Machine Learning Predicts Cardiovascular Events in Patients With Diabetes: The Silesia Diabetes-Heart Project

Katarzyna Nabrdalik, MD, PhD[a,b*], Hanna Kwiendacz, MD, PhD[a], Karolina Drożdż, MD[a], Krzysztof Irlik[c], Mirela Hendel[c], Agata M. Wijata, PhD[d], Jakub Nalepa, PhD, DSc[e], Elon Correa, PhD[b], Weronika Hajzler[f], Oliwia Janota, MD[g], Wiktoria Wójcik[c], Janusz Gumprecht, MD, PhD[a], and Gregory Y.H. Lip, MD, PhD[b,h]

From the [a] Department of Internal Medicine, Diabetology and Nephrology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland, [b] Liverpool Centre for Cardiovascular Science at University of Liverpool, Liverpool John Moores University and Liverpool Heart and Chest Hospital, Liverpool, UK, [c] Students' Scientific Association by the Department of Internal Medicine, Diabetology and Nephrology in Zabrze, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland, [d] Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland, [e] Faculty of Automatic Control, Electronics and Computer Science, Department of Algorithmics and Software, Silesian University of Technology, Gliwice, Poland, [f] Doctoral School, Department of Pediatric Hematology and Oncology in Zabrze, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland, [g] Doctoral School, Department of Internal Medicine, Diabetology and Nephrology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland and [h] Department of Clinical Medicine, Aalborg University, Aalborg, Denmark.

The work is a part of Statutory Work of Medical University of Silesia (KNW-1-007/N/8/K and KNW-1-175-K/9/K). AMW and JN were supported by the Silesian University of Technology funds through the grant for maintaining and developing research potential. JN was also supported by the Silesian University of Technology Rector's grant (02/080/RGJ22/0026).
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
*Corresponding author: Katarzyna Nabrdalik, Department of Internal Medicine, Diabetology and Nephrology, Faculty of Medical, Sciences in Zabrze, Medical University of Silesia, 3 Maja St 13-15, Katowice, 41-800, Poland. E-mail: knabrdalik@sum.edu.pl

Curr Probl Cardiol 2023;48:101694
0146-2806/$ – see front matter
https://doi.org/10.1016/j.cpcardiol.2023.101694

Abstract: We aimed to develop a machine learning (ML) model for predicting cardiovascular (CV) events in patients with diabetes (DM). This was a prospective, observational study where clinical data of patients with diabetes hospitalized in the diabetology center in Poland (years 2015-2020) were analyzed using ML. The occurrence of new CV events following discharge was collected in the follow-up time for up to 5 years and 9 months. An end-to-end ML technique which exploits the neighborhood component analysis for elaborating discriminative predictors, followed by a hybrid sampling/boosting classification algorithm, multiple logistic regression (MLR), or unsupervised hierarchical clustering was proposed. In 1735 patients with diabetes (53% female), there were 150 (8.65%) ones with a new CV event in the follow-up. Twelve most discriminative patients' parameters included coronary artery disease, heart failure, peripheral artery disease, stroke, diabetic foot disease, chronic kidney disease, eosinophil count, serum potassium level, and being treated with clopidogrel, heparin, proton pump inhibitor, and loop diuretic. Utilizing those variables resulted in the area under the receiver operating characteristic curve (AUC) ranging from 0.62 (95% Confidence Interval [CI] 0.56-0.68, $P < 0.01$) to 0.72 (95% CI 0.66-0.77, $P < 0.01$) across 5 nonoverlapping test folds, whereas MLR correctly determined 111/150 (74.00%) high-risk patients, and 989/1585 (62.40%) low-risk patients, resulting in 1100/1735 (63.40%) correctly classified patients (AUC: 0.72, 95% CI 0.66-0.77). ML algorithms can identify patients with diabetes at a high risk of new CV events based on a small number of interpretable and easy-to-obtain patients' parameters. (Curr Probl Cardiol 2023;48:101694.)

Introduction

Although there is a huge advancement in medical treatment options of cardiovascular (CV) events including antiplatelet agents and statins, adverse CV events still remain a significant threat to patients with diabetes.[1] Cardiovascular disease (CVD) is

associated with high mortality among patients with type 2 diabetes (T2DM) where it may account for more than half of deaths due to T2DM and is also the leading cause of mortality in type 1 diabetes (T1DM).[2] Developed risk models apply for the general population and, separately, for people with diabetes,[3,4] but these models often do not generalize well when applied to other populations.[5] Indeed, CVD risk prediction scores based on traditional risk factors could not identify individuals who experienced a CV event in 10 years of follow-up among patients with T2DM, whereas all 22 evaluated models had a comparable and modest discriminative ability.[6]

Due to high mortality related to CVD among patients with diabetes and suboptimal prediction of risk with the traditional clinical risk prediction scales, there is a continuing need for implementing new techniques, including data-driven machine learning (ML) approaches, for predicting the risk of CVD. Identifying patients with a high risk of developing new CV events is important for treatment intensification and personalization what can result in minimizing the risk and improved patients' survival. Therefore, it is of pivotal importance to build CVD risk stratification tools which could be exploited in day-to-day clinical practice.

Preventive CV medicine increasingly implements modelling techniques to estimate the individual's absolute risk of a CV event. Modern ML techniques extract useful patterns from large datasets to answer clinical questions and have demonstrated significant promise for risk stratification across various populations.[7,8] ML can help in identifying predictors and relationships between them that may not be identified by traditional models,[9] thus new risk factors may emerge. In our recent work, we demonstrated that a ML approach can accurately identify metabolic-associated fatty liver disease (MAFLD) patients with prevalent CVD based on the easy-to-obtain patient parameters.[10]

In relation to risk prediction, administrative and survey data can be used to develop a tool for identifying the incidence of six chronic diseases (ie, congestive heart failure, diabetes, obstructive pulmonary disease, lung cancer, myocardial infraction, and stroke).[11] However, few studies were devoted to the development of risk predictors specific for a diabetes population.[12] Cho et al. introduced a model using different ML algorithms based on medical data to identify the onset of the diabetic nephropathy.[13] Unsupervised ML clustering techniques have been, on the other hand, exploited to understand the patients' profiles which may be used to identify the unique characteristics of the T2DM population.[14]

To the best of our knowledge, there has been no study performed which utilized ML approaches to predict CV events in hospitalized patients with diabetes based on clinical, laboratory, and demographical parameters. In this study, we address this research gap and aimed to introduce a ML processing chain for this task.

Research Design and Methods

Study Design and Participants

The Silesia Diabetes-Heart Project is a single center, observational, prospective cohort study performed in patients with diabetes, hospitalized in the diabetology ward in Zabrze, Poland (January 2015-September 2020). The patients were followed for up to 5 years and 9 months in order to collect patients' CV status following discharge. The study has been registered on ClinicalTrials.gov (NCT05626413).

The eligibility criteria were as follows: patients with T1DM or T2DM We excluded patients with a terminal stage of cancer and those who died during hospital stay. Every patient signed a suitable informed consent for in-hospital treatment on admission, and no additional consent related to this analysis was necessary since only anonymized registry data has been analyzed. The study protocol has been approved by the Medical University of Silesia Ethics Committee (PCN/0022/KB/126/20) and the need for informed consent was waived by this Ethics Committee. All methods were performed in accordance with relevant regulations and the study was conducted in accordance with the Declaration of Helsinki.

Anthropometric parameters, including height and weight were measured at discharge by standard methods, and the body mass index (BMI) was calculated as weight/height^2. All blood pressure measurements during hospital stay were recorded and the mean blood pressure of all the measurements was calculated. Arterial hypertension was defined as a systolic blood pressure $\geq$140 mm Hg and/or a diastolic blood pressure $\geq$90 mm Hg or previous treatment with antihypertensive medications. The obesity was diagnosed when BMI $\geq$30 whereas the overweight was diagnosed when BMI $\geq$25 but $<$30. T2DM has been diagnosed based on a known history of this disease.

Hypercholesterolemia was recognized when a patient had this diagnosis present in the documented medical history and/or there was newly recognized plasma total cholesterol $\geq$3.8 mmoL/L ($\geq$150 mg/

dL) and/or patient was on statin therapy. Low-density lipoprotein was not measured routinely during hospital stay that is why it could not be considered as a diagnostic parameter of hypercholesterolemia. Hypertriglyceridemia was recognized when a patient had this diagnosis present in the documented medical history and/or there was newly recognized plasma triglyceride $\geq$1.7 mmoL/L and/or patient was on fibrate therapy. Chronic kidney disease (CKD) was recognized when a patient had this diagnosis present in the documented medical history (defined as persistently reduced estimated glomerular filtration rate (eGFR) $<$60 mL/min per 1.73 m^2, or persistently elevated urine albumin excretion (UAE) $\geq$30 mg/g, or both, for more than 3 months).

MAFLD was diagnosed if there was evidence of steatosis acquired by the hepatic ultrasonography. Ultrasonography examination was performed with the use of the ARIETTA 750 ultrasound system (Hitachi) equipped with a C253 transducer.

Diabetic foot disease (DFD) was defined based on the presence of infection, ulceration, or destruction of tissues of the foot. Diabetic peripheral neuropathy diagnosis was based on feet examination and was defined based on the presence of symptoms of nerve dysfunction manifested by inability to sense vibration, temperature, or touch. Diabetic retinopathy was defined based on fundus examination during the hospital stay or medically documented history of diabetic retinopathy.

Hyperuricemia was considered when a patient had medically documented history of hyperuricemia and/or was treated with xanthine oxidase inhibitor and/or uric acid concentration exceeded that of 6 mg/dL (360 μmoL/L) in women and 7 mg/dL (420 μmoL/L) in men.

Heart failure (HF) was diagnosed if there was a medically documented history of heart failure diagnosis or a new onset heart failure was diagnosed based on signs, symptoms and structural or functional impairment of the heart assessed with echocardiography during the hospital stay. Echocardiography was performed with the use of the ARIETTA 750 ultrasound system (Hitachi) equipped with a S121 transducer. The presence of CVD was defined as one or more of the following: angiography-confirmed coronary artery disease (CAD); myocardial infarction; coronary bypass grafting; stroke; carotid stenosis of at least 50% in diameter; and/or angiography-confirmed, clinically significant, lower extremities artery stenosis (peripheral artery disease).

Follow-up was performed between March 2021 and September 2021 through a phone contact with the patient or patient's family member to

verify whether any new CV event occurred since the hospital discharge (the shortest period between the hospital discharge and the phone contact was 6 months and the longest was 5 years and 9 months). A new CV event was defined as a new occurrence of nonfatal myocardial infarction, nonfatal stroke, hospitalization for unstable angina, heart failure, atrial fibrillation, and death due to a CV reason.

Biochemical Methods

Blood samples were placed in the lithium heparin or ethylenediaminetetraacetic acid tubes and urine samples were collected and analyzed on the day of hospital admission. The details regarding biochemical methods are discussed in the supplementary material.

Data Preprocessing and Multifold Cross-Validation

The primary end point of interest for this study was the prediction of the occurrence of a new CV event during the follow-up time. Overall, 81 patients' parameters collected at baseline were considered as predictor variables (Table 1). Each predictor was independently standardized (z-scored) before training a classifier, and the missing values were imputed using the factorial analysis.[15] To quantify the generalization of the ML classifiers, we followed a multifold cross-validation procedure, in which a model is trained and tested multiple times.

Predicting new Cardiovascular Events Using ML

To predict whether a patient with diabetes is likely to develop a new CV event, we investigated demographic (2 parameters), clinical (diabetes-related: 3, CV-related: 10, diabetic complications: 3, general: 3, concomitant diseases: 6), laboratory (30), and pharmacotherapy-related (24) parameters (81 parameters in total) (Table 1). The most discriminative predictors were selected using the neighborhood component analysis (NCA).[16] Due to a high imbalance between the patients without and with a new CV event (imbalance ratio of 10.6), we utilized a hybrid sampling/boosting algorithm (RUSBoost) for learning from skewed data, with the hyperparameter values as suggested in.[17] Additionally, we fitted the multiple logistic regression (MLR) model over the most discriminative predictors using the full dataset (without splitting it into folds). We investigate sensitivity and specificity of the classifier, and percentage of correctly classified (CC) patients with and without a CV event. Receiver operating characteristic (ROC) curves and area under them (AUC) were

TABLE 1. Patient parameters

Parameter	Patients without event (n = 1585)	Patients with event (n = 150)	*P*-value
	Demographic parameters		
Age [years]	58.38 ± 17.67 (61.00)	66.37 ± 11.27 (66.50)	<0.001
Men, n (%)	732 (46.18%)	84 (56.00%)	0.021
	Clinical parameters		
	Diabetes-related		
BMI [kg/m^2]	30.33 ± 7.16 (30.08)	31.21 ± 7.01 (31.07)	0.201
Duration of diabetes [years]	11.04 ± 9.14 (10.00)	12.85 ± 8.74 (10.00)	0.011
Type of diabetes [% of type 1]	362 (22.84%)	8 (5.33%)	<0.001
	Cardiovascular-related		
Atrial fibrillation	144 (9.35%)	15 (10.34%)	0.711
Carotid arteries stenosis	24 (1.51%)	5 (3.33%)	0.970
Coronary artery disease	**525 (33.12%)**	**88 (58.67%)**	**0.001**
Heart failure	**273 (17.22%)**	**44 (29.33%)**	**<0.001**
Hypertension	1139 (71.91%)	127 (84.67%)	0.001
Mean diastolic blood pressure [mm Hg]	76.36 ± 7.48 (77.00)	75.72 ± 7.67 (76.00)	0.367
Mean heart rate [bpm]	80.43 ± 15.36 (80.00)	79.00 ± 13.08 (80.00)	0.591
Mean systolic blood pressure [mm Hg]	127.94 ± 14.91 (127.00)	130.16 ± 15.19 (130.00)	0.065
Peripheral artery disease	**66 (4.176%)**	**15 (10.00%)**	**0.001**
Stroke	**123 (7.76%)**	**21 (14.00%)**	**0.008**
	Diabetic complications		
Diabetic foot disease	**40 (2.53%)**	**10 (6.67%)**	**0.004**
Diabetic peripheral neuropathy	142 (8.96%)	10 (6.67%)	0.343
Retinopathy	572 (36.09%)	56 (37.33%)	0.762
	General		
Current smoker [% of yes]	289 (18.23%)	29 (19.33%)	0.739
Emergency admission [% of yes]	420 (26.52%)	45 (30.00%)	0.355
Number of days of hospital stay	7.27 ± 2.80 (7.00)	7.30 ± 2.87 (7.00)	0.907
	Concomitant diseases		
Chronic kidney disease	**275 (17.41%)**	**47 (31.33%)**	**<0.001**
	597 (37.69%)	58 (38.67%)	0.809

(*continued*)

TABLE 1. (*continued*)

Parameter	Patients without event (n = 1585)	Patients with event (n = 150)	*P*-value
Degenerative disease of the spine			
Hypercholesterolemia	1009 (63.66%)	104 (69.33%)	0.166
Hypertriglyceridemia	592 (37.35%)	59 (39.33%)	0.632
Hyperuricemia	445 (28.08%)	59 (39.33%)	0.004
MAFLD	894 (56.40%)	97 (64.67%)	0.051
Laboratory parameters			
Alanine aminotransaminase [U/L]	33.86 ± 68.52 (22.80)	32.31 ± 37.93 (20.15)	0.104
Aspartate aminotransaminase [U/L]	32.83 ± 88.07 (22.10)	33.72 ± 41.54 (21.70)	0.253
Basophil count [10^9/L]	0.04 ± 0.06 (0.03)	0.03 ± 0.03 (0.02)	0.073
Creatinine [mmoL/L]	91.19 ± 39.37 (80.00)	107.14 ± 55.76 (91.50)	<0.001
CRP [mg/L]	19.30 ± 54.48 (3.40)	21.77 ± 40.35 (4.09)	0.006
eGFR [mL/min/1.73m^2]	80.24 ± 32.07 (79.57)	69.60 ± 30.46 (65.87)	<0.001
Eosinophil count [10^9/L]	**0.18 ± 0.18 (0.15)**	**0.33 ± 1.31 (0.15)**	**0.643**
HbA1c [%]	9.12 ± 2.38 (8.80)	9.02 ± 2.29 (8.79)	0.928
HCT [%]	40.07 ± 5.86 (40.60)	38.33 ± 6.23 (38.80)	0.001
Hgb [g/dL]	13.64 ± 2.13 (13.80)	12.97 ± 2.28 (13.10)	<0.001
Ketones - urine sample	333 (21.35%)	22 (15.28%)	0.066
Lymphocyte count [10^9/L]	2.28 ± 3.24 (2.05)	2.01 ± 0.95 (1.84)	0.025
MCH [pg]	30.66 ± 2.76 (30.60)	30.86 ± 3.11 (30.70)	0.777
MCHC [g/dL]	33.94 ± 1.30 (34.00)	33.76 ± 1.31 (33.73)	0.072
MCV [fL]	90.24 ± 6.59 (90.00)	91.35 ± 7.90 (90.55)	0.178
Mean fast. glycemia [mg/dL] first day	195.25 ± 82.69 (180.00)	202.08 ± 86.98 (176.00)	0.521
Mean fast. glycemia [mg/dL] last day	135.70 ± 36.06 (132.00)	138.19 ± 37.58 (132.50)	0.251
Mean post. glycemia [mg/dL] first day	176.22 ± 66.27 (164.50)	179.69 ± 63.71 (166.00)	0.593
Mean post. glycemia [mg/dL] last day	139.07 ± 29.73 (136.00)	142.95 ± 33.42 (139.00)	0.251
Monocyte count [10^9/L]	0.65 ± 0.54 (0.55)	0.64 ± 0.35 (0.58)	0.603
Neutrophil count [10^9/L]	5.92 ± 4.01 (4.99)	6.16 ± 3.67 (5.25)	0.142
Platelet count [10^9/L]	249.24 ± 91.70 (239.00)	249.57 ± 96.94 (240.50)	0.809
Potassium [mmol/L]	**4.59 ± 0.57 (4.55)**	**4.72 ± 0.65 (4.68)**	**0.008**
Protein - urine sample	638 (40.66%)	70 (48.61%)	0.127

(*continued*)

TABLE 1. (*continued*)

Parameter	Patients without event (n = 1585)	Patients with event (n = 150)	*P*-value
Red blood cell count [10^{12}/L]	4.48 ± 0.74 (4.53)	4.23 ± 0.75 (4.33)	<0.001
Sodium [mmol/L]	139.53 ± 33.44 (139.00)	138.63 ± 4.50 (139.00)	0.600
Total cholesterol [mmol/L]	4.65 ± 1.51 (4.49)	4.35 ± 1.40 (4.24)	0.015
Triglyceride [mmol/L]	1.87 ± 1.73 (1.50)	1.95 ± 1.38 (1.50)	0.285
Uric acid [mmol/L]	325.50 ± 112.94 (310.00)	350.84 ± 119.59 (343.50)	0.010
White blood cell count [10^9/L]	8.95 ± 5.02 (8.00)	9.12 ± 4.25 (8.20)	0.327
	Pharmacotherapy		
ACEi/ARB	821 (51.80%)	90 (60.00%)	0.055
Allopurinol	338 (21.32%)	48 (32.00%)	0.003
Alpha blocker	164 (10.35%)	20 (13.33%)	0.256
Amiodarone	10 (0.63%)	0 (0.00%)	0.329
ASA	748 (47.19%)	103 (68.67%)	<0.001
Beta blocker	803 (50.66%)	108 (72.00%)	<0.001
Calcium blocker	430 (27.13%)	43 (28.67%)	0.686
Clopidogrel	**60 (3.79%)**	**21 (14.00%)**	**<0.001**
Digoxin	28 (1.77%)	0 (0.00%)	0.101
DPP-4 inhibitors	255 (16.09%)	18 (12.00%)	0.189
Fibrate	30 (1.89%)	2 (1.33%)	0.626
GLP-1 agonist	30 (1.89%)	5 (3.33%)	0.230
Heparin	**78 (4.92%)**	**16 (10.67%)**	**0.003**
Insulin	1248 (78.74%)	118 (78.67%)	0.984
PPI	**447 (28.20%)**	**68 (45.33%)**	**<0.001**
Loop diuretic	**448 (28.26%)**	**67 (44.67%)**	**<0.001**
Metformin	764 (48.20%)	76 (50.67%)	0.564
NOAC	98 (6.18%)	7 (4.67%)	0.457
Nonloop diuretics	256 (16.15%)	17 (11.33%)	0.121
Potassium-sparing diuretics	153 (9.65%)	20 (13.33%)	0.150
SGLT-2 inhibitor	186 (11.74%)	17 (11.33%)	0.884
Statin	812 (51.23%)	98 (65.33%)	0.001
Sulfonylureas	425 (26.81%)	48 (32.00%)	0.173
VKA	48 (3.03%)	5 (3.33%)	0.836

Abbreviations: ASA, acetylsalicylic acid; ACEi, angiotensin-converting-enzyme inhibitors; ARB, angiotensin receptor blockers; CRP, c-reactive protein; DPP-4, dipeptidyl peptidase-4; HCT, hematocrit; Hgb, hemoglobin; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; NOAC, novel oral anticoagulants; VKA, vitamin K anticoagulant.

For each parameter (if applicable), we report its mean ± standard deviation, together with the median (in parentheses).

The *P*-values were calculated using either χ^2 or Mann-Whitney U-test, as appropriate.

The most discriminative (12) predictors selected using the neighborhood component analysis are boldfaced.

calculated for the classifiers. The clinical utility of the models was analyzed in the decision curve analysis. Furthermore, we performed unsupervised cluster analysis of all patients (without splitting them into folds) and executed hierarchical clustering operating on all and selected predictors. The optimal number of clusters was identified using the Calinski-Harabasz quality criterion.[18]

GraphPad Prism 9.4.1 was exploited for statistical processing, whereas MATLAB R2022b for feature selection, classification and clustering. To visualize the clustering results in the high-dimensional feature spaces, we used t-distributed stochastic neighbor embedding (t-SNE).[19]

Results

We have identified 2115 eligible patients of which 1735 were enrolled into the study (53% female, mean±SD age of 59.1±17.4 years; mean± SD duration of diabetes: 11.2±9.1 years) with diabetes (21.32% T1DM and 78.68% T2DM) (Supplementary Figure S1). For all parameters with missing values (44 in total), the mean (median) percentage of patients with missing data was 1.80% (0.75%), with the maximum of 13.66% for the duration of diabetes. All 1735 patients were split into 5 nonoverlapping stratified folds which maintain the original distribution of the patients with and without a new CV event during the follow-up. Each fold becomes a test fold exactly once, whereas the remaining four folds constitute the training set.

In total, 150 patients (8.66%) had a new CV event during the follow-up (incidence 9 per 100 patients/7 months of follow-up). Feature selection was performed for each fold separately, with 12 predictors selected consistently for all folds (Table 1). We investigated the generalization capabilities of the ML model across all test folds, and the metrics were also aggregated to elaborate their mean and median values across 5 test folds. The classification performance, quantified as specificity, sensitivity, percentage of all CC patients, and percentage of CC patients with and without CV events, is reported in Table 2. For the selected most discriminative predictors (CAD, heart failure, peripheral artery disease, stroke, diabetic foot, CKD eosinophil count, serum potassium level, and being treated with clopidogrel, heparin, proton pump inhibitor, and loop diuretic), the sensitivity scores ranged from 0.50 (fold 2) to 0.63 (fold 1), with the corresponding specificity of 0.67 and 0.62, respectively. In fold 2, 227 of 347 (65.42%) of all patients were correctly classified as those with or without a new CV event, whereas 15 of 30 (50.00%) and 212 of 317 (66.88%) high- and low-risk patients were correctly identified. For

TABLE 2. Classification performance of the ML model

Metric	Features	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Median
Specificity	Selected (12)	**0.62**	**0.67**	**0.68**	**0.67**	**0.67**	**0.66**	**0.67**
	All (81)	0.57	0.65	0.63	**0.67**	0.66	0.64	0.65
Sensitivity	Selected (12)	0.63	**0.50**	**0.60**	**0.60**	**0.60**	**0.59**	**0.60**
	All (81)	**0.70**	**0.50**	0.57	0.53	0.57	0.57	0.57
CC with event [%]	Selected (12)	63.33	**50.00**	**60.00**	**60.00**	**60.00**	**58.67**	**60.00**
	All (81)	**70.00**	**50.00**	56.67	53.33	56.67	57.33	56.67
CC without event [%]	Selected (12)	**62.46**	**66.88**	**68.14**	66.88	**66.88**	**66.25**	**66.88**
	All (81)	56.78	64.98	62.78	**67.19**	66.25	63.60	64.98
CC All [%]	Selected (12)	**62.54**	**65.42**	**67.44**	**66.28**	**66.28**	**65.59**	**66.28**
	All (81)	57.93	63.69	62.25	65.99	65.42	63.05	63.69

The best metrics are boldfaced.

fold 1, 217 of 347 (62.54%) of all patients were correctly classified, with 19 of 30 (63.33%) and 198 of 317 (62.46%) high- and lowrisk diabetic patients were correctly identified.

In virtually all cases, the classification performance of RUSBoost exploiting the most discriminative predictors outperformed the model trained over all patients' parameters. The highest AUC for the model utilizing all predictors was 0.68, 95% CI 0.63-0.74 (fold 3), and the lowest AUC for this classifier amounted to 0.63, 95% CI 0.57-0.69 (fold 2) (Fig 1A). On the other hand, the ROC analysis of the models fitted over the selected parameters revealed that the highest AUC amounted to 0.72, 95% CI 0.66-0.77 (fold 1), whereas the lowest AUC was 0.62, 95% CI 0.56-0.68 (fold 2) (Fig 1B). A RUSBoost model operating over 30 patients' parameters which are statistically significantly different ($P < 0.05$) across the patients with and without a CV event (Table 1) resulted in AUC ranging from 0.62, 95% CI 0.56-0.68 to 0.70 95% CI 0.65-0.76. Finally, we fitted the MLR model over the selected (12) predictors (Fig 1C), and extracted the optimal cut-point value using the Index of Union method.[20] This model correctly determined 111 of 150 (74.00%) high-risk patients, and 989 of 1585 (62.40%) low-risk patients, resulting in 1100 of 1735 (63.40%) correctly classified patients, with AUC of 0.72.

The clinical utility of RUSBoost trained over all predictors and the most discriminative patients' parameters was investigated in Figure 1D and Figure 1E, respectively. In general, the model exploiting the selected features had significantly better clinical utility (above the probability threshold of 7% and below the 15%) in terms of net benefit than the 2 alternative treatment strategies, ie, treat all or none. Such clinical utility can be observed for the model operating over all patients' parameters as

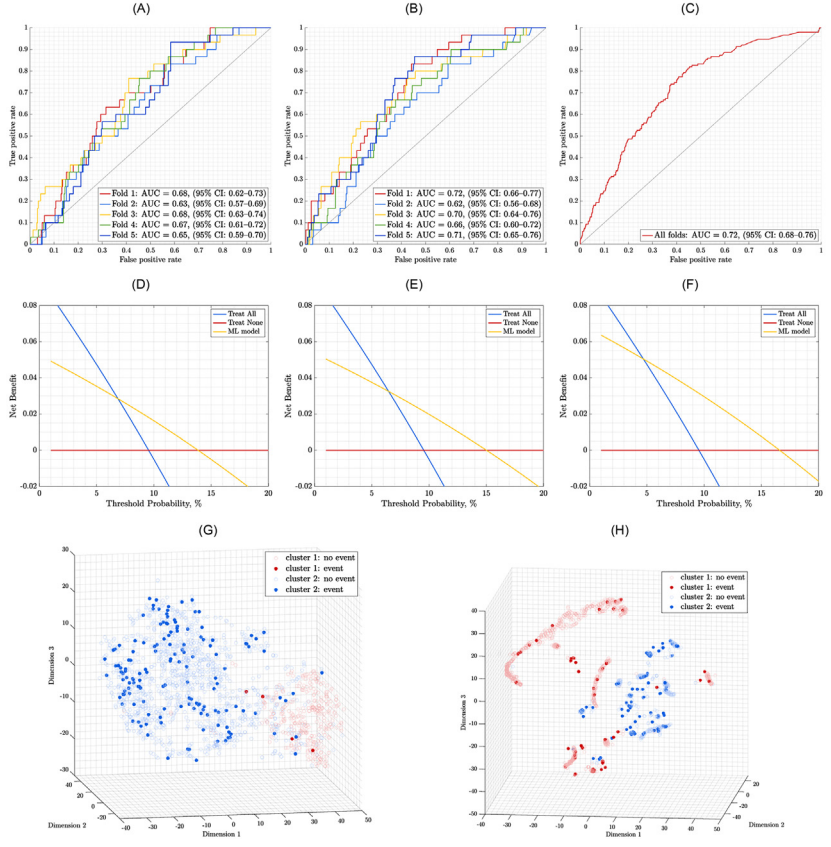

FIG 1. The ROC curves obtained using (A) the RUSBoost model trained over *all* (81) patient parameters, (B) the RUSBoost model trained over *the subset of the most discriminative* (12) patient parameters, and (C) the multiple logistic regression model fitted over *the subset of the most discriminative* (12) patient parameters (for the entire dataset), together with the decision curve analysis showing clinical utility of using (D) the RUSBoost model trained *all* (81) patient parameters, (E) the RUSBoost model trained over *the subset of the most discriminative* (12) patient parameters, and (F) the multiple logistic regression model fitted over the subset of *the most discriminative* (12) patient parameters. In (G) and (H), we show the t-SNE visualization of 2 patient clusters obtained using hierarchical clustering over *all* (81) and the subset of *the most discriminative* (12) patient parameters, respectively. For the ROC curves, the 45° curve through the origin shows the discriminatory ability of the classifier not better than chance.

well (Fig 1D), but the number of features contributing to the predictions is 6.75 × larger (12 most discriminative parameters elaborated using NCA vs 81 all patients' parameters). Also, the MLR model fitted over the selected predictors offers notable clinical utility (above the probability threshold 5% and below 17%) (Fig 1F).

Hierarchical clustering performed over all patients using the most discriminative parameters (Table 2) indicated 2 groups of patients (blue and red dots in Figure 1H, with the high-risk patients annotated with the filled dots), with the 1 group encompassing the majority of high-risk patients (96/150, 64.00%; in this case, the classification metrics would amount to sensitivity: 0.64 and specificity: 0.64; the nonbinary predictors were scaled to the 0-1 range). The importance of feature selection is shown in Figure 1G, where we presented the results of clustering performed over all (81) predictors (similarly, 2 clusters were indicated as optimal using the Calinski-Harabasz criterion). Here, the t-SNE visualization showed that clustering was of poorer quality with respect to the low- and highrisk patients—although the majority of high-risk patients were included in 1 cluster (146/150, 97.33%), 1241 low-risk patients were included in the same cluster, resulting in an extremely large number (1241) of false positives (sensitivity: 0.97, specificity: 0.22). Table 3 gathers the feature values calculated for the patients included in both clusters obtained using hierarchical clustering (with and without feature selection), further confirming significant differences across the 2-cluster patients.

Discussion

The principal findings of our study are 3-fold: (*1*) we determined the most discriminative patients' parameters which can be exploited to build supervised and unsupervised ML models for identifying diabetic patients with a high risk of a CV event, (*2*) we showed, following a rigorous multifold cross validation, that a ML algorithm can generalize well over unseen patients while exploiting only 12 interpretable and easy-to-obtain predictors (CAD, heart failure, peripheral artery disease, stroke, DFD, CKD, eosinophil count, serum potassium level, and being treated with clopidogrel, heparin, proton pump inhibitor, and loop diuretic), and (*3*) we proved better clinical utility of the ML models when compared to the "treat all" and "no treatment" strategies. Determining such high- and low-risk patients is extremely important in relation to precision medicine to discriminate patients that should be treated immediately for avoiding future CV events.

The ML supervised RUSBoost model operating on 12 most discriminative features achieved high stability across five nonoverlapping test folds (Fig 1B), with the AUC values ranging from 0.62 to 0.72. This model outperformed its counterpart exploiting all patients' parameters (which would be much more challenging to capture and analyze in the clinical settings), obtaining AUC between 0.63 and 0.68 (Fig 1A), clearly

TABLE 3. Feature values (% of yes) for the patients in both clusters after elaborated using hierarchical clustering with and without feature selection

	With feature selection (12 predictors)			*Without feature selection* (81 predictors)		
Parameter	**Cluster 1**	**Cluster 2**	***P*-value**	**Cluster 1**	**Cluster 2**	***P*-value**
Coronary artery disease	1.03%	90.25%	<0.001	1.44%	43.84%	<0.001
Chronic kidney disease	9.78%	32.68%	<0.001	2.02%	22.76%	<0.001
Clopidogrel	0.47%	11.39%	<0.001	0.57%	5.70%	<0.001
Diabetic foot disease	3.09%	2.55%	0.510	0.86%	3.39%	0.012
Eosinophil count [10^9/L]	0.19 ± 0.52 (0.14)	0.19 ± 0.20 (0.15)	0.033	0.17 ± 0.13 (0.14)	0.19 ± 0.47 (0.15)	0.778
Heart failure	0.09%	47.38%	<0.001	0.57%	35.33%	<0.001
Heparin	3.46%	8.55%	<0.001	0.57%	6.63%	<0.001
PPI	20.22%	44.83%	<0.001	7.18%	35.33%	<0.001
Loop diuretic	12.92%	56.52%	<0.001	1.43%	36.77%	<0.001
Peripheral artery disease	3.37%	6.75%	0.001	0.29%	5.77%	<0.001
Potassium [mmol/L]	4.55 ± 0.54 (4.51)	4.67 ± 0.61 (4.64)	<0.001	4.55 ± 0.53 (4.51)	4.61 ± 0.58 (4.58)	0.031
Stroke	1.12%	19.79%	<0.001	0.86%	10.17%	<0.001
Event [% of yes]	***0.05%***	***14.39%***	***<0.001***	***1.15%***	***10.53%***	***<0.001***

For each parameter (if applicable), we report its mean ± standard deviation, together with the median (in parentheses). The *P*-values were calculated using either χ^2 or Mann-Whitney U-test, as appropriate.

indicating the importance of appropriate feature selection. This is further manifested while building a RUSBoost model operating over 30 patients' parameters which are significantly different ($P < 0.05$) across the patients with and without a CV event (Table 1)—exploiting such a model does not bring any improvements when compared to the one built upon 12 predictors, with AUC ranging from 0.62 to 0.70. The MLR model fitted over 12 predictors offered high discriminative power with AUC of 0.72 over all patients, and it is unlikely to overfit while exploiting such a small number of predictors (Fig 1C). Unsupervised clustering further demonstrated that the selected features allow for grouping the diabetic patients into high- and lowrisk ones (Fig 1H), characterized by significantly different parameter values in such separate clusters (Table 3).

Patients at a higher risk of CV events are those who present with history of CVD, namely CAD, peripheral artery disease, stroke and heart failure, which is observed at the level of individual feature analysis (Table 1). CAD, peripheral artery disease, stroke and heart failure are significantly different for the groups of patients with and without an event ($P < 0.05$). This clearly shows that the highest risk is present in patients who are affected with atherosclerotic vascular disease at each main vascular site (heart, brain, lower extremities) and heart failure as well.

Diabetes *per se* is an important risk factor for HF and often HF is the first CVD diagnosed in patients with T2DM.[21] In the considered cohort, significantly more patients with T2DM are present in the group with a CV event, when compared to those without it. Moreover, a concomitant disease, namely CKD, became a significant discriminator of patients with future CV events. CKD is one of the risk factors which should be systematically assessed (at least annually) in all patients with diabetes for prevention and management of both atherosclerotic CVD and HF.[22]

Among several diabetes-related parameters such as HbA1c, mean glycemia, diabetic retinopathy, neuropathy, and diabetes duration, only DFD was distinguished by NCA. However, patients with DFD has been recently found to be at the highest risk of future fatal events,[23] and there exist estimates proving that the life expectancy of patients with DFD is similar to people with cancers like colon or breast.[24]

The inflammatory background of coronary atherosclerosis is suggested by higher eosinophil counts which was discriminative to distinguish patients with future CV events. High eosinophil counts were associated with an increased serum fibrinogen and platelet counts, and an increased risk and severity of coronary atherosclerosis.[25] This is confirmed in our cluster analysis, where the eosinophil count was higher in patients in the high-risk group (Table 3).

Maintaining potassium within a reference range is very important especially in relation to new cardioprotective and renoprotective therapies that can promote potassium retention.[26] Indeed, there is a continuous U-shaped relationship between serum potassium and all-cause mortality in the total population and in patients with HF, CKD, diabetes, as well as all 3 diseases and even in patients without these diseases.[27] In a population-based analysis, serum potassium concentration ≥5.0 mEq/L was associated with all-cause mortality, CVD death, and non-CVD death, and of note, all-cause mortality was also increased among patients with serum potassium levels within the normal range, being 4.0-4.9 mEq/L.[28] In our study, the concentration of potassium significantly differentiates diabetic patients with and without an event, and the potassium levels were one of the most discriminative predictors by NCA.

Pharmacotherapy has a discriminative effect. Heparin, clopidogrel, loop diuretics, and proton pump inhibitors (PPI) were used more frequently by patients with high risk of CV events. Though no causal inference can be made from this study, utilizing those drugs most likely does not increase a risk *per se*, but were rather drugs that are prescribed more often for patients with more comorbidities. For example, patients who recently had myocardial infarction and underwent percutaneous coronary intervention often take both clopidogrel and PPI. Likewise, those with HF are treated with loop diuretics. Moreover, these drugs are inseparably related to CAD or HF, and we demonstrate their utility as predictors of CV events.

There are important limitations of our study. This was a single center study so we cannot generalize the outcomes for the whole population of patients with diabetes. The study is observational in its design. Diagnosis of new onset heart failure was made according to the European Society of Cardiology guidelines,[29] yet it was somewhat limited due to inaccessibility of natriuretic peptides measurements. This concerned especially heart failure with preserved ejection which cases might have been under recognized. Despite the fact that heart failure is a clinical syndrome with distinct phenotypes, we did not categorize patients based on ejection fraction in this study.

Conclusions

The ML models operating on a small subset of the most discriminative, interpretable and easy to obtain patients' parameters could help disentangle the heterogeneity of population of patients with diabetes in terms of CV events, and can be used to tailor more efficient prevention and

therapeutic strategies. This gives an opportunity to move from a "one-size fits-all" strategy to precision CV prevention approaches.

Ethics Approval and Consent to Participate

The study protocol has been approved by the Medical University of Silesia Ethics Committee (PCN/0022/KB/126/20) and the need for informed consent was waived by this Ethics Committee.

Consent for Publication

Not applicable.

Availability of Data and Materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributors

KN, HK, JG, GYHL—substantial contribution to the conception and design of the work; KD, KI, MH, WH, OJ, WW— collected the data; HK—prepared the data set for statistical analysis. JN, AMW, EC—designed the machine learning algorithms; AMW—implemented and verified the machine learning algorithms; AMW, JN—performed the computational experiments; AMW, JN – performed the data analysis; AMW, JN—prepared tables and figures; KN, JN, AMW—drafted the manuscript; JG, GYHL—substantively revised the work. KN and JN are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All of the authors have read, corrected and approved the submitted version of the paper.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.cpcardiol.2023.101694.

REFERENCES

1. Low Wang CC, Hess CN, Hiatt WR, et al. Clinical update: cardiovascular disease in diabetes mellitus. *Circulation* 2016;133:2459–502. https://doi.org/10.1161/CIRCULATIONAHA.116.022194.

2. Rawshani A, Rawshani A, Franzén S, et al. Mortality and cardiovascular disease in Type 1 and Type 2 diabetes. *N Engl J Med* 2017;376:1407–18. https://doi.org/10.1056/NEJMoa1608664.
3. Chamnan P, Simmons RK, Sharp SJ, et al. Cardiovascular risk assessment scores for people with diabetes: a systematic review. *Diabetologia* 2009;52:2001–14. https://doi.org/10.1007/s00125-009-1454-0.
4. Coronary risk prediction for those with and without diabetes. *Eur J Cardiovasc Prev Rehabil* 2006;13:30–6. https://doi.org/10.1097/00149831-200602000-00005.
5. Kengne AP, Patel A, Colagiuri S, et al. The Framingham and UK Prospective Diabetes Study (UKPDS) risk equations do not reliably estimate the probability of cardiovascular events in a large ethnically diverse sample of patients with diabetes: the action in diabetes and vascular disease: pretera. *Diabetologia* 2010;53:821–31. https://doi.org/10.1007/s00125-010-1681-4.
6. Dziopa K, Asselbergs FW, Gratton J, et al. Cardiovascular risk prediction in type 2 diabetes: a comparison of 22 risk scores in primary care settings. *Diabetologia* 2022;65:644–56. https://doi.org/10.1007/s00125-021-05640-y.
7. Ross EG, Jung K, Dudley JT, et al. Predicting future cardiovascular events in patients with peripheral artery disease using electronic health record data. *Circ Cardiovasc Qual Outcomes* 2019;12:e004741. https://doi.org/10.1161/CIRCOUTCOMES.118.004741.
8. Cho S-Y, Kim S-H, Kang S-H, et al. Pre-existing and machine learning-based models for cardiovascular risk prediction. *Sci Rep* 2021;11:8886. https://doi.org/10.1038/s41598-021-88257-w.
9. Ward A, Sarraju A, Chung S, et al. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. npj Digit Med 2020 31. *NPJ Digit Med* 2020;3:1–7. https://doi.org/10.1038/s41746-020-00331-1.
10. Drożdż K, Nabrdalik K, Kwiendacz H, et al. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach. *Cardiovasc Diabetol* 2022;21:240. https://doi.org/10.1186/s12933-022-01672-9.
11. Ng R, Sutradhar R, Wodchis WP, et al. Chronic disease population risk tool (CDPoRT): a study protocol for a prediction model that assesses population-based chronic disease incidence. *Diagnostic Progn Res* 2018;2:19. https://doi.org/10.1186/s41512-018-0042-5.
12. van Dieren S, Beulens JWJ, Kengne AP, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart* 2012;98:360–9. https://doi.org/10.1136/heartjnl-2011-300734.
13. Hossain ME, Uddin S, Khan A. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Syst Appl* 2021;164:113918. https://doi.org/10.1016/j.eswa.2020.113918.
14. Carrillo-Larco RM, Castillo-Cara M, Anza-Ramirez C, et al. Clusters of people with type 2 diabetes in the general population: unsupervised machine learning approach using national surveys in Latin America and the Caribbean. *BMJ open diabetes Res care* 2021;9:1–8. https://doi.org/10.1136/bmjdrc-2020-001889.
15. Audigier V, Husson F, Josse J. A principal component method to impute missing values for mixed data. *Adv Data Anal Classif* 2016;10:5–26. https://doi.org/10.1007/s11634-014-0195-1.

16. Wei Yang, Kuanquan, Wang WZ. Neighborhood component feature selection for high-dimensional data. *JCP* 2012;7:161–8.
17. Seiffert C, Khoshgoftaar TM, Hulse J Van, et al. RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Trans Syst Man, Cybern - Part A Syst Humans.* 2010;40:185–97. https://doi.org/10.1109/TSMCA.2009.2029559.
18. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
19. Oliveira FHM, Machado ARP, Andrade AO. On the use of t-distributed stochastic neighbor embedding for data visualization and classification of individuals with parkinson's disease. *Comput Math Methods Med* 2018;2018:8019232. https://doi.org/10.1155/2018/8019232.
20. Unal I. Defining an optimal cut-point value in ROC analysis: an alternative approach. *Comput Math Methods Med* 2017;2017:3762651. https://doi.org/10.1155/2017/3762651.
21. Groenewegen A, Rutten FH, Mosterd A, et al. Epidemiology of heart failure. *Eur J Heart Fail* 2020;22:1342–56. https://doi.org/10.1002/ejhf.1858.
22. 10. cardiovascular disease and risk management: standards of medical care in diabetes-2022. *Diabetes Care* 2022;45:S144–74. https://doi.org/10.2337/dc22-S010.
23. Mader JK, Haas W, Aberer F, et al. Patients with healed diabetic foot ulcer represent a cohort at highest risk for future fatal events. *Sci Rep* 2019;9:10325. https://doi.org/10.1038/s41598-019-46961-8.
24. Armstrong DG, Boulton AJM, Bus SA. Diabetic foot ulcers and their recurrence. *N Engl J Med* 2017;376:2367–75. https://doi.org/10.1056/NEJMra1615439.
25. Niccoli G, Ferrante G, Cosentino N, et al. Eosinophil cationic protein: a new biomarker of coronary atherosclerosis. *Atherosclerosis* 2010;211:606–11. https://doi.org/10.1016/j.atherosclerosis.2010.02.038.
26. Sica DA, Struthers AD, Cushman WC, et al. Importance of potassium in cardiovascular disease. *J Clin Hypertens* 2002;4:198–206. https://doi.org/10.1111/j.1524-6175.2002.01728.x.
27. Collins AJ, Pitt B, Reaven N, et al. Association of serum potassium with all-cause mortality in patients with and without heart failure, chronic kidney disease, and/or diabetes. *Am J Nephrol* 2017;46:213–21. https://doi.org/10.1159/000479802.
28. Hughes-Austin JM, Rifkin DE, Beben T, et al. The relation of serum potassium concentration with cardiovascular events and mortality in community-living individuals. *Clin J Am Soc Nephrol* 2017;12:245–52. https://doi.org/10.2215/CJN.06290616.
29. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the heart failure association (HFA) of the ESC. *Eur Heart J* 2021;42:3599–726. https://doi.org/10.1093/EURHEARTJ/EHAB368.