scientific reports

Check for updates

OPEN

Predicting 1-year mortality of patients with diabetes mellitus in Kazakhstan based on administrative health data using machine learning

Aidar Alimbayev[1,2], Gulnur Zhakhina[2], Arnur Gusmanov[2], Yesbolat Sakko[2], Sauran Yerdessov[2], Iliyar Arupzhanov[1], Ardak Kashkynbayev[3], Amin Zollanvari[1] & Abduzhappar Gaipov[2]✉

Diabetes mellitus (DM) affects the quality of life and leads to disability, high morbidity, and premature mortality. DM is a risk factor for cardiovascular, neurological, and renal diseases, and places a major burden on healthcare systems globally. Predicting the one-year mortality of patients with DM can considerably help clinicians tailor treatments to patients at risk. In this study, we aimed to show the feasibility of predicting the one-year mortality of DM patients based on administrative health data. We use clinical data for 472,950 patients that were admitted to hospitals across Kazakhstan between mid-2014 to December 2019 and were diagnosed with DM. The data was divided into four yearly-specific cohorts (2016-, 2017-, 2018-, and 2019-cohorts) to predict mortality within a specific year based on clinical and demographic information collected up to the end of the preceding year. We then develop a comprehensive machine learning platform to construct a predictive model of one-year mortality for each year-specific cohort. In particular, the study implements and compares the performance of nine classification rules for predicting the one-year mortality of DM patients. The results show that gradient-boosting ensemble learning methods perform better than other algorithms across all year-specific cohorts while achieving an area under the curve (AUC) between 0.78 and 0.80 on independent test sets. The feature importance analysis conducted by calculating SHAP (SHapley Additive exPlanations) values shows that age, duration of diabetes, hypertension, and sex are the top four most important features for predicting one-year mortality. In conclusion, the results show that it is possible to use machine learning to build accurate predictive models of one-year mortality for DM patients based on administrative health data. In the future, integrating this information with laboratory data or patients' medical history could potentially boost the performance of the predictive models.

The burden of diabetes is a rising concern in healthcare worldwide. According to the estimates of the International Diabetes Federation (IDF), in 2021, there were 537 million adults living with diabetes, which is 6.79% of the world's population[1]. According to the global epidemiological data from 2017, the predicted number of diabetes will be 693 million by 2030, while the most recent study projects a rise up to 783 million by 2045. IDF study reports that 75% of cases live in low- and middle-income countries, and there were 6.7 million deaths worldwide in 2021[2]. In Kazakhstan, overall 472,950 people were listed with Type 1 and Type 2 Diabetes in inpatient and outpatient registries in the same period[3].

Diabetes mellitus (DM) affects the quality of life and leads to disability, high morbidity, and premature mortality. DM is a risk factor for cardiovascular, neurological, and renal diseases, and places a major burden on healthcare systems globally. At the same time, populations around the world are rapidly aging, and that further

[1]Department of Electrical and Computer Engineering, School of Engineering and Digital Sciences, Nazarbayev University, Kabanbay Batyr Avenue 53, Astana, Kazakhstan. [2]Department of Medicine, School of Medicine, Nazarbayev University, Kerey and Zhanibek Khans Street 5/1, Astana, Kazakhstan. [3]Department of Mathematics, Nazarbayev University, Kabanbay Batyr Avenue 53, Astana, Kazakhstan. ✉email: abduzhappar.gaipov@nu.edu.kz

contributes to a higher number of incident DM cases. Elderly people can have multiple comorbidities and complications along with DM, which elevates mortality rates[4]. In this regard, predicting the one-year mortality of patients with DM is at the core of health management systems as it can help clinicians tailor treatments to improve the survival of DM patients.

In recent years, constructing data-driven predictive models using machine learning has found various applications in health care[5,6]. In[7], Random Forest (RF) algorithm was used for the early prediction of diabetes using a number of variables such as regular and ultralente insulin dose, socio-demographic factors, and hypoglycemic symptoms, to just name a few. In another study[8], different predictive models were examined to predict diabetes based on several factors such as glucose level, blood pressure, and insulin. Moreover, machine learning techniques were used to predict the mortality of diabetes patients based on HbA1c and lipid parameters[9]. With successful applications of machine learning in disease and mortality prediction, it is highly anticipated that it can be used to predict the one-year mortality of patients with DM based on ordinary clinical variables. Some mortality prognostic models have been developed using machine learning approaches on clinical and administrative data[9–11]. Furthermore, several studies have attempted to predict mortality for DM patients in an intensive care unit (ICU)[12–15]. However, predicting the one-year mortality of DM patients based solely on administrative health data including diagnoses, comorbidities, procedures, and demographics have not been used before. This is in sharp contrast with the previous studies where additional information including the results of laboratory tests or vital signs (e.g., ICU admission) were used for prediction.

In this regard, we used the Unified National Electronic Health System (UNEHS) of Kazakhstan to collect ordinary clinical data for a large cohort of DM patients who registered in hospitals across the country between January 2014 and December 2019. The detailed description of database is given elsewhere[16]. The collected data was then divided into four subcohorts to predict mortality within a year (starting from 2016) based on collected clinical data up to the end of the preceding year. We then develop a comprehensive machine learning platform to construct one predictive model of one-year mortality for each subcohort. Our study points to the feasibility and robustness of the developed machine learning (ML) platform for predicting the one-year mortality of DM patients in Kazakhstan using aggregated nationwide administrative healthcare data. We also identify and rank the importance of clinical variables that were used by the constructed predictive models of mortality.

To our knowledge, there is a lack of models that can distinguish high-risk populations and forestall the mortality of individuals with diabetes in Central Asian countries. The development of a prognostic model for one-year mortality in diabetes mellitus has the potential to assist healthcare practitioners in devising individualized treatment plans and interventions that can mitigate adverse consequences. Furthermore, this could aid in the allocation of resources, as patients who are deemed high-risk may necessitate more frequent monitoring or follow-up care.

Results

Data description. The objective of this study is to predict one-year mortality in DM patients based on administrative health data. In this regard we collected clinical data for patients diagnosed with DM from UNEHS[3], which is a nationwide electronic health record repository of patients admitted to hospitals across Kazakhstan between mid-2014 and December 2019. After excluding patients with the missing outcome, which is the mortality with possible values being dead or alive, the data was divided into four yearly-specific cohorts to predict mortality within a specific year based on clinical information collected up to the end of the preceding year. Hereafter, these subcohorts are referred to as 2016-, 2017-, 2018-, and 2019-cohorts and contain 262,212, 301,563, 337,846, and 370,807 patients, respectively. For example, the cohort of 2018 contains only patients who have been admitted to the hospital and were alive on or prior to 31st December 2017 and, at the same time, the value of the outcome variable in 2018 is known (see Supplementary materials for more details). The data is highly imbalanced with the ratio of death to alive being, 10,490:251,722, 11,568:289,995, 13,168:324,678, 13,534:357,273, for 2016-, 2017-, 2018-, and 2019-cohorts, respectively. The clinical variables used as predictors of mortality in the collected cohorts are listed in Table 1 (more information in the Supplementary Table S1). The missing values of numeric and categorical predictors were imputed based on the median and mode of those variables in training data, respectively. We used a stratified random split to divide each yearly-specific cohort with

Feature	Description	Unit	Type
Age	Age at the first hospitalisation recorded in the database	Years	Numeric
Sex	Gender: female or male	Binary	Categorical
Ethnicity	Three major categories: Kazakhs, Russians and others	Tertiary	Categorical
Type of Diabetes	Type of diabetes: T1D or T2D, and other types	Tertiary	Categorical
CHD	Comorbidity for diabetes (yes/no)	Binary	Categorical
CVA	Comorbidity for diabetes (yes/no)	Binary	Categorical
Neoplasms	Comorbidity for diabetes (yes/no)	Binary	Categorical
Hypertension	Comorbidity for diabetes (yes/no)	Binary	Categorical
Hospitalisation	Number of hospitalizations during observation period	Frequency	Numeric
Duration of Diabetes	Follow-up time until December 31 the preceding year of prediction	Years	Numeric

Table 1. Name and description of predictors (features) used in yearly-specific cohorts. CHD, coronary heart disease; CVA, cerebrovascular accident; T1D, type 1 diabetes; T2D, type 2 diabetes.

an 80/20 ratio into training and test sets. The stratification is performed to keep the proportion of samples that appear in training and test sets the same as the full cohort. The training set is used for predictive model (classifier) training and selection, while the test set is used for model evaluation. Each year-specific constructed predictive model classifies the status of the patient (dead or alive) within the next year after collecting the patients' clinical data.

Training and selecting yearly-specific classifier of one-year mortality—model training and selection. We deployed nine classifiers, namely, Gaussian Naïve Bayes (GNB)[17], K-nearest neighbors (KNN)[17], logistic regression with L_2 ridge penalty (LRR)[18], random forest (RF)[18], AdaBoost with decision trees (ADB)[19], gradient boosting with regression trees (GBRT)[20], XGBoost (XGB)[21], linear discriminant analysis (LDA)[22], and perceptron (PER)[17] (see the Materials and Methods section for more details on the rationale behind the selecting these classifiers). The candidate hyperparameter space for each classifier is discussed in the Materials and Methods. The developed ML platform performs model selection (including hyperparameter tuning) using each yearly-specific training set by calculating the area under the curve (AUC) performance metric using stratified 5-fold cross-validation (5-CV). Table 2 shows the 5-fold CV estimate of the AUC for each classifier. As observed in Table 2, GBRT achieved the highest AUC for the years 2016 and 2017, while XGB showed the highest AUC for 2018 and 2019. That being said, both classifiers are from the class of gradient boosting ensemble learning. This shows the superiority of gradient-boosting ensemble learning compared with other algorithms in our application.

Evaluating year-specific classifier of one-year mortality—model evaluation. The best year-specific classification algorithm and the values of its hyperparameters that were identified in the model selection phase were used to train one final year-specific classifier on the entire training set. Then each of these trained classifiers is evaluated on the corresponding (year-specific) test set using several performance metrics including AUC, balanced accuracy, sensitivity, specificity, and the geometric mean of sensitivity and specificity (G-mean). Figure 1 shows the entire process of model selection and evaluation. The results of the model evaluation are shown in Table 3. The confusion matrices across all year-specific test sets are presented in Supplementary Materials (Supplementary Tables S2–S5). All classifiers achieved an AUC greater than 0.78, which is ranked 'fair' (close to 'good') as per objective metrics of diagnostic tests (see[23] for performance guidance based on AUC). At the same time, the estimated AUCs on test sets are quite close to the AUCs previously achieved using a 5-fold CV. This observation per se shows the robustness of developed classifiers. The results also show that the developed classifiers have a higher sensitivity than specificity. In the trade-off between sensitivity and specificity of our developed classifiers, this is indeed a desirable feature for our application, because the cost of not detecting (and no intervention thereof) a patient who will die within a year is (much) higher than a patient who is labeled as "death" but will truly survive.

Impact direction and importance of each feature for predicting one-year mortality. We performed a SHAP[24] (short for SHapley Additive exPlanations) analysis to: (1) infer the direction of impact of each feature on mortality prediction made by the year-specific model; and (2) measure the overall importance of each feature on outcome prediction. In this regard, we estimated SHAP values for the year-specific classifier that was selected in the model selection stage; that is to say, for 2016 and 2017, they were estimated for the GBRT classifier, and for 2018 and 2019, they were estimated for the XGB classifier. Furthermore, SHAP values were computed for all variables in the training dataset as no feature selection has been performed (see Discussion). Figure 2 shows the SHAP summary dot plot and mean absolute SHAP bar plot for the 2016-specific cohort. Similar plots for other year-specific classifiers are presented in Supplementary Materials. From Fig. 2a, we also

Classifier	AUC			
	2016	2017	2018	2019
GNB	0.705 ± 0.009	0.697 ± 0.009	0.695 ± 0.009	0.703 ± 0.003
KNN	0.602 ± 0.003	0.585 ± 0.003	0.588 ± 0.003	0.584 ± 0.004
LRR	0.756 ± 0.006	0.744 ± 0.003	0.743 ± 0.002	0.749 ± 0.005
RF	0.687 ± 0.004	0.681 ± 0.009	0.682 ± 0.003	0.689 ± 0.003
ADB	0.770 ± 0.003	0.758 ± 0.005	0.755 ± 0.005	0.764 ± 0.005
XGB	0.793 ± 0.007	0.784 ± 0.002	**0.791 ± 0.003**	**0.797 ± 0.006**
GBRT	**0.795 ± 0.005**	**0.786 ± 0.007**	0.787 ± 0.002	0.794 ± 0.002
LDA	0.755 ± 0.007	0.742 ± 0.005	0.741 ± 0.005	0.748 ± 0.002
PER	0.593 ± 0.013	0.605 ± 0.012	0.641 ± 0.011	0.596 ± 0.010

Table 2. The classifier-specific estimated AUC mean ± standard deviation over 5 folds of 5-fold cross-validation obtained on the training set. The year-specific highest AUC mean and standard deviation is identified in bold. GNB, Gaussian Naive Bayes; KNN, K-nearest neighbors; LRR, logistic regression; RF, random forest; ADB, Adaboost with decision trees; GBRT, gradient boosting with regression trees; XGB, XGBoost; LDA, linear discriminant analysis; PER, perceptron.

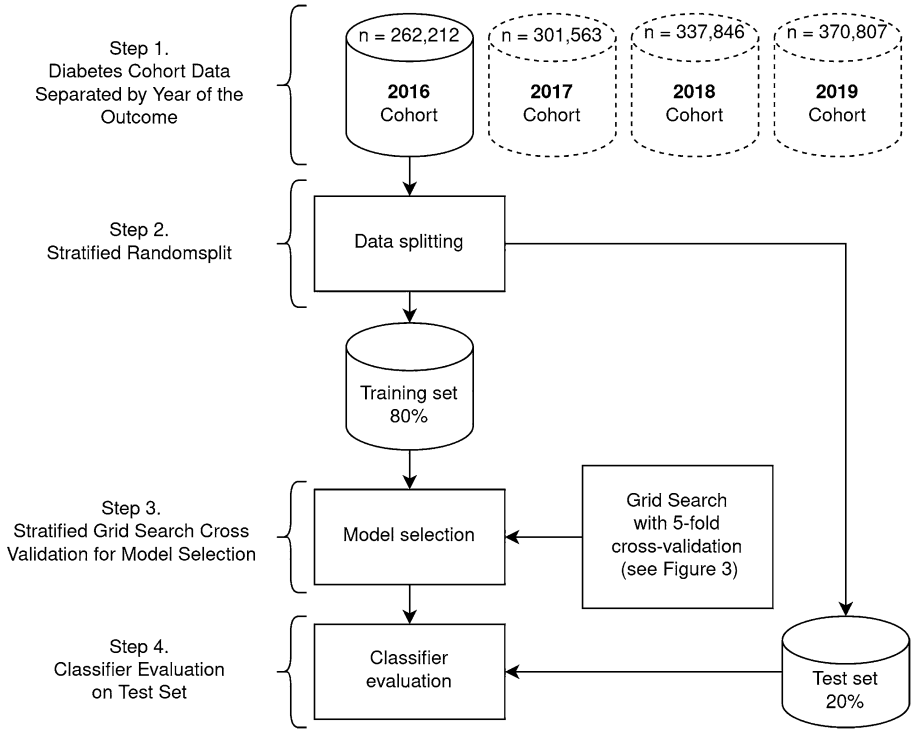

Figure 1. A schematic diagram of the developed machine learning platform.

Classifier	AUC	Balanced accuracy	Sensitivity	Specificity	G-mean
2016-classifier (GBRT)	0.791	0.698	0.876	0.520	0.722
2017-classifier (GBRT)	0.787	0.690	0.879	0.502	0.779
2018-classifier (XGB)	0.787	0.674	0.899	0.499	0.775
2019-classifier (XGB)	0.799	0.644	0.937	0.352	0.777

Table 3. Performance metrics of year-specific selected classifier evaluated on the test data for the same year. GBRT, gradient boosting with regression trees; XGB, XGBoost.

observe that age and duration of diabetes are directly proportional to higher mortality. Considering binary values of hypertension, the results show that the lack of hypertension (encoded as 0) is associated with higher mortality. To summarize the results of SHAP values across all four cohorts, we determined the *average* of the mean absolute SHAP value (AMAS) for each feature across all years. This result ranks the features (from highest to lowest importance) as age, duration of diabetes, hypertension, sex, neoplasms, hospitalisation, CHD, CVA, type of diabetes, and ethnicity with AMAS being 0.0129, 0.0097, 0.0041, 0.0034, 0.0003, 0.0018, 0.0013, 0.0010, 0.0008, 0.0006, and 0.0003, respectively. Based on these findings age, duration of diabetes, hypertension, and sex are the top four most important features for predicting one-year mortality.

Discussion

The results in Table 3 show that all trained yearly-specific classifiers achieved a predictive performance in the range of 0.78–0.799 in terms of AUC. At the same time, as per objective metrics of diagnostic tests, an estimated AUC in the range of 0.7–0.8 is generally considered a 'fair' predictive capacity for the test[23].

Several studies have predicted the mortality of DM patients using a combination of clinical and administrative data. For instance, a recent study[12] predicted the mortality of diabetic patients admitted to the ICU using nine classifiers including LR, RF, AB, XGB, GBM, artificial neural network (ANN) and majority voting. XGB and majority voting showed the best performance with an AUC of 0.867 and 0.867, respectively. Similarly, another study[13] predicted the mortality of critically ill patients with DM using the Charlson comorbidity index (CCI), Elixhauser comorbidity index, the diabetes complications severity index (DSCI), RF, and LR as the main prediction models. The LR achieved an AUC of 0.785, while RF achieved an AUC of 0.787.

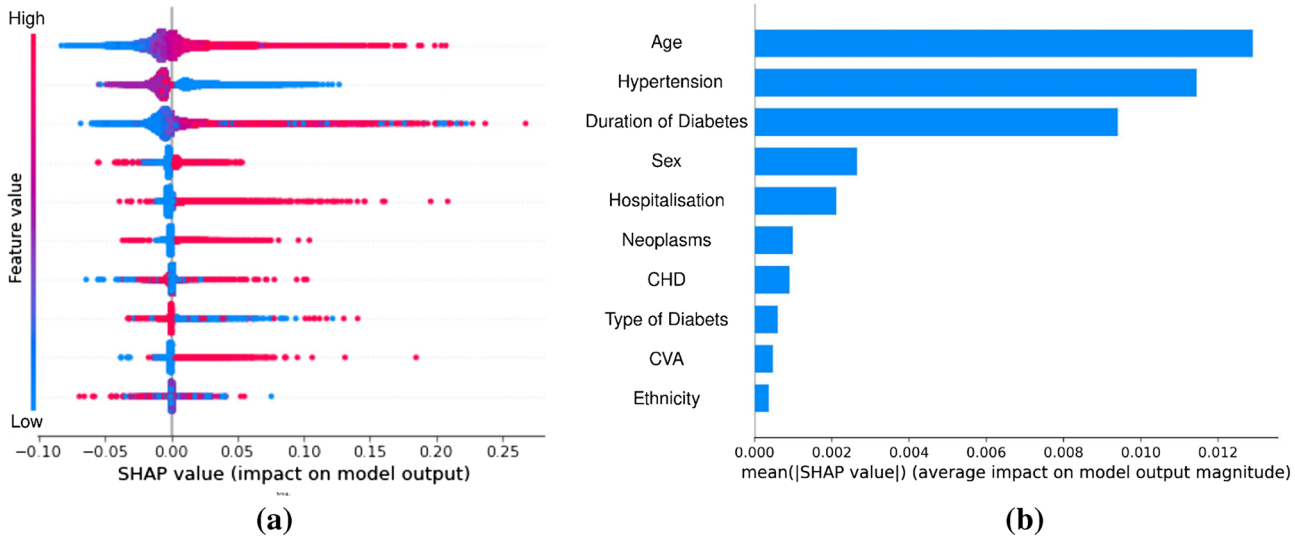

Figure 2. SHAP analysis of 2016-specific cohort: (**a**) SHAP summary dot plot for the 2016-specific cohort. A red dot shows a high value of the feature for a patient, whereas a blue dot shows a low value. The likelihood of mortality increases (decreases) for a positive (negative) SHAP value. Positive SHAP values for red dots show a direct dependence on the feature and the outcome, whereas the same values for blue dots imply an inverse dependence. (**b**) The mean absolute SHAP value bar plots for the 2016-specific cohort. The plot shows the feature importance on outcome prediction made by the model (a longer bar shows a more important feature).

In another study[14], the mortality of heart failure patients with diabetes was predicted using nine classifiers, including LR, RF, SVM, KNN, DT, GBM, XGBoost, LightGBM, and Bagging. The RF algorithm outperformed other algorithms, achieving an AUC of 0.92. Mortality prediction of patients with diabetes and sepsis in ICU using five classifiers were investigated in another study[15]. Authors used LR with lasso regularization, Bayes LR, decision tree, RF, and XGBoost. Out of five classifiers, the RF model showed the best performance, achieving an accuracy of 0.883. In another investigation[9], Random Survival Forest (RSF) was used to predict the mortality of patients with diabetes and study the hazardous effects of HbA1C and lipid variability. The RSF model achieved an AUC of 0.866. Table 4 provides a summary of studies on predicting mortality of DM patients.

Although our identified models have a 'fair' predictive capacity (close to 'good'), their estimated AUC is generally lower than the previous studies[9,12,14,15]. This state of affairs can be attributed in part to the availability

Study	Dataset	Patients amount			Algorithm	Performance by AUC	Important predictors
		Alive	Dead	Imbalance proportion			
Lee, S. et al.[9]	Dataset from Hong Kong hospitals	25,186	12,372	0.491	Regularized and Weighted RSF	0.8663	Age, chronic kidney disease, baseline hemoglobin, heart failure
Barsasella, D. et al.[10]	Dataset from Taiwan NHIRD	28,510	883	0.03	RF	0.97	Displacement of lumbar intervertebral disc, cerebral artery occlusion, age, hearing loss
Ye, J. et al.[12]	MIMIC-III	8790	1164	0.132	XGBoost Majority Voting	0.86	DCSI sum, Elixhauser sum, CCI sum, mean glucose
Anand, R. S. et al.[13]	MIMIC-III	3729	382	0.102	RF and LR	0.787	Mean glucose, mean HbA1c, type of admission, severity scores
Yang, B. et al.[14]	MIMIC-IV	2815	395	0.140	RF	0.92	APS III, SOFA, urine output minimum, lactate_max
Qi, J. et al.[15]	MIMIC-IV	5000	896	0.142	RF	0.883	Lactate, age, oxygen saturation, systolic blood pressure
	eICU-CRD	715	727	0.145			
	dtChina	390	69	0.177			

Table 4. Comparison of studies and their main predictors (important features), and model performance in terms of AUC. MIMIC, Medical Information Mart for Intensive Care; SOFA, Sequential Organ Failure Assessment; APS III, Acute Physiology Score (APS) III; eICU-CRD, eICU Collaborative Research Database; dtChina, a large critical care database in China; NHIRD, National Health Insurance Research Database; RF, random forest.

of clinical information regarding the laboratory tests and vital signs that were used in the previous investigations, whereas in our study none of these information were used.

The results of our study show that the top four most important features for predicting one-year mortality are age, duration of diabetes, hypertension, and sex. It is worthwhile to mention that throughout the work, the "importance" value of a feature is a measure of "association" between the feature and the mortality (rather than a notion of "causality"). Nonetheless, from Fig. 2a (and other similar figures in Supplementary Materials), it is observed that the age and the duration of diabetes are directly proportional to a higher mortality. Furthermore, the results show that the lack of hypertension is associated with higher mortality. In this case, hypertension has a paradoxical protective effect[3]. It can be partly explained by the reverse epidemiological phenomenon of standard risk factors in chronic diseases and chronic infections such as HIV/AIDS[25,26]. Previous studies[9–15] have reported various predictors of mortality in diabetes; however, the identified factors have not been consistently replicated across studies, as summarized in Table 4. Age is the only predictor that was consistently shown to be significant in several studies, as well as in ours.

The association between age and diabetes mortality has been extensively studied in the literature. A number of studies have reported that increasing age is associated with a higher risk of diabetes-related mortality[27–29]. Research based on nationwide registers in Denmark showed that individuals who are diagnosed with diabetes at an older age have a higher mortality risk within the first two years after diagnosis[30]. On the other hand, another study showed that individuals diagnosed with type 2 diabetes at a younger age had a greater likelihood of mortality compared to those diagnosed at an older age[31]. Our findings indicate that elderly age at diabetes diagnosis is associated with an elevated risk of mortality.

The association between gender and diabetes mortality has been a topic of interest in recent studies, particularly in the context of gender influence on diabetes management and outcomes. The systematic review and meta-analysis conducted by Wang and colleagues showed that women with diabetes have generally a higher risk of coronary heart disease and all-cause mortality compared to men with the same condition. Specifically, women with diabetes have a 58% greater risk of CHD and a 13% greater risk of all-cause mortality[32]. Another systematic review stated that the additional likelihood of developing cancer and the higher risk of death that comes with having diabetes are slightly more pronounced in women than in men[33]. Although the majority of studies show that women with diabetes have higher risk of mortality than men with the same condition, our results indicates the opposite, which is supported by several studies[34,35]. A study from Germany found that men had a higher mortality rate associated with total T2D compared to women due to a greater relative mortality associated with undiagnosed T2D in men compared to women[35]. One possible explanation for this gender difference could be that women in Germany receive a diagnosis for T2D earlier in the course of the disease than men, which could lead to better management and outcomes. This explanation may also apply to our study, as women in Kazakhstan have greater awareness of the diabetes condition. Moreover, among older people in Kazakhstan, women had significantly higher rates of DM control (31.8%) compared to men (22.6%)[36].

The studies from Scotland and Sweden found that among diabetic patients, women with congestive heart failure (CHF) as a comorbidity have higher mortality rate compared to men with a similar condition[37,38]. It can be related to differences in diabetes management and access to care, as well as biological factors such as hormonal changes. Moreover, type 2 diabetes is associated with a two to four-fold increase in the risk of developing CHF and ischemic stroke[39]. Numerous studies show that patients with diabetes and CHF had a significantly higher risk of all-cause mortality compared to those without CHF, even after adjusting for various clinical and demographic factors[40,41]. The increased mortality risk in patients with both CHF and diabetes may be related to impaired cardiac function, insulin resistance, and chronic inflammation. The results of the current study are consistent with the literature.

Although research shows that comorbid hypertension increases the mortality among diabetes population[42–44], the results of this study indicate the opposite. The management of hypertension in individuals with diabetes can reduce the mortality risk by reducing the risk of developing complications related to both conditions. Effective management of hypertension can help prevent or slow the progression of damage to the blood vessels, reducing the risk of heart attack, stroke, and other cardiovascular complications[45–47]. Studies have shown that good blood pressure control can reduce the risk of cardiovascular disease and mortality in individuals with diabetes. In fact, a blood pressure goal of less than 130/80 mmHg is recommended for individuals with diabetes in order to reduce their risk of cardiovascular complications[48,49]. More profound research on this issue is needed.

The longer duration of hospitalization was significantly associated with severe complications and mortality in the Korean diabetic cohort[50]. A similar tendency was shown in the results of the current study.

Considering the relatively limited number of features (10 attributes presented in Table 1) and their administrative types, the reported range of AUC for the constructed classifiers is indeed a considerable achievement for predicting the one-year mortality of DM patients. That being said, there are a few limitations in our analysis.

From a clinical perspective, one limitation is that our data neither includes laboratory data nor patients' medical history. In addition, the database lacks information on important comorbidities and anthropometric indices such as Alzheimer's disease, renal diseases, amputations, and BMI. Collecting and using this information would potentially boost the performance of our predictive models. Nonetheless, including this information would require running and retraining all our predictive models. At the same time, collecting further detailed patients' medical history from clinical notes available through UNEHS calls for advanced natural language processing. From a machine learning perspective, one limitation of our developed machine learning pipeline is the lack of a feature selection stage. Although this is not a critical stage in the current study due to the large sample size and a small number of features, adding laboratory data and/or the patient's medical history would possibly add a number of additional features. In that case, having a feature selection would be generally expected and help due to the curse of dimensionality in pattern recognition[17] (also known as the peaking phenomenon[51]). We leave these investigations for future studies.

Despite the limitations of the study, there are some advantages that are noteworthy. To begin with, the data utilized in this study was derived from a population-based registry, which provides a substantial amount of information that is representative of a population of roughly half a million data points. Additionally, the data collection period was sufficiently long to encompass prevalent diabetes cases. Additionally, this study is the first of its kind in Central Asia to anticipate the one-year mortality of diabetes patients, and thus contributes significant information to the existing body of literature on this topic. The analysis took into account comorbidities as well as demographic factors. These findings can help in the development of improved protocols and strategies to manage diabetes in healthcare settings, while also considering socio-demographic factors and cultural variations. Moreover, the results may aid in increasing community awareness campaigns and promoting healthy lifestyles to prevent diabetes mortality. Lastly, these results may be useful in initiating further research on the cost-effectiveness of diabetes management in order to assess the economic burden of the disease.

Conclusion

This study developed a comprehensive machine learning platform to predict one-year mortality in patients with DM based on administrative health data. The results of the study showed that the constructed data-driven models can predict one-year mortality in DM patients with an AUC of more than 0.78, which is considered 'fair' (close to 'good') as per objective metrics of diagnostic tests. The study identified age, duration of diabetes, hypertension and sex as the top most important features. These findings could be used to develop better treatment protocols for diabetes patients that take into account socio-demographic and cultural factors. Additionally, the results would help increase community awareness campaigns and promote healthy lifestyles to prevent diabetes mortality.

Overall, this study demonstrates the potential for using machine learning to build accurate predictive models of one-year mortality in DM patients based solely on administrative health data. This focus is warranted because it can help healthcare practitioners to develop individualized treatment plans and interventions to mitigate adverse consequences for high-risk patients. Furthermore, it could aid in resource allocation, as high-risk patients may require more frequent monitoring or follow-up care. Integrating our findings with further information such as laboratory data, patients' medical history, and information on important comorbidities and anthropometric indices could potentially improve the performance of the predictive models in the future.

Materials and methods

Study population. In this dataset, patients with Type 1, Type 2, and other types of diabetes were included. The database was extracted from UNEHS based on International Classification of Diseases 10 (ICD-10) codes for diabetes (Type 1 DM: E10; Type 2 DM: E11). The UNEHS collects individual inpatient and outpatient electronic registries with clinical data. All of these patients were registered between 2014 and 2019. The study involved secondary data that was derived from the UNEHS. Therefore, the requirement for informed consent from study participants was waived by the Nazarbayev University Institutional Review Ethics Committee (NU-IREC 490/18112021). All methods were carried out in accordance with the "Reporting of studies conducted using observational routinely-collected health data" (RECORD) guideline. After cleaning and preprocessing the initial dataset, the final cohort consisted of 472,950 DM patients.

Comorbidity selection. There are several key comorbidities that can affect diabetes mortality. Diabetes can lead to the development of cardiovascular diseases[52,53], cerebrovascular accident (CVA), also known as stroke[54,55] and chronic kidney disease[56]. In addition, diabetes is associated with obesity[57] and hypertension[42,43] with modifiable and non-modifiable risk factors. The UNEHS databases for hypertension[58], CVA[59], coronary heart disease and neoplasms were merged using patients' unique population registry numbers to define comorbid conditions. Diabetes Mellitus (DM) and neoplasms, or tumors, have a complex relationship. While there is evidence to suggest that individuals with DM are at an increased risk for certain types of neoplasms, the underlying mechanisms are not yet fully understood. According to Zhu and Qu[60], the risk of cancers appears to be increased in both type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM). Cancer was also reported to be the second most common cause of death for people with T1DM.

Model training, selection, and evaluation. We used nine classifiers: (1) Gaussian Naïve Bayes (GNB); (2) K-nearest neighbors (KNN); (3) logistic regression with L_2 ridge penalty (LRR); (4) random forest (RF); 95) AdaBoost with decision trees (ADB); (6) gradient boosting with regression trees (GBRT); (7) XGBoost (XGB); (8) linear discriminant analysis (LDA); and (9) perceptron (PER). Table 5 shows the candidate values of hyperparameters that were used in the model selection phase for these classifiers.

In this study, the choice of prediction models was based on several principles. First, we selected model types that cover five commonly known groups: ensemble, Gaussian process, nearest neighbor, linear models, and discriminant analysis. Second, these models have been used extensively in previous studies to predict comorbidities of diabetes, preliminary diagnosis of diabetes, and mortality rate.

Many of our models were used previously for predicting ICU admissions of COVID-19 patients[61]. LDA has been deployed for predicting diabetes through fatty biomarkers in blood[62]. KNN was used to predict diabetes risk of de-identified patients from the Vanderbilt University Medical Center (VUMC) through the use of the Medical Information Mart for Intensive Care III (MIMIC-III) dataset[63]. GBM, XGBoost, AdaBoost, LR, and RF were utilized to predict one-year mortality rate in heart transplantation patients, including those with diabetes mellitus[64]. Similarly, other researchers used random forest and logistic regression to predict mortality rate in diabetic ICU patients[13].

Studies based on the Istituto Clinico Scientifico Maugeri in Italy predicted diabetes complications using LR, NB, and RF[65]. A similar problem was addressed by other researchers that showed the superiority of XGBoost[12].

Classifiers	Hyperparameter	Candidate hyperparameter space
Gaussian Naive Bayes (GNB)	–	–
K Neighbors Classifier (KNN)	Number of neighbours	3, 5
Logistic Regression (LRR)	Penalty	L_2
	Regularization parameter C	100, 10, 1.0, 0.1, 0.01
Random Forest Classifier (RFC)	Number of estimators	10, 100, 1000
	Maximum depth	2, 5, 10, 20, 50
	Maximum features	'auto', 'sqrt', 'log2'
Ada Boost Classifier (ADB)	Number of estimators	10, 100, 1000
	Learning rate	0.001, 0.01, 0.1
Gradient Boost Classifier (GBC)	Number of estimators	10, 100, 1000
	Learning rate	0.001, 0.01, 0.1
XGBoost Classifier (XGB)	Maximum depth	5, 10, 100
	Number of estimators	10, 100, 1000
	Learning rate	0.001, 0.01, 0.1
Linear Discriminant Analysis (LDA)	Solver	'svd', 'lsqr', 'eigen'
	Tolerance	0.00001, 0.0001, 0.0003
Perceptron (PER)	Alpha	0.0001, 0.001, 0.01
	Penalty	L_2, L_1, None

Table 5. Hyperparameters space for grid search with cross-validation model selection.

The XGBoost itself is considered as one of the best predictive models for tabular data, and it has been widely used in Kaggle competitions[66]. In our case, XGBoost and GBRT showed the best performance.

The yearly-specific model selection was performed using stratified 5-fold cross-validation (5-CV) applied to each yearly-specific training set. Figure 3 shows a schematic diagram of the model selection procedure using a 5-fold CV. The stratification is performed to keep the proportion of samples that appear in each fold the same as the original data. This practice gives a better view of the classifier performance in situations when the prior probability of classes is the same as their proportions within the data at hand. Furthermore, in each iteration of 5-CV, we standardize each feature based on the training data used in that iteration (i.e., full training data with one fold excluded). In this regard, we subtract the mean of that feature and divide it by its standard deviation. This way the feature vector is centered around zero and will have a standard deviation of one. The statistics obtained from the iteration-specific training data are then used to normalize the held-out data in the excluded fold. The selection of a year-specific classifier is based on the AUC metric, which is independent of any specific decision threshold used in the classifier[67]. As a result, the decision threshold of selected classifiers is further tuned using training data to maximize the geometric mean of sensitivity and specificity (G-mean). This is in contrast with the usual practice of relying on a classifier "default" decision threshold, which may lead to low G-mean values for highly imbalanced datasets such as ours.

The best year-specific classification rule and the values of its hyperparameters that were identified from the 5-CV model selection were used to train one final year-specific classifier on the entire training set after normalization. To normalize the entire training set, the same normalization that was used in each iteration of 5-CV was used. For prediction and evaluation on the test set, the statistics that were obtained on the training set are used to normalize each observation in the test set before using it as the input for the classifier.

Software and packages. The computations were performed using a virtual server with an AMD Opteron Processor 6174-2.19 GHz with 22 processors, 200 GB of RAM, and storage 3.9 TB, running Windows Server 2019 Standard (64 bit) operating system. The main program was implemented in Python (version 3.10; Python Software Foundation) using open-source packages including sklearn (version 1.1.2), xgboost, numpy, pandas, seaborn, matplotlib, and shap.

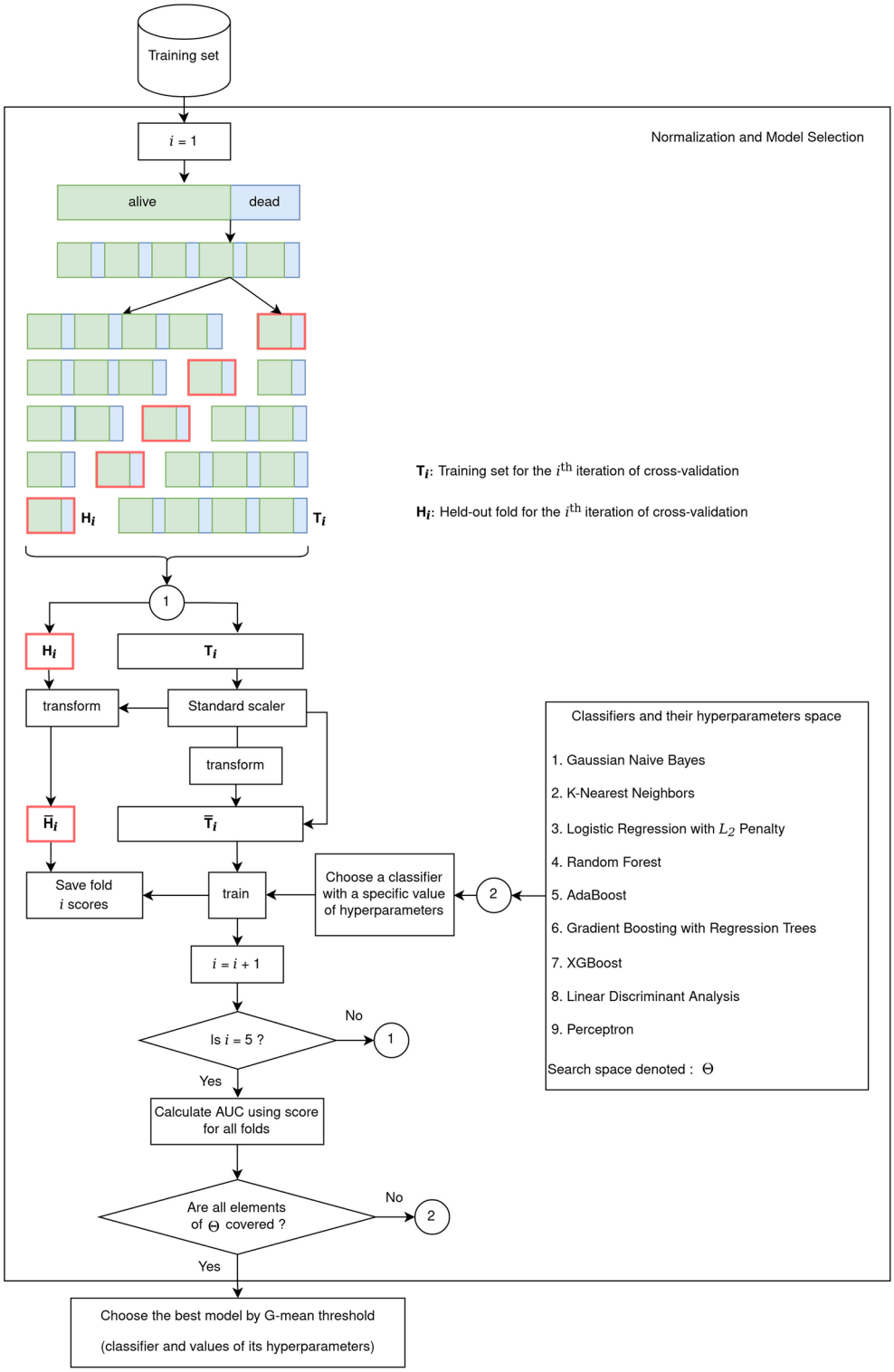

Figure 3. A schematic diagram of the implemented model selection with 5-fold cross-validation.

Ethics approval and consent to participate. The study was approved by the Nazarbayev University Institutional Review Ethics Committee (NU-IREC 203/29112019), with exemption from informed consent.

Data availability

The data that support the findings of this study are available from the Republican Center for Electronic Health of the Ministry of Health of the Republic of Kazakhstan but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from

the corresponding author, Gaipov A., upon reasonable request and with permission of the Ministry of Health of the Republic of Kazakhstan.

Received: 25 December 2022; Accepted: 19 May 2023
Published online: 24 May 2023

References

1. Federation, I. *IDF Diabetes Atlas, Tenth* (International Diabetes, 2021).
2. Cho, N. H. *et al.* IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* **138**, 271–281. https://doi.org/10.1016/j.diabres.2018.02.023 (2018).
3. Gaipov, A. *et al.* Epidemiology of type 1 and type 2 diabetes mellitus in Kazakhstan: Data from unified national electronic health system 2014–2019. https://www.researchsquare.com/article/rs-1432205/v3 (2022).
4. Chentli, F., Azzoug, S. & Mahgoun, S. Diabetes mellitus in elderly. *Indian J. Endocrinol. Metab.* **19**, 744. https://doi.org/10.4103/2230-8210.167553 (2015).
5. Wiens, J. & Shenoy, E. S. Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clin. Infect. Dis.* **66**, 149–153. https://doi.org/10.1093/cid/cix731 (2018).
6. Shailaja, K., Seetharamulu, B. & Jabbar, M. A. Machine Learning in Healthcare: A Review. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* 910–914 (IEEE, 2018). https://doi.org/10.1109/ICECA.2018.8474918
7. VijiyaKumar, K., Lavanya, B., Nirmala, I. & Caroline, S. S. Random Forest Algorithm for the Prediction of Diabetes. In *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)* 1–5 (IEEE, 2019). https://doi.org/10.1109/ICSCAN.2019.8878802.
8. Mujumdar, A. & Vaidehi, V. Diabetes prediction using machine learning algorithms. *Procedia Comput. Sci.* **165**, 292–299. https://doi.org/10.1016/j.procs.2020.01.047 (2019).
9. Lee, S. *et al.* Glycemic and lipid variability for predicting complications and mortality in diabetes mellitus using machine learning. *BMC Endocr. Disord.* **21**, 94. https://doi.org/10.1186/s12902-021-00751-4 (2021).
10. Barsasella, D. *et al.* A machine learning model to predict length of stay and mortality among diabetes and hypertension inpatients. *Med. (B Aires)* **58**, 1568. https://doi.org/10.3390/medicina58111568 (2022).
11. De Silva, K. *et al.* Clinical notes as prognostic markers of mortality associated with diabetes mellitus following critical care: A retrospective cohort analysis using machine learning and unstructured big data. *Comput. Biol. Med.* **132**, 104305. https://doi.org/10.1016/j.compbiomed.2021.104305 (2021).
12. Ye, J., Yao, L., Shen, J., Janarthanam, R. & Luo, Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Med. Inform. Decis. Mak.* **20**, 295. https://doi.org/10.1186/s12911-020-01318-4 (2020).
13. Anand, R. S. *et al.* Predicting mortality in diabetic ICU patients using machine learning and severity indices. *AMIA Jt. Summits Transl. Sci. Proc.* **2017**, 310–319 (2018).
14. Yang, B., Zhu, Y., Lu, X. & Shen, C. A novel composite indicator of predicting mortality risk for heart failure patients with diabetes admitted to intensive care unit based on machine learning. *Front. Endocrinol. (Lausanne)* **13**, 917838. https://doi.org/10.3389/fendo.2022.917838 (2022).
15. Qi, J. *et al.* Machine learning models to predict in-hospital mortality in septic patients with diabetes. *Front. Endocrinol. (Lausanne)* **13**, 1034251. https://doi.org/10.3389/fendo.2022.1034251 (2022).
16. Gusmanov, A. *et al.* Review of the research databases on population-based Registries of Unified electronic Healthcare system of Kazakhstan (UNEHS): Possibilities and limitations for epidemiological research and Real-World Evidence. *Int. J. Med. Inform.* **170**, 104950. https://doi.org/10.1016/j.ijmedinf.2022.104950 (2023).
17. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification* (Willey, 2001).
18. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2009).
19. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class AdaBoost. *Stat. Interface* **2**, 349–360. https://doi.org/10.4310/SII.2009.v2.n3.a8 (2009).
20. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 258. https://doi.org/10.1214/aos/1013203451 (2001).
21. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). https://doi.org/10.1145/2939672.2939785.
22. Anderson, T. W. Classification by multivariate analysis. *Psychometrika* **16**, 31–50. https://doi.org/10.1007/BF02313425 (1951).
23. Pines, J. M., Carpenter, C. R., Raja, A. S. & Schuur, J. D. *Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules*. (Wiley, 2012). https://doi.org/10.1002/9781118482117.
24. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Curran Associates, Inc., 2017). https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
25. Kalantar-Zadeh, K., Block, G., Humphreys, M. H. & Kopple, J. D. Reverse epidemiology of cardiovascular risk factors in maintenance dialysis patients. *Kidney Int.* **63**, 793–808. https://doi.org/10.1046/j.1523-1755.2003.00803.x (2003).
26. Kopple, J. D. The phenomenon of altered risk factor patterns or reverse epidemiology in persons with advanced chronic kidney failure. *Am. J. Clin. Nutr.* **81**, 1257–1266. https://doi.org/10.1093/ajcn/81.6.1257 (2005).
27. Tang, O. *et al.* Mortality implications of prediabetes and diabetes in older adults. *Diabetes Care* **43**, 382–388. https://doi.org/10.2337/dc19-1221 (2020).
28. Forbes, A. Reducing the burden of mortality in older people with diabetes: A review of current research. *Front. Endocrinol. (Lausanne)* **11**, 133. https://doi.org/10.3389/fendo.2020.00133 (2020).
29. Kelly, P. J. *et al.* Predicting mortality in people with Type 2 diabetes mellitus after major complications: A study using Swedish National Diabetes Register data. *Diabetic Med.* **31**, 954–962. https://doi.org/10.1111/dme.12468 (2014).
30. Carstensen, B., Rønn, P. F. & Jørgensen, M. E. Prevalence, incidence and mortality of type 1 and type 2 diabetes in Denmark 1996–2016. *BMJ Open Diabetes Res. Care* **8**, e001071. https://doi.org/10.1136/bmjdrc-2019-001071 (2020).
31. Nanayakkara, N. *et al.* Impact of age at type 2 diabetes mellitus diagnosis on mortality and vascular complications: Systematic review and meta-analyses. *Diabetologia* **64**, 275–287. https://doi.org/10.1007/s00125-020-05319-w (2021).
32. Wang, Y. *et al.* Sex differences in the association between diabetes and risk of cardiovascular disease, cancer, and all-cause and cause-specific mortality: A systematic review and meta-analysis of 5,162,654 participants. *BMC Med.* **17**, 136. https://doi.org/10.1186/s12916-019-1355-0 (2019).
33. Ohkuma, T., Peters, S. A. E. & Woodward, M. Sex differences in the association between diabetes and cancer: A systematic review and meta-analysis of 121 cohorts including 20 million individuals and one million events. *Diabetologia* **61**, 2140–2154. https://doi.org/10.1007/s00125-018-4664-5 (2018).
34. Kanaya, A. M., Grady, D. & Barrett-Connor, E. Explaining the sex difference in coronary heart disease mortality among patients with type 2 diabetes mellitus. *Arch. Internal Med.* **162**, 1737. https://doi.org/10.1001/archinte.162.15.1737 (2002).

35. Röckl, S. *et al.* All-cause mortality in adults with and without type 2 diabetes: Findings from the National Health Monitoring in Germany. *BMJ Open Diabetes Res. Care* **5**, 1. https://doi.org/10.1136/bmjdrc-2017-000451 (2017).
36. Supiyev, A. *et al.* Diabetes prevalence, awareness and treatment and their correlates in older persons in urban and rural population in the Astana region, Kazakhstan. *Diabetes Res. Clin. Pract.* **112**, 6–12. https://doi.org/10.1016/j.diabres.2015.11.011 (2016).
37. MacDonald, M. R. *et al.* Discordant short- and long-term outcomes associated with diabetes in patients with heart failure: Importance of age and sex. *Circ. Heart Fail.* **1**, 234–241. https://doi.org/10.1161/CIRCHEARTFAILURE.108.794008 (2008).
38. Andersson, C. *et al.* Long-term impact of diabetes in patients hospitalized with ischemic and non-ischemic heart failure. *Scand. Cardiovasc. J.* **44**, 37–44. https://doi.org/10.3109/14017430903312438 (2010).
39. Bertoluci, M. C. & Rocha, V. Z. Cardiovascular risk assessment in patients with diabetes. *Diabetol. Metab. Syndr.* **9**, 25. https://doi.org/10.1186/s13098-017-0225-1 (2017).
40. Targher, G. *et al.* In-hospital and 1-year mortality associated with diabetes in patients with acute heart failure: Results from the ESC-HFA Heart Failure Long-Term Registry. *Eur. J. Heart Fail.* **19**, 54–65. https://doi.org/10.1002/ejhf.679 (2017).
41. Dauriz, M. *et al.* Association between diabetes and 1-year adverse clinical outcomes in a multinational cohort of ambulatory patients with chronic heart failure: Results from the ESC-HFA heart failure long-term registry. *Diabetes Care* **40**, 671–678. https://doi.org/10.2337/dc16-2016 (2017).
42. de Boer, I. H. *et al.* Diabetes and hypertension: A position statement by the American Diabetes Association. *Diabetes Care* **40**, 1273–1284. https://doi.org/10.2337/dci17-0026 (2017).
43. Ohishi, M. Hypertension with diabetes mellitus: Physiology and pathology. *Hypertens. Res.* **41**, 389–393. https://doi.org/10.1038/s41440-018-0034-4 (2018).
44. Strange, G. *et al.* Threshold of pulmonary hypertension associated with increased mortality. *J. Am. Coll. Cardiol.* **73**, 2660–2672. https://doi.org/10.1016/j.jacc.2019.03.482 (2019).
45. Mannucci, E., Dicembrini, I., Lauria, A. & Pozzilli, P. Is glucose control important for prevention of cardiovascular disease in diabetes?. *Diabetes Care* **36**, S259–S263. https://doi.org/10.2337/dcS13-2018 (2013).
46. Cameron, A. C., Lang, N. N. & Touyz, R. M. Drug treatment of hypertension: Focus on vascular health. *Drugs* **76**, 1529–1550. https://doi.org/10.1007/s40265-016-0642-8 (2016).
47. Petrie, J. R., Guzik, T. J. & Touyz, R. M. Diabetes, hypertension, and cardiovascular disease: Clinical insights and vascular mechanisms. *Can. J. Cardiol.* **34**, 575–584. https://doi.org/10.1016/j.cjca.2017.12.005 (2018).
48. Whelton, P. K. *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. *J. Am. Coll. Cardiol.* **71**, e127–e248. https://doi.org/10.1161/HYP.0000000000000066 (2018).
49. Leung, A. A. *et al.* Hypertension Canada's 2017 guidelines for diagnosis, risk assessment, prevention, and treatment of hypertension in adults. *Can. J. Cardiol.* **33**, 557–576. https://doi.org/10.1016/j.cjca.2017.03.005 (2017).
50. Yoo, H., Choo, E. & Lee, S. Study of hospitalization and mortality in Korean diabetic patients using the diabetes complications severity index. *BMC Endocr. Disord.* **20**, 122. https://doi.org/10.1186/s12902-020-00605-5 (2020).
51. Zollanvari, A., James, A. P. & Sameni, R. A theoretical analysis of the peaking phenomenon in classification. *J. Classif.* **37**, 421–434. https://doi.org/10.1007/s00357-019-09327-3 (2020).
52. Amiel, S. A. *et al.* Hypoglycaemia, cardiovascular disease, and mortality in diabetes: Epidemiology, pathogenesis, and management. *Lancet Diabetes Endocrinol.* **7**, 385–396. https://doi.org/10.1016/S2213-8587(18)30315-2 (2019).
53. Glovaci, D., Fan, W. & Wong, N. D. Epidemiology of diabetes mellitus and cardiovascular disease. *Curr. Cardiol. Rep.* **21**, 21. https://doi.org/10.1007/s11886-019-1107-y (2019).
54. Tun, N. N., Arunagirinathan, G., Munshi, S. K. & Pappachan, J. M. Diabetes mellitus and stroke: A clinical update. *World J. Diabetes* **8**, 235. https://doi.org/10.4239/wjd.v8.i6.235 (2017).
55. Lau, L., Lew, J., Borschmann, K., Thijs, V. & Ekinci, E. I. Prevalence of diabetes and its effects on stroke outcomes: A meta-analysis and literature review. *J. Diabetes Investig.* **10**, 780–792. https://doi.org/10.1111/jdi.12932 (2019).
56. Winocour, P. H. Diabetes and chronic kidney disease: An increasingly common multi-morbid disease in need of a paradigm shift in care. *Diabetic Med.* **35**, 300–305. https://doi.org/10.1111/dme.13564 (2018).
57. Boles, A., Kandimalla, R. & Reddy, P. H. Dynamics of diabetes and obesity: Epidemiological perspective. *Biochim. Biophys. Acta BBA Mol. Basis Dis.* **1026–1036**, 2017. https://doi.org/10.1016/j.bbadis.2017.01.016 (1863).
58. Yerdessov, S. *et al.* Epidemiology of arterial hypertension in Kazakhstan: Data from unified nationwide electronic healthcare system 2014–2019. *J. Cardiovasc. Dev. Dis.* **9**, 52. https://doi.org/10.3390/jcdd9020052 (2022).
59. Zhakhina, G. *et al.* Incidence and mortality rates of strokes in Kazakhstan in 2014–2019. *Sci. Rep.* **12**, 16041. https://doi.org/10.1038/s41598-022-20302-8 (2022).
60. Zhu, B. & Qu, S. The relationship between diabetes mellitus and cancers and its underlying mechanisms. *Front. Endocrinol. (Lausanne)* **13**, 800995. https://doi.org/10.3389/fendo.2022.800995 (2022).
61. Subudhi, S. *et al.* Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *NPJ Digit. Med.* **4**, 87. https://doi.org/10.1038/s41746-021-00456-x (2021).
62. Yi, L. Z. *et al.* Plasma fatty acid metabolic profile coupled with uncorrelated linear discriminant analysis to diagnose and biomarker screening of type 2 diabetes and type 2 diabetic coronary heart diseases. *Metabolomics* **4**, 30–38. https://doi.org/10.1007/s11306-007-0098-7 (2008).
63. Mani, S., Chen, Y., Elasy, T., Clayton, W. & Denny, J. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu. Symp. Proc.* **2012**, 606–615 (2012).
64. Zhou, Y. *et al.* Prediction of 1-year mortality after heart transplantation using machine learning approaches: A single-center study from China. *Int. J. Cardiol.* **339**, 21–27. https://doi.org/10.1016/j.ijcard.2021.07.024 (2021).
65. Dagliati, A. *et al.* Machine learning methods to predict diabetes complications. *J. Diabetes Sci. Technol.* **12**, 295–302. https://doi.org/10.1177/1932296817706375 (2018).
66. Brownlee, J. *XGBoost with Python: Gradient Boosted Trees with XGBoost and Scikit-learn* (Machine Learning Mastery, 2018).
67. Saarela, M., Ryynänen, O.-P. & Äyrämö, S. Predicting hospital associated disability from imbalanced data using supervised learning. *Artif. Intell. Med.* **95**, 88–95. https://doi.org/10.1016/j.artmed.2018.09.004 (2019).

Acknowledgements

We thank all staff from the Republican Center of Electronic Healthcare for providing data and consultancy.

Author contributions

A.A. implemented the machine learning framework and helped draft the manuscript. G.Z. provided the clinical background and helped draft the manuscript. A.G. participated in data collection and preprocessing. Y.S. participated in data collection. S.Y. participated in data collection. I.A. helped draft the manuscript and implement the machine learning framework. A.K. participated in experimental design and coordination. A.Z. designed the machine learning framework, drafted the manuscript, and participated in experimental design and coordination.

A.G. conceived the study, provided the clinical background, and helped coordinate and draft the manuscript; all authors approved the final version of the manuscript.

Funding
This study was supported by grants from the Nazarbayev University Faculty Development Research Grant Program FDCRGP 2020–2022 (Funder Project Reference: 240919FD3913, title: Aggregation and utilization of the large-scale administrative health data in Kazakhstan for population health research and surveillance) and Nazarbayev University Faculty Development Research Grant Program FDCRGP 2023–2025 (Funder Project Reference: 20122022FD4104, title: In-depth epidemiology and modeling of the 10-year trends of cardiovascular diseases and their complications in Kazakhstan using aggregated big data from the Unified National Electronic Healthcare System). The funder had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript. A.G. is a PI of the project.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-35551-4.

Correspondence and requests for materials should be addressed to A.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.