from T2D. Consequently, one of the major challenges of modern care is to develop and sustain a person-centered management of T2D that relies on interdisciplinary work, communication, data collection, continuous monitoring, and processing and well as reduction of costs. However, several Italian Regions are independently designing their own models for chronic management and reorganization of territoriality care, with inevitable inhomogeneities. Further barriers to optimal care include limited appointment times, lack of easy access to patient information, and fragmentation of data between healthcare providers. Optimizing the use of the EHRs by configuring a CDSS, changing workflow patterns to include team management, and implementing a

structured patient education can improve the management of T2D in primary care and in successive levels [158, 159].

A wider adoption of EHRs would reduce health care costs, medical errors [18, 20], healthcare disparities, patient complications in hospitals and mortality [21, 22, 160]. Moreover, sharing EHRs among healthcare professionals will decrease the use of unnecessary services, such as repeated laboratory tests every time the patient changes hospital and office visits [18, 19]. In this scenario, in order to foster the digitalization and sharing of health data in an easy and accessible way, as well as to coordinate data flows, the NMI has been established, in cooperation between FIMMG (the largest Italian federation of GPs) and Federsanità ANCI (the Italian federation of Public Health Agencies). This was done with the final goal to offer HIT services to GPs at national level.

Ermakova et al. [161] surveyed the use of cloud computing technology in healthcare. The survey pointed out the importance, for GPs, to share healthcare data under a common standard system [162, 163, 164]. On this consideration, and considering the central role of GPs in effective chronic-disease diagnosis and management strategies, a platform for GPs data sharing and unified T2D patient management was developed, guaranteeing the interoperability (e.g., using EHRs data standards) of the platform with other healthcare databases.

5.1.1. Contributions

This work overcomes solutions in the literature (Sec. 5.2), by developing a novel framework with relevant contributions:

- The proposed solution is based on the standard EHR structure used by GPs in their daily activities, ensuring large-scale use;
- A novel set of quality indicators for shared data and T2D care process quality is presented;
- The framework is equipped with a ML-based CDDS, analyzing the shared EHRs for T2D screening.
- The architecture involves quality-care evaluation by a second ML approach, with manual annotation on five quality classes;
- The CDSS testing was performed on 41237 T2D patients, one order of magnitude larger than the dataset presented in the closer work to ours [54], with real data collected from about 800 GPs;
- A quality-based economic incentive model is proposed to foster GPs empowerment. Up to our knowledge, this is one of the first real applications of quality measures to a standard chronic care model.

The proposed framework is currently used in the NMI cloud with a Software as a Service (SaaS) design that ensure scalability and real-time performances with a direct access from GPs ambulatory software in Italy. All Web Services Description Languages (WSDLs) of the proposed EHRs data standard used in this work are publicly available[1] and can be considered one of the most comprehensive data standard for GP's EHR in Europe.

5.2. Related work

5.2.1. EHR use and sharing

The adoption of EHRs can represent a possible solution to integrate data provided by different information sources transforming them into useful shared knowledge. This allows to define metrics and assessment of clinical performance as well as to take corrective actions to support better decision-making based on a set of clinical indicators defined to manage the intervention of patients with diabetes.

In a feasibility study within an Italian regional environment, Pecoraro et al. [159] show the applicability of a shared EHR in a clinical governance framework. The use of EHRs has the advantage of managing standardized data already integrated in several health infrastructures. An Austrian study [158] underlines as the continuity of care in chronic diseases has a positive impact during the patient follow-up. Yamaguchi et al. [165] investigated the effectiveness of cooperation of medical experts using data from EHRs in a medium-sized local hospital. Sharing information and electronic clinical path are the main factors in promoting inter-professional work. In addition, introducing an electronic information technology tool, the authors have reported a challenging strategy to improve T2D integrated management. T2D is a chronic and transversal disease related to many other pathologies. Therefore, the integrated management should possess a wide scalability and must be able to discriminate and evaluate other chronic complications that result from T2D [166]. In another work [167], an EHR architecture is used to discriminate prevalence and incidence of cardiovascular disease (CVD) in T2D patients. The large availability of EHR data allows to extract information relevant to favourable or unfavourable long-term strategies related to specific glucose-lowering therapies.

5.2.2. EHR analysis for CDDS

EHR-CDSS based have great potential to improve the diabetes care. A systematic review presents the potential clinical, social and economic benefits that a CDDS could add to an already existing healthcare system [168, 169]. Clinical guidelines for optimal management of diabetes are widely available, yet adherence to these guidelines

[1] `http://cloud.fimmg.org/wsdl.php?wsdl`

remains variable [170]. CDSS systems is designed to guide optimal medical therapy based on individual patient characteristics extracted from the EHR [168]. CDSS tools have been developed to provide reminders for routine laboratory testing, recommendations for specific medication choices, and alerts for potential drug-drug interactions. Electronic clinical reminders have evidenced an increased adherence to recommended pharmacotherapy and screening [171]. Holbrook et al. [172] showed that when the decision support is shared between physician and patient through a web-based interface, significant improvements in clinical diabetes care can be achieved. In a research conducted by Modafar Ati et al. [173], a knowledge based system is created and then integrated with a EHRs database as part of the national E-Health infrastructure. This is used to create a system based on Service Oriented Architecture that is able to predict or monitor the condition of any diabetic patient based on a certain number of features defined by the health authorities.

A widely adopted approach for identifying subjects with and without T2D is to involve experienced physicians that manually design algorithms based on their experience and examination of EHR data [174, 175, 176, 177]. However, such strategies increasingly prove to be limited and not scalable [174, 175, 177] due to the laborious process of human intervention and rule abstraction capabilities of experts. Furthermore, expert algorithms are often designed with conservative identification strategy, thus may fail to identify complex (e.g., borderline) subjects and miss a significant number of potential T2D cases [50]. Thus, recent work in addition to EHRs has introduced a CDSS integrated with a ML based framework [2, 3].

ML and data mining models are increasingly utilized in diabetes related research from EHR data. These studies have primarily focused on mining T2D-related EHR data for clinical purposes. For instance, some studies aimed at forecasting clinical risk of diabetes from EHR [24, 4]. Wang et al. explain as the use of a shared decision-making (SDM) process in antihyperglycemic medication strategy decisions is necessary due to the complexity of the conditions of diabetes patients. Knowledge of guidelines is used as decision aids in clinical situations, and during this process no patient health conditions are considered. It is proposed a SDM system framework for T2D patients that not only contains knowledge abstracted from guidelines but also employs a multilabel classification model that uses class-imbalanced EHR data and that aims to provide a quality care model to help physicians and patients having a SDM conversation [54, 178] and to improve chronic care models.

5.3. Methods

In this section, a CDDS framework for T2D is introduced as well as the dataset used for evaluation. The framework is depicted in Figure 5.1 and comprises five main components:

- Clinical data collection of T2D patients from EHRs and data sharing in a cloud infrastructure (Sec. 5.3.1;
- T2D patients enrollment (Sec. 5.3.2);
- Data indicators and features (Sec. 5.3.3);
- Enrolled patient management (screening and follow-up): Self-Audit & Data Quality (Sec. 5.3.4);
- Quality score for economic incentives (Sec. 5.4.4).

The framework is comprehensively evaluated on the a T2D dataset collected for this work. The details of the data collection and ground truth labeling are also discussed (Sec. 5.4.1).

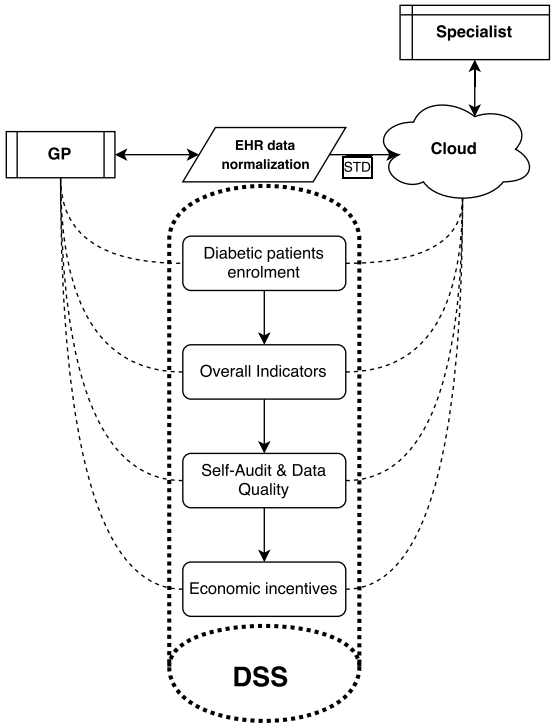


Figure 5.1.: GP's workflow in T2D Integrated Management Care.

GP membership to the system is free and there are no sanctions for GPs that do not intend to attend. GPs involved and patients cooperate to apply scientific guidelines. The flowchart in Fig. 5.2 shows a CDSS for T2D patients integrated management care.

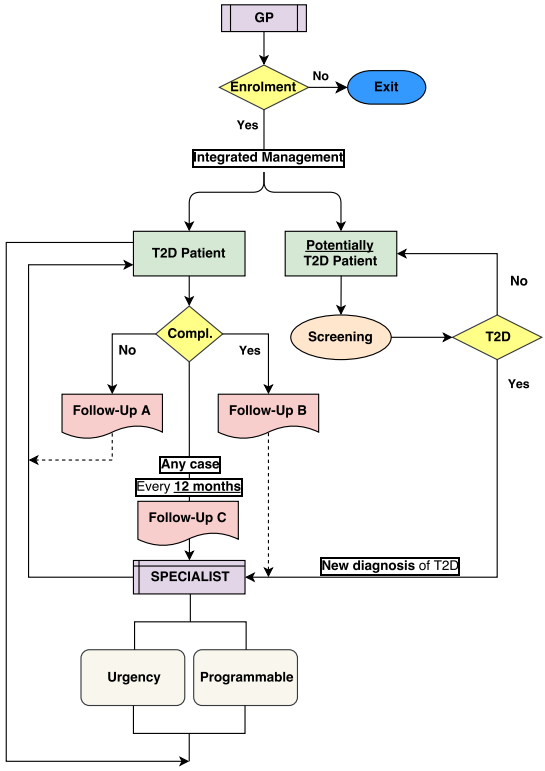


Figure 5.2.: CDSS for T2D patients integrated management care managing patient enrolment and treatment in the same conceptual flow and using the same data features

5.3.1. Data collection of T2D patients from EHRs and data sharing in a cloud infrastructure

The NMI platform manages a cloud computing project that, through the integration of GPs' EHR databases, is able to realise: network medicine, audit process, data reporting, integrated management programs between GPs and specialists for treating chronic pathologies. NMI aggregates databases available from GPs in a unique standardized language and share them in a cloud platform. The database is available for transversal interoperability with other GPs, and for a vertical interoperability with several healthcare professionals of the NHS. The architectural features of the system meets stringent requirements: security and privacy, high reliability, ease of access and wide interoperability through the availability of flexible interfaces and standard communication protocols. Additionally, the system is equipped with services and tools which make it useful and usable by general practitioners, satisfying his/her needs for practice of the daily-life profession.

Authentications and authorizations Data are security protected through encryption during both transfer and storage. Data access is strictly allowed to only those that have the required permissions [150]. In particular, all connections and accesses are tracked and are subject to verification of the credentials and the possessed permits. A second level of access is regulated by a further 32-byte long key. The key is issued directly by Netmedica Italia to access different services according to the needs of the user.

Interoperability via Web Services Interface The frameowrk allows a high degree of interoperability with other applications. It has an API interface to intercede directly with the cloud database. Most of the features of the Netmedica Cloud are provided by a Web Service Interface that exposes various functionalities through the publication of precise methods to be invoked. The name of the main interface is FIMMGwsdl. It is based on the Simple Object Access Protocol over HyperText Transfer Protocol and the default style is Remote Procedure Call.

The features offered by the web services can be grouped into the following macro categories: Access, Writing, Consultation, and Other Services. Other services include features such as: notifications, patient report and delete record. These features and the relative methods are used by the extractor program that is specifically designed to receive the complete encrypted patient card. All patient cards are potentially analyzed, uploaded and stored, still encrypted, to the unified and normalized database.

Data extractor The data centralization procedure of Netmedica Cloud is based on the specially developed automatic extractor software called NetDesk. All EHRs are first encrypted with the GP's secret key and then transferred to the Netmedica cloud. The GP can install automatically the program with a wizard. Data are collected from the outpatient database, by applications that allow the extraction of clinical features of outpatients. The process of extraction normalizes the database according to a record layout defined in XML. After extracting and standardizing data into XML, through Web Services, data are forwarded to the cloud, where they are aggregated into a normalized database.

The extraction process takes place in 2 phases: first massive data extraction, successive extractions according to incremental logic. The GP can arrange the timing of the extraction (every 10, 30 ... minutes), even differing in a daily time when he is not using the PC. In clinics where more general practitioners work, it is possible to install the extractor on a network server that accesses the medical management software and the cloud database in multi-user mode. A GUI interface is available for managing user authorizations, scheduling the process, timing and extraction type.

Database platform Through the WSDL, many services and applications communicate with the database. The database allows on-line sharing of care data, even among

professionals that normally use different ambulatory management software. The organization of the data takes place in a patient centric manner and its structure has been designed as flat as possible. The figure 5.3 shows the available EHR field in the database.

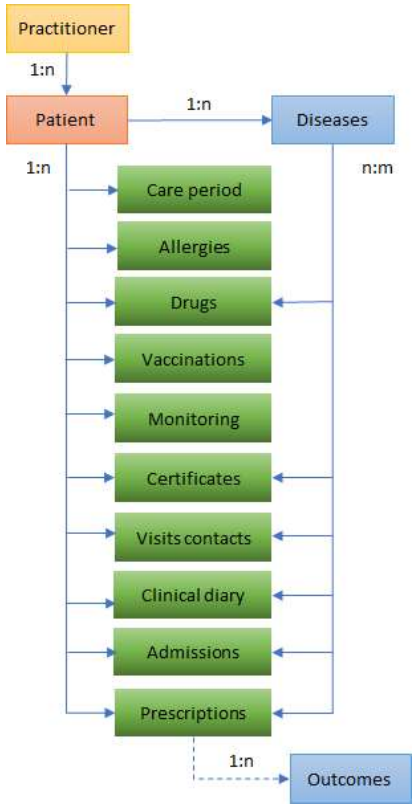


Figure 5.3.: The available EHR fields in the database

A main patient registry table contains all the patient's id informations. These informations are appropriately encrypted according to the key assigned to the physician.

5.3.2. T2D patients enrollment

Once the GP is logged in with his/her own credentials, the system automatically proposes a list of T2D patients or potentially T2D (Table 5.1) extracted from the EHRs. Only T2D patients without major (uncontrolled) complications are automatically enrolled.

Complications that lead to patient exclusions are:

- The presence of, at least, one of these pathologies: coronary heart disease, ictus,

Table 5.1.: T2D patients' enrolment CDSS output. Examples with real (anonymized) data.

					Update Indicators		Add Patient	
ID Patient	**Type I**	**Type II**	**Exemptions**	**Drugs**	**Checks**	**Complications**	**Enrolment**	**Actions**
A0001					*	*		Add
A0002					*	*		Remove
A0003		*					*	Remove
A0004		*	*		*		*	Remove
A0005			*				*	Remove
A0006					*			Add
A0007		*		*			*	Remove
A0008		*					*	Remove
A0009		*	*		*		*	Remove
A0010		*	*		*		*	Remove
A0011		*	*		*		*	Remove
...	...	...	...	...	...	...	...	...

peripheral arterial disease, diabetic or hypertensive retinopathy;

- The presence of uncompensated diabetes: estimated in the presence of HbA1c $> 8\%$ in the last year;

- Insulin treatment in the last year.

Potential T2D patients may be automatically added by GPs. Potential T2D patients are subjects who, while not presenting the NHS code for T2D, have a high chance of developing T2D due to ongoing pharmacological treatment or specific clinical checks. With the "Add Patient" button, the physician starts collecting and maintaining informed all those patients who adhere to the integrated management path.

Table 5.1 shows an example of the output of the enrollment that suggests patients to be added or removed. The final decision is always performed by the GP.

5.3.3. Data indicators and features

For a proper and correct implementation of a T2D integrated management system, it is necessary to define overall indicators for monitoring and evaluating treatment results. The care-quality indicators used to allow GPs to monitor and improve the care process of diabetic patients were taken from the international literature and the most important international guidelines on diabetes management. In order to evaluate the quality of the assistance provided and the conformity with the standards defined in the guidelines, it is crucial to identify process and outcome indicators to measure the achievement of the set goals. Specifically, the proposed indicators allows to control the activities of chronic care model and ultimately evaluate the capability of the integrated management pathway. Thus, these represent salient information to verify if and in which entity the totality of results has reached the set goals for improving the chronic care quality model.

The proposed care-quality indicators are:

- # 1.1 Indicator *"T2D Patients"* highlights the correspondence of the percentage of T2D patients with respect to total assisted patients enclosed in GP's EHR. From here on, the following ratios will be calculated only respect to the total of enrolled T2D patients that have express the consent for adhering to integrated management system . The GP will schedule the visit of patients with T2D already under treatment and in according to the integrated follow up will update the EHR with indicator data required. The physician requires the determination of each indicator and records the value in the EHR.

- # 2.1 Indicator *"Diabetics with annual HbA1c"* is obtained from the ratio between T2D patients with HbA1c monitored minimum during last 12 months and the total number of T2D patients. This indicator expresses an adequate follow up of the patient.

- # 2.2 Indicator *"Diabetics with annual lipid profile"* is obtained from the ratio between T2D patients with lipid profile monitored minimum during last 12 months and the total number of T2D patients. It is demonstrated that LDL cholesterol reduction in diabetic patients reduces severe cardiovascular risks. The measurement cannot be calculated for triglyceride values > 200 mg/dl.

- # 2.3 Indicator *"Diabetics with annual AP"* is obtained from the ratio between T2D patients with arterial pressure measured minimum during last 12 months and the total number of T2D patients. It is evidenced that the average prevalence of hypertension in diabetes is about 50%.

- # 2.4 Indicator *"Diabetics with BMI"* is obtained from the ratio between T2D patients with Body Mass Index measured and the total number of T2D patients. This indicator is indispensable to evaluate the effectiveness of therapy and can suggest a cardiovascular risk factor. The physician processes BMI and records the values in the EHR and, finally, performs an educational reinforcement.

- # 2.5 Indicator *"Diabetics with Waist"* is obtained from the ratio between T2D patients with waist measured and the total number of T2D patients. This indicator is indispensable to evaluate the effectiveness of therapy and can suggest a cardiovascular risk factor. The physician processes this value and records it in the EHR and, finally, performs an educational reinforcement.

- # 2.6 Indicator *"Diabetics with annual Microalbuminuria"* is obtained from the ratio between T2D patients with microalbuminuria measured minimum during last 12 months and the total number of T2D patients. Microalbuminuria is an early marker of diabetic nephropathy when there is still hope for reversibility or arresting progression.

- # 2.7 Indicator *"Diabetics with annual Creatinine"* is obtained from the ratio between T2D patients with creatinine measured minimum during last 12 months and the total number of T2D patients. This indicator is a very sensitive and specific index of glomerular insufficiency. It is important not only to diagnose kidney failure, but also for any contraindications to the use of nephrotoxic drugs.

- # 3.1 Indicator *"Diabetics with HbA1c $\leq$ 6.5%"* is obtained from the ratio between the number of T2D patients with latest registered value of HbA1c $\leq$ 6.5% and the total number of T2D patients. Values below 6.5% prevent the onset of complications.

- # 3.2 Indicator *"Diabetics with LDL $\leq$ 130 mg/dl"* is obtained from the ratio between the number of T2D patients with latest registered value of LDL cholesterol $\leq$ 130 mg/dl and the total number of T2D patients. Reduction in LDL cholesterol values reduces cardiovascular risk. Physician reinforces life fitness education, evaluates therapeutic strategy after stratification of cardiovascular risk.

- # 3.3 Indicator *"Hyp diabetics with AP $\leq$ 130/80 mmHg"* is obtained from the ratio between T2D and hypertensive patients with AP registered value $\leq$ 130/80 mmHg and the total number of T2D patients. Antihypertensive therapy in diabetic subjects, if effectively conducted, reduces micro and macrovascular complications. The GP monitors the values of the Arterial Pressure and, eventually, modifies the therapy.

Table 5.2 shows aggregated data collection under the evaluation period. For each indicator, the (*"Ratio"*) achieved by the GP is shown. The *"Ratio"* estimates the correlation between the single indicator and the entire patient population.

The percentage Ratio achieved by the physician for each indicator is compared with the expected *"Target"* established by the diabetes project, and its positive or negative *"Distance"* from the target is calculated. At this step, in order to assign the overall *"Acceptable Level of Performance (LAP)"* score, only if the target is exceeded by the ratio, the LAP score provided by each indicator is assigned. The *"Mean"* value of the indicator of all GPs participating in the project is also reported.

Every GP can also consult the following tables that report for information purposes the use of antidiabetic drugs (Table 5.3) and the detection of complications(Table 5.4). For each indicator, the GP can compare his performance with correspondent average value reached by all GPs participating in the project. In particular, the information reported in Tables 3 and 4 refers to only one GP. However, the "Mean" value represents the average of all GPs participating in the project. Although these Tables represent standard medical information related to antidiabetic drugs subministration and complications in act, it allows the comparison between the incidence of each indicator for each GP and the average value reached by all GPs participating in the project. Data

Table 5.2.: Care quality Indicators under evaluation for improving the clinical performance. Data from a single GP.

#	Description	Num	Den	Ratio	Target	Distance	Mean	LAP
1.1	T2D Patients	74	1501	4.93%	3%	64.33%	6.17 %	150
2.1	Diabetics with annual HbA1c	43	74	58.11%	70%	-16.99%	66.41 %	100
2.2	Diabetics with annual lipid profile	56	74	75.68%	60%	26.13%	46.64 %	100
2.3	Diabetics with annual AP	24	74	32.43%	90%	-63.96%	63.35 %	100
2.4	Diabetics with annual BMI	5	74	6.76%	70%	-90.35%	72.73 %	50
2.5	Diabetics with annual waist	0	74	0	50%	-100%	53.95 %	50
2.6	Diabetics with annual microalbuminuria	10	74	13.51%	50%	-72.97%	31.00 %	100
2.7	Diabetics with annual creatinine	59	74	79.73%	60%	32.88%	41.25 %	50
3.1	Diabetics with HbA1c $\leq$ 6.5%	35	74	47.3%	25%	89.19%	51.95 %	100
3.2	Diabetics with LDL $\leq$ 130 mg/dl	43	74	58.11%	20%	190.54%	37.44 %	100
3.3	Hyp diabetics with AP $\leq$ 130/80 mmHg	5	50	10%	20%	-50%	37.68 %	100

indicators were used specifically to determine the economic incentives, but globally they may help every single GP to improve diabetes care by focusing on specific complications and drugs that have a greater incidence for them.

Table 5.3.: Antidiabetics drugs subministration. Data from a single GP.

Description	Num	Den	Ratio	Mean
Diet treatment	49	75	65.33%	20.67%
Insuline treatment	2	75	2.67%	19.46%
Metformin treatment	21	75	28%	44.71%
Sulfaminide treatment	6	75	8%	13.44%
Acarbose treatment	1	75	1.33%	5.03%
Pre cost treatment	3	75	4%	24.01%

Table 5.4.: Complications in act. Data from a single GP.

Description	Num	Den	Ratio	Mean
Ischemic cardiopaty	3	75	4%	17.27%
AMI	1	75	1.33%	2.84%
Revascularisation	1	75	1.33%	1.82%
Claudicatio	2	75	2.67%	0.37%
TIA	1	75	1.33%	1.74%

These indicators, together with a subset of the EHRs, are used to perform a machine-learning-based evaluation of chronic care quality (Subsection 5.4.2).

5.3.4. Enrolled patient management (screening and follow-up)

The patient management comprises the following steps: Potentially diabetes patient screening, New diagnosis of T2D, Follow-up A of T2D patient without complications, Follow-up B of T2D patient with stabilized complications, and Follow-up C of all patients with T2D.

Potentially diabetes patient screening (Figure 5.2) To enroll patients that are suspected to develop T2D, GPs inspect and record lifestyle habits (eating habits, alcohol, smoking, physical activity, work activity), measure and record weight, height, BMI, AP, waist and calculate and record the cardiovascular risk score [179]). The screening of potentially diabetic patients was performed periodically by evaluating the fasting plasma glucose test (more cost-effective than HbA1c and OGTT) in subjects over 45 years. GPs perform a fasting plasma glucose test to discriminate diabetes in subjects with BMI $>$ 25 kg/m^2 and at least one or more of the following conditions in subjects under 45 years:

- Physical inactivity;
- 1st degree familiarity with T2D;
- Belonging to a high-risk ethnic group;
- Arterial hypertension ($\geq$ 140/90 mmHg) or antihypertensive therapy in act;
- HDL cholesterol $<$ 35 mg/dl and/or triglycerides $>$ 250 mg/dl;
- Past diagnosis of gestational diabetes or infant birth with $>$ 4 kg weight;
- Previous diagnosis of Impaired Glucose Tolerance (IGT) or Impaired Fasting Glucose (IFG), HbA1c 42-48 mmol/mol;
- Insulin resistance;
- Clinical evidence of cardiovascular disease (AMI, stroke, claudicatio, etc.) according to a cardiovascular risk score [179].

In the absence of the previous criterion, screening should start at the age of 45 years.

If the blood glucose is not diagnostic for diabetes ($<$126 mg/dl), screening should be repeated at least three years, considering a more frequent test for subjects with dysglycemia ($>$ 100 and $<$ 126 mg/dl). In addition to diabetes, other dysglycemia patterns are known. To define these conditions, however, the use of the term "prediabetes" may be misleading and thus not recommended. Hence, the following values of the main glycemic parameters should be considered, as they identify subjects at risk of diabetes and cardiovascular disease [180, 181, 182]:

- Fasting blood glucose 100-125 mg/dl (IFG)
- 2-hour glucose after OGTT 140-199 mg/dl (IGT)
- HbA1c 42-48 mmol/mol (only with IFCC aligned assay)

New diagnosis of T2D (Figure 5.2) GP makes a general visit and prescribes the first indications on lifestyle (diet, physical activity, smoking abolition, etc.). Moreover, GP considers the opportunity to initiate drug therapy (metformin, if not contraindicated) and to send the patient to the dietician. Finally, GP requires investigations for the first diagnostic check-up by the specialist:

- HbA1c, total cholesterol, HDL, LDL, triglycerides; creatinine, AST, ALT, GGT, blood count;
- Microalbuminuria;
- Full urine examination;
- ECG (and cardiologic examination at discretion);
- Fundus oculi.

Then GP sends the patient to diabetes center to perform:

- Diagnostic overview;
- Specialists clinical staging and any complications;
- Certification for diabetes exemption;
- Compilation, if necessary, of the therapeutic plan, assessment of care criticality, individual or group therapeutic education planning.

Finally, depending on the clinical condition, the specialist:

- Starts not complicated T2D patients' follow-up (follow-up A);
- In agreement with GP, approves the care plan for insulin-dependent diabetes and/or complications and/or inadequate control (follow-up B).

Follow-up A of T2D patient without complications (Figure 5.2) The care quality is also based on specific follow up for every enrolled patients. The proposed NMI system requests every GP to register data relevant to the follow up process, which are automatically retrieved when requested by the GP. GP conducts a general medical examination: history to detect urinary, visual, cardiovascular and neurological disorders (erectile dysfunction, muscle cramps, paraesthesia, skin disorders, etc.); peripheral wrists, vascular soffits, heart rate, tendon reflexes, tactile sensitivity examination, skin and feet examination.

Every 3 months within GP's dedicated outpatient clinic:

- Body weight, BMI and waist;
- AP;

- Evaluation of the blood glucose control performed by the patient.

GP each year prescribes: HbA1c, blood glucose and any other examinations based on clinical judgment and/or how agreed with the diabetic specialist, full urine examination, microalbuminuria, clearance, creatinine, total cholesterol, HDL, triglycerides, ECG.

GP every 2 years prescribes fundus oculi and record results in the EHR.

Follow-up B of T2D patient with stabilized complications (Figure 5.2)

Every 6 months GP sends patients with stabilized complications to the diabetes center:

- Activities suggested by Follow-up A;
- In relation to clinical needs, diabetic pathology specialist (including examination aimed at finding lesions of the feet).

Depending on the intervals programmed for insulin-treated diabetics and/or with evolving complications and/or inadequate control, GP sends the patient to the diabetes center in case of:

- Periodic inspection, if provided by the individual care path, agreed with the diabetic team;
- Social-welfare criticisms that lead to erroneous or non-therapeutic adherence;
- Failure to maintain agreed therapeutic goals, especially if present:
  - Severe and/or repeated hypoglycemia;
  - Rapidly evolving neurological, renal, ocular or macrovascular complications;
  - Diabetic foot (ulceration or infection);
  - Pregnancy in diabetes, gestational diabetes.

Moreover, diabetic center can:

- Perform further specialist examinations (ecocolordoppler, angiographic exams, percutaneous oximetry, electromyography, retinography, etc.);
- Activate additional therapeutic treatments;
- Agree with GP for any personalized clinical-therapeutic-assistance plan (in the case of diabetes with evolving complications);
- Manage with a multidisciplinary approach, and according to organizational resources, patients who have:

  - Severe metabolic instability;
  - Neurological, renal, ocular or macrovascular complications that are rapidly evolving;
  - Diabetic foot (ulceration or infection);
  - Erectile dysfunction;
  - Pregnancy in diabetes, gestational diabetes.

Follow-up C of all patients with T2D (Figure 5.2) Every 12 months, GP sends patient to the diabetic centre to allow annual screening, sharing all the available data. If the clinical conditions are stable, the annual renewal of the therapeutic plan will be reported directly by GP, otherwise a new one will be planned.

The integrated management provides a specialist's visit in the following cases (beyond the new diagnosis and annual screening):

- Urgency:
  - Acute metabolic deficit;
  - Repeated episodes of hypoglycaemia;
  - Pregnancy in diabetic women;
  - Appearance of foot ulcer or ischemic and/or infectious lesions at the lower extremities.
- Programmable:
  - Repeated glycemic fasting $>$ 180 mg/dl;
  - HbA1c $>$ 6.5% in two consecutive determinations;
  - Appearance of clinical signs related to complications.

The NMI infrastructure makes data processing possible for individual physicians or diabetic team for every one of the previously described follow up. Data processing is focused on audit tools for improving the use of the medical tool by the physician and on specific local projects aimed at the treatment of chronic diseases or prescriptive appropriateness. The processing system is built in such a way as to maintain historical memory of past elaborations for reporting data or for checking trends. Using a standard data format, regardless of the record software used by the physician, facilitates the collection, processing, and sharing of information.

5.3.5. Self-Audit & Data Quality

Through self-audit, the physician can evaluate his performance compared to colleagues on a set of standard indicators that can be subdivided into four areas:

i. *Recording completeness:* The accuracy of the collected data, the presence of the main outpatient data (AP, BMI, etc.) and the recording of laboratory results in numerical format are evaluated;

ii. *Adherence to prevalence:* Distances from the prevalence of the major chronic conditions are shown to the physician. Patients potentially affected by the pathologies under investigation are reported by examining therapeutic prescriptions, examinations carried out, exemptions granted etc.;

iii. *Treatment of chronic diseases:* The physician is evaluated on the main indicators of fitness identified by the international guidelines as compared to the main chronic diseases;

iv. *Contact intensity:* In addition to the quality of the recordings, the amount of these records is also measured. The purpose is to document the activities of the physician.

Table 5.5 shows the indicators under LAP score evaluation for each patient. For the indicators: HbA1c, LDL cholesterol and Pressure (only if the patient is hypertensive), the cell assumes green or red colour, as the result for the examination falls within the thresholds established by project. For indicators: BMI, waist, microalbuminuria and creatinine, the star symbol in green is displayed if the data is recorded with a coherent numeric value within the period indicated by the project. The system finally has the complete set of features for every T2D patient (Table A.1) where all the indicators required by the project are displayed. If a data used in the dataset for the care quality and economic incentives evaluation is not registered, the relative row is shown in grey. Process indicators are continuously monitored. If a data is not collected, the physician is alerted by the system to understand the nature of the problem. If a data is collected incompletely, a different type of notification is sent proactively to the physician.

Table 5.5.: Self Audit & Data Quality aggregated visualization. Sample data for a GP. Red show warning data for a particular feature of an enrolled patient.

Patient	HbA1c	LDL	Press.	BMI	Waist	Micr.	Creat.
Patient 1	6.54	150					★
Patient 2	8.46	120	80/150				★
Patient 3	7.27	117		★			★
Patient 4		123					★
Patient 5	6.82	144				★	★
...	...	...	...	...	...	...	...

5.4. CDSS analysis on a real-use case for quality care evaluation

The results of care quality evaluation relevant to 2018-2019 are presented. The quality of care was evaluated for every GP based on T2D Patient's feature set (see Table A.1), with the main purpose to foster care-delivery quality improvement. The dataset annotation was firstly described in Section 5.4.1, while Section 5.4.2 described the ML approach and results. Finally, Section 5.4.4 reported the impact of the proposed CDSS in terms of economic incentives.

5.4.1. Dataset annotation

A subset of the dataset comprised of 1780 patients was extracted from the entire dataset (41237 patients) and was manually annotated by experts. This distribution was equally balanced across the five follow-up phases (19% Potentially diabetes patient screening, 21% New diagnosis of T2D, 20% Follow-up A of T2D patient without complications, 22% Follow-up B of T2D patient with stabilized complications, 18% Follow-up C of all patients with T2D). Each of the follow-up phase (i.e. Potentially diabetes patients screening, a new diagnosis of T2D, Follow-up A of T2D patient without complications, Follow-up B of T2D patients with stabilized complications and Follow-up C of all patients with T2D) described in Section 5.3.4 was manually annotated by a team of 10 experts (5 from GPs leading group and 5 from diabetic centers). The experts evaluated the chronic care quality according to a 5-Likert ordinal scale [183] ranging from level 1 (Excellent) to level 5 (Poor). The labels may be affected by the inter-observer/expert variability: the experts can evaluate the chronic care quality in a different way based on their different motivation, experience and background knowledge. For this reason, this problem was alleviated by averaging the response of the ten expert GPs according to a majority vote approach. As future work, it may possible to rank the label according to a confidence level [184] and to further minimize the inter-rater variability by developing a Multi-task learning approach and maximizing a consensus among annotators.

The input data of the classifier were represented by the T2D patient's feature set described in Table A.1. The majority vote of the expert ratings represented the chronic care quality ground-truth. The final dataset was comprised of a total of 1780 observations balanced across the five follow-up phases. Two years interval (2018-2019) was considered for learning and evaluating the ML model while the data of the following 6 months were used to evaluate the improvement of the economic incentives.

5.4.2. Machine learning approach

The Random Forest (RF) [98] was employed for classifying chronic care quality. RF is a variant of bagging proposed by [98] and consists of an ensemble of decision trees (DT) (i.e., n° of DT) generated by an independent identically distributed random vectors. RF is developed by sampling from the samples, from the features (i.e., n° of features to be selected) and by changing two tree-parameters (i.e., max n° of splits and max n° of size) [99]. The splitting features for each node was computed according to the Gini index metric.

The model was built using Azure Machine Learning Studio and was deployed as web services on the proposed Service-oriented architecture. The 10 cross-validation (CV) procedure was implemented, dividing all datasets into 10 folds and selecting iteratively nine folds for training and one fold for testing. This procedure was stratified across the five follow-up phase. The optimization of the RF hyperparameters (i.e., n° of RT, max n° of splits, n° of features to select at random for each decision split) was performed implementing a grid-search and optimizing the macro-recall in a nested Fivefold Cross-Validation.

5.4.3. Machine learning results

The RF achieved an overall accuracy (averaged over the 10 fold) of $98\% \pm 2\%$ and macro-recall of $96\% \pm 1\%$. This result suggests how there is a close dependency between the indicators displayed in Table A.1 and the chronic care quality ground-truth. Moreover, these indicators are informative for each of the follow-up phases of T2D patients. Accordingly, the proposed CDDS might be exploited to support all GPs over time by providing incentives for moving from one class to another (i.e. from Poor to Excellent) with the main objective of improving the chronic care quality.

Furthermore, the extracted results pointed out how the proposed CDDS, along with the economic incentives, brought to a significant improvement of class A (i.e. an increasing number of patients in class A [Follow-up A of T2D patient without complications]), with more than 12% of the increase in the first 6 months. This result refers to a six-month prospective outcome of the selected samples (1780 patients). The total incentive costs are irrelevant if compared with the impact of the care quality and reduction of T2D consequences over time and their social costs.

5.4.4. Impact of the proposed CDDS: Quality care evaluation for economic incentives

By selecting the LAP column, GP observes the LAP schedule and the incentives calculated for the reference period (Table 5.6). By selecting the Diabetics column, the enrolled patients' schedule is shown. The GP visualizes the overall indicators that determine the LAP score. Moreover, by selecting the patient's name, GP access to every

Table 5.6.: Data delivered consultation

Date	Diabetics	LAP	Enrollement inc.	LAP inc.	Total inc.
2019	72	500	3. 600,00 €	2.160,00 €	5.760,00 €

detailed indicator.

GPs receive the remuneration of 50 € per patient enrolled. In addition to standard remuneration, the GP annually receives a bonus related to the LAP score achieved. For determining the overall "LAP Score", GP's performance is compared with the project "Target" for each indicator. If the target is reached or exceeded, the "LAP" for the indicator is assigned (see Table 5.2).

Thus, the LAP score determines a further economic remuneration (LAP inc.) that can be 30, 40 or 50 € per patient/year, as showed in Table 5.7.

Table 5.7.: LAP score incentives

LAP	LAP inc. per patient
From 300 to 599	30 €
From 600 to 5799	40 €
From 800 to 1000	50 €

The bonus is accumulated over the test period and can be used in similar cases both from the economic point of view or as a performance indicator that can be transformed into different loyalty programs.

5.5. Discussion

The proposed CDDS laid the foundation for enhancing the sharing of information among other GPs by allowing a more planning clinical diagnosis and analysis and continuity of assistance to patients who need it. The experimental results show how the ML model is able to support the GP while accurately predicting the chronic care quality based on specific indicators selected by GPs. However one important limitation of the proposed pilot study may be the specific focus on diabetes care quality. In medicine, all the guidelines consider the management of the pathology in standard conditions or with the most frequent and known comorbidities; then it is up to the physician's preparation and awareness to adapt the recommendations of the guidelines to the specificity of the individual patient, also considering the infinite variability of individual clinical conditions. Future work may be addressed to (i) validate the proposed CDDS for the management of different chronic diseases and (ii) generalize and standardize these quality indicators for the prediction of chronic care quality related to different pathology.

Chapter 6.

Short-term kidney disease evolution

Kidney disease (KD) may hide complex causes and is associated with a tremendous socio-economic impact. A timely identification and management from the first level of medical care represent the most effective strategy to address the growing global burden sustainably. Clinical practice guidelines suggest utilizing estimated Glomerular Filtration Rate (eGFR) for routine evaluation within a screening purpose. Accordingly, the analysis of Electronic Health Records (EHRs) using Machine Learning (ML) techniques offers great opportunities to monitor and predict the eGFR trend over time. This work aims to propose a novel Semi-Supervised Multi-Task Learning (SS-MTL) approach for predicting short-term KD evolution on multiple General Practitioners' EHR data. the SS-MTL approach was demonstrated to be able to (i) capture the eGFR temporal evolution by imposing a temporal relatedness between consecutive time-windows and (ii) exploit useful information from unlabeled patients when labeled patients are less numerous with a gain of up to 4.1 % in terms of *Recall*. This situation reflects the real-case scenario, where available labeled samples are limited, but those unlabeled much more abundant. The SS-MTL approach, also given the high level of interpretability, might be the ideal candidate in general practice to get integrated within a decision support system for KD screening purposes.

6.1. Introduction

Kidney disease, often incautiously underestimated as a comorbidity of diabetes or hypertension, may hide complex causes and is associated with a tremendous socio-economic impact [185, 186]. Worldwide, 19 million disability-adjusted life-years were directly attributable to a reduced GFR, which measures the health-state of kidney functionality [187]. According to WHO recommendations, if KD is early-diagnosed and an effective screening strategy is adopted, the worsening of kidney function can be slowed or averted by inexpensive interventions [188]. Thus, the timely identification and management of chronic KD (CKD) from the first level of medical care (e.g., general practice) represent the most effective strategy to address the growing global

burden sustainably. Most recent clinical practice guidelines suggest utilizing eGFR for routine evaluation within a screening purpose, rather than a GFR measure, needed when an accurate assessment is required [189]. The 6 CKD stages strictly based on eGFR values serve to assess the kidney functionalities [190]. Accordingly, the analysis of EHRs using Machine Learning (ML) techniques offers a great opportunity to monitor the eGFR trend over time and predict its value in the short-term period. Unfortunately, in a real-case scenario, EHRs collected by GPs include several challenges such as multi-source and non-standardized data, incomplete or missing values, registration errors, data sparsity, privacy-preserving, etc [191].

Patients are followed over some time by GPs, which, at each visit, store a large variety of clinical events (i.e., exam prescriptions, medications, pathologies, lab tests, etc). Thus, eGFR evolution can be modeled using Multi-Task Learning (MTL) approach [192, 193], where the prediction of the eGFR status at a single time point is considered as a task and the predictive models at different time points may be similar because temporally related. Differently from the intensive care unit EHR datasets [194], in GPs scenario the limited availability of i) patients (i.e., spatial-transversal data) and/or ii) patients' medical history (i.e., time-longitudinal data) precludes an adequate labeled sample size (i.e., annotation of eGFR status over time) for exploiting a robust and representative supervised learning strategy. Usually, labeled data are expensive to collect and unlabeled data are abundant. Accordingly, also in the best-case scenario where a large amount of transversal and longitudinal data is available, the label might be sparsely distributed over time. This point is a crucial issue in the clinical-use case, where data labeling is prohibitive (especially for the healthy subjects) and possibly captures only the most important events of pathological subjects, and besides, unlabeled data are abundant.

Starting from these motivations, the work aims to propose a novel Semi-Supervised Multi-task Learning (SS-MTL) approach for predicting short-term KD evolution on multiple GPs' EHR data. The SS-MTL approach combines a Semi-Supervised Learning (SSL) strategy with an MTL procedure to i) impose a temporal relatedness between consecutive time windows to predict the eGFR status over time and ii) exploit both labeled and unlabeled samples in the learning procedure for capturing high-discriminative temporal patterns. Thus, two research questions (RQs) are formulated to measure the effectiveness of the proposed approach for state-of-the-art approaches:

- ***RQ1:*** *Is the MTL approach capable to capture the eGFR temporal evolution?*

- ***RQ2:*** *Is the SS-MTL approach capable to capture useful information from unlabeled patients?*

6.2. Related work

Machine Learning techniques have been already adequately proven to be effective in dealing with sequential temporal data in many applicative research areas, including especially healthcare scenarios. In particular, EHR data have been largely exploited to accomplish predictive tasks such as stages of chronic diseases, disease complications, intensive care unit clinical events, etc. These approaches spread from standard ML models such as Logistic Regression (LR) [122, 195, 196], Decision Tree (DT) [197], Random Forest (RF) [122, 196], Gradient Boosting Tree (Boosting) [122], Support Vector Machine (SVM) [2, 197] to more appealing and complex Deep Learning (DL) frameworks, mostly based on feedforward [198], Long-Short Term Memory (LSTM) [122], and Convolutional Neural Network [122] architectures.

The MTL approach is a well-known and consolidated learning paradigm to address health informatics and clinician prediction tasks, capable of extracting useful information from multiple related tasks and improve the overall generalization performance [199]. In [198, 200] authors tried to answer when MTL improved prediction performance for different clinical tasks using EHR data. Multi-task feedforward [198] and multi-task LSTM networks [200] were compared with baseline single task networks and LR models. Most related to our work is the paper [193], where a temporal MTL was adopted to stratify the risk of renal function deterioration. In fact, the different clinical tasks do not differ by their intrinsic nature (i.e., eGFR prediction), but from their temporal evolution (i.e., time windows). Differently from [193], in our work, this problem is modeled as an SSL scenario, where the label is sparse over time. Additionally, the whole raw EHR data is used rather than performing a feature selection for each task, so as to potentially avoid a lack of relevant information to detect hidden patterns.

As mentioned before, in GPs EHR data, labeled data are expensive to collect and unlabeled data are abundant. Moreover, even if originally a huge amount of labeled data is available, during a real-case scenario usually happens that after the preprocessing stage (e.g., inclusion/exclusion criteria) a considerable amount of labeled data is going to be reduced [4, 193]. Thus, the precondition of collecting a huge labeled sample size is necessary, but not easily satisfied especially in the GP scenario where large and publicly available datasets are limited. In [130, 201] the training labeled sample size was augmented using GANs and conditional GANs, respectively, without considering unlabeled data. Given this operational necessity to retrieve labeled information, MTL could be combined with SSL, leading to Semi-Supervised Multi-Task Learning (SS-MTL) paradigm, where a training set of each task consists of both labeled and unlabeled data to exploit useful information contained in the unlabeled data in order to further improve the MTL performance. A similar rationale was proposed in [202], where a multi-task setting based on an SSL technique, named Positive and Unlabeled learning (PU), was implemented for addressing a disease gene prioritization

problem. A different SSL technique (i.e., Label Propagation [LP]), which constructs a similarity graph over all input data, was proposed in [203] to generate personalized drug recommendations by leveraging patient similarity and drug similarity analytics. In our proposed approach, the Self-Learning Algorithm (SLA) inspired from [204] is utilized as SSL paradigm, which during the training stage (i.e., negative and positive samples), iteratively assigns pseudo-labels to the set of unlabeled training samples that have their margin above a threshold automatically achieved from this bound.

After evaluating the state-of-the-art, our proposed SS-MTL approach represents the first attempt to combine the SSL paradigm in an MTL scenario where the main goal is to predict the eGFR evolution based on EHR data.

Therefore, the applicative theoretical novelty of this work actively contributes to the biomedical informatics field when a large number of unlabeled samples and a temporal relatedness between consecutive tasks are involved. In this work, the SS-MTL approach, capable to predict and explain the short-term KD evolution, contributes to improve the KD management especially at an early stage. Thus, in general practice, the SS-MTL approach may be integrated in a decision support system for screening purposes.

6.3. Clinical data: mFIMMG dataset

The publicly available *m*FIMMG dataset[1], which is extracted from the standardized FIMmG Netmedica Cloud computing infrastructure [150, 1], stores a 10-year ($2010-2019$) activity collected by 6 GPs, and consists of 14175 patients and 6 main fields. The demographic field is composed by age and gender. The monitoring field (i.e., diastolic and systolic blood pressure, height, weight, and waist) contains only continuous predictors, as well as the lab tests field where all the laboratory outcomes (e.g., eGFR) are stored. The remaining fields (i.e., pathologies, drugs, exam prescriptions [exams]) are all categorical.

6.3.1. Preprocessing

Figure 6.1 shows all the preprocessing procedure: i) *eGFR*, ii) *Labeled samples*, and iii) *Temporal data*.

6.3.1.1. eGFR

The eGFR index was calculated by the authors using a unique the CKD-EPI formula [205, 206], as a combination of 4 factors:

$$eGFR = f(creatinin, age, gender, race) \tag{6.1}$$

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

This rationale mitigated the inter-laboratory variability. Among all patients, let call labeled samples (#5812) the subset of patients whose at least a single eGFR index is known, unlabeled samples (#8363) the remaining.

6.3.1.2. Labeled samples

Let t_g the time-stamp of the last eGFR observation, the previous 1-year time-stamp is defined as:

$$t_t = t_g - 12\ months \tag{6.2}$$

Among all labeled samples only those which satisfy the following criteria were selected:

- At least a single observation of all fields (#5494);
- At least 2-year eGFR medical history before t_t, that must include 2 or more eGFR observations (#2176).

Table 6.1 shows the eGFR distribution at t_g time-stamp of the *selected* samples (i.e., from now on mentioned as labeled samples) in according with the CKD stages [190]. The remaining samples named *discarded samples* (#3636) from now on were merged with unlabeled samples and named as such (#11999).

Additionally, for each field, only features whose appearances are less than 5% of the total of labeled samples were excluded. Regarding the monitoring field, only the blood pressure feature is over cut, then was grouped with lab tests field because of the same continuous nature. From now on, all the included features were named predictors.

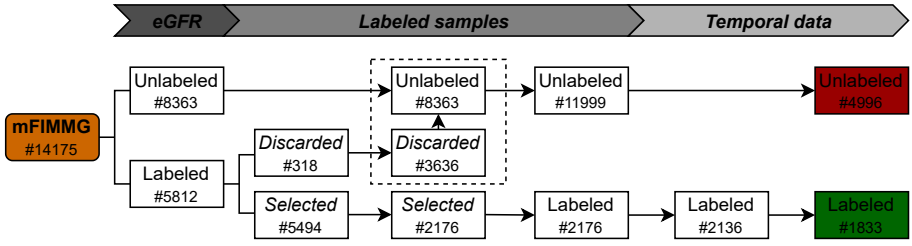


Figure 6.1.: mFIMMG dataset preprocessing: labeled and unlabeled samples.

6.3.1.3. Temporal data

Following [193], 6-month granularity was chosen to define a time-window. For each labeled sample, only the first five consecutive non-overlapping time-windows (i.e., 2.5-year medical history) before t_t were chosen. If few time-windows were chosen the eGFR temporal evolution could not be caught by the predictive model; on the other hand, the model would risk overfitting because the more observations the patient

Table 6.1.: Distribution of eGFR for the labeled samples (#2176) in according with the CKD stages.

CKD stage	**eGFR** [ml/min/1.73m^2]	**%**
I	$\geq$ 90: normal	19.35
II	60–89: mild reduction	53.31
IIIa	45–59: mild-moderate reduction	16.59
IIIb	30–44: moderate-severe	7.49
IV	15–29: sever reduction	2.85
V	$<$ 15: kidney failure	0.41

would have, the more the patient would tend to have chronic kidney complications (i.e., low eGFR values) [207]. Thus, for each field, all the patients that did not contain observations in any of the selected five time-windows were deleted (#2136).

The information about eGFR before t_t was deleted only from the lab tests field (i.e., eGFR continuous values) because already indirectly present through the predictors used in CKD-EPI formula (6.1), while from exams field was left (i.e., times of eGFR examination prescription). Finally, a supplementary field named 'Overall'- which consists of the aggregation of drugs, exams, lab tests predictors only if they were fully shared by the same patient - was provided (#1833 samples and 494 predictors). Additionally, Overall* field included also demographic predictors (i.e., gender and age).

On the contrary, for each field of unlabeled samples, five random consecutive time-windows were chosen if patients shared at least a single observation of the same predictors extracted from the labeled samples, by obtaining the final Overall and Overall* fields (#4996 samples).

Both categorical and continuous features were appropriately standardized during the preprocessing stage. The one-hot encoding was used on categorical features (i.e., pathologies, exams, drugs), while the z-score was used on continuous features (i.e., lab tests) by removing the mean and scaling to unit variance. Thus, categorical fields reflect the presence or the absence of a given pathology, drug, or exam without displaying any missing values. On the other hand, the continuous field (lab tests) may present missing values or outliers. For that reason, an outlier detection strategy based on scaled median absolute deviation and an extra-values imputation of missing values was performed for both labeled and unlabeled samples of the lab tests field. Table 6.2 shows the final configuration of the mFIMMG dataset after the preprocessing stage.

6.4. Methods

The binary classification task consists in predicting the short-term (1-year) eGFR evolution. Given the longitudinal information of each patient, according to Table 6.1 the assumption is to predict CKD stage I (e.g., negative or normal samples, y^-) from the others (e.g., positive or risky samples, y^+).

Table 6.2.: Final configuration of the mFIMMG dataset after the preprocessing stage.

	Pathologies	Drugs	Exams	Lab tests	**Overall**	**Overall***
Predictors	38	309	135	50	494	496
Total samples	5660	9533	9530	7479	6829	6829
Labeled samples	707	1853	1887	1877	1833	1833
Unlabeled samples	4953	7680	7643	5602	4996	4996

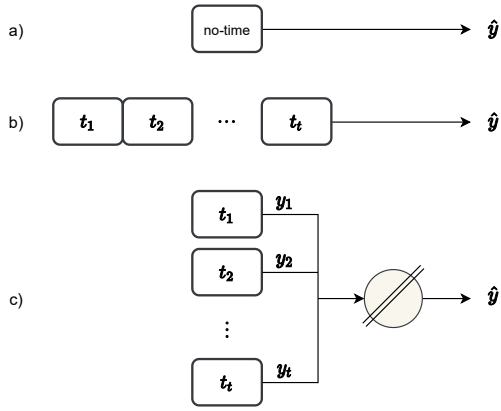


Figure 6.2.: Three different approaches. a) No-temporal: the temporal information is averaged across all time-windows; b) Stacked-temporal: the temporal information is preserved by concatenating longitudinally all the time-windows; and c) Multitask-temporal: each time-window is treated as a separate task.

6.4.1. Notations

The main mathematical notations used in the following Sec. 6.4 were summarized in Table 6.3.

6.4.2. Baseline approaches

No-temporal In this approach, the continuous predictors were averaged across all time windows, while the categorical ones were aggregated. Even if the temporal information has vanished, this approach handles the challenge of irregular sampling and missing values.

Stacked-temporal In this approach temporal information was preserved by concatenating longitudinally all the time windows. This approach can capture temporal information across time windows, but it may suffer from overfitting, considering the increasing number of predictors which is directly proportional to the number of time windows.

Table 6.3.: Notations.

Symbol	Description
n	# of samples
d	# of predictors
t	# of tasks (time-windows)
$X \in \mathbb{R}^{n \times d}$	Observations
- $x \in Z_l$	- labeled
- $x^{'} \in V_u$	- unlabeled
- $\widetilde{x}^{'} \in \widetilde{Z}_u$	- pseudolabeled
- $\widetilde{x} \in \widetilde{Z}$	- labeled and pseudolabeled
$W \in \mathbb{R}^{d \times t}$	Weights
$Y \in \mathbb{R}^{n \times t}$	Targets
- $y \in Z_l$	- labeled
- $y^{'} \in V_u$	- unlabeled
- $\widetilde{y}^{'} \in \widetilde{Z}_u$	- pseudolabeled
- $\widetilde{y} \in \widetilde{Z}$	- labeled and pseudolabeled
$\hat{y} \in \mathbb{R}^{n \times 1}$	Target predictions

6.4.3. Semi-Supervised Multi-Task Learning (SS-MTL)

In the following subsection the SS-MTL approach is introduced by providing: i) multi-task temporal Lasso formulation (see Sec. 6.4.3.1), ii) Self-Learning Algorithm formulation (see Sec. 6.4.3.2), and iii) SS-MTL approach implementation (see Sec. 6.4.3.3).

6.4.3.1. MTL: multi-task temporal Lasso

Multitask-temporal In this approach (see Figure 6.2c) the temporal information was handled as a MTL problem. Each time-window was treated as a separate task and then, the resulting intermediate outputs $(y_1, y_2, ..., y_t)$ were combined to obtain the final prediction $\hat{y}$.

Considering the following MTL problem with t tasks (time-windows), n samples, and d predictors, the model encodes the temporal information using regularization terms. Let $\{x_1, ..., x_n\}$ be the input data and $\{y_1, ..., y_n\}$ be the targets, where each $x_i \in \mathbb{R}^d$ represents a sample, and $y_i \in \mathbb{R}^t$ is the corresponding target at different time-windows. $X = [x_1, ..., x_n]^T \in \mathbb{R}^{n \times d}$ is denoted as the data matrix, $Y = [y_1, ..., y_n]^T \in \mathbb{R}^{n \times t}$ as the target matrix, and $W = [w_1, ..., w_t] \in \mathbb{R}^{d \times t}$ as the weight matrix. The whole formulation of multitask temporal Lasso is given by [192, 208]:

$$\min_{W} \quad L(W) + \rho_1 \|W\|_F^2 + \rho_2 \sum_{i=1}^{t-1} \|W_i - W_{1+1}\|_F^2 + \rho_3 \|W\|_{2,1} \tag{6.3}$$

where $L(W)$ is the loss function and ρ_1, ρ_2, ρ_3, represent the regularization penalties: the first penalty controls the complexity of the model; the second penalty couples the neighbor tasks, encouraging every two neighbour tasks to be similar (temporal

smoothness); and the third penalty induces the grouped sparsity, which performs the joint feature selection on the tasks at different time-windows (longitudinal feature selection). The temporal information is modeled as a type of graph regularization A, where neighbour tasks are coupled via edges.

A is the structure matrix which encodes the task relatedness. In the temporal group Lasso formulation, A is defined as an $(t-1)\times t$ sparse matrix, in which $A_{i,i}=1$ and $A_{i,i+1}=-1$; and thus, the formulation can be written in a simpler form:

$$\min_{W} \quad L(W)+\rho_1\|W\|_F^2+\rho_2\|WA\|_F^2+\rho_3\|W\|_{2,1} \tag{6.4}$$

However, this formulation assumes that for each sample a predictor is simultaneously selected or not at all time-windows. The convex fused sparse group Lasso (CFG) formulation overcomes this issue [208]:

$$\min_{W} \quad L(W)+\rho_1\|W\|_1+\rho_2\|AW^T\|_1+\rho_3\|W\|_{2,1} \tag{6.5}$$

Accordingly, the CFG with Logistic loss model solves the CFG regularized multi-task Logistic regression problem:

$$\begin{aligned}\min_{W,c} \quad & \sum_{i=1}^{t}\sum_{j=1}^{n_i}\log\Big\{1+exp\Big[-Y_{i,j}\Big(W_j^T X_{i,j}+c_i\Big)\Big]\Big\}+ \\ & +\rho_1\|W\|_1+\rho_2\|AW^T\|_1+\rho_3\|W\|_{2,1}\end{aligned} \tag{6.6}$$

where ρ_3 controls group sparsity for joint feature selection, while ρ_1, which controls element-wise sparsity and ρ_2 which controls the fused regularization represent the parameters for the fused Lasso.

6.4.3.2. SSL: Self-Learning Algorithm (SLA)

Semi-supervised learning, also referred as learning with partially labeled data, concerns the case where a prediction function is learned on both labeled and unlabeled training samples. Unlabeled training samples may contain valuable information on the prediction problem at hand which exploitation may lead to a performant prediction function. For a binary classification scenario, a set of labeled training samples $Z_l=\{(x_i,y_i)\mid i=1,...,l\}$ and a set of unlabeled training samples $V_u=\{x_i^{'}\mid i=l+1,...,l+u\}$ are defined.

Considering learning algorithms that work in a fixed hypothesis space H of binary classifiers and given the whole training set $S=Z_l\cup V_u$, the task of the learner $h\in H$ is to choose a posterior distribution Q over H such that the Q-weighted majority vote classifier B_Q (i.e., Bayes classifier) will have the smallest possible risk on samples of V_u. Defining the Bayes classifier:

$$B_Q(x)=sign[E_{h\sim Q}h(x)] \tag{6.7}$$

Its empirical error over the unlabeled set V_u, called the transductive risk, can be defined as:

$$R_u(B_Q) = \frac{1}{u} \sum_{x' \in V_u} [B_Q(x') \neq y'] \tag{6.8}$$

The corresponding Gibbs classifier, G_Q, is randomly chosen from the hypothesis space H according to the posteriror distribution Q and its transductive risk over the unlabeled training set is defined by:

$$R_u(G_Q) = \frac{1}{u} \sum_{x' \in V_u} E_{h \sim Q}[h(x') \neq y'] \tag{6.9}$$

Note that these risks cannot be estimated as the labels of unlabeled examples are unknown.

In [209, Ch. 3] the margin of a Bayes classifier was shown to be an indicator of confidence respecting the cluster assumption in semi-supervised learning which stipulates that the decision boundary passes through low density regions. Supposing to have a tight upper bound $R_u^{\delta}(G_Q)$ over the risk of the Gibbs classifer G_Q which holds with probability $1-\delta$, [204] showed that it is possible to bound the transductive risk of the Bayes classifier with high probability.

This result follows from a bound on the joint Bayes risk depending on a threshold θ :

$$R_{u \wedge \theta}(B_Q) = \frac{1}{u} \sum_{x' \in V_u} [B_Q(x') \neq y' \wedge m_Q(x') > \theta] \tag{6.10}$$

where $m_Q(\cdot) = |E_{h \sim Q} h(\cdot)|$ is the absolute value output of the Bayes classifier, denoted as the unsigned margin function.

This bound over the joint Bayes risk can be estimated by considering the distribution of unsigned margins regarding the threshold θ and it constitutes the working hypothesis of the margin-based Self-Learning Algorithms (SLA). This algorithm first trains a classifier on the labeled training set. The output of the learner can then be used to assign pseudolabels to unlabeled examples (denoted by the set $\widetilde{Z}_u$ in what follows) having a margin above a certain threshold θ and the supervised method is repeatedly retrained upon the set of the initial labeled and unlabeled examples that have been classified in the previous steps. The threshold θ is iteratively estimated at each step of the algorithm as the one which minimizes the conditional Bayes error defined as:

$$R_{u|\theta}(B_Q) = P_u(B_Q(x') \neq y' \mid m_Q(x' > \theta) = \frac{R_{u \wedge \theta}(B_Q)}{P_u(m_Q(x' > \theta)} \tag{6.11}$$

In practice, the upper bound $R_Q^{\delta}(G)$ of the risk of the Gibbs classifier which is involved in the computation of θ in equation (6.8) is fixed to its worst value 0.5.

Algorithm SLA

Input: Labeled and Unlabeled training sets: Z_l, V_u
Initialize
Train a classifier H on Z_l
Set $\widetilde{Z}_u \leftarrow \varnothing$
repeat
Compute the margin threshold θ from (6.8)
$S \leftarrow \{ (x', y') \mid x' \in V_u; m_Q(x' \geq \theta \wedge y' = sign(H(x')) \}$
$\widetilde{Z}_u \leftarrow \widetilde{Z}_u \cup S, V_u = V_u \backslash S$
Learn a classifier H by optimizing a global loss function on Z_l and $\widetilde{Z}_u$
until V_u is empty or no adds to $\widetilde{Z}_u$;
Output: The final $\widetilde{Z} = Z_l \cup \widetilde{Z}_u$

6.4.3.3. Implementation of SS-MTL

The training experimental procedure adopted by our proposed method is shown in Figure 6.3.

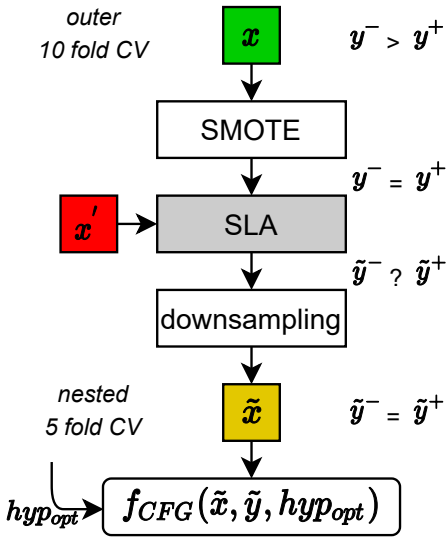


Figure 6.3.: SS-MTL: training experimental procedure.

At the beginning of the outer 10-fold cross-validation (10-CV) procedure, negative labeled samples y^- were around four times more numerous than those positive y^+, thus SMOTE [55] was utilized to balance the labeled samples ($y^- = y^+$). In the experiments, two different Bayes classifiers, which are Decision Trees (DT) and SVM, were considered in the SLA algorithm for Overall and Overall* fields, respectively. On the contrary for the other single fields only DT model was used. This rationale is justified by the fact that after having tested all the possible combinations of classifiers (i.e., LR, DT, RF, Boosting, SVM) within the SLA algorithm, in terms of predictive performance, the SVM resulted the best classifier for Overall* field, while the DT

classifier for all the others.

During each SLA iteration, every candidate pseudolabel is chosen only if selected (i.e., above threshold θ) for all time-windows. After that, the final prediction associated to the pseudolabel was selected by testing 3 different strategies (i.e., majority voting (majvot), unanimous, and Gibbs).

From the final SLA output $\in \widetilde{Z}$, the imbalance ratio between $\widetilde{y}^-$ and $\widetilde{y}^+$ is unknown ($\widetilde{y}^-$? $\widetilde{y}^+$), thus random downsampling over the pseudolabel majority class was performed in order to achieve again a balanced condition. The hyperparameters tuning was performed by implementing a grid-search and maximizing the *macro-Recall* within a nested 5-fold cross validation (5-CV) procedure. The rationale behind the optimization of the *Macro-recall* in the validation set is justified by the fact of achieving an objective that is more clinical relevant for a screening purpose. Thus, the authors, following this rationale, preferred to minimize the false negatives and achieve a trade-off between sensitivity and specificity. This choice has been also performed according to the most recent state-of-the-art approaches in predictive medicine scenario [2, 4]. The optimal hyperparameters (*hyp opt*), $\widetilde{x}$, and $\widetilde{y}$ were fed to the MTL model (i.e., CFG) for the training stage. The final prediction of the SS-MTL was computed by averaging the margin outputs of each single t task and then taking the decision based on the *sign* function:

$$\hat{y}_i = sign\Big(\frac{\sum_{i=1}^{t} \widetilde{x}^T w_i + c_i}{t}\Big) \tag{6.12}$$

The code to replicate the SS-MTL approach is publicly released by the authors.

6.4.4. Experimental comparisons

Our proposed SS-MTL approach was compared with baseline approaches (i.e., no-temporal, stacked-temporal) and with the MTL approach. Moreover, to better contextualize the proposed SS-MTL in the Semi-Supervised Learning (SSL) literature, the Self-Learning Algorithm (SLA) procedure was also compared with other existing SSL techniques, such as Positive and Unlabeled learning and Label Propagation. These approaches adopted as ML models those employed in the state-of-the-art closer to our setting (see Sec. 6.2), such as LR [122, 195, 196] with Lasso regularizer; DT [197]; RF [122, 196]; Boosting [122]; and SVM [197, 2] with Lasso regularizer. Experimental results were provided both for single (i.e., pathologies, drugs, exams, lab tests) and Overall/Overall* fields, by utilizing or not (i.e., noSLA) the SLA procedure. The same ML model adopted externally for the 10-CV was utilised also within the SLA procedure.

The predictive performance was evaluated according to the following standard metrics for classification task defined in Section 2.4.4: *Accuracy*, *Precision*, *Recall*, *F1* and *AUC*.

Table 6.4.: Range of hyperparameters (hyp) for each model: Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), Support Vector Machine (SVM) with Lasso regularizer, and Convex Fused Group Lasso (CFG) with Logistic regression model.

Model	**Hyp**	**Range**
LR [122, 195, 196]	Lambda	$\{10^{-5}, 10^{-4}10^{-3}, 10^{-2}, 0.1, 1\}$
DT [197]	max # of splits	$\{100, 200, 300, 400, 500\}$
RF [122, 196]	# of DT # of predictors to select	$\{25, 50, 75, 100, 125, 150\}$ $\{\frac{all}{4}, \frac{all}{3}, \frac{all}{2}, all\}$
Boosting [122]	max # of splits learning rate	$\{50, 100, 150, 200\}$ $\{10^{-2}, 0.1, 1\}$
SVM [197, 2]	Lambda	$\{10^{-5}, 10^{-4}10^{-3}, 10^{-2}, 0.1, 1\}$
CFG [192, 208]	ρ_1 ρ_2 ρ_3	$\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1\}$ $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1\}$ $\{10^{-3}, 10^{-2}, 0.1\}$

Table 6.4 summarizes the range of the hyperparameters optimized in all the experiments.

6.5. Experimental results

The experimental results of the SS-MTL approach are shown as predictive performance comparison with baseline approaches (i.e., no-temporal [Sec. 6.5.2], stacked-temporal [Sec. 6.5.3]) and with the MTL approach (Sec. 6.5.4). For the baseline approach (i.e., no-temporal), the SLA procedure (i.e., the SSL technique from which our proposed approach is originated) is firstly compared with other SSL techniques, such as PU and LP (Sec. 6.5.1).

In particular, Section 6.5.4 shows the trend of the predictive performance in relation to different portions of labeled training samples. This rational is due to the intention to measure the reliability of SS-MTL for dealing with a higher portion of unlabeled samples as excepted in a real-case scenario. Finally, the experimental results of the SS-MTL approach are shown in terms of pattern localization (Sec. 6.5.5) to measure the importance of the predictors.

6.5.1. State-of-the-art comparison: Semi-Supervised Learning (SSL)

Table 6.5 shows the comparison of the experimental of the SLA procedure with other SSL techniques (i.e., PU, LP). The comparison was performed only for the Overall* field of the baseline (i.e., no-temporal) approach. The predictive performance of all

ML models that used the SLA procedure is clearly superior to the other SSL techniques (i.e., PU, LP), thus the SLA procedure was selected as the SSL paradigm for the proposed SS-MTL approach.

Table 6.5.: Experimental results comparison of the Self-Learning Algorithm (SLA) procedure with other Semi-Supervised Learning (SSL) techniques (i.e., Positive and Unlabeled learning [PU], Label Propagation [LP]). The comparison was performed only for the Overall* field of the baseline (i.e., no-temporal) approach. The best result in terms of *Recall* was highlighted in bold.

SLA	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.744	0.629	0.620	**0.660**	0.741
DT	0.792	0.677	0.670	**0.697**	0.693
RF	0.838	0.730	0.731	**0.734**	0.827
Boosting	0.849	0.687	0.760	**0.660**	0.847
SVM	0.716	0.627	0.623	**0.685**	0.749
LP	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.651	0.575	0.582	0.632	0.724
DT	0.698	0.583	0.592	0.618	0.616
RF	0.788	0.687	0.644	0.692	0.811
Boosting	0.813	0.646	0.707	0.640	0.829
SVM	0.598	0.559	0.576	0.655	0.710
PU	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.759	0.610	0.611	0.607	0.680
DT	0.795	0.553	0.594	0.538	0.532
RF	0.816	0.598	0.638	0.646	0.721
Boosting	0.811	0.516	0.662	0.508	0.692
SVM	0.705	0.601	0.609	0.640	0.729

6.5.2. State-of-the-art comparison: No-temporal

Table 6.6 shows the comparison results for the no-temporal approach. Considering the SS-MTL an evolution of standard LR model, the comparison of the SS-MTL approach with the LR model would represent the most fair and straight comparison. The SS-MTL approach performance ($Recall = 0.737 \pm 0.054$) for Overall* field was greater than no-temporal (LR: $Recall = 0.660 \pm 0.048$) and stacked-temporal (LR: $Recall = 0.657 \pm 0.042$) in SLA configuration. Again for Overall field, the SS-MTL approach performance ($Recall = 0.668 \pm 0.053$) was greater than no-temporal (LR: $Recall = 0.616 \pm 0.062$) and stacked-temporal (LR: $Recall = 0.588 \pm 0.034$) in SLA configuration.

Nevertheless, if a global overview is considered, the best performance ($Recall = 0.734 \pm 0.051$) for no-temporal approach was obtained by the RF model for Overall* field in SLA configuration, but still lower than the best ones obtained by MTL ($Recall = 0.742 \pm 0.060$) approach and SS-MTL ($Recall = 0.737 \pm 0.054$) approach.

Table 6.6.: **No-temporal:** Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), and Support Vector Machine (SVM) with Lasso regularizer. In SLA procedure the same classifier adopted externally in 10-CV was used. Overall* indicates that also gender and age were included as predictors. Best result in terms of *Recall* was highlighted in bold for each field.

	noSLA					SLA				
Pathologies	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.519	0.456	0.529	0.556	0.557	0.528	0.452	0.516	0.530	0.568
DT	0.614	0.499	0.531	0.553	0.576	0.652	0.511	0.531	0.556	0.549
RF	0.628	0.506	0.530	0.554	0.608	0.642	0.517	0.537	0.562	0.579
Boost	0.652	0.523	0.544	**0.572**	0.585	0.651	0.518	0.539	0.563	0.580
SVM	0.488	0.437	0.524	0.546	0.563	0.501	0.445	0.532	0.559	0.568
Drugs	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.638	0.557	0.573	0.618	0.643	0.631	0.552	0.570	0.612	0.650
DT	0.694	0.540	0.540	0.549	0.559	0.628	0.543	0.561	0.598	0.602
RF	0.759	0.561	0.568	0.559	0.618	0.724	0.542	0.542	0.543	0.538
Boost	0.781	0.557	0.580	0.554	0.608	0.767	0.538	0.552	0.537	0.596
SVM	0.594	0.536	0.574	0.625	0.645	0.597	0.541	0.579	**0.633**	0.659
Exams	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.607	0.534	0.559	0.594	0.647	0.610	0.537	0.562	0.600	0.643
DT	0.707	0.552	0.551	0.559	0.557	0.670	0.550	0.553	0.574	0.604
RF	0.778	0.548	0.576	0.546	0.663	0.774	0.543	0.569	0.541	0.602
Boost	0.798	0.526	0.600	0.531	0.662	0.797	0.528	0.593	0.533	0.639
SVM	0.593	0.536	0.571	0.617	0.667	0.606	0.547	0.579	**0.629**	0.670
Lab tests	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.645	0.559	0.572	0.611	0.661	0.656	0.559	0.566	0.599	0.656
DT	0.743	0.574	0.574	0.576	0.576	0.710	0.574	0.572	0.588	0.588
RF	0.806	0.630	0.661	**0.619**	0.761	0.789	0.605	0.622	0.598	0.731
Boost	0.815	0.498	0.565	0.522	0.759	0.815	0.514	0.649	0.529	0.743
SVM	0.633	0.550	0.565	0.602	0.657	0.668	0.567	0.571	0.603	0.653
Overall	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.703	0.582	0.579	0.607	0.676	0.706	0.587	0.584	0.616	0.683
DT	0.741	0.576	0.574	0.581	0.569	0.711	0.597	0.593	0.629	0.598
RF	0.803	0.640	0.654	0.632	0.762	0.777	0.615	0.620	0.612	0.695
Boost	0.821	0.532	0.673	0.541	0.770	0.816	0.542	0.635	0.546	0.783
SVM	0.651	0.583	0.601	**0.665**	0.709	0.677	0.592	0.599	0.654	0.706
Overall*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.739	0.622	0.615	0.650	0.727	0.744	0.629	0.620	0.660	0.741
DT	0.796	0.658	0.660	0.662	0.666	0.792	0.677	0.670	0.697	0.693
RF	0.830	0.717	0.716	0.722	0.854	0.838	0.730	0.731	**0.734**	0.827
Boost	0.847	0.678	0.761	0.650	0.875	0.849	0.687	0.760	0.660	0.847
SVM	0.693	0.613	0.617	0.683	0.747	0.716	0.627	0.623	0.685	0.749

Instead for Overall field, the best performance ($Recall = 0.665 \pm 0.062$) was obtained by the SVM in noSLA configuration. This result is comparable with the one extracted for the SS-MTL approach ($Recall = 0.668 \pm 0.053$).

6.5.3. State-of-the-art comparison: Stacked-temporal

Table 6.7.: **Stacked-temporal:** Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), and Support Vector Machine (SVM) with Lasso regularizer. In SLA procedure the same classifier adopted externally in 10-CV was used. Overall* indicates that also gender and age were included as predictors. Best result in terms of *Recall* was highlighted in bold for each field.

	noSLA					SLA				
Pathologies	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Logistic	0.658	0.506	0.521	0.537	0.561	0.682	0.534	0.543	0.566	0.562
DT	0.607	0.500	0.533	0.563	0.578	0.703	0.539	0.546	**0.568**	0.563
RF	0.550	0.468	0.524	0.549	0.571	0.545	0.467	0.529	0.558	0.554
Boost	0.680	0.526	0.536	0.557	0.561	0.685	0.512	0.522	0.533	0.541
SVM	0.646	0.511	0.529	0.550	0.570	0.726	0.539	0.544	0.555	0.557
Drugs	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Logistic	0.679	0.533	0.534	0.544	0.623	0.653	0.553	0.561	0.591	0.641
DT	0.693	0.540	0.540	0.550	0.548	0.613	0.532	0.554	0.588	0.567
RF	0.750	0.549	0.554	0.547	0.610	0.743	0.561	0.562	0.562	0.529
Boost	0.758	0.555	0.565	0.554	0.620	0.701	0.538	0.537	0.545	0.617
SVM	0.587	0.535	0.579	**0.633**	0.665	0.582	0.530	0.574	0.624	0.666
Exams	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Logistic	0.665	0.540	0.543	0.560	0.603	0.648	0.539	0.546	0.567	0.616
DT	0.686	0.528	0.528	0.535	0.530	0.648	0.530	0.537	0.554	0.544
RF	0.764	0.551	0.565	0.548	0.611	0.748	0.531	0.540	0.531	0.561
Boost	0.791	0.480	0.515	0.504	0.629	0.787	0.487	0.512	0.507	0.611
SVM	0.624	0.547	0.566	**0.605**	0.648	0.615	0.540	0.562	0.599	0.646
Lab tests	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Logistic	0.651	0.552	0.560	0.590	0.645	0.657	0.555	0.562	0.591	0.642
DT	0.723	0.554	0.553	0.556	0.554	0.675	0.541	0.542	0.557	0.537
RF	0.785	0.573	0.602	0.566	0.711	0.777	0.578	0.593	0.573	0.670
Boost	0.818	0.496	0.699	0.521	0.731	0.811	0.490	0.585	0.516	0.713
SVM	0.611	0.545	0.572	0.617	0.663	0.638	0.563	0.580	**0.628**	0.671
Overall	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Logistic	0.727	0.571	0.567	0.579	0.660	0.721	0.576	0.572	0.588	0.669
DT	0.727	0.570	0.567	0.577	0.567	0.690	0.573	0.573	0.601	0.586
RF	0.795	0.625	0.638	0.617	0.745	0.775	0.602	0.611	0.598	0.660
Boost	0.817	0.509	0.646	0.526	0.751	0.816	0.503	0.633	0.522	0.738
SVM	0.661	0.583	0.595	0.652	0.713	0.668	0.590	0.600	**0.659**	0.713
Overall*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Logistic	0.755	0.611	0.606	0.621	0.706	0.762	0.638	0.630	0.657	0.742
DT	0.772	0.638	0.633	0.648	0.622	0.769	0.648	0.639	0.668	0.619
RF	0.816	0.696	0.691	0.704	0.834	0.805	0.678	0.675	0.684	0.777
Boost	0.840	0.632	0.745	0.610	0.857	0.835	0.622	0.733	0.602	0.835
SVM	0.745	0.651	0.641	0.700	0.782	0.751	0.659	0.647	**0.709**	0.787

Table 6.7 shows the comparison results for the stacked-temporal approach. Focusing on the comparison of the LR model, the SS-MTL approach performance ($Recall = 0.737 \pm 0.054$) for Overall* field was greater than stacked-temporal (LR: $Recall =$

0.657±0.042) in SLA configuration. Again for Overall field, the SS-MTL approach performance ($Recall = 0.668 \pm 0.053$) was greater than stacked-temporal (LR: $Recall = 0.588 \pm 0.034$) in SLA configuration.

Nevertheless, if a global overview is considered, the best performance ($Recall = 0.709 \pm 0.057$) was obtained by the SVM model for Overall* field in SLA configuration. Accordingly for Overall field, the best performance ($Recall = 0.659 \pm 0.047$) was still obtained by the SVM model in SLA configuration. These results were lower than those extracted by the SS-MTL approach for Overall* ($Recall = 0.737 \pm 0.054$) and Overall field ($Recall = 0.668 \pm 0.053$).

6.5.4. Multitask-temporal comparison

Table 6.8.: **Multitask-temporal:** Decision Tree (DT) classifier was used to select pseudolabels in SLA procedure, except for Overall* where Support Vector Machine (SVM) with Lasso regularizer was used. Overall* indicates that also gender and age were included as predictors. Fraction (f) represents the amount of labeled samples used in the training stage. The table depicts the SS-MTL majvot configuration. Best result in terms of *Recall* was highlighted in bold for each field.

f = 100%	**MTL**					**SS-MTL**				
	Accuracy	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Pathologies	0.766	0.510	0.542	0.510	0.517	0.782	0.523	0.541	**0.522**	0.511
Drugs	0.568	0.526	0.584	**0.640**	0.680	0.581	0.533	0.582	0.638	0.690
Exams	0.580	0.532	0.579	**0.631**	0.677	0.573	0.522	0.568	0.611	0.673
Lab tests	0.622	0.561	0.587	**0.643**	0.687	0.621	0.560	0.587	0.642	0.687
Overall	0.625	0.568	0.598	0.664	0.720	0.636	0.575	0.601	**0.668**	0.713
Overall*	0.765	0.681	0.668	**0.742**	0.816	0.750	0.670	0.661	0.737	0.820
f = 30%	**MTL**					**SS-MTL**				
	Accuracy	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Pathologies	0.730	0.512	0.516	0.515	0.542	0.600	0.457	0.542	**0.549**	0.564
Drugs	0.584	0.531	0.575	0.626	0.671	0.640	0.566	0.584	**0.635**	0.680
Exams	0.625	0.550	0.569	**0.610**	0.651	0.626	0.539	0.556	0.587	0.630
Lab tests	0.662	0.570	0.576	0.616	0.656	0.660	0.575	0.584	**0.629**	0.658
Overall	0.682	0.590	0.593	0.642	0.686	0.707	0.612	0.610	**0.662**	0.700
Overall*	0.758	0.655	0.644	0.692	0.784	0.746	0.665	0.657	**0.731**	0.811

Figure 6.4 compares the performance trend (i.e., *Recall*) over the fraction of labeled training samples $x, y \in Z_l$ for MTL and SS-MTL approaches considering both Overall and Overall* fields. Starting from a total of 4996 unlabeled samples $x' \in V_u$, figure 6.5 shows the trend of the pseudolabels samples $\widetilde{x}', \widetilde{y}' \in \widetilde{Z}_u$ selected by the SS-MTL approach (after random downsampling) over the fraction of labeled training samples. Table 6.8 shows more in detail the predictive performance for MTL and SS-MTL approaches. In particular, two configurations were highlighted, where both the full amount (f=100%) and a specific portion (f=30%) of labeled samples was utilised in the training stage. Comparable performance were obtained by the MTL ($Recall =$

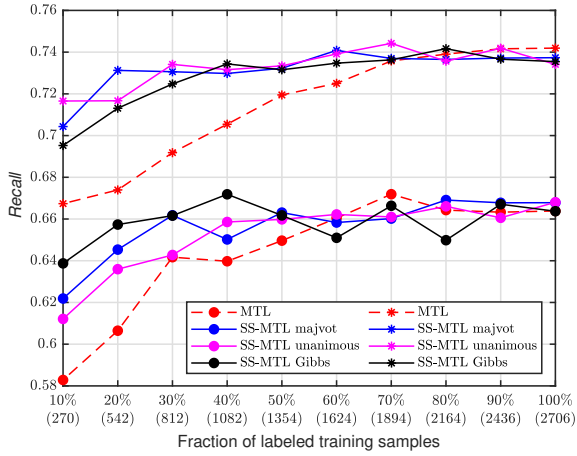


Figure 6.4.: MTL and SS-MTL approaches: Recall trend over fraction of labeled training samples $x, y \in Z_l$. In the legend, stars indicate that gender and age were included as predictors (Overall*), filled circles were not (Overall).

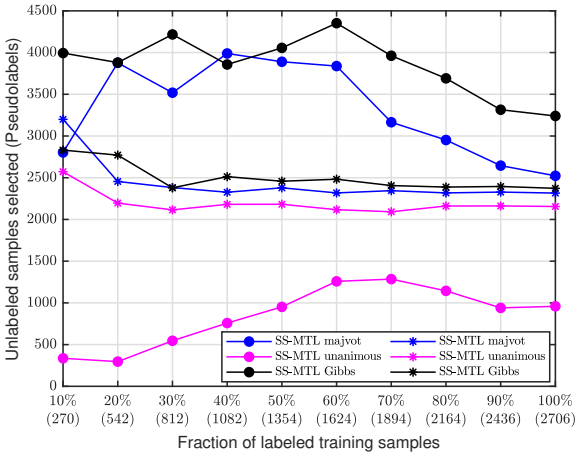


Figure 6.5.: Pseudolabel samples $\widetilde{x}', \widetilde{y}' \in \widetilde{Z}_u$ selected from SLA procedure (after random downsampling) over fraction of labeled training samples $x \in Z_l$. In the legend, stars indicate that gender and age were included as predictors (Overall*), filled circles were not (Overall).

0.742±0.060) approach and SS-MTL ($Recall = 0.737 \pm 0.054$) for Overall* field with f=100%. On the contrary, if f=30% the best performance ($Recall = 0.731 \pm 0.049$) was obtained by the SS-MTL approach with an important gain of 4.1% with respect to MTL ($Recall = 0.692 \pm 0.035$). The rationale to emphasize this result was due to the fact that the performance of SS-MTL remained stable from f=100% to f=30% while

the performance of MTL decreased (see Figure 6.4).

6.5.5. Pattern localisation

Table 6.9 explains which predictors in SS-MTL majvot (f=30%) configuration were more decisive to predict the next 1-year eGFR state. The final percentage weight of each predictor showed in Table 6.9 was calculated by averaging the weights of the model over 10 folds, and then, over the t tasks.

Table 6.9.: Top-10 predictors for SS-MTL majvot approach with f =30 %. Overall* indicates that also gender and age were included as predictors. D=Drugs; E=Exam; M=Monitoring.

	Overall			**Overall***		
Rank	Field	Predictors	W [%]	Field	Predictors	W [%]
1)	D	Valsartan and diuretics	3.78	M	Age	44.85
2)	D	Colecalciferol (vitamin D3)	3.59	D	Furosemide	3.26
3)	D	Levothyroxine	3.17	D	Metformin	2.42
4)	D	Alfuzosin	3.16	D	Amlodipine	1.40
5)	D	Lansoprazole	3.08	D	Ramipril and amlodipine	1.38
6)	D	Furosemide	2.89	D	Valsartan and diuretics	1.32
7)	D	Acetylsalicylic acid	2.78	D	Pravastatin	1.28
8)	D	Pantoprazole	2.67	D	Atorvastatin	1.27
9)	E	Interview and evaluation	2.51	D	Bisoprolol	1.19
10)	D	Nebivolol	2.28	D	Omeprazole	1.16
		Others	70.09		*Others*	40.47

6.6. Discussion

This work has mainly contributed to the biomedical informatics field for the following points:

- Introduction of the SS-MTL paradigm for predicting short-term KD evolution. The proposed high-interpretable approach seeks to learn from labeled and unlabeled samples while imposing a temporal relatedness between consecutive tasks (i.e., time windows).

- Measurement and demonstration of the effectiveness of the SS-MTL approach with respect to the state-of-the-art in real-use case scenario (i.e., GP EHR dataset). The benefits in terms of predictive performance are particularly pronounced the more numerous the unlabeled samples are than those labeled. This condition reflects the real clinical use case where the observations of each patient lack annotation or are only partially labeled.

The impact of the predictive performance and pattern localization experimental results will be thoroughly discussed in Section 6.6.1 and Section 6.6.2. Then, limitations and future work will be argued in Section 6.6.3 and Section 6.6.4.

6.6.1. Predictive performance

In the following section the two RQs formulated in Sec.6.1 will be discussed.

6.6.1.1. RQ1: Is the MTL approach capable to capture the eGFR temporal evolution?

The MTL approach as showed in Table 6.8 was capable to capture the eGFR temporal evolution, because for Overall* configuration in terms of $Recall = 0.742 \pm 0.060$ was superior than the best competitors for no-temporal (Table 6.6) and stacked-temporal (Table 6.7) approaches (RF: $Recall = 0.722 \pm 0.036$; RF: $Recall = 0.704 \pm 0.067$, respectively). Instead, if age and gender were not considered (i.e., Overall), the MTL performance ($Recall$ = 0.664 ± 0.048) was close to SVM ($Recall$ = 0.665 ± 0.062) for no-temporal approach but superior than SVM ($Recall$ = 0.652 ± 0.053) for stacked-temporal approach. However, the performance of the MTL approach remained greater than the baseline LR model for both Overall* and Overall configurations. These outcomes highlighted the importance to include the temporal evolution of the predictors in the ML model. Moreover, experimental results suggested how demographic information was highly discriminative in terms of predictive performance.

The single fields of the MTL approach that mostly affected the predictive performance were drugs and lab tests, while exams seemed to impact less. For instance, the single lab tests field in MTL reached a $Recall$ until 0.643 ± 0.039, much more superior than the other competitors. On the contrary, the pathologies field obtained very poor results and for this reason, was excluded from the Overall and Overall* fields. Results evidenced how the predictive performance of MTL and no-temporal approaches were globally superior to one of the stacked-temporal approaches, which encapsulated the temporal information by aggregating longitudinally the time windows, and this aspect may suffer much the high temporal data sparsity. However, the MTL approach was capable of modeling and interpreting through the regularization strategy the progression of the temporal information, otherwise lost in the no-temporal approach.

6.6.1.2. RQ2: Is the SS-MTL approach capable to capture useful information from unlabeled patients?

The SS-MTL approach was mostly capable to gain useful information from unlabeled patients, in terms of predictive performance concerning to MTL, when labeled patients were less numerous than those unlabeled. This situation commonly reflects the real-case general practice scenario, where available labeled samples size is limited, while

unlabeled samples are much more abundant.

Specifically (see Figure 6.4), the SS-MTL approach did not add an important gain compared to MTL in predictive performance both for Overall and Overall* fields when the full fraction (f=100%) of labeled training sample size was considered. But, if f was progressively decreased (i.e., both for MTL and SS-MTL), the predictive performance kept on being still similar until f=70% for Overall* and until f=60% for Overall. After these cut points, the more f decreases, the more the spread between SS-MTL and MTL increased due to an MTL predictive performance worsening. This finding suggested that our proposed SS-MTL approach was convenient since at least unlabeled samples (# 4996) were almost 2.5 times more numerous than labeled samples (# 1894 at f=70%). Additionally, the SS-MTL predictive performance until f=30% remained almost constant if compared to f=100%, while the MTL performance decreased much earlier as seen before. This further finding proved how the SS-MTL approach was reliable in dealing with unlabeled information.

Basically, for the Overall* field, the *Recall* trend across SLA majvot, unanimous, and Gibbs seemed to be more stable. On the contrary, for the Overall field, the *Recall* trend was more fluctuating and it appeared that SLA unanimous was less performing than the others. These considerations may be fully explained in Figure 6.5, from which it has emerged that the number of pseudo-labels selected by SLA directly interfered with the SS-MTL working stability. Indeed, the most stable performance of the SS-MTL for the Overall* field was influenced by almost constant pseudo-labels selected by SLA. However, even if for the Overall field at f=30% more pseudo-labels were selected by SS-MTL Gibbs and SS-MTL majvot than Overall*, the gap between the predictive performance remained fairly constant across different f thresholds (see figure 6.4). These findings suggested how an increase of almost $2K$ pseudo-labels between Overall and Overall* fields was not related to an increase in predictive performance. Indeed, the pseudo-labels may not be necessarily informative enough to improve the generalization performance of the ML model.

We demonstrated that all the models used for SSL techniques obtained the best predictive performance with the SLA procedure. A central hypothesis in SSL, based on which discriminant models are developed, is the *low density separation* assumption (**H**) [210] which stipulates that the decision boundary should pass through low-density regions. In this sense, contrary to PU [202], the negative class has a central role in finding the decision boundary. In this sense, SLA follows assumption **H** which is also shown to be effective in our experiments. Instead, graphical models, as LP [203], are based on manifold assumption and construct a graph where the nodes represent training examples and the edges reflect similarities between them. The class label of each labeled node is then propagated to its neighbors using label spreading techniques. The similarity between the two observations is based on their Euclidean distance in the feature space, and due to the curse of dimensionality when the dimension of the space is high - as in our case - the Euclidean distance does not reflect well the proximity

between examples.

6.6.2. Clinical significance

The proposed SS-MTL approach was high-interpretable and this aspect assumes an important relevance in the general practice scenario. In fact, obtaining only satisfactory predictive results might be useless if then the results cannot be interpreted by GPs, which need to understand and explain which factors have mostly determined a prediction. From experimental results, it has turned out how gender and age may play a key role compared to other predictors for forecasting the next 1-year eGFR state. In fact, the predictive performance of the SS-MTL approach for the Overall* field (f=30%) was much greater than the one for the Overall field (see Table 6.8). This finding was fully clarified in Table 6.9, from which it emerged that age was the leading predictor with importance of 44.85%, while gender did not appear as a discriminant factor. Although age has already been adequately demonstrated to be one of the major factors in kidney functionality, a prediction merely based on age provided inferior predictive performance, as proven also in [196]. The remaining predictors belonged to the drugs field and this aspect suggested how highly discriminative the past patient's pharmacological treatment might be. In particular, the best contenders such as furosemide and metformin are strictly correlated to variations of eGFR value. Furosemide treatment reduces kidney functionalities for patients with cardiovascular pathologies [211, 212], while metformin administration in patients suffering from moderate CKD is associated with clinical outcome improvements [213]. The creatinin, even if it has been used in Eq. 2.1 for the calculation of the CKD-EPI formula, did not appear as one of the best top-10 predictors. Since the demographic predictors (i.e., gender and age) are included in the eGFR formula (see Eq. 2.1) the performance of the predictive model improved in the Overall* experiment. On the other hand, if demographic information (i.e., gender and age) were not considered (i.e., Overall and single modality experiments) to discover further discriminative predictors besides demographic information, there was no predominant predictor over others, but the pharmacological pathway remained still decisive with respect to the other fields.

6.6.3. Limitations

In this work, the Overall/Overall* fields did not account for the pathologies' information, which caused a predictive performance worsening. In fact, the pathologies field contains much more static information than the others, and it may have found difficult to offer discriminative temporal information to the predictive model. Perhaps, the exclusion of pathologies among the predictors may limit the global contextualization of the clinical problem. To better combine and make coexist heterogeneous feature sets consisting of various EHR fields (e.g., pathologies, exams, drugs, lab tests) of different data types (e.g., categorical, continuous), multi-view learning approaches [214]

may be explored as an intriguing future direction.

We used linear models, which assume linear relationships between the variables, and the outcome and we did not take into account the non-linear combination of different predictors that could potentially affect the outcome. In this context, we may explore non-linear models with different features map in order to discover new hidden high-discriminative temporal patterns.

6.6.4. Future work

Future work may be addressed to explore interesting directions by including different experimental procedures, task definitions, and data processing.

It would be interesting to apply the SS-MTL approach considering only patients enclosed within a specific range of CKD stages and/or, unlike our strategy, predict CKD stages I and II from the others. Alternatively, binary classification could be applied to the prediction of the variation in time of the eGFR value above a certain experimental threshold [193]. Other very promising and attractive solutions could be to extend the current SS-MTL binary classification problem to a multiclass classification problem [215] or to learning to rank approach (i.e., learning the risk prediction using an ordinal structure of all CKD stages).

For what concernc the data processing, the strong class imbalance may be addressed using more advanced data imputation strategies rather than SMOTE, median/mean imputation [3] and KNN [216]. For instance, the missing values of the EHR field may be imputed by using conditional GAN [217] across different temporal windows and different spatial views (i.e., EHR fields).

Chapter 7.

COVID-19 Complications

Accurate risk stratification of patients with coronavirus disease 2019 (COVID-19) is a critical point to optimize resources allocation and deploy targeted interventions. The objective of this study is to predict the improvement or worsening of the Sequential Organ Failure Assessment (SOFA) score among COVID-19 patients admitted to the Intensive Care Unit (ICU). A prognostic observational cohort study with 5 days of follow-up from the ICU admission was designed. 96 ICUs participated at the RIsk Stratification in COVID-19 patients in the Intensive Care Unit (RISC-19-ICU) registry. The study was performed from March 17^{th} to October 31^{st}, 2020. Data were analyzed from November 5^{th} to November 30^{th}, 2020. RISC-19-ICU registry included patients diagnosed with COVID-19 admitted to the ICU with absolute SOFA change greater than 1 after 5 days from the ICU admission. Worsening or improvement of SOFA defined, respectively, as SOFA ≥ 2 points or SOFA ≤ 2 points, was predicted using the eXtreme Gradient Boosting (XGBoost) model. Among the 1613 COVID-19 patients in the RISC-19-ICU registry, 675 patients satisfied necessary inclusion conditions, with a median age of 64 (interquartile range (IQR) 56-63) and time from symptoms onset to ICU admission of 8 days (IQR 6-11). At the ICU admission, SOFA was 11 (IQR 6-14) with a pO2/FiO2 ratio of 121.6 (IQR 80.9-170.9), and 86% of patients were mechanically ventilated. The SOFA worsening was correctly predicted among 320 (83%) of the 385 while the SOFA improvement among 210 (72%) of the 290 COVID-19 patients. The corresponding area under the mean ROC curve was 0.86. The model selected the Glasgow coma scale, state of shock, use of vasopressors, and bilirubin concentration as the most relevant features to determine the SOFA modifications. Machine Learning (ML)-based prediction model could support physicians to accurately identify patients more likely to worsen or improve their conditions at the time of ICU admission in the following 5 days. Implementation of the model could help in the optimization of available resources and identification of early treatments to be adopted.

7.1. Introduction

The COVID-19 outbreak represents one of the most critical global health emergencies in modern times, reaching almost 1.2 million deaths worldwide at the end of October 2020. The COVID-19 pandemic poses an unprecedented challenge for policymakers across the world, given the pace at which its effects are unfolding and its potential to cause critical illness with single- and multi-organ failure. ICUs capacity has been rapidly exceeded in several regions around the world during the first weeks of the COVID-19 outbreak. The recent surge of COVID-19 cases in Europe is already challenging national healthcare systems. The ability to predict patients' complications and outcomes by analyzing the medical records of patients in ICU is hampered by numerous challenges such as difficulty in finding structured clinical data, missing values, and datasets collecting a sufficient amount of patients. Under these conditions, predicting the risk of a particular patient to develop complications associated with COVID-19 or to improve his conditions is relevant and may help both healthcare organization and clinical management of the patients defining a personalized risk profile and optimizing the appropriateness of care. The RISC-19-ICU registry was launched on March 17^{th}, 2020 and created to provide near real-time assessment of patients developing critical illness due to COVID-19 [34]. It includes up to 96 centers from 15 different countries with a continuously expanding number of critically ill COVID-19 patients, encompassing 1613 individual admissions at the end of October 2020. The analytical capability of ML methods has proven to be extremely accurate and in some cases superior to classical statistical approaches. This improvement has been also confirmed by recent work in this field aimed to propose ML methodologies for providing the prediction of risk conditions and complications related to chronic diseases [2, 4, 218]. Therefore, a prediction model could be trained using the parameters collected in the RISC-19-ICU registry at ICU admission, to estimate the worsening or improvement of critically ill COVID-19 patients within the first 5 days. The SOFA score was used to evaluate disease severity. The SOFA score is used to track the patient's status during the stay in ICU to determine the extent of a person's organ function or rate of failure. The score is based on six different scores, one each for the respiratory, cardiovascular, hepatic, coagulation, renal and neurological systems. The SOFA score can be measured daily on all patients admitted to ICU to determine the level of acuity and mortality risk. The accurate prediction of SOFA score may be relevant to the clinical scenario to provide risk profiles of individual patients from which a different intensity of care can be deduced, with consequent modification of the control time according to the needs.

7.1.1. Problem formulation

the SOFA change on the fifth day from ICU admission for individual patients was predicted. Worsening or improvement of SOFA was defined, respectively, as an increase

in SOFA ≥ 2 points or a decrease of ≤ 2 points.

Thus, the objective of the work aims to predict and manage the complications which COVID-19 patients develop, by designing a clinical decision support system (CDSS) based on ML algorithms to provide:

- "Risk profiles" of the COVID-19 patients in ICU from which a different intensity of care can be deduced, providing a shortening of the waiting time and increase the appropriateness of care
- "Prediction of risk of short term complications" of the COVID-19 patients that will activate personalized prevention systems directly addressed to the patient: from targeted recalls to targeted motivational and training activities.

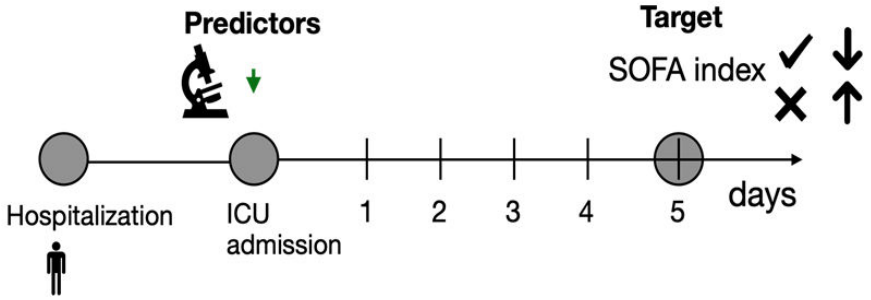


Figure 7.1.: Experimental setup of the proposed method.

7.2. Clinical data: RISC-19-ICU registry

The RISC-19-ICU registry [34] aims to collect an anonymized dataset to characterize patients that develop life-threatening critical illness due to COVID-19 and make it accessible to collaborative analysis. The data collected may be composed of a core dataset and/or an extended dataset. The core dataset consists of a basic set of parameters, of which many are commonly generated during treatment of critically ill patients with COVID-19 in an intensive care unit. The extended dataset consists of parameters that may be measured during treatment of critically ill patients with COVID-19 in an ICU, depending on clinical practice, indication, and availability of the measurement method. The data accumulating in the registry as the pandemic or subsequent waves develop are made available to the collaborators to support an optimal response to the pandemic threat. The information gained on the initial characteristics and disease course via the RISC-19-ICU registry may contribute to a better understanding of the risk factors for developing critical illness due to COVID-19 and for an unfavorable disease course, and thus support informed patient triage and management decisions.

The target population consists of i) patients admitted to a collaborating center with COVID-19 infection confirmed according to the WHO guidelines; ii) collaborating centers contributing anonymized data to the registry may be one of the following: ICU centers, non-ICU centers treating patients with high-flow oxygen therapy or non-invasive ventilation.

Patient characteristics, laboratory, and physiological parameters at the time of ICU admission were used as predictors (Figure 7.1, Table A.2). A feature with a number of missing values greater than 70 % was excluded from the model.

7.3. Methods

The XGBoost method [219] was applied as a prediction model in consideration of its characteristics of high generalization performance and the low risk of overfitting that outperforms other data mining methods widely used for solving predictive medicine tasks (see Figure 7.2) [2]. The gradient tree boosting algorithms extend the concept of adaptive boosting by sequentially adding predictors and correcting previous models using the gradient descent algorithm.

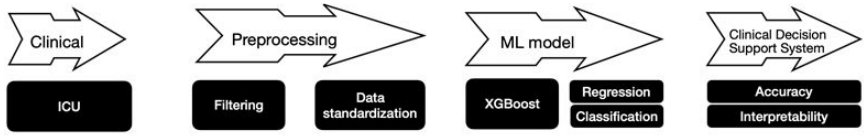


Figure 7.2.: Flowchart of the proposed method.

7.3.1. Experimental procedure

The model was built using the extended RISC-19-ICU registry after excluding patients with SOFA change ≤ 1 and patients with missing SOFA at the admission or at day 5. Among the 1613 patients included in the RISC-19-ICU registry, 1030 have stayed in the ICU for five or more days and had valid SOFA scores both at ICU admission and at day 5. The model was tested using a 10 fold cross-validation (CV-10) procedure. CV-10 was implemented by dividing all subjects into ten folds, by selecting nine folds for training and one fold for testing. The optimization of the hyperparameters was performed by implementing a grid-search and optimizing the macro-recall in a nested 5-fold CV. Hence, each split of the outer loop was trained with the optimal hyperparameters tuned in the inner loop. Although this procedure is computationally expensive, it allows obtaining an unbiased and robust performance evaluation. The model was applied to the entire RISC-19-ICU cohort consisting of 675 patients with an absolute SOFA score change of ≥ 2 points between the two-time points. The median age in the study cohort was 64 (interquartile range (IQR) 56-63), 74% were

males, and median body mass index (BMI) of 27.8 (IQR 25.4-31.6). At the ICU admission, median SOFA was 11 (IQR 6-14), median time from symptoms onset was 8 days (IQR 6-11), median pO2/FiO2 ratio 121.6 (IQR 80.9-170.9), and 86% of patients were mechanically ventilated (see Table 7.1).

Table 7.1.: Descriptive statistics.

Description	**X1st.Qu.**	**Median**	**Mean**	**X3rd.Qu.**
n° patients	675	675	675	675
Age	56	64	62.817	72
SOFA	6	11	10.319	14
Gender	0	1	0.745	1
BMI	25.391	27.766	29.029	31.561
Symptoms to ICU	6	8	9.316	11
pO2/FiO2 ratio	80.883	121.619	153.963	170.875
Mechanical ventilation	1	1	0.858	1

7.4. Experimental results

The model correctly predicted SOFA worsening among 320 (83%) of the 385 and SOFA improvement among 210 (72%) of the 290 COVID-19 patients (Table 7.2) with a corresponding area under the mean ROC curve of 0.86 (Figure 7.3). As expected, the most relevant features selected to determine the SOFA modifications were mainly related to the SOFA items such as the Glasgow coma scale, state of shock, use of vasopressors, and bilirubin concentration (Figure 7.4). However, other well-known features related to the patient's outcome such as type of respiratory support, comorbid conditions, the Acute Physiology And Chronic Health Evaluation (APACHE), and the Simplified Acute Physiology Score (SAPS) scores contributed to the correct prediction. Notably, the presence/absence of diabetes mellitus was among the most relevant conditions playing a role in the prediction of SOFA modifications at day five.

Table 7.2.: Confusion matrices (rows are the true classes) of the XGBoost algorithm for solving the classification task.

	Worsening	*Improvement*
Worsening	0.83	0.17
Improvement	0.28	0.72

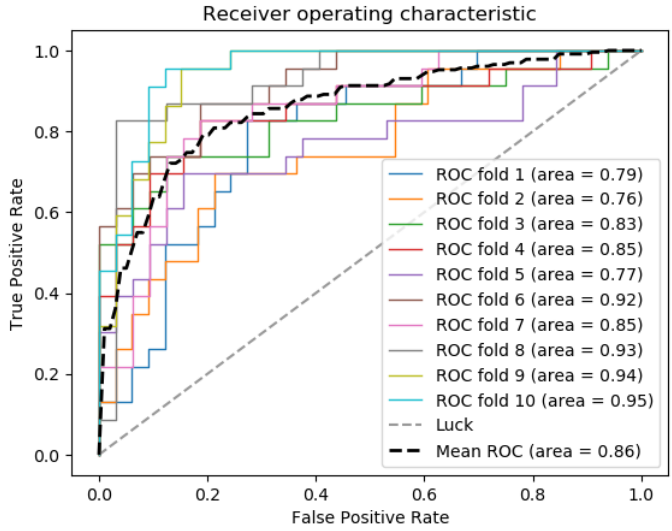


Figure 7.3.: Receiver operating characteristic (ROC) analysis over each fold.

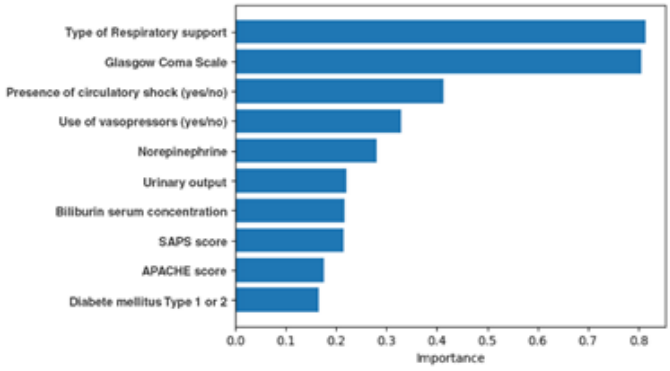


Figure 7.4.: Visualisation of the top 10 most discriminative features for predicting the SOFA score according to the XGBoost algorithm.

7.5. Discussion

Findings provide encouraging evidence regarding the use of a prediction model using advanced ML algorithms to discriminate patients likely to worsen or improve their clinical condition. The ML algorithm represents the main core of a CDSS that provides the risk profiles of the individual patients in the ICU in terms of prediction of risk of short-term complications. The implementation of the actual model may provide the creation of a strategic tool for the optimization of intensity able to assist physicians to make differentiated and personalized treatment decisions for critically ill COVID-19 patients. With the rapid increase of the RISC-19-ICU cohort, the aim is

to predict more clinically relevant outcomes such as the risk for endotracheal intubation, need for renal replacement therapy, and mortality. Moreover, the extension of the participation to the RISC-19-ICU registry to non-ICU departments will provide relevant information to identify patients with a high probability of failing a non-invasive ventilation trial and at high risk of ICU admission for non-respiratory complications. Additionally, the proposed model can ensure high interpretability by discovering relevant features related to the development of complications associated with COVID-19. The integration of the model into the RISC-19-ICU registry may allow predicting the risk profile of morbidity and resource consumption of the patients according to their clinical features. In conclusion, the use of advanced prediction models will ensure the more appropriate use of resources for those patients who need them most by increasing the appropriateness of care and activating a personalized prevention system directly addressed to the patient.

Chapter 8.

Conclusions

In the following chapter the conclusions of each single work presented in the thesis are shown. Then, the final considerations of the thesis are discussed.

Table 8.1 summarizes the ML challenges, listed in Section 1.4, which were faced by the ML approaches proposed for each work (ch. 2 ÷ ch. 7).

Table 8.1.: Machine Learning (ML) challenges (see Section 1.4) faced by the ML methodologies proposed in each chapter.

ML Methodologies	High-dimensional & heterogeneous data	Unbalance setting	Sparse labeling	Temporal ambiguity	Explainability/ Interpretability	Generalization
[2]T2D discovering (ch.2)	**x**	**x**	**x**		**x**	**x**
[3]IR: clinical factors (ch.3)			**x**	**x**	**x**	**x**
[4]IR: T2D early stage (ch.4)		**x**	**x**	**x**	**x**	**x**
[1]CDSS for T2D evaluation care (ch.5)	**x**		**x**		**x**	**x**
[5]KD early stage risk (ch.6)	**x**	**x**	**x**	**x**	**x**	**x**
[6]COVID-19 complications (ch.7)	**x**		**x**		**x**	**x**

In chapter 2 was presented the SB-SVM approach for discovering T2D in the novel collected EHR dataset, named FIMMG dataset. The SB-SVM approach was proven to be the best compromise between predictive performance and computation time, with respect to other ML and DL EHR-based approaches widely employed in the state-of-the-art for solving this task. The SB-SVM was able to manage high dimensional data, by increasing the model interpretability and finding the most relevant features while dealing with the usual unbalanced class distribution. The SB-SVM approach may be embedded in a CDSS to aid the physician in discovering and preventing T2D at an early stage, offering an adequate T2D integrated management system and patient's follow-up.

In chapter 3 was presented the TyG-er approach, a high-interpretable ensemble regression model for providing knowledge about the identification of clinical factors correlated with a T2D risk condition. In particular, the sparse nature of the employed publicly available FIMMG_*obs* dataset reflects the clinical use-case where not all laboratory exams are prescribed regularly over time. The TyG-er approach was able to carry information about the identification of the TyG index, strictly correlated with the IR condition while extracting the most relevant non-glycemic features from routine laboratory exams.

In chapter 4 was presented the MIL-Boost approach for the early prediction of T2D risk condition (low vs high T2D risk) using past EHR temporal information collected from a single GP. As demonstrated by the high predictive performances, the model interpretability, and the capability to handle variability in the number of observations, the MIL-Boost is a reliable approach that may represent the main core of a CDSS.

In chapter 5 was presented and tested on a real clinical use-case scenario and over time a comprehensive framework for supporting GPs during diabetes early detection and enrollment stage. In particular, was proposed an integrated chronic care model based on ML and data sharing between GPs and diabetes centers, as the main core of a CDSS. The quality care evaluation in a clinical use-case scenario demonstrated how the empowerment of the GPs through the use of the platform (integrating the proposed CDSS), along with the economic incentives, may speed up the improvement of care. The National Sanitary System and its regional agencies are funding the GP's incentives. The overall investment is based on the concept that prevention is less expensive than intervention. Chronic care models contribute to a long-term positive balance on the overall care strategy and budget, both in economic value and in quality of life.

In chapter 6 was presented the SS-MTL approach for predicting short-term KD evolution on multiple GPs' EHR data, named the mFIMMG dataset. The SS-MTL approach was capable to capture the eGFR temporal evolution by imposing a temporal relatedness between consecutive time-windows and was most capable to capture useful information from unlabeled patients when labeled patients are less numerous than those unlabeled. This situation reflects commonly the real-case general practice scenario, where available labeled samples are limited, but those unlabeled are much more abundant. The SS-MTL approach, exhibiting also a high level of interpretability, might be the ideal candidate in general practice to get integrated within a CDSS for CKD screening purposes.

In chapter 7 was presented the XGBoost approach for predicting the risk of developing complications for COVID-19 patients in ICU. In particular, utilising only the patient's information at ICU admission, the XGBoost model was capable to correctly predict the improvement or the worsening of the patient's clinical condition at 5 days from ICU admission. The proposed approach may help physicians to identify in advance COVID-19 patients admitted to the ICU at high risk of worsening their conditions. This information could benefit the therapeutic interventions with a more precise care strategy. With the power of analytics and prediction, we can advance to prescriptive medicine, effectively controlling the response of our patients starting with the earliest phases of an incipient critical illness and extending throughout the course of their care. With the prediction, prescription, and prevention of severe illness and organ dysfunction, the most common and vexing problems of critical care medicine may be reduced.

8.1. Final considerations

In the emerging era of PPM, the management of diseases is evolving into a more personalized approach, as more multi-source data are available from high-dimensional and heterogeneous EHR fields. PPM is capable to extend basic risk scores and predictions into personalized screening and monitoring, personalized forecasting of biomarkers, disease trajectories and clinical outcomes, and personalized estimation of treatment effects over time. PPM improves patients' quality of life and mitigates the increased cost of frequently necessary life-long treatments.

The application of ML methodologies to big data is facilitating the evolution of traditional medicine to intelligent healthcare ecosystems (see Figure 1.1). These healthcare ecosystems are already transforming the traditional analytic medicine - which describes what happened - to an emerging PM - which accurately predicts when an event will occur and understands why happened. In the near future, these healthcare ecosystems, in which science, biomedical informatics, incentives, and lifestyle education are aligned for continuous improvement and innovation, will look at prescriptive paradigms, which will control events or make events happen. Thus, these healthcare ecosystems will make easier the transition of a clinic or hospital from a traditional healthcare environment to a smarter learning healthcare system, which sets the foundation for personalised medicine on an international scale to ensure the best and the most timely response for patient care.

This last aspect may assume particular relevance in ICUs, where medical care is often time-sensitive because high-stakes decisions are made with incomplete information and imperfect knowledge. In critical care, PM may create opportunities such as predicting arrhythmias or cardiac arrest in minutes, respiratory or renal failure in hours, hospital complications and readmissions in months, etc. The fullest expression of these healthcare systems requires both additional scientific discovery and the real-time integration and analysis of these large-scale data to overcome the human limitations of information overload and cognitive processing. Instead, in general practice, PM will be fundamental to manage chronic pathologies (e.g., T2D, KD, cardiovascular disease, etc.) for preventive, screening, diagnostic, and treatment purposes.

With targeted investments, these healthcare ecosystems could be used for a variety of purposes at a local (e.g. clinic, hospital, etc.), regional, national, or even international scale. These healthcare ecosystems could turn EHRs from a tedious data-entry infrastructure into a powerful tool that empowers healthcare providers and patients alike—especially given the fact that a great deal of potentially relevant information is lost because it cannot be appreciated by clinicians, whereas ML predictions could make such data salient.

8.1.1. Open challenges

These healthcare ecosystems are not yet being used at a large scale to empower healthcare professionals and patients. The removal of several obstacles could accelerate this transformation process.

The first obstacle is the preprocessing stage (i.e., data cleaning, preparation, and standardization). In almost every case, a data scientist in collaboration with a healthcare professional will need to review, clean, and prepare data to be understood by ML frameworks. In the healthcare scenario, methods and protocols for maintaining and sharing EHR vary between countries and regions; furthermore, EHR data types vary between locations, organizations, and even individuals. EHR data need to be recorded and shared in a consistent, high-quality, and easy-to-read manner across the healthcare ecosystem.

The second obstacle is the ongoing need to build layers of abstraction that permit various users to interact with ML frameworks at their own knowledge level. This includes creating intuitive, easy-to-use interfaces that enable end-users (i.e., physicians) to describe clinical tasks they want to solve. The physician's objective should be understandable also by other users (i.e., data scientists) operating in different layers. Additionally, the trade-off between humans and the ML framework should be smooth, but at the same time, the distinct roles well-defined. For example, physicians and data scientists may need to formulate specific clinical tasks together at the beginning of the process and evaluate results at the end; by contrast, the ML framework can generally handle tasks in the middle, such as model construction, prediction, and treatment estimation.

The third obstacle is the commonization of components. User-friendly ML frameworks should be designed as modular blocks and selected depending on the objective of the clinical task. These blocks should also speak a standardized language that allows them to operate seamlessly when added to a framework.

In the short term, the removal of the obstacles outlined above would be an important step toward the realization of an ambitious and meaningful long-term vision, consisting of the creation of a novel healthcare ecosystem in which ML can understand, collaborate with, and empower humans. This healthcare ecosystem would require that, on the one hand, the ML methodologies would analyze EHR datasets and offer predictions and recommendations for the patient (i.e., personalized screening, monitoring, diagnosis, early diagnosis) while also providing interpretable results and uncertainty estimates associated with the various predictions and recommendations. On the other hand, would provide recommendations tailored to individual decision-makers, such as clinicians, administrators, or researchers. This would be based on a deeper, more fundamental understanding of human behavior and decision-making, combined with informed judgements regarding which tasks can be automated, which tasks can be recommended (but not fully automated), and which tasks should be left entirely to humans. While this healthcare ecosystem requires a level of cognition that is arguably

beyond the current abilities of ML, this vision would not be realized at a large scale in the near future. Whether AI systems are smarter than human practitioners makes for a stimulating ethical debate — but is largely irrelevant. Combining ML "software" with the best human clinician "hardware" will permit delivery of care that outperforms what either can do alone. Research encourages to publicly use and provide every information and data resource to consistently improve our collective health.

Appendix A.

Table A.1.: Type 2 diabetes patient's features set. Example from a real anonymous patient.

PATIENT DETAILS		
Glycated hemoglobin		Not registered
Total cholesterol		Not registered
HDL cholesterol		Not registered
Triglycerides		Not registered
LDL cholesterol		Not registered
Blood pressure	21/01/2016	80/130
Weight	21/01/2016	70
Height	21/01/2016	156
Body mass index	21/01/2016	28.76
Waist	21/01/2016	98
Microalbuminuria		Not registered
Creatinine		Not registered
Exemptions:		
E00		
013_R		
E10		
Pharmacological treatments:		
Glucophage * 30 pills 500 mg	14/01/2016	Metmorfin
Lifestyles:		
Smoking habit		Not detected
Cigarettes per day		Not detected
Alcohol consumption		Not detected
Alcohol type		Not detected
Physical activity		Not detected
Macrovascular Complications:		
Ischemic cardiopathy		Not affected
Acute myocardial infarction		Not affected
Revascularization		Not affected

Table A.1.: Type 2 diabetes patient's features set. Example from a real anonymous patient.

PATIENT DETAILS		
Claudicatio		Not affected
Transient ischemic attack		Not affected
Stroke		Not affected
Angina		Not affected
Eye Examination:		
Eye examination	28/02/2012	Performed
Retinopathy		Not affected
Blindness		Not affected
Examination of the Foot:		
Ulcers		Not affected
Amputations		Do not suffer
Renal complications:		
Nephropathy		Not affected
Dialysis		Not subjected
Educational Reinforcement:		
Power		Not done
Motor activity		Not done
Self control		Not done
Foot prevention		Not done
Verified glycemic control		Not done
Findings:		
Foot inspection		Not done
Uricemia		Unregistered
AST		Not registered
GGT		Not registered
Blood count formula		Not done
Cardiovascular risk		Not registered
Specializations:		
ECG		Not done
Diabetic visit		Not done
Cardiological visit		Not done
Eye examination	28/02/2012	Performed
Fundus oculi		Not done
Nephrological visit		Not done
Neurological visit		Not done
Date enrolled:	16/09/2015	

Table A.2.: Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%.

Features	**Missing*, %**
Age	0.00
Weight	2.99
Height	5.53
Gender	0.15
Adm_smoke	12.41
Initial Symptoms	
Rinorrhea	0.00
Headache	0.00
Dry Cough	0.00
Sore throat	0.00
Coloured Sputum production	0.00
Fatigue	0.00
Shortness of Breath	0.00
Nausea or Vomiting	0.00
Diarrhea	0.00
Myalgia or Arthralgia	0.00
Hemoptysis	0.00
Chest pain	0.00
Dizziness	0.00
Chills	0.00
Anorexia	0.00
Abdominal pain	0.00
Fever	0.00
Conjunctivitis	0.00
Tachypnea	0.00
Apnea	0.00
Syncope	0.00
Olfactory disorder	0.00
Taste disorder	0.00
Other neurological disorder	0.00
Fever before admission	67.26
Signs of Infection previous to ICU admission	
Throat Congestion	0.00
Tonsil Swelling	0.00
Enlargement of Lymph nodes	0.00
Rash	0.00
Rales	0.00

Table A.2.: Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%.

Features	Missing*, %
Comorbidities	
Myocardial Infarction / Ischemic Heart Disease	0.00
Chronic Heart Failure	0.00
Peripheral Vascular Disease	0.00
Chronic Arterial Hypertension	0.00
Diabetes mellitus Type 1 or 2	0.00
Diabetes mellitus with End Organ Damage	0.00
Cerebrovascular Disease	0.00
Dementia	0.00
Hemiplegia	0.00
Chronic Obstructive Pulmonary Disease	0.00
Chronic Restrictive Pulmonary Disease	0.00
Pulmonary Hypertension	0.00
Connective Tissue Disease	0.00
Peptic Ulcer Disease	0.00
Mild Liver Disease	0.00
Moderate to Severe Liver Disease	0.00
Moderate to Severe Chronic Kidney Disease	0.00
Solid Tumor (Localized)	0.00
Solid Tumor (Metastatic)	0.00
Leukemia	0.00
Lymphoma	0.00
HIV	0.00
Chronic Hepatitis B	0.00
Chronic Hepatitis C	0.00
Immunosupression for any reason	0.00
Previous Medication	
ACE-Inhibitors	0.00
Angiotensin II receptor blockers	0.00
Other cardiovascular medication (not including lipid lowering drugs)	0.00
Lipid lowering drugs	0.00
Platelet aggregation inhibitors	0.00
Oral anticoagulants	0.00
Inhalative Corticosteroids	0.00
Systemic Corticosteroids	0.00
Other immunosuppressive therapies	0.00
Oral antidiabetic drugs	0.00

Table A.2.: Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%.

Features	**Missing*, %**
Insulin	0.00
Non-steroidal anti-inflammatory drugs used to treat COVID-19 symptoms	0.00
Antiepileptics	0.00
Exposure to COVID-19 focus in past 14 days	
None Known	0.00
COVID-19 infected person (family or friends)	0.00
COVID-19 infected person	0.00
Travel to an Endemic Region up to 03-2020 (China, Italy, Iran)	0.00
Previous Therapies (before ICU Admission)	
Antibiotic Therapy	0.00
Specific Anti-SARS-CoV-2 Therapies (off-label)	0.00
Additional Oxygen (Nasal Canula, Mask, NIV, High-Flow)	0.00
Additional Parameters at Admission	
temperature	41.11
hematocrit	42.75
sodium	2.69
potassium	2.84
Symptoms duration	
sympt_to_hosp	5.38
sympt_to_dg	4.33
hosp_to_icu	19.43
Patient's indexes	
bmi	5.83
saps	0.15
apache	0.00
sofa.x	0.00
Clinical Assessment	
Wheezing	0.00
Rales/ Crackles	0.00
Pleural friction rub	0.00
Silent Chest	0.00
cyanosis_yn	69.96
GCS	4.93
estimated_urine_output	6.43
patient_shock_yn	64.42
Temperature [°C]	62.03
Heart Rate [1/min]	8.22

Table A.2.: Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%.

Features	**Missing*, %**
Respiratory Rate [1/min]	14.65
Systolic Arterial Pressure [mmHg]	62.03
Mean Arterial Pressure [mmHg]	12.41
Diastolic Arterial Pressure [mmHg]	62.03
Central Venous Pressure [mmHg]	69.06
Percutaneous Arterial Oxygen Saturation [%]	39.46
Vasopressors and Inotropes	
Norepinephrine Dose [µg/min]	40.51
Was Anti-Hypertensive Medication needed?	67.86
Arterial and central venous	
Arterial oxygen saturation [%]	65.17
Arterial pO2 [kPa]	5.23
Arterial pCO2 [kPa]	5.08
Arterial pH	5.23
Arterial HCO3- [mmol/L]	5.68
Arterial Base Excess [mmol/L]	66.37
Glucose [mmol/L]	7.32
Arterial Lactate [mmol/L]	1.20
FiO2 [%]	10.46
Need for Rescue Measures?	
Prone Positioning	0.00
Inhalative Nitric Oxide	0.00
Extracorporeal CO2 Removal	0.00
Continuous Renal Replacement Therapy or Hemodialysis of any form	0.00
vv-ECMO	0.00
va-ECMO (incl. vv-a, v-va, vv-va cannulation)	0.00
Abnormalities in Thorax Radiography	
Circumscribed infiltrate	0.00
Unilateral patchy infiltrates	0.00
Bilateral patchy infiltrates	0.00
Signs of interstitial abnormalities	0.00
Emphysema	0.00
Pneumothorax	0.00
Pleural Effusions	0.00
Signs of fungal infection (nodular aspect)	0.00
Abnormalities in Thorax CT	
Circumscribed infiltrate	0.00

Table A.2.: Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%.

Features	**Missing*, %**
Unilateral patchy infiltrates	0.00
Bilateral patchy infiltrates	0.00
Interstitial abnormalities	0.00
Emphysema	0.00
Pneumothorax	0.00
Pleural Effusions	0.00
Signs of fungal infection (nodular aspect, reverse Halo Sign)	0.00
Crayz Paving	0.00
Pulmonary Embolism	0.00
Valvulopathy in TTE or TEE	
None	0.00
Moderate to severe mitral regurgitation	0.00
Moderate to severe tricuspid regurgitation	0.00
Aortic valve stenosis	0.00
Other	0.00
Laboratory Parameters	
Leucocyte Count [109/L]	10.46
Neutrophil Count [109/L]	27.65
Lymphocyte Count [109/L]	27.50
Thrombocyte Count [109/L]	22.27
Hemoglobin [g/L]	66.67
Hematocrit [%	66.67
D-Dimer [µg/L]	41.55
CRP [mg/L]	18.09
PCT [µg/L]	38.71
LDH [U/L]	49.33
Serum Ferritin [µg/L]	68.16
Total Bilirubin [µmol/L]	39.91
Albumin [g/L]	69.81
Creatinine [µmol/L]	12.11
Urea [mmol/L]	32.44
Creatine Kinase [U/L]	57.70
High Sensitivity Troponin [ng/l]	56.80
Neuromuscular Relaxation	
Were Neuromuscular blocking agents used in the chart day?	67.86
Rocoronium	0.00
Vecuronium	0.00

Table A.2.: Percentage [%] of missing values for each feature. *The cut-off of missing values used to exclude a feature from the model was 70%.

Features	**Missing*, %**
Cisatracurium	0.00
Atracurium	0.00
Mivacurium	0.00
Other	0.00
time	0.00
Computed values	
pf_ratio	13.60
vent_ratio	18.39

Bibliography

[1] E. Frontoni, L. Romeo, M. Bernardini, S. Moccia, L. Migliorelli, M. Paolanti, A. Ferri, P. Misericordia, A. Mancini, and P. Zingaretti, "A Decision Support System for Diabetes Chronic Care Models based on General Practitioner Engagement and EHR Data Sharing," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 8, pp. 1–12, 2020.

[2] M. Bernardini, L. Romeo, P. Misericordia, and 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*, vol. 24, no. 1, pp. 235–246, 2020.

[3] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, and 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*, vol. 112, p. 103358, 2019.

[4] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, and L. Burattini, "Early temporal prediction of Type 2 Diabetes Risk Condition from a General Practitioner Electronic Health Record: A Multiple Instance Boosting Approach," *Artificial Intelligence in Medicine*, p. 101847, 2020.

[5] M. Bernardini, L. Romeo, E. Frontoni, and M. R. Amini, "A Semi-Supervised Multi-Task Learning approach for predicting short-term Kidney Disease evolution," *IEEE Journal of Biomedical and Health Informatics*, 2020 [submitted].

[6] J. Montomoli, L. Romeo, S. Moccia, M. Bernardini, L. Migliorelli, A. Donati, A. Carsetti, P. D. W. Garcia, T. Fumeaux, P. Guerci, R. A. Schuepbach, E. Frontoni, and M. P. Hilty, "Predicting 5-day SOFA score at ICU admission in COVID-19 patients: a proof-of-concept study using prospectively collected data from 1613 patients in the RISC-19-ICU registry," *Journal of the American Medical Association*, 2020 [submitted].

[7] D. M. Roden, "Cardiovascular pharmacogenomics: current status and future directions," *Journal of Human Genetics*, vol. 61, no. 1, pp. 79–85, 2016.

[8] R. Ramaswami, R. Bayer, and S. Galea, "Precision medicine from a public health perspective," *Annual Review of Public Health*, vol. 39, pp. 153–168, 2018.

[9] S. Mathur and J. Sutton, "Personalized medicine could transform healthcare," *Biomedical Reports*, vol. 7, no. 1, pp. 3–5, 2017.

[10] G. Pravettoni and S. Triberti, "A "P5" Approach to Healthcare and Health Technology," in *P5 eHealth: An Agenda for the Health Technologies of the Future*. Springer, Cham, 2020, pp. 3–17.

[11] C. Tuena, M. Semonella, J. Fernández-Álvarez, D. Colombo, and P. Cipresso, "Predictive precision medicine: Towards the computational challenge," in *P5 eHealth: An Agenda for the Health Technologies of the Future*. Springer, Cham, 2020, pp. 71–86.

[12] M. Y. Jen and M. Varacallo, "Predictive medicine," *StatPearls*, 2020.

[13] D. R. Leff and G.-Z. Yang, "Big data for precision medicine," *Engineering*, vol. 1, no. 3, pp. 277–279, 2015.

[14] J. J. Chen, T.-P. Lu, Y.-C. Chen, and W.-J. Lin, "Predictive biomarkers for treatment selection: statistical considerations," *Biomarkers in Medicine*, vol. 9, no. 11, pp. 1121–1135, 2015.

[15] R. Hodson, "Precision medicine," *Nature*, vol. 537, no. 7619, p. S49, 2016.

[16] D. Blumenthal, "Stimulating the adoption of health information technology," *New England Journal of Medicine*, vol. 360, no. 15, pp. 1477–1479, 2009.

[17] D. Blumenthal and M. Tavenner, "The "meaningful use" regulation for electronic health records," *New England Journal of Medicine*, vol. 2010, no. 363, pp. 501–504, 2010.

[18] B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, and P. G. Shekelle, "Systematic review: impact of health information technology on quality, efficiency, and costs of medical care," *Annals of Internal Medicine*, vol. 144, no. 10, pp. 742–752, 2006.

[19] C. Chen, T. Garrido, D. Chock, G. Okawa, and L. Liang, "The Kaiser Permanente Electronic Health Record: transforming and streamlining modalities of care," *Health Affairs*, vol. 28, no. 2, pp. 323–333, 2009.

[20] R. Kaushal, K. G. Shojania, and D. W. Bates, "Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review," *Archives of Internal Medicine*, vol. 163, no. 12, pp. 1409–1416, 2003.

[21] R. Amarasingham, L. Plantinga, M. Diener-West, D. J. Gaskin, and N. R. Powe, "Clinical information technologies and inpatient outcomes: a multiple hospital study," *Archives of Internal Medicine*, vol. 169, no. 2, pp. 108–114, 2009.

[22] S. T. Parente and J. S. McCullough, "Health information technology and patient safety: evidence from panel data," *Health Affairs*, vol. 28, no. 2, pp. 357–360, 2009.

[23] A. E. Anderson, W. T. Kerr, A. Thames, T. Li, J. Xiao, and 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*, vol. 60, pp. 162 – 168, 2016.

[24] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," *Artificial Intelligence in Medicine*, vol. 41, no. 3, pp. 251–262, 2007.

[25] C. J. Patel, N. Pho, M. McDuffie, J. Easton-Marks, C. Kothari, I. S. Kohane, and P. Avillach, "A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey," *Scientific Data*, vol. 3, p. 160096, 2016.

[26] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

[27] "Healthcare Cost and Utilization Project. Agency for Healthcare Research and Quality, Rockville, MD," https://www.ahrq.gov/data/hcup/index.html, content last reviewed November 2019.

[28] D. Teodoro, E. Sundvall, M. J. Junior, P. Ruch, and S. M. Freire, "ORBDA: An openEHR benchmark dataset for performance assessment of electronic health record servers," *PloS One*, vol. 13, no. 1, p. e0190028, 2018.

[29] T. Walley and A. Mantgani, "The uk general practice research database," *The Lancet*, vol. 350, no. 9084, pp. 1097–1099, 1997.

[30] A. Bourke, H. Dattani, and M. Robinson, "Feasibility study and methodology to create a quality-evaluated database of primary care data," *Journal of Innovation in Health Informatics*, vol. 12, no. 3, pp. 171–177, 2004.

[31] J. Hippisley-Cox, D. Stables, and M. Pringle, "QRESEARCH: a new general practice database for research," *Journal of Innovation in Health Informatics*, vol. 12, no. 1, pp. 49–50, 2004.

[32] A. Clegg, C. Bates, J. Young, R. Ryan, L. Nichols, E. A. Teale, M. A. Mohammed, J. Parry, and T. Marshall, "Development and validation of an electronic frailty index using routine primary care electronic health record data," *Age and Ageing*, vol. 45(3), pp. 353–360, 2016.

[33] S. A. Johannesdottir, E. Horváth-Puhó, V. Ehrenstein, M. Schmidt, L. Pedersen, and H. T. Sørensen, "Existing data sources for clinical epidemiology: The Danish national database of reimbursed prescriptions," *Clinical Epidemiology*, vol. 4, pp. 303–313, 2012.

[34] P. D. W. Garcia, T. Fumeaux, P. Guerci, D. M. Heuberger, J. Montomoli, F. Roche-Campo, R. A. Schuepbach, and M. P. Hilty, "Prognostic factors associated with mortality risk and disease progression in 639 critically ill patients with COVID-19 in Europe: Initial report of the international RISC-19-ICU prospective observational cohort," *EClinicalMedicine*, vol. 25, p. 100449, 2020.

[35] L. J. Frey, E. V. Bernstam, and J. C. Denny, "Precision medicine informatics," 2016.

[36] C.-J. Hsiao, E. Hing, and J. Ashman, *Trends in Electronic Health Record System Use Among Office-based Physicians, United States, 2007-2012*. US Department of Health and Human Services, 2014, no. 75.

[37] Z. Obermeyer and E. J. Emanuel, "Predicting the future-big data, machine learning, and clinical medicine," *The New England Journal of Medicine*, vol. 375, no. 13, p. 1216, 2016.

[38] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches," *Medical Care*, pp. S106–S113, 2010.

[39] S. Mezzatesta, C. Torino, P. D. Meo, G. Fiumara, and A. Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis," *Computer Methods and Programs in Biomedicine*, 2019.

[40] J. Su, J. Hu, J. Jiang, J. Xie, Y. Yang, B. He, J. Yang, and Y. Guan, "Extraction of risk factors for cardiovascular diseases from chinese electronic medical records," *Computer Methods and Programs in Biomedicine*, vol. 172, pp. 1 – 10, 2019.

[41] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *Medical Research Methodology*, vol. 19, no. 1, p. 64, 2019.

[42] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[43] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490*, 2016.

[44] K. G. M. M. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation." *Diabetic Medicine*, vol. 15 7, pp. 539–53, 1998.

[45] International Diabetes Federation., *IDF Diabetes Atlas, 8th edition. Brussels, Belgium,* International Diabetes Federation, 2017.

[46] WHO *et al.*, *Global report on diabetes*. World Health Organization, 2016.

[47] G. Sheikhi and H. Altınçay, "The Cost of Type ii Diabetes Mellitus: A Machine Learning Perspective," in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*. Springer, 2016, pp. 824–827.

[48] I. Kamkar, S. K. Gupta, D. Phung, and S. Venkatesh, "Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso," *Journal of Biomedical Informatics*, vol. 53, pp. 277–290, 2015.

[49] B. H. Cho, H. Yu, K.-W. Kim, T. H. Kim, I. Y. Kim, and S. I. Kim, "Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods," *Artificial Intelligence in Medicine*, vol. 42, no. 1, pp. 37–53, 2008.

[50] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," *International Journal of Medical Informatics*, vol. 97, no. Supplement C, pp. 120–127, 2017.

[51] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[52] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Procedia Computer Science*, vol. 82, pp. 115–121, 2016.

[53] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, p. 16, 2010.

[54] Y. Wang, P. F. Li, Y. Tian, J. J. Ren, and J. S. Li, "A shared decision-making system for diabetes medication choice utilizing electronic health record data," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 5, pp. 1280–1287, 2017.

[55] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[56] K.-M. Jung, "Support vector machines for unbalanced multicategory classification," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[57] J. Bi, Y. Chen, and J. Z. Wang, "A sparse support vector machine approach to region-based image categorization," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 1121–1128.

[58] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[59] H. Masnadi-Shirazi, N. Vasconcelos, and A. Iranmehr, "Cost-sensitive support vector machines," *arXiv preprint arXiv:1212.0975*, 2012.

[60] G. Lee and C. Scott, "Nested support vector machines," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1648–1660, 2010.

[61] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

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

[63] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[64] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines." in *ICML*, vol. 98, 1998, pp. 82–90.

[65] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[66] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.

[67] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[68] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for Machine Learning*. The MIT Press, 2011.

[69] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.

[70] M. A. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[71] J. C. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999.

[72] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[73] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," vol. 2, pp. 42–47, 01 2012.

[74] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classif.*, vol. 10, 06 2000.

[75] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *Advances in neural information processing systems*, 1998, pp. 507–513.

[76] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Medical concept representation learning from electronic health records and its application on heart failure prediction," *arXiv preprint arXiv:1602.03686*, 2016.

[77] H. Li, X. Li, M. Ramanathan, and A. Zhang, "Identifying informative risk factors and predicting bone disease progression via deep belief networks," *Methods*, vol. 69, no. 3, pp. 257–265, 2014.

[78] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[79] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[80] L. Baldassarre, M. Pontil, and J. Mourão-Miranda, "Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding," *Frontiers in neuroscience*, vol. 11, p. 62, 2017.

[81] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[82] S. Tateishi, H. Matsui, and S. Konishi, "Nonlinear regression modeling via the lasso-type regularization," *Journal of Statistical Planning and Inference*, vol. 140, no. 5, pp. 1125–1134, 2010.

[83] K. H. Mikkelsen, F. K. Knop, M. Frost, J. Hallas, and A. Pottegård, "Use of antibiotics and risk of type 2 diabetes: a population-based case-control study," *The Journal of Clinical Endocrinology & Metabolism*, vol. 100, no. 10, pp. 3633–3640, 2015.

[84] B. Boursi, R. Mamtani, K. Haynes, and Y.-X. Yang, "The effect of past antibiotic exposure on diabetes risk," *European Journal of Endocrinology*, pp. EJE–14, 2015.

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

[86] J. Tuomilehto and P. E. Schwarz, "Preventing diabetes: Early versus late preventive interventions," *Diabetes Care*, vol. 39, no. Supplement 2, pp. 115–120, 2016.

[87] American Diabetes Association, "Classification and diagnosis of diabetes: Standards of medical care in diabetes—2018," *Diabetes Care*, vol. 41, no. Supplement 1, pp. 13–27, 2018.

[88] B. Antuna-Puente, E. Disse, R. Rabasa-Lhoret, M. Laville, J. Capeau, and J.-P. Bastard, "How can we measure insulin sensitivity/resistance?" *Diabetes & Metabolism*, vol. 1849, no. 3, pp. 169–264, 2011.

[89] L. E. Simental-Mendía, M. Rodríguez-Morán, and 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*, vol. 6, no. 4, pp. 299–304, 2008.

[90] F. Guerrero-Romero, L. E. Simental-Mendía, M. González-Ortiz, E. Martínez-Abundis, M. G. Ramos-Zavala, S. O. Hernández-González, O. Jacques-Camarena, and 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*, vol. 95, no. 7, pp. 3347–3351, 2010.

[91] D. Navarro-González, L. Sánchez-Íñigo, J. Pastrana-Delgado, A. Fernández-Montero, and 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*, vol. 86, pp. 99 – 105, 2016.

[92] B. Dorcely, K. Katz, R. Jagannathan, S. S. Chiang, B. Oluwadare, I. J. Goldberg, and M. Bergman, "Novel biomarkers for prediabetes, diabetes, and associated complications," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, vol. 10, pp. 345–361, 2017.

[93] A. Talaei-Khoei and 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*, vol. 119, pp. 22 – 38, 2018.

[94] F. Castiglione, P. Tieri, A. De Graaf, C. Franceschi, P. Liò, B. Van Ommen, C. Mazzà, A. Tuchel, M. Bernaschi, C. Samson, T. Colombo, G. C. Castellani, M. Capri, P. Garagnani, S. Salvioli, V. A. Nguyen, I. Bobeldijk-Pastorova, S. Krishnan, A. Cappozzo, M. Sacchetti, M. Morettini, and M. Ernst, "The onset of type 2 diabetes: Proposal for a multi-scale model," *Journal of Medical Internet Research*, vol. 15, no. 10, 2013.

[95] M. Maniruzzaman, N. Kumar, M. M. Abedin, M. S. Islam, H. S. Suri, A. S. El-Baz, and J. S. Suri, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Computer Methods and Programs in Biomedicine*, vol. 152, pp. 23–34, 2017.

[96] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[97] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: a critical evaluation," *Medical Informatics and Decision Making*, vol. 16, no. 3, p. 74, 2016.

[98] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[99] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2004.

[100] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[101] S. Janitza, E. Celik, and A.-L. Boulesteix, "A computationally fast variable importance test for random forests for high-dimensional data," *Advances in Data Analysis and Classification*, vol. 12, no. 4, pp. 885–915, 2018.

[102] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[103] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[104] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis: Communications in statistics," *Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[105] "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," *Decision Support Systems*, vol. 51, no. 1, pp. 141 – 154, 2011.

[106] H. Gylling, M. Hallikainen, J. Pihlajamäki, P. Simonen, J. Kuusisto, M. Laakso, and T. A. Miettinen, "Insulin sensitivity regulates cholesterol metabolism to a greater extent than obesity: lessons from the METSIM Study," *Journal of Lipid Research*, vol. 51, no. 8, pp. 2422–2427, 2010.

[107] E. Krishnan, B. J. Pandya, L. Chung, A. Hariri, and O. Dabbous, "Hyperuricemia in young adults and risk of insulin resistance, prediabetes, and diabetes: A 15-year follow-up study," *American Journal of Epidemiology*, vol. 176, no. 2, pp. 108–116, 2012.

[108] M. A. de Vries, A. Alipour, B. Klop, G.-J. M. van de Geijn, H. W. Janssen, T. L. Njo, N. van der Meulen, A. P. Rietveld, A. H. Liem, E. M. Westerman, W. W. de Herder, and M. C. Cabezas, "Glucose-dependent leukocyte activation in patients with type 2 diabetes mellitus, familial combined hyperlipidemia and healthy controls," *Metabolism*, vol. 64, no. 2, pp. 213 – 217, 2015.

[109] X. Liu, O.-P. R. Hamnvik, J. P. Chamberland, M. Petrou, H. Gong, C. A. Christophi, D. C. Christiani, S. N. Kales, and C. S. Mantzoros, "Circulating alanine transaminase (ALT) and γ-glutamyl transferase (GGT), but not fetuin-A, are associated with metabolic risk factors, at baseline and at two-year follow-up: The prospective Cyprus Metabolism Study," *Metabolism*, vol. 63, no. 6, pp. 773 – 782, 2014.

[110] L. Pinnaduwage, C. Ye, A. J. Hanley, P. W. Connelly, M. Sermer, B. Zinman, and R. Retnakaran, "Changes over time in hepatic markers predict changes in insulin sensitivity, β-Cell function, and glycemia," *The Journal of Clinical Endocrinology & Metabolism*, vol. 103, no. 7, pp. 2651–2659, 04 2018.

[111] L. Bozkurt, C. S. Göbl, A. Tura, M. Chmelik, T. Prikoszovich, L. Kosi, O. Wagner, M. Roden, G. Pacini, A. Gastaldelli *et al.*, "Fatty liver index predicts further metabolic deteriorations in women with previous gestational diabetes," *PloS One*, vol. 7, no. 2, p. e32710, 2012.

[112] D.-J. Lee, J.-S. Choi, K.-M. Kim, N.-S. Joo, S.-H. Lee, and K.-N. Kim, "Combined effect of serum gamma-glutamyltransferase and uric acid on Framingham risk score," *Archives of Medical Research*, vol. 45, no. 4, pp. 337–342, 2014.

[113] S. Riaz, "Study of protein biomarkers of diabetes mellitus type 2 and therapy with vitamin B1," *Journal of Diabetes Research*, vol. 2015, 2015.

[114] M. Y. Bertram and T. Vos, "Quantifying the duration of pre-diabetes," *Australian and New Zealand journal of Public Health*, vol. 34, no. 3, pp. 311–314, 2010.

[115] R. Taylor, "Insulin resistance and type 2 diabetes," *Diabetes*, vol. 61, no. 4, pp. 778–779, 2012.

[116] M. E. Cerf, "Beta cell dysfunction and insulin resistance," *Frontiers in Endocrinology*, vol. 4, p. 37, 2013.

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

[118] K. Sikka, A. Dhall, and M. S. Bartlett, "Classification and weakly supervised pain localization using multiple segment representation," *Image and Vision Computing*, vol. 32, no. 10, pp. 659–670, 2014.

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

[120] J. Guo, X. Yuan, X. Zheng, P. Xu, Y. Xiao, and B. Liu, "Diagnosis labeling with disease-specific characteristics mining," *Artificial Intelligence in Medicine*, vol. 90, pp. 25 – 33, 2018.

[121] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, p. 26094, 2016.

[122] J. Zhao, Q. Feng, P. Wu, R. A. Lupu, R. A. Wilke, Q. S. Wells, J. C. Denny, and W.-Q. Wei, "Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction," *Scientific Reports*, vol. 9, no. 1, p. 717, 2019.

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

[124] J. Zhao, S. Gu, and A. McDermaid, "Predicting outcomes of chronic kidney disease from emr data based on random forest regression," *Mathematical Biosciences*, vol. 310, pp. 24–30, 2019.

[125] A. Shknevsky, Y. Shahar, and R. Moskovitch, "Consistent discovery of frequent interval-based temporal patterns in chronic patients' data," *Journal of Biomedical Informatics*, vol. 75, pp. 83–95, 2017.

[126] 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.

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

[128] S. Mani, Y. Chen, T. Elasy, W. Clayton, and 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.

[129] A. Pimentel, A. V. Carreiro, R. T. Ribeiro, and H. Gamboa, "Screening diabetes mellitus 2 based on electronic health records using temporal features," *Health Informatics Journal*, vol. 24, no. 2, pp. 194–205, 2018.

[130] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and 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.

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

[132] E. W. Steyerberg, "Missing values," in *Clinical Prediction Models*. Springer, 2019, pp. 127–155.

[133] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

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

[135] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

[136] 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*, vol. 18, no. 5, pp. 808–818, 2014.

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

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

[139] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[140] J. Friedman, T. Hastie, R. Tibshirani *et al.*, "Additive logistic regression: a statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

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

[142] Y. Chevaleyre and 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.

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

[144] L. Rasmy, Y. Wu, N. Wang, X. Geng, W. J. Zheng, F. Wang, H. Wu, H. Xu, and 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*, vol. 84, pp. 11 – 16, 2018.

[145] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *Journal of Biomedical Informatics*, vol. 69, pp. 218 – 229, 2017.

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

[147] V. H. Telle-Hansen, K. B. Holven, and S. M. Ulven, "Impact of a healthy dietary pattern on gut microbiota and systemic inflammation in humans," *Nutrients*, vol. 10, no. 11, 2018.

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

[149] 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*, vol. 11, no. 10, p. e0163721, 2016.

[150] E. Frontoni, A. Mancini, M. Baldi, M. Paolanti, S. Moccia, P. Zingaretti, V. Landro, and P. Misericordia, "Sharing health data among general practitioners: The Nu. Sa. project," *International Journal of Medical Informatics*, vol. 129, pp. 267–274, 2019.

[151] G. Vazquez, S. Duval, D. R. Jacobs Jr, and K. Silventoinen, "Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis," *Epidemiologic Reviews*, vol. 29, no. 1, pp. 115–128, 2007.

[152] J. L. Harding, M. E. Pavkov, D. J. Magliano, J. E. Shaw, and E. W. Gregg, "Global trends in diabetes complications: a review of current evidence," *Diabetologia*, vol. 62, no. 1, pp. 3–16, 2019.

[153] International Diabetes Federation, "Diabetes epidemic in Europe," https://www.euro.who.int/en/health-topics/noncommunicable-diseases/diabetes/news/news/2011/11/diabetes-epidemic-in-europe,Accessed:11-9-2020.

[154] Osservatorio ARNO Diabete, "Rapporto ARNO Diabete 2019," http://www.siditalia.it/news/2547-21-11-2019-rapporto-arno-diabete-2019, Accessed: 10-9-2020.

[155] S. C. Bolge, N. M. Flores, and J. H. Phan, "The burden of poor mental well-being among patients with type 2 diabetes mellitus: examining health care resource use and work productivity loss," *Journal of Occupational and Environmental Medicine*, vol. 58, no. 11, p. 1121, 2016.

[156] Bruno, Ed., *Guida alla legislazione regionale sul diabete in Italia*. Società Italiana di Diabetologia, 2012.

[157] Ministero della Salute, "Piano nazionale della malattia diabetica," 1985.

[158] C. Rinner, S. K. Sauter, G. Endel, G. Heinze, S. Thurner, P. Klimek, and G. Duftschmid, "Improving the informational continuity of care in diabetes mellitus treatment with a nationwide Shared EHR system: Estimates from austrian claims data," *International Journal of Medical Informatics*, vol. 92, pp. 44–53, 2016.

[159] F. Pecoraro, D. Luzi, and F. L. Ricci, "Secondary uses of ehr systems: A feasibility study," in *E-Health and Bioengineering Conference, 2013*. IEEE, 2013, pp. 1–6.

[160] Z. Bi, M. Wang, L. Ni, G. Ye, D. Zhou, C. Yan, X. Zeng, and J. Chen, "A practical electronic health record-based dry weight supervision model for hemodialysis patients," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, pp. 1–9, 2019.

[161] T. Ermakova, B. Fabian, and R. Zarnekow, "Security and privacy system requirements for adopting cloud computing in healthcare data sharing scenarios," vol. 4, 08 2013.

[162] C. He, X. Jin, Z. Zhao, and T. Xiang, "A cloud computing solution for hospital information system," in *Intelligent Computing and Intelligent Systems, on IEEE International Conference*, vol. 2, 2010, pp. 517–520.

[163] Q. Huang, L. Ye, M. Yu, F. Wu, and R. Liang, "Medical information integration based cloud computing," in *IEEE International Conference on Network Computing and Information Security,*, vol. 1, 2011, pp. 79–83.

[164] H. Löhr, A.-R. Sadeghi, and M. Winandy, "Securing the e-health cloud," in *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, 2010, pp. 220–229.

[165] H. Yamaguchi and Y. Ito, "Improving the effectiveness of interprofessional work teams using ehr-based data in the treatment of chronic diseases: An action research study," in *Management of Engineering & Technology (PICMET), 2014 Portland International Conference on*. IEEE, 2014, pp. 3492–3497.

[166] C. Chang, K. Liao, Y. Chen, S. Wang, M. Jan, and G. Wang, "Radial pulse spectrum analysis as risk markers to improve the risk stratification of silent myocardial ischemia in type 2 diabetic patients," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1–9, 2018.

[167] M. T. Korytkowski, E. K. French, M. Brooks, D. DeAlmeida, J. Kanter, M. Lombardero, V. Magaji, T. Orchard, and L. Siminerio, "Use of an electronic health record to identify prevalent and incident cardiovascular disease in type 2 diabetes according to treatment strategy," *BMJ Open Diabetes Research and Care*, vol. 4, no. 1, p. e000206, 2016.

[168] T. J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R. R. Coeytaux, G. Samsa, V. Hasselblad, J. W. Williams, M. D. Musty *et al.*, "Effect of clinical decision-support systems: a systematic review," *Annals of Internal Medicine*, vol. 157, no. 1, pp. 29–43, 2012.

[169] V. Patel, M. E. Reed, and R. W. Grant, "Electronic health records and the evolution of diabetes care: a narrative review," *Journal of Diabetes Science and Technology*, vol. 9, no. 3, pp. 676–680, 2015.

[170] E. A. McGlynn, S. M. Asch, J. Adams, J. Keesey, J. Hicks, A. DeCristofaro, and E. A. Kerr, "The quality of health care delivered to adults in the United States," *New England Journal of Medicine*, vol. 348, no. 26, pp. 2635–2645, 2003.

[171] T. D. Sequist, T. K. Gandhi, A. S. Karson, J. M. Fiskio, D. Bugbee, M. Sperling, E. F. Cook, E. J. Orav, D. G. Fairchild, and D. W. Bates, "A randomized trial of electronic clinical reminders to improve quality of care for diabetes and coronary artery disease," *Journal of the American Medical Informatics Association*, vol. 12, no. 4, pp. 431–437, 2005.

[172] A. Holbrook, L. Thabane, K. Keshavjee, L. Dolovich, B. Bernstein, D. Chan, S. Troyan, G. Foster, H. Gerstein, C. I. Investigators *et al.*, "Individualized electronic decision support and reminders to improve diabetes care in the community: Compete ii randomized trial," *Canadian Medical Association Journal*, vol. 181, no. 1-2, pp. 37–44, 2009.

[173] M. Ati, W. Omar *et al.*, "Knowledge based system framework for managing chronic diseases based on service oriented architecture," in *2012 8th International Conference on Information Science and Digital Content Technology (ICIDT2012)*, vol. 1. IEEE, 2012, pp. 20–23.

[174] A. N. Kho, M. G. Hayes, L. Rasmussen-Torvik, J. A. Pacheco, W. K. Thompson, L. L. Armstrong, J. C. Denny, P. L. Peissig, A. W. Miller, W.-Q. Wei *et al.*, "Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 212–218, 2011.

[175] R. Kudyakov, J. Bowen, E. Ewen, S. L. West, Y. Daoud, N. Fleming, and A. Masica, "Electronic health record use to classify patients with newly diagnosed versus preexisting type 2 diabetes: infrastructure for comparative effectiveness research and population health management," *Population Health Management*, vol. 15, no. 1, pp. 3–11, 2012.

[176] M. L. Ho, N. Lawrence, C. van Walraven, D. Manuel, E. Keely, J. Malcolm, R. D. Reid, and A. J. Forster, "The accuracy of using integrated electronic health care data to identify patients with undiagnosed diabetes mellitus." *Journal of Evaluation in Clinical Practice*, vol. 18 3, pp. 606–11, 2012.

[177] W.-Q. Wei, C. L. Leibson, J. E. Ransom, A. N. Kho, and C. G. Chute, "The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects," *International Journal of Medical Informatics*, vol. 82 4, pp. 239–47, 2013.

[178] G. Fico, A. Fioravanti, M. T. Arredondo, J. Gorman, C. Diazzi, G. Arcuri, C. Conti, and G. Pirini, "Integration of personalized healthcare pathways in an ict platform for diabetes managements: A small-scale exploratory study," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 1, pp. 29–38, Jan 2016.

[179] M. F. Piepoli, A. W. Hoes, S. Agewall, C. Albus, C. Brotons, A. L. Catapano, M.-T. Cooney, U. Corra, B. Cosyns, C. Deaton *et al.*, "2016 European guidelines on cardiovascular disease prevention in clinical practice: The sixth joint task force of the european society of cardiology and other societies on cardiovascular disease prevention in clinical practice," *European Heart Journal*, vol. 37, no. 29, pp. 2315–2381, 2016.

[180] UK Prospective Diabetes Study Group and others, "Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes," *The Lancet*, vol. 352, no. 9131, pp. 837–853, 1998.

[181] Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group, "Retinopathy and nephropathy in patients with type 1 diabetes four years after a trial of intensive therapy," *New England Journal of Medicine*, vol. 342, no. 6, pp. 381–389, 2000.

[182] "American Diabetes Association Standards of Medical Care in Diabetes," 2018.

[183] J. D. Brown, "Likert items and scales of measurement," *Statistics*, vol. 15, no. 1, pp. 10–14, 2011.

[184] T. Heikkilä, L. Dalgaard, and J. Koskinen, "Designing autonomous robot systems-evaluation of the r3-COP decision support system approach," in *SAFECOMP 2013-Workshop DECS of the 32nd International Conference on Computer Safety, Reliability and Security*, 2013, p. NA.

[185] V. A. Luyckx, M. Tonelli, and J. W. Stanifer, "The global burden of kidney disease and the sustainable development goals," *Bulletin of the World Health Organization*, vol. 96, no. 6, p. 414, 2018.

[186] A. Levin, M. Tonelli, J. Bonventre, J. Coresh, J.-A. Donner, A. B. Fogo, C. S. Fox, R. T. Gansevoort, H. J. Heerspink, M. Jardine *et al.*, "Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy," *The Lancet*, vol. 390, no. 10105, pp. 1888–1917, 2017.

[187] N. J. Kassebaum, M. Arora, R. M. Barber, Z. A. Bhutta, J. Brown, A. Carter, D. C. Casey, F. J. Charlson, M. M. Coates, M. Coggeshall *et al.*, "Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015," *The Lancet*, vol. 388, no. 10053, pp. 1603–1658, 2016.

[188] World Health Organization *et al.*, "Tackling NCDs:'best buys' and other recommended interventions for the prevention and control of noncommunicable diseases," World Health Organization, Tech. Rep., 2017.

[189] A. S. Levey, J. Coresh, H. Tighiouart, T. Greene, and L. A. Inker, "Measured and estimated glomerular filtration rate: current status and future directions," *Nature Reviews Nephrology*, pp. 1–14, 2019.

[190] P. E. Stevens and A. Levin, "Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline," *Annals of Internal Medicine*, vol. 158, no. 11, pp. 825–830, 2013.

[191] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, and Z. He, "Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review," *Journal of the American Medical Informatics Association*, 2020.

[192] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 814–822.

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

[194] S. Bailly, G. Meyfroidt, and J.-F. Timsit, "What's new in ICU in 2050: big data and machine learning," *Intensive Care Medicine*, vol. 44, no. 9, pp. 1524–1527, 2018.

[195] P. Fraccaro, S. van der Veer, B. Brown, M. Prosperi, D. O'Donoghue, G. S. Collins, I. Buchan, and N. Peek, "An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK," *BMC Medicine*, vol. 14, no. 1, p. 104, 2016.

[196] S. Ravizza, T. Huschto, A. Adamov, L. Böhm, A. Büsser, F. F. Flöther, R. Hinzmann, H. König, S. M. McAhren, D. H. Robertson *et al.*, "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," *Nature Medicine*, vol. 25, no. 1, pp. 57–59, 2019.

[197] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Computers in Biology and Medicine*, vol. 109, pp. 101–111, 2019.

[198] D. Y. Ding, C. Simpson, S. Pfohl, D. C. Kale, K. Jung, and N. H. Shah, "The effectiveness of Multitask Learning for Phenotyping with Electronic Health Records Data." in *Pacific Symposium on Biocomputing*. World Scientific, 2019, pp. 18–29.

[199] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[200] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, pp. 1–18, 2019.

[201] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional gans," *arXiv preprint arXiv:1706.02633*, 2017.

[202] F. Mordelet and J.-P. Vert, "ProDiGe: Prioritization of Disease Genes with multitask machine learning from positive and unlabeled examples," *BMC Bioinformatics*, vol. 12, no. 1, p. 389, 2011.

[203] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: leveraging patient similarity and drug similarity analytics," *Summits on Translational Science Proceedings*, vol. 2014, p. 132, 2014.

[204] M. R. Amini, N. Usunier, and F. Laviolette, "A transductive bound for the voted classifier with an application to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2009, pp. 65–72.

[205] A. S. Levey, L. A. Stevens, C. H. Schmid, Y. L. Zhang, A. F. Castro, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene *et al.*, "A new equation to estimate glomerular filtration rate," *Annals of Internal Medicine*, vol. 150, no. 9, pp. 604–612, 2009.

[206] L. A. Inker, C. H. Schmid, H. Tighiouart, J. H. Eckfeldt, H. I. Feldman, T. Greene, J. W. Kusek, J. Manzi, F. Van Lente, Y. L. Zhang *et al.*, "Estimating glomerular filtration rate from serum creatinine and cystatin C," *New England Journal of Medicine*, vol. 367, no. 1, pp. 20–29, 2012.

[207] A. K. Bello, M. Alrukhaimi, G. E. Ashuntantang, S. Basnet, R. C. Rotter, W. G. Douthat, R. Kazancioglu, A. Köttgen, M. Nangaku, N. R. Powe *et al.*, "Complications of chronic kidney disease: current state, knowledge gaps, and strategy for action," *Kidney International Supplements*, vol. 7, no. 2, pp. 122–129, 2017.

[208] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1095–1103.

[209] M.-R. Amini and N. Usunier, *Learning with Partially Labeled and Interdependent Data*. Springer, 2015.

[210] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[211] D. Singh, K. Shrestha, J. M. Testani, F. H. Verbrugge, M. Dupont, W. Mullens, and W. W. Tang, "Insufficient natriuretic response to continuous intravenous furosemide is associated with poor long-term outcomes in acute decompensated heart failure," *Journal of Cardiac Failure*, vol. 20, no. 6, pp. 392–399, 2014.

[212] Y. Okuhara, S. Hirotani, Y. Naito, A. Nakabo, T. Iwasaku, A. Eguchi, D. Morisawa, T. Ando, H. Sawada, E. Manabe *et al.*, "Intravenous salt supplementation

with low-dose furosemide for treatment of acute decompensated heart failure," *Journal of Cardiac Failure*, vol. 20, no. 5, pp. 295–301, 2014.

[213] M. J. Crowley, C. J. Diamantidis, J. R. McDuffie, C. B. Cameron, J. W. Stanifer, C. K. Mock, X. Wang, S. Tang, A. Nagi, A. S. Kosinski *et al.*, "Clinical outcomes of metformin use in populations with chronic kidney disease, congestive heart failure, or chronic liver disease: a systematic review," *Annals of Internal Medicine*, vol. 166, no. 3, pp. 191–200, 2017.

[214] M. Kandemir, A. Vetek, M. Goenen, A. Klami, and S. Kaski, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, 2014.

[215] V. Feofanov, E. Devijver, and M.-R. Amini, "Transductive bounds for the multi-class majority vote classifier," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3566–3573.

[216] C. Ye, T. Fu, S. Hao, Y. Zhang, O. Wang, B. Jin, M. Xia, M. Liu, X. Zhou, Q. Wu *et al.*, "Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning," *Journal of Medical Internet Research*, vol. 20, no. 1, p. e22, 2018.

[217] A. Doinychko and M.-R. Amini, "Biconditional generative adversarial networks for multiview learning with missing views," in *European Conference on Information Retrieval*. Springer, 2020, pp. 807–820.

[218] L. Romeo, G. Armentano, A. Nicolucci, M. Vespasiani, G. Vespasiani, and E. Frontoni, "A novel spatio-temporal multi-task approach for the prediction of diabetes-related complication: a cardiopathy case of study."

[219] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, pp. 1–4, 2015.