IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 18, 2021, accepted November 6, 2021, date of publication November 10, 2021,
date of current version November 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3127274

A Clinical Decision Support System to Stratify the Temporal Risk of Diabetic Retinopathy

MICHELE BERNARDINI[1], LUCA ROMEO[1,2], ADRIANO MANCINI[1], AND EMANUELE FRONTONI[1], (Member, IEEE)

[1]Department of Information Engineering, Università Politecnica delle Marche, 60131 Ancona, Italy
[2]Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, 16163 Genova, Italy

Corresponding author: Michele Bernardini (m.bernardini@pm.univpm.it)

This work was supported in part by a research agreement between Meteda srl company and the Department of Information Engineering, Università Politecnica delle Marche; and in part by the Microsoft Grant Award: "AI for Health Covid-19."

ABSTRACT Diabetic Retinopathy (DR) is the most common and insidious microvascular complication of diabetes, and can progress asymptomatically until a sudden loss of vision occurs. Although DR is prevalent nowadays, its prevention remains challenging. The multiple aim of this study was to predict the risk of developing DR as diabetic complication (task 1) and, subsequently, temporally stratify the DR risk (task 2) using electronic health records data. To perform these objectives, a novel preprocessing procedure was designed to select both control and pathological patients, and moreover, a novel fully annotated/standardized 120K dataset from multiple diabetologic centers was provided. Globally, although the Extreme Gradient Boosting model offers satisfying predictive performance, the Random Forest model obtained the best predictive performance to solve task 1 and task 2, reaching the best Area Under the Precision-Recall Curve of 72.43 % and 84.38 %, respectively. Also the features importance extracted from the best Machine Learning (ML) models is provided. The proposed Artificial Intelligence-based solution was proven to be capable of generalizing across different diabetologic centers while ensuring high-interpretability. Moreover, the proposed ML solution is currently being adopted as a Clinical Decision Support System in several diabetologic centers for DR screening and follow-up purposes.

INDEX TERMS Predictive medicine, diabetic retinopathy, machine learning, electronic health records.

I. INTRODUCTION

The diabetic retinopathy (DR) is the most common and insidious microvascular complication of diabetes, and can progress asymptomatically until a sudden loss of vision occurs [1]. Almost all patients with type 1 diabetes mellitus and 60% of patients with type 2 diabetes mellitus will develop DR during the first 20 years from onset of diabetes [1]. With the rising prevalence of diabetes and increasing numbers of people with diabetes living longer, the number of people with DR and visual impairment due to this disease is rising worldwide [2]. Early diagnosis of diabetic patients and appropriate timely treatment has gradually become an effective measure to prevent DR disease, with a positive economic impact on patients and the healthcare system. Although DR is prevalent nowadays, its prevention remains challenging. Physicians typically diagnose the presence and severity of DR through visual assessment of the fundus images by direct evaluation. Given the large number of diabetes patients globally, this process is expensive and time consuming [3], [4]. Furthermore, 75% of worldwide DR patients live in underdeveloped areas, where sufficient specialists and adequate medical infrastructures for this purpose are unavailable [5]. Consequently, millions of persons continue to experience vision impairment without proper predictive diagnosis and eye care.

Screening tests are performed in asymptomatic persons to assess for the presence of a particular disease or the risk of that disease. An effective screening test program should reduce morbidity and mortality in a population by detecting disease at a stage at which treatment will make a difference. Global screening programs have been created to counter the proliferation of preventable eye diseases, but DR exists at large a scale and its detection on individual basis is scarcely effective. Additionally, given also the current cost-conscious era of healthcare, a policy to reduce unnecessary screening for several retinal diseases is becoming necessary [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente García-Díaz.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

The availability and the huge amount of Electronic Health Record (EHR) data is exponentially growing, thus early diagnosis of patients with diabetes using EHR data has gradually become an effective measure to prevent DR disease [7]. Using routine EHR data (i.e., demographics information, lab tests, exams, pathologies), it is possible to predict whether diabetic patients are going to develop DR. Comparing this approach with conventional gold standard DR diagnosis (e.g., fundus images), it does not only eliminate the time and money costs, but also maintains an acceptable accuracy [8]. Conventional diagnosis of DR requires a professional medical facility to obtain the fundus image by advanced medical equipment and then a professional physician to evaluate the single case of study. Early diagnosis of DR with convenient, easy-to-access, free, routine EHR data may result a simple and convenient alternative to manage and treat diabetic patients.

The multiple aim of this study was to i) predict the risk of developing DR as diabetic complication (i.e., presence/absence of DR) [i.e., task 1] and, subsequently only for DR patients, ii) temporally stratify the DR risk (i.e., in 0-2 years or in 2-5 years) [i.e., task 2] using EHR data.

The purpose of this work is to bridge the gap between the ML research applied to EHR data and the development of a Clinical Decision Support System (CDSS) while keeping humans at the center of the design and evaluation process [9]–[11]. It is worth noting that the task 1 predicts the presence/absence of DR, while the task 2, if the presence of DR was predicted by the task 1, predicts when the patients will develop DR (i.e., in 0-2 years or in 2-5 years). The proposed two-stage hierarchical ML procedure can be seen as a sequential predictive process within a Clinical Decision Support System (CDSS) application. In particular, main contributions to the biomedical informatics field can be summarized as follows: (i) the employment of the novel fully annotated/standardized 120k dataset from multiple diabetologic centers, thus leading to a higher clinical impact procedure, (ii) the design of a novel preprocessing procedure for selecting control and pathological patients (iii) the application of a two-stage ML procedure to firstly predict the presence/absence of DR and secondly to stratify the temporal risk of the disease for each patient, (iv) the effectiveness of the proposed experimental procedure for generalizing across different diabetologic centers.

The rest of the paper is organized as follows: Section II gives an overview of the state-of-the-art approaches for risk prediction and risk stratification of DR; Section III describes the 120K dataset and the preprocessing procedure; Section IV describes the experimental and validation procedure of the ML models comparison; Section V shows the predictive performance and features importance results; Section VI and Section VII discuss the experimental findings and conclude the paper.

II. RELATED WORK

The gold standard of DR diagnosis is represented by the assessment of fundus images. Thus, among all diabetes-related complications, the DR is the most studied field based on DL imaging techniques. Thus several DL models based on fundus images were designed to identify DR from a non-pathological condition [5], classify different DR stages [4], and predict future DR progression [1]. Diagnostic studies for DR based on EHR data have been still poorly explored, but something more has already been attempted in terms of screening and risk prediction. Targeted screening intervals based on EHR data analysis was adopted to detect DR and to reduce the burden of unnecessary screening examinations [12]. Increasing the interval between screening visits for DR beyond 1 year in low-risk patients is reasonable, since the data showed little difference between the 1-year and 2-year screening frequency with respect to clinical outcomes [13]; additionally, extending the time interval between screening visits to every 3 or 4 years on the basis of retinopathy status and glycated hemoglobin level might effectively decrease the rates of screening adherence in the population [6].

Individualized risk assessments were studied using both epidemiologic and clinical data, including the type and duration of diabetes, glycated hemoglobin or mean blood glucose levels, blood pressure, and the presence and grade of retinopathy [14]. Almost a 30-year period of diabetic patients' fundus images were analysed to simplify individualized risk assessments: an accurate assessment of the risk of proliferative DR or clinically significant macular edema was possible with the use of only the patient's current retinopathy status and glycated hemoglobin levels [12]. Moreover, also a recommended time until the next eye examination on the basis of these two factors was estimated.

Focusing on ML-based solutions using EHR data, in [8] the predictive performance of several ML models (i.e., Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB)) were compared with the aim to identify DR adopting features engineering techniques. However, differently from our proposed work, the temporal risk stratification task was not provided. Moreover, in [8] the classes of DR and control patients were perfectly balanced, and this aspect, beyond that favoring the predictive ML model, unfortunately never reflects the real clinical scenario. In [15] was presented a robust end-to-end ML-based SaaS framework, consisting of a ridge regularized survival SVM with a clinical kernel, coupled with Chi-square distance-based feature selection, to uncover relevant DR risk factors associated with disease outcomes by exploiting the weak correlations in EHRs. In our work we use neither features selection [15] nor features engineering [8] because these strategies require human effort and affects the standardized EHR data not ensuring reproducibility and scalability. In [16] a data-driven survival analysis approach was presented to predict when a patient will develop complications after the initial T2D diagnosis and to rank the associated risk. Moreover, to better capture the correlations of time-to-events of multiple complications, a further multi-task version of the survival model was developed. However, as in [8], [15], the

observational temporal window of interest (TWOI) for DR patients was selected considering only the DR diagnosis code, but not also the non-DR diagnosis code. Instead to the best of our knowledge, we did the first attempt to define a TWOI for DR patients based on a double-check of DR and non-DR codes. This conservative choice to avoid a misleading TWOI for DR patients is motivated by the presence of frequent typos, physician's transcription errors, and EHR framework faults or anomalies in real-world EHR scenario. On the contrary, differently from DL imaging techniques [1], [4], [5], all the studies based on EHR data [8], [15], [16] gave much more attention to model interpretability and pattern localization.

Previous breakthrough research findings rely on DL techniques to diagnose DR in patients with medical imaging. Although the medical imaging achieves reasonable recognition accuracy, the application of mass, easy-to-obtain and routine EHR data can make an early diagnosis of the DR more convenient and suitable.

III. MATERIALS

A. DATASET

The 120K dataset, provided by Regione Marche, was collected by aggregating patients of several Italian diabetologic centers. The dataset consists of 120K diabetic patients and was organized in the following 3 different fields:

- The *demographics field* stores the patient's identificative number (ID patient), gender, year of birth, and diabetes diagnosis date. In particular, the first name and the surname of the patients were anonymized and associated to a random numeric ID patient.
- The *pathological field* stores the ID patient, the pathology codes, and the pathology diagnosis date.
- The *lab tests field* stores ID patient, the lab tests codes, the lab tests values, and the lab tests prescription date.

B. PREPROCESSING

In accordance with the diabetologist, all the pathology codes associated with DR were identified and summarized in Table 1. The first pathology code (i.e., -3001) indicates a non-DR condition, while all the remaining codes indicate a DR condition.

All the pathology codes that were not included in Table 1 were removed from pathological field. Then, for each patient, both pathology codes and lab tests codes were removed if pathology diagnosis date and lab tests prescription date were earlier than diabetes diagnosis date. Finally, the inclusion criteria to select the time-widow of interest (TWOI) were presented for both control patients and retino patients as depicted in Figure 1.

1) CONTROL PATIENTS - TWOI

A control patient must have at least 2 pathology codes (i.e., -3001) of non-DR and none of the remaining pathology codes available in Table 1. A TWOI of a control patient (see Figure 1 - upper side) is delimited by the earliest pathology code of non-DR and the latest code of non-DR.

TABLE 1. Pathology codes associated to diabetic retinopathy (DR).

Code	Description
-3001	Non-DR
-3002	Non-proliferating DR
-3003	Preproliferating DR
-3004	Proliferating DR
-3005	Complicated proliferating DR
-3006	Diabetic maculopathy
-3007	Laser-treated DR
-3008	Ipertensive DR
-3009	Glaucoma
-3011	Blindness
-3013	Blindness - other causes
-3016	Pathological fluorangiography
-3201	Laser-treated non proliferating DR
-3202	Laser-treated proliferating DR
-3203	Advanced diabetic ophthalmopathy
-3204	Advanced diabetic ophthalmopathy - not treated
-3205	Blindness - from diabetes - monoocular
-3206	Blindness - from diabetes - biocular
-3207	Blindness - other causes - monoocular
-3208	Blindness - other causes - biocular
-3256	Laser-treated preproliferating DR

2) RETINO PATIENTS - TWOI

A control patient must have at least a pathology code (i.e., -3001) of non-DR and at least one of the remaining pathology codes available in Table 1. A TWOI of a retino patient (see Figure 1 - bottom side) is delimited by the earliest pathology code (i.e., -3001) of non-DR and the earliest pathology code of DR available in Table 1. A patient was included in the study only if the date of the earliest non-DR code (i.e., -3001) was before the earliest date of DR code.

C. TASKS DEFINITION

During the task definition stage, the matrices X_1 and X_2 fed to the ML model were defined both for task 1 and task 2, as well as the ground-truth vectors Y_1 and Y_2

1) TASK 1

The task 1, defined as the prediction between control and DR patients, was evaluated by taking the average of all the lab tests values enclosed in the range of the TWOI. The $X_1 = M_1 \times N_1$ matrix was obtained (see Figure 2), which was composed by $m_1 = 1, 2, \ldots, M_1$ patients and $n_1 = 1, 2, \ldots, N_1 - 3$ unique lab tests codes (i.e., predictors). In addition to the already existing predictors (i.e., unique lab tests codes), also the information of gender, age, and duration of diabetes was added, by obtaining the final $M_1 \times N_1$ matrix. All the missing values of X_1 matrix was filled with an extra-values imputation (i.e., -999). Several standard data imputation techniques (i.e., median, mean, KNN) were tested, but extra values imputation guaranteed the best predictive performance. This benefit can be explained by the fact that the extra value imputation allows to properly track the missing value mechanism, thus exploiting a correlation between the occurrences of missing values and the dependent variables.

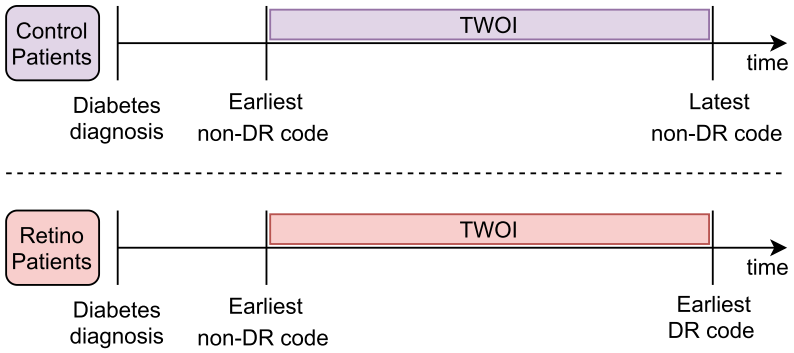


FIGURE 1. Preprocessing: Observational time window of interest (TWOI) for control and retino patients.

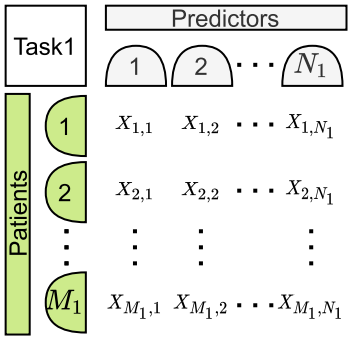


FIGURE 2. Task 1: $X_1 = M_1 \times N_1$ matrix, composed by $m_1 = 1, 2, \ldots, M_1$ patients and $n_1 = 1, 2, \ldots, N_1$ predictors.

TABLE 2. Task 1 statistics.

Description	Statistics
Total patients	40555
Control:	31611
Retino:	8944
Gender	
Male:	56%
Female:	44%
Age $(years)$	$68(\pm 12)$
Diabetes duration $(years)$	$12(\pm 8)$

Figure 3 shows the missing values distribution of the X_1 matrix. The ground-truth vector Y_1 of size $M_1 \times 1$ is composed of control patients, labelized as negative and retino patients, labelized as positive.

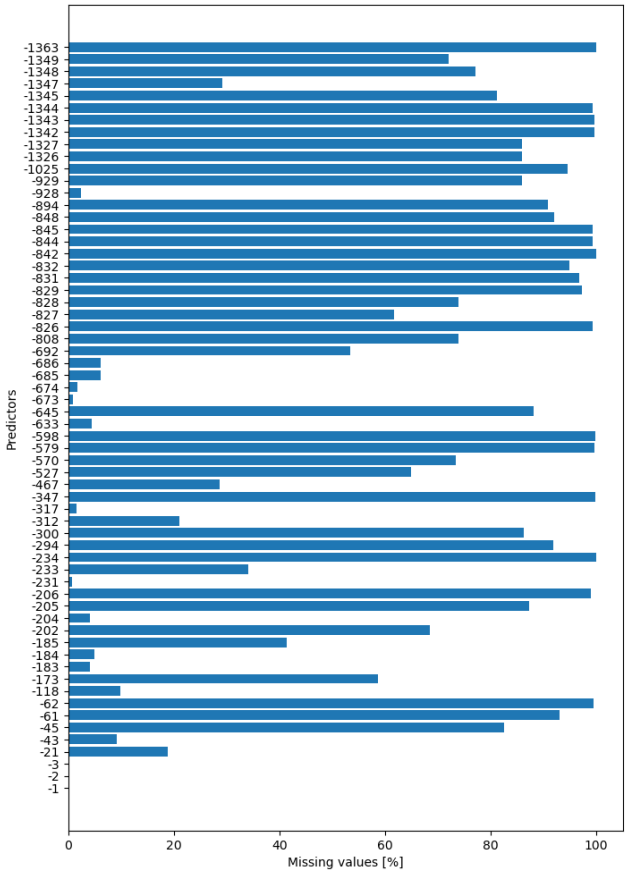


FIGURE 3. Missing values distribution of the predictors of X_1 matrix. The longer the bar is, the more missing values are present.

2) TASK 2

The task 2, defined as the temporal stratification of the DR risk, was evaluated only among the retino patients. For each patient, the unique lab tests prescription dates enclosed in the TWOI were pointed out. Each of those represented an observation of the patient. Thus, for each patient, starting from the earliest observation close to the lower boundary of the TWOI, the mobile averages of all the lab tests values inside the range of the dynamic time-windows were taken, observation by observation, until the latest observation close to the upper boundary of the TWOI. A $M_2 \times N_2$ matrix was obtained (see Figure 4), which was composed by $m_2 = 1, 2, \ldots, M_2$ total observations of all patients and $n_2 = 1, 2, \ldots, N_2 - 4$ unique lab tests codes. In addition to the already existing predictors (i.e., unique lab tests codes), also the information of gender, age, duration of diabetes, and incremental number of observations (seq) per patient was added, by obtaining the final $M_2 \times N_2$ matrix. All the missing values of X_2 matrix was filled with

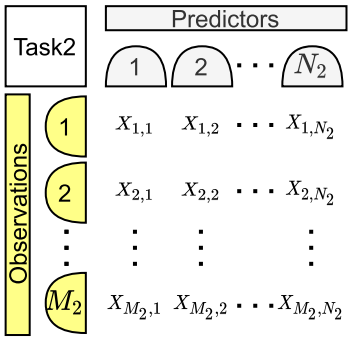


FIGURE 4. Task 2: $X_2 = M_2 \times N_2$ matrix, composed by $m_2 = 1, 2, \ldots, M_2$ observations and $n_2 = 1, 2, \ldots, N_2$ predictors.

TABLE 3. Task 2 statistics.

Description	Statistics
Retino patients	8944
n° of total observations in $0-5$ years	109976
n° of observations per patient in $0-5$ years	$12(\pm 8)$
Gender	
Male:	54%
Female:	46%
Age $(years)$	$68(\pm 11)$
Diabetes duration $(years)$	$13(\pm 9)$

an extra-values imputation (i.e., -999). The task 2 consisted in the prediction of the temporal distance between the date of each patient's observation and the date of DR diagnosis. The risk was defined "high" if the temporal distance is within the range of $0-2$ years, otherwise was defined mid-low if within the range of $2-5$ years. The ground-truth vector Y_2 of size $M_2 \times 1$ is composed of short-term risk patients and long-term risk patients. Retino patients whose temporal distance was greater than 5 years are excluded from the study.

IV. METHOD

A. EXPERIMENTAL PROCEDURE

To perform the task 1, a Tenfold Cross-Validation (CV-10) experimental procedure was chosen. CV-10 was implemented dividing all patients in ten folds, by selecting nine folds for training and one fold for testing. During the training stage of the CV-10 procedure, SMOTE [17] was utilized to equally balance DR patients with respect to control patients. CV-10 procedure was implemented without considering the temporal evolution of predictors, providing an overall average of the patient's clinical history.

On the contrary, to perform the task 2, a Tenfold Cross-Validation Over Patients (CVOP-10) was chosen. CVOP-10 was implemented dividing all observations grouped by patients in ten folds, by selecting nine folds for training and one fold for testing. CVOP-10 procedure was implemented considering the temporal evolution of the patient's predictors and allowed to generalize across unseen patients.

TABLE 4. Laboratory tests codes (i.e., predictors) of the 120K dataset used in task 1 and task 2 experiments. Each code of the 120K dataset can be univocally associated to each ICD-9 (9th revision of the International Statistical Classification of Diseases) code.

Code	Description	Uom
-1363	Post-prandial glycaemia	mg/dl
-1349	Albumin to creatinine ratio (ACR)	mg/mmol
-1348	Creatinine clearance	ml/min
-1347	LDL cholosterol (calc)	mg/dl
-1345	Creatininuria	mg/dl
-1344	HbA1c [lab.3]	%
-1343	HbA1c [lab.2]	%
-1342	HbA1c [lab.1]	%
-1327	Left Winsor index	Null
-1326	Right Winsor index	Null
-1025	ACR (calc)	mg/mmol
-929	Microalbuminuria	mg/24h
-928	BMI	Kg/m²
-894	Urine culture (1=neg 2=pos)	Null
-848	Potassium (uri)	mEq/l
-845	AER III	mcg/min
-844	AER II	mcg/min
-842	Microalbuminuria (II)	mg/l
-832	Pre-prandial glycaemia	mg/dl
-831	Pre-dinner glycaemia	mg/dl
-829	Glycaemia h 23	mg/dl
-828	Post-prandial glycaemia	mg/dl
-827	Post-breakfast glycaemia	mg/dl
-826	Post-dinner glycaemia	mg/dl
-808	Creatinine clearance (calc)	ml/min
-692	Urine ketones	mg/dl
-686	Diastolic pressure	mmHg
-685	Systolic pressure	mmHg
-674	Height	cm
-673	Weight	kg
-645	Urea	mg/dl
-633	12-hour fasting triglycerides	mg/dl
-598	Sodium (uri)	mEq/L
-579	Albuminuria/creatinuria ratio	Null
-570	Proteines (uri)	mg/dl
-527	Blood plates	1000/mm3
-467	Microalbuminuria	mg/l
-347	Glicosuria	G/L
-317	Fasting glycaemia	mg/dl
-312	Gamma-glutamyl transferase (GGT)	UI/L
-300	Alkaline Phosphatase	UI/L
-294	Fibrinogen (sie)	mg/dl
-234	A1 hemoglobin (tot)	Null
-233	Hemoglobin	g/dl
-231	Glycated hemoglobin (HbA1c)	%
-206	Creatinine clearance (uri) F	ml/min
-205	Creatinine clearance (uri) M	ml/min
-204	Creatinine	mg/dl
-202	Creatine phosphokinase (sie)	UI/L
-185	LDL cholesterol	mg/dl
-184	HDL cholesterol	mg/dl
-183	Cholesterol (tot)	mg/dl
-173	Weist	cm
-118	Serum glutamic-oxaloacetic transaminase (SGOT)	UI/L
-62	Amylase (uri)	U/L
-61	Amylase	UI/L
-45	Albumin excretion rate (AER)	mcg/min
-43	Alanine aminotransferase test (GPT)	UI/L
-21	Uric acid	mg/dl
-3	Gender	Null
-2	Age	years
-1	Diabates duration	years
0	Seq	Null

B. METRICS

The proposed task 1 and task 2 were evaluated by considering the following metrics for the classification task:

- *Accuracy*: the percentage of correct predictions;
- *Macro-precision* (Precision): the *Precision* is calculated for each class and then the unweighted mean is taken;
- *Macro-recall* (Recall): the *Recall* is calculated for each class and then the unweighted mean is taken;
- *Macro-F1* (F1): the harmonic mean of precision and recall averaged over all output categories;
- *Area Under the receiver operating characteristic Curve* (AUC): the AUC represents the probability that the classifier will rank a randomly chosen positive sample higher than a randomly chosen negative one;
- *Area Under the Precision-Recall Curve* (PRAUC): The PRAUC can be interpreted as the relationship between precision and recall (sensitivity) and is considered more informative than the AUC plot when evaluating binary classifiers on imbalanced data [18].

C. VALIDATION PROCEDURE

For what concern the CV-10 and CVOP-10 experimental procedures, the optimization of the hyperparameters of the ML models was performed implementing a grid-search and optimizing the *Recall* in a nested Fivefold Cross-Validation. *Recall* was preferred over other optimization objectives, because the minimization of false negatives has the most clinical relevance for the task 1 experiment. Hence, each split of the outer loop was trained with the optimal hyperparameters tuned in the inner loop. Although this procedure was computationally expensive, it allowed to obtain an unbiased and robust performance evaluation [19]. In according with the work in [8], the predictive performances of several ML models such as XGBoosting (XGB), LR, DT, RF, SVM, and NB were compared. Table 5 summarizes the range of the hyperparameters optimized during validation stage for each ML model. The features importance of the XGB model was extracted in according to the logic of showing the number of times the feature is used to split data, while for the RF model in according to the logic of averaging the decrease in impurity over trees. We decided to not explore model-agnostic methods for showing the importance of each feature, because we aimed to emphasize the model intrinsic dependency. For that reason we show a global feature importance that is intrinsic within the designed ensemble-based white box models. Future work could be explored in order to provide further interpretability of the proposed approach by exploiting post-hoc explainable AI methodology, specifically tailored to clinician point of view. This methodology includes the possibility (i) to provide local feature importance (SHAP [20] and LIME [21]), (ii) to unraveled rules and laws (features interaction) and (iii) to provide transparent risk equations (model non-linearity) [22].

TABLE 5. Range of hyperparameters (Hyps) for each machine learning model: XGBoosting (XGB), logistic regression (LR), decision tree (DT), random forest (RF), support vector machine (SVM), and naive Bayes (NB).

Model	Hyps	Range
XGB	n° of estimators	$\{100, 150, 200\}$
	max n° of splits	$\{25, 50, 75, 100\}$
	learning rate	$\{10^{-2}, 0.1, 1\}$
LR	regularization coefficient	$\{0.01, 0.1, 1, 10\}$
	l1 ratio	$\{0, 0.25, 0.50, 0.75, 1\}$
DT	max n° of splits	$\{25, 50, 75, 100\}$
RF	max n° of splits	$\{5, 10, 15, 20, 25\}$
	n° of estimators	$\{50, 100, 150, 200, 250\}$
SVM	box constraint	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$
NB	variance smoothing	$\{10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}\}$

V. EXPERIMENTAL RESULTS

Table 6 shows the predictive performance experimental results and it is evident as RF and XGB have proved to be best models for both task 1 and task 2. In accordance also with [8], RF1 is the best model for task 1 (Recall: 73.91, AUC: 86.96, and PRAUC: 72.43); while for task 2, XGB2 obtained the best predictive performance in terms of Recall (75.66) and RF2 in terms of AUC (86.76) and PRAUC (84.38).

Figure 5 shows the top-10 discriminant predictors for task 1, while Figure 6 for task 2. Only the two best models (i.e., XGB and RF) were considered in this evaluation.

VI. DISCUSSION

A. PREDICTIVE PERFORMANCE

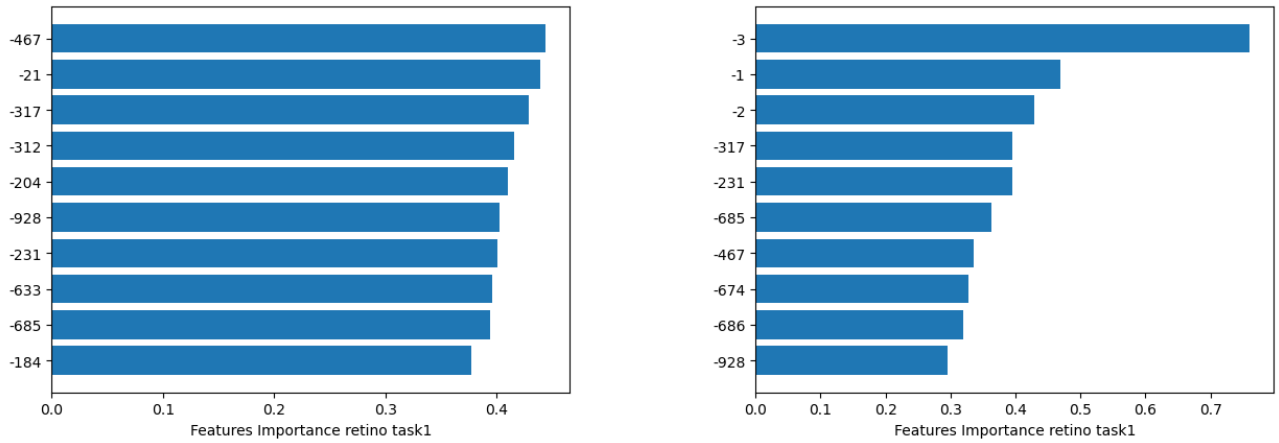
Globally, XGB and RF models obtained the best predictive performance to solve task 1 and task 2. Tree-based models (i.e., DT, RF, XGB) obtained the best predictive performance in both tasks and considerably overcome the other ML models. However, the simple DT1 did not guarantee robustness against class imbalance, because PRAUC (57.30) was significantly inferior than RF1 (72.43) and XGB1 (71.26). Thus, only the more complex tree-based models (i.e., RF, XGB) achieved the best result to solve task1 and task 2. Differently from [8] we designed the experimental setup using raw EHR data without employing features engineering techniques. This aspect assumes a great relevance in terms of data interpretability and algorithm scalability. Thus, the possibility of extracting raw features from EHRs permits the clinician to appreciate and interpret each single features contribution, and moreover, this scenario could be more easily scalable and transferable to other standardized clinician domains avoiding an hand-crafted human intervention.

B. CLINICAL SIGNIFICANCE

The contribution of the single predictors changes every time in relation to different tasks and different ML models (i.e., inter-task and inter-model variability). The common intersection regards the strong correlation between features importance and its associated missing value rate (mvr). The higher the quality and the completeness of the data, the more

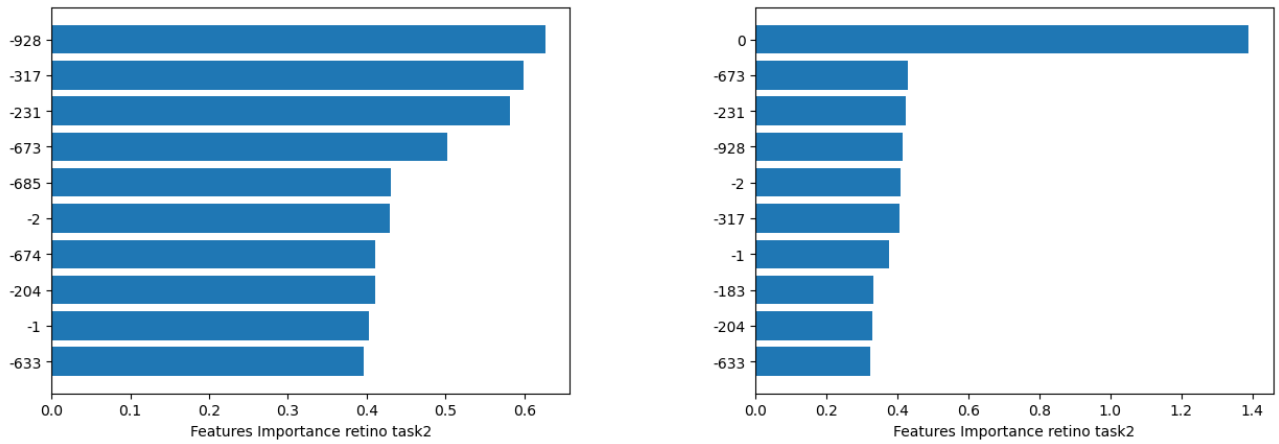
TABLE 6. Experimental results (i.e., task1 and task2) for each machine learning model: XGBoosting (XGB), logistic regression (LR), decision tree (DT), random forest (RF), support vector machine (SVM), and naive Bayes (NB). Predictive performance and standard deviation are reported for each metric.

	Accuracy	F1	Precision	Recall	AUC	PRAUC
XGB1	85.07 ± 0.33	75.56 ± 0.63	80.10 ± 0.61	73.02 ± 0.66	85.69 ± 0.50	71.26 ± 0.88
LR1	57.30 ± 1.38	54.60 ± 1.13	59.40 ± 0.67	63.57 ± 0.99	67.30 ± 1.72	32.45 ± 2.47
DT1	76.66 ± 0.70	68.50 ± 0.79	67.53 ± 0.78	70.17 ± 0.86	70.10 ± 0.87	57.30 ± 1.24
RF1	84.77 ± 0.41	75.86 ± 0.61	78.86 ± 0.77	73.91 ± 0.56	86.96 ± 0.56	72.43 ± 0.98
SVM1	58.64 ± 0.58	56.09 ± 0.46	60.84 ± 0.37	65.72 ± 0.51	71.28 ± 0.51	39.33 ± 1.05
NB1	44.38 ± 2.79	44.14 ± 2.49	58.10 ± 0.75	59.50 ± 1.43	65.53 ± 1.16	33.00 ± 1.19
XGB2	76.71 ± 0.78	75.93 ± 0.85	76.58 ± 0.79	75.66 ± 0.86	86.08 ± 0.69	83.87 ± 0.76
LR2	61.17 ± 0.97	56.61 ± 1.22	60.56 ± 1.22	57.90 ± 1.00	66.09 ± 0.93	57.03 ± 1.48
DT2	71.30 ± 0.92	70.70 ± 0.97	70.75 ± 0.95	70.60 ± 0.98	70.60 ± 0.97	73.73 ± 1.08
RF2	76.76 ± 1.08	75.60 ± 1.20	77.25 ± 1.09	75.15 ± 1.18	86.76 ± 0.88	84.38 ± 1.12
SVM2	62.38 ± 0.75	58.11 ± 1.08	62.08 ± 1.01	59.63 ± 0.89	69.95 ± 0.77	59.20 ± 1.33
NB2	$58..10 \pm 1.75$	57.76 ± 1.88	58.60 ± 1.44	58.50 ± 1.51	61.96 ± 1.45	52.34 ± 1.59



(a) XGB1 (b) RF1

FIGURE 5. Features importance for task 1: XGB1 and RF1. The legend of the features is reported in Table 4.



(a) XGB2 (b) RF2

FIGURE 6. Features importance for task 2: XGB2 and RF2. The legend of the features is reported in Table 4.

important the predictor will assume. This suggests how a high-quality data collection and representation always positively affects every AI-based solution.

Starting from task 1, the gender, diabetes duration, and age (mvr $= 0\%$) represented the most important predictors for RF1, while microalbuminuria (mvr $\approx 30\%$), uric acid

(mvr ≈ 20%), and fasting glycaemia (mvr ≈ 0%) for XGB1. The common most important predictors selected by both XGB1 and RF1 are the following: microalbuminuria (mvr ≈ 30%), fasting glycaemia (mvr ≈ 0%), BMI (mvr ≈ 0%), glycated hemoglobin (mvr ≈ 0%), and systolic pressure (mvr ≈ 5%).

Moving to task 2, the sequential number of observations per patient (mvr = 0%), weight (mvr ≈ 0%), and glycated hemoglobin represented the most important predictors for RF2, while BMI, fasting glycaemia, and glycated hemoglobin for XGB2. The temporal information as the sequential number of observations per patient appears predominant in RF2, while the 3 most important predictors in XGB2 were appeared in top-10 predictors in XGB1. The XGB model tends to give importance to the same cluster of predictors, on the contrary the RF model was able to capture the very important temporal information represented by sequential number of observations per patient to perform the task 2.

Both ML models (i.e., RF and XGB), adopting the proposed novel preprocessing procedure for selecting control and pathological patients, were effective capable to generalize across different diabetologic centers, a necessary requirement to transfer the AI-based solution to other clinical scenarios.

The proposed AI-based solution represents the core of a CDSS. In fact, Meteda srl, leading company in Italy for innovative software solutions for diabetes, has already efficiently integrated the proposed ML-based CDSS into the EHR architecture of some diabetic centers for a pilot study. The first release of the predictive medicine tool is focused on DR and can be addressed to primary care levels or specialists for screening and follow-up purposes. Clinicians are currently adopting the proposed ML-based CDSS trained on 120K dataset on new unseen patients from other different diabetic centers. The preliminary outcomes are proving that our proposed solution is generalizable also across heterogeneous clinical scenarios.

C. FUTURE WORK

Future work may be oriented to extend the operating range and provide to predict also other diabetic complications (i.e., cardiopathy, nefropathy, neuropathy, and vasculopathy). In this direction, a full-version of the proposed CDSS will be released to the whole clinical ecosystem. Another relevant aspect to focus on could be enhancing the quality of EHR data collection in clinical scenario; thus, the proposed ML-based CDSS could help each diabetic centers to reach baseline standards in the collection, completeness, and quality of the data. Diabetic centers located below a certain quality threshold could be alerted by the ML-based CDSS to bridge the gap. Moreover, diabetic centers should be pushed to achieve targeted data quality indicators (e.g., mvr, prescription of a particular lab test, repetition of a particular lab test, etc.), both to facilitate the predictive system and to improve the diabetic patient's quality of life. Future work may be addressed to measure the generalizability of the model across different healthcare infrastructure and lifestyle by collecting and including diabetic patients from different geographical areas.

VII. CONCLUSION

This work proposed a two-stage ML procedure as the core of a CDSS to firstly predict the presence/absence of DR and secondly to temporally stratify the risk of the disease for each patient. For this objective, a novel preprocessing procedure was designed to select both control and pathological patients, and moreover, the novel fully annotated/standardized 120K dataset from multiple diabetologic centers was provided. The proposed AI-based solution was proven to be capable of generalizing across different diabetologic centers and was utilised as pilot study in several diabetologic centers for DR screening and follow-up purposes.

ACKNOWLEDGMENT

The authors would like to give special thanks to Dr. Giacomo Vespasiani and Prof. Antonio Nicolucci.

REFERENCES

[1] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, and M. Prunotto, "Deep learning algorithm predicts diabetic retinopathy progression in individual patients," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–9, Dec. 2019.

[2] J. W. Y. Yau, S. L. Rogers, R. Kawasaki, E. L. Lamoureux, J. W. Kowalski, T. Bek, S.-J. Chen, J. M. Dekker, A. Fletcher, J. Grauslund, and S. Haffner, "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, vol. 35, no. 3, pp. 556–564, 2012.

[3] M. N. Ozieh, K. G. Bishu, C. E. Dismuke, and L. E. Egede, "Trends in health care expenditure in US adults with diabetes: 2002–2011," *Diabetes Care*, vol. 38, no. 10, pp. 1844–1851, Oct. 2015.

[4] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. Ahmed Khan, and W. Jadoon, "A deep learning ensemble approach for diabetic retinopathy detection," *IEEE Access*, vol. 7, pp. 150530–150539, 2019.

[5] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[6] J. B. Rosenberg and I. Tsui, "Screening for diabetic retinopathy," *New England J. Med.*, vol. 376, no. 16, pp. 1587–1588, 2017.

[7] J. Benbassat and B. C. P. Polak, "Reliability of screening methods for diabetic retinopathy," *Diabetic Med.*, vol. 26, no. 8, pp. 783–790, 2009.

[8] Y. Sun and D. Zhang, "Diagnosis and analysis of diabetic retinopathy based on electronic health records," *IEEE Access*, vol. 7, pp. 86115–86120, 2019.

[9] P. Watkinson, D. Clifton, G. Collins, P. McCulloch, and L. Morgan, "DECIDE-AI: New reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence," *Nature Med.*, vol. 27, no. 2, pp. 186–187, Feb. 2021.

[10] 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 J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 235–246, Jan. 2020.

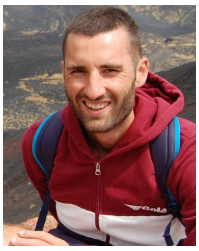
[11] M. Bernardini, L. Romeo, E. Frontoni, and M.-R. Amini, "A semi-supervised multi-task learning approach for predicting short-term kidney disease evolution," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3983–3994, Oct. 2021.

[12] D. M. Nathan, I. Bebu, D. Hainsworth, R. Klein, W. Tamborlane, G. Lorenzi, R. Gubitosi-Klug, and J. M. Lachin, "Frequency of evidence-based screening for retinopathy in type 1 diabetes," *New England J. Med.*, vol. 376, no. 16, pp. 1507–1516, Apr. 2017.

[13] S. Taylor-Phillips, H. Mistry, R. Leslie, D. Todkill, A. Tsertsvadze, M. Connock, and A. Clarke, "Extending the diabetic retinopathy screening interval beyond 1 year: Systematic review," *Brit. J. Ophthalmol.*, vol. 100, no. 1, pp. 105–114, 2016.

[14] T. Aspelund, Ó. Þórisdóttir, E. Ólafsdottir, A. Gudmundsdottir, A. B. Einarsdóttir, J. Mehlsen, S. Einarsson, Ó. Pálsson, G. Einarsson, T. Bek, and E. Stefánsson, "Individual risk assessment and information technology to optimise screening frequency for diabetic retinopathy," *Diabetologia*, vol. 54, no. 10, pp. 2525–2532, Oct. 2011.
[15] P. Chakraborty and F. Farooq, "A robust framework for accelerated outcome-driven risk factor identification from EHR," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 1800–1808.
[16] B. Liu, Y. Li, Z. Sun, S. Ghosh, and K. Ng, "Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 1–14, 2018.
[17] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.
[18] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0118432.
[19] 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," *Comput. Biol. Med.*, vol. 112, Sep. 2019, Art. no. 103358.
[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
[21] M. Tulio Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," 2016, *arXiv:1606.05386*.
[22] A. M. Alaa and M. van der Schaar, "Demystifying black-box models with symbolic metamodels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 11304–11314.

MICHELE BERNARDINI received the M.Sc. degree in electronic engineering from Università Politecnica delle Marche (UNIVPM). He is currently a Postdoctoral Researcher at the Department of Information Engineering (DII), UNIVPM. His main research interest includes machine learning applied to predictive medicine scenarios using electronic health records data.



LUCA ROMEO received the Ph.D. degree in computer science from the Department of Information Engineering (DII), Università Politecnica delle Marche (UNIVPM), in 2018. His Ph.D. thesis was on "Applied Machine Learning for Human Motion Analysis and Affective Computing." He is currently a Postdoctoral Researcher with the DII, UNIVPM, and is also affiliated with the Unit of Cognition, Motion and Neuroscience and Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genova. His research interests include machine learning applied to biomedical applications, affective computing, and motion analysis.

ADRIANO MANCINI received the Ph.D. degree in intelligent artificial systems from the Department of Information Engineering (DII), Universitá Politecnica delle Marche, in 2010. He currently holds an Assistant Professor position with the DII, Universitá Politecnica delle Marche. His research interests include mobile robotics also for assisted living, machine learning, image processing, and geographic information systems. He is a coauthor of more than 120 international papers in his research fields and is involved in different EU projects and technological transfer projects.

EMANUELE FRONTONI (Member, IEEE) was born in Fermo, Italy, in 1978. He is a Professor of computer vision and deep learning with the Department of Information Engineering (DII), Universitá Politecnica delle Marche (UNIVPM). His main research interests include artificial intelligence, computer vision, human behavior analysis, augmented reality, and sensitive spaces. He is a coauthor of more than 150 international papers in his research fields and is involved in different EU projects and technological transfer projects with national and international companies.