Healthcare Analytics 5 (2024) 100292

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients

Masoomeh Zeinalnezhad [a,*], Saman Shishehchi [b]

[a] Department of Industrial Engineering, West Tehran Branch, Islamic Azad University, Tehran, Iran
[b] Department of Electrical and Computer Engineering, Buein Zahra Technical University, Buein Zahra, Qazvin, Iran

ARTICLE INFO

Handling Editor: Madijd Tavana

Keywords:
Genetic algorithms
Data mining
Meta-heuristic
Healthcare industry
Hospital readmission
Diabetes

ABSTRACT

Reducing hospital readmission rate is a significant challenge in the healthcare industry for managers and policymakers seeking to improve healthcare and lower costs. This study integrates data mining and meta-heuristic techniques to predict the early readmission probability of diabetic patients within 30 days of discharge. The research dataset was obtained from the UC Irvine Machine Learning Repository, including 101765 instances with 50 features representing patient and hospital outcomes, collected from 130 US hospitals. After data preprocessing, including cleansing, sampling, and normalization, a Chi-square analysis is done to confirm and rank the 20 identified factors affecting the readmission risk. As the algorithms' performance could vary based on the features' characteristics, several classification algorithms, including a Random Forest (RF), Neural Network (NN), and Support Vector Machine (SVM), are employed. Moreover, the Genetic Algorithm (GA) is integrated into the SVM algorithm, called GA-SVM, for hyper-parameter tuning and increasing the prediction accuracy. The performance of the models was evaluated using accuracy, recall, precision, and f-measure metrics. The results indicate that the accuracy of RF, GA-SVM, SVM, and NN are calculated respectively as 74.04 %, 73.52 %, 72.40 %, and 70.44 %. Using GA to adjust c and gamma hyper-parameters led to a 1.12 % increase in SVM prediction accuracy. In response to increasing demand and considering poor hospital conditions, particularly during epidemics, these findings point out the potential benefits of a more tailored methodology in managing diabetic patients.

1. Introduction

Healthcare is one of the most significant challenges in recent decades. Various chronic diseases have emerged; one is diabetes. It is anticipated that the number of diabetic patients will rise to around 643 million by 2030 and up to 783 million in 2045 [1]. Diabetes is common among hospitalized patients and contributes to augmented length of stay, readmission risk, and inferior outcomes [2]. Much research has been done to diagnose or predict it. Applying data mining algorithms in medical records has a significant impact on the field of healthcare. Various classification techniques such as support vector machine, convolutional/artificial neural networks, random forest, logistic regression, and decision tree are employed in the medical field successfully [3]. Recently, Ejiyi et al. [1] proposed a robust model for predicting diabetes mellitus. They employed "Shapley Additive Explanation" to extract feature significance and establish the essential features for fitting random forest, extra tree, xgboost, and adaboost models. Similarly, Islam et al. [4], employing the publicly available Pima Indian Diabetes dataset and applying K-fold cross-validation, hyper-parameter tuning, and various ML classifiers, designed a decision support system for diabetes prediction.

Despite high-quality evidences showing improved clinical outcomes for diabetic patients, many do not receive various preventive and therapeutic interventions. Unexpected readmission and death are persistent among diabetic patients [5]. This could be because of arbitrary diabetes management in hospitals that fail to control glycemic [6]. Failure to deliver appropriate diabetes care increases hospital costs and impacts the mortality and morbidity of the patients who face complications related to diabetes. Hospital readmission raises hospitalization costs and mortality risk [7], influences the hospitals' reputation negatively, and consumes a large amount of medical resources [8]. Predicting readmission possibilities in the early stages prompts more attention to high-risk patients, powers the healthcare system, and reduces healthcare expenditure [9]. It reduces the financial and medical problems of the

* Corresponding author.
E-mail address: m.zeinalnezhad@gmail.com (M. Zeinalnezhad).

https://doi.org/10.1016/j.health.2023.100292
Received 14 August 2023; Received in revised form 5 November 2023; Accepted 15 December 2023
Available online 16 December 2023

healthcare industry to some extent.

The challenge is how to predict the readmission risk with an acceptable accuracy. We face a large volume of data in the healthcare industry, with a wide gap between data collection and interpretation [10]. Data mining could be a great solution [11]. However, we should consider the sensitivity of medicine and its close association with human life, confusion in the definition of data mining, privacy and confidentiality of health data, and the difficulty of changing the habits of healthcare providers from traditional medicine to evidence-based medicine.

The related research confirms that data mining algorithms have successfully predicted the readmission risk. Masip et al. [12] designed a linear logistic regression model to assess mortality and readmission rates and identify predictive features for discharged patients from a first heart failure-associated admission. Sundararaman [10] proposed the hospital readmissions for heart failure using feature selection from unstructured data with a class imbalance in discharge summary notes. Xue et al. [11], using 1170 samples, of which 310 were readmitted within 30 days, designed accurate predictors to capture detailed temporal trends of variables for 30-day readmission risk.

Limited studies have explored data mining techniques for readmission risk prediction, especially those related to diabetes. Hammoudeh et al. [9] combined data engineering and neural networks' ability to predict readmission risk for diabetes patients. They assumed feature selection as part of the deep learning model. They concluded that by feeding the attributes to the deep learning model, the neural networks learn the influence of each attribute and assign the weights consequently. They show that this combination has outperformed compared to other algorithms. SMOTE was chosen in that research to increase the size of the samples. Shancheng Jiang and coworkers [13] developed a risk prediction framework for hospital readmission on various diseases by combining feature selection algorithms and data mining models.

Furthermore [14], reported the performance of several neural and statistical prediction models on a large real dataset. They used a weighted K-nearest neighbors (KNN) algorithm and a regression tree-based rule system. Neto [15] used a data mining algorithm to predict the risk of readmission. The final results revealed that the most efficient algorithm was Random Forest, with 0.898 accuracy. Alajmani [16] compared the performance of several data mining algorithms to predict the risk of diabetes readmission, including logistic regression (LR), multi-layer perceptron (MLP), Naïve Bayesian (NB) classifier, decision tree (DT), and support vector machine (SVM). They concluded that SVM performs best, while the NB classifier and LR analysis are the worst. Maximizing the margin between the support vectors and hyperplane in SVM leads to better classification performance [17]. Im et al. [18] proposed discriminative pattern-based features to improve readmission prediction. They consider associations between features and do not focus on individual features of health conditions. Cui et al. [19] focused on pre-discharge intervention since many readmissions stem from low care quality during hospitalization and a weak discharge process. In medicine, scoring methods are usually used to predict a patient's readmission risk. Scoring methods generally consist of several characteristics associated with readmission. By ranking the patient characteristics, one can classify patients into several groups, such as high-risk and low-risk patients, and then physicians can predict the probability of future readmission. The most well-known scoring methods are the LACE score and the HOSPITAL score. Many studies use these two methods to predict the risk of readmission for a particular disease. Although these methods are more straightforward for clinical doctors, they do not often have satisfactory accuracy [19].

To improve the prediction accuracy of SVM, which is easily affected by the Kernel performance and parameter selection, Cui et al. [19] have compared the performance of several different Kernel functions and have used the Genetic algorithm to tune the sensitive parameters. They have compared the results with other methods such as the LACE score, Decision Tree, Bayesian methods, Logistic regression, and neural network [19]. Bhuvan [20] concluded that the readmission of patients was due to inadequate care during the initial hospitalization. Applying the Bayesian network, Random Forest, AdaBoost, Bayesian algorithm, and neural networks, they found that random forest had the highest accuracy.

Motivated by the current revolution of data mining in numerous fields, especially in the healthcare industry, this research aims to.

1. Integrate data mining algorithms and meta-heuristic techniques to develop a more accurate classifier;
2. Predict the early readmission risk of diabetic patients;
3. Employ a Chi-square analysis to identify/confirm and rank the significant factors affecting the early readmission risk;
4. Apply several classification algorithms such as random forest, neural network, and support vector machine to introduce the best-fitted model; and
5. Adjust the hyper-parameters of the support vector machine using a Genetic Algorithm to increase the prediction performance.

In this study, a publicly available dataset, named "Diabetes 130-US hospitals for years 1999–2008″, taken from the UC Irvine Machine Learning Repository, is analyzed, considering balanced sampling, data preprocessing, normalization, feature selection, and feature ranking.

The rest of this paper is organized as follows: Section 2 introduces the research methodology, including data preprocessing and modeling. Section 3 reports the research findings. Section 4 discusses the results and the contributions of this study. Section 5 concludes the research and suggests future avenues to extend this work and resolve some of its limitations.

2. Materials and methods

The research methodology is presented in Fig. 1 in detail.

The following subsections describe the various research steps regarding Fig. 1.

2.1. Dataset description

The research dataset was obtained from the UC Irvine Machine Learning Repository donated by Strack et al. [6], which can be accessed at https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008. It represents ten years of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes who underwent laboratory medications and stayed up to 14 days. The database consists of 101765 instances collected from 1999 to 2008, including 50 features representing patient and hospital outcomes (See Appendix 1). It contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab tests performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization. They belong to 4 major sections: basic patient information, medical records file of the patient, medication used by the patient, and readmission status of the patient, which is utilized as the target feature.

2.2. Data preprocessing

The preprocessing and data preparation for subsequent data mining steps are among the most critical [21]. Data preprocessing includes various steps, such as data cleansing, data reduction, sampling, and imputing missing data, which will be discussed here.

2.2.1. Data cleansing

The presence of poor-quality data in the dataset, missing values, or even noise may reduce the accuracy of the results, cause incorrect prediction of the readmission status of patients, and slow the computations

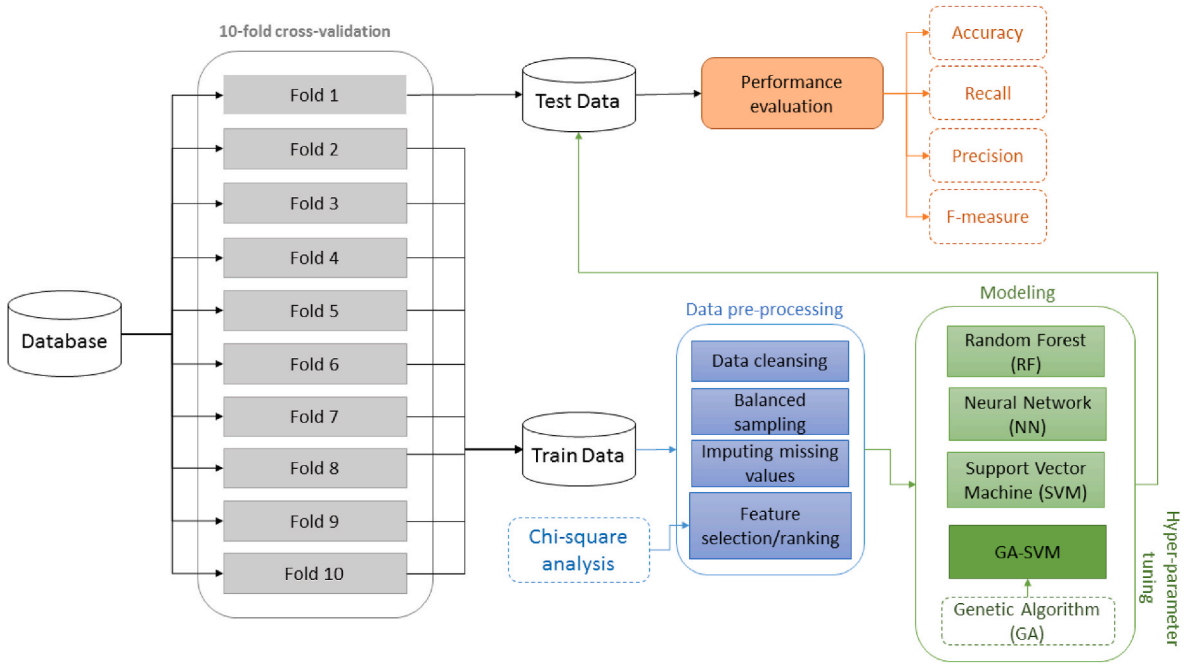

Fig. 1. The research methodology.

[22]. This research aims to determine the readmission status of patients within 30 days of discharge. Among the 101765 patients, almost 35 % are readmitted after 30 days, 11 % are readmitted under 30 days, and 54 % are never readmitted. Therefore, by omitting patients who are readmitted until 30 days after first discharge, we have almost 65 % of the database; namely, our population size is 66221 instances.

2.2.2. *Sampling*

The imbalanced nature of the dataset poses a threat to training any ML-based models [23]. The SMOTE (Synthetic Minority Oversampling TEchnique) is a popular oversampling algorithm to solve the class distribution imbalance problem. However, when we oversample the minority-class samples via SMOTE, new samples are randomly generated between two sampling points and have no controllability, which is expected to generate more abnormal samples [24].

SMOTE generates synthetic data samples similar to those in the dataset without duplicating them. It is a statistical technique to increase the number of cases in the dataset artificially. It is used when there is not enough actual data available. For example, Ray et al. [23] had 377 observations as a minority class and the rest of the 19969 as a majority class. So, using SMOTE, they generated 15080 *synthetic data samples* to have a balanced training dataset comprising 15,457 (44 %) minority and 56 % majority observations.

Fonseca and Bacao [25] expressed that there are three main approaches to addressing imbalanced training data, including cost-sensitive solutions that attribute a higher misclassification cost to the minority class, algorithmic-level solutions to improve the training of the minority class, and re-sampling solutions that generate synthetic minority class observations and remove majority class observations. Re-sampling could be done through under-sampling, over-sampling, or a hybrid approach [25]. While SMOTE performs well on many datasets, it has the drawback of generating noisy samples [26,27] and needs modification for "Nominal" and "Continuous" features [25]. According to the characteristics of the database, from the 66221 available observations, we randomly selected 2500 observations related to the diabetic patients who were readmitted and 2500 related to those not readmitted in 30 days. We had enough *actual* observations to select randomly and have balanced training data. Because of this unique advantage, our designed models are less likely to over-fit and more realistic.

2.2.3. *Missing values management*

The presence of missing values is a regular occurrence in most datasets. Missing values usually significantly affect the results, and various methods exist to resolve this problem. One of these methods is replacing missing values with the mean or mode of that feature, which is used in this research [19]. Two percent of the values of the "Race" feature were missing, so we replaced them with the mode of that feature.

2.3. *Feature selection/confirmation/ranking*

It is unusual that all the variables in the dataset, especially the publicly available ones, help build a model. Redundant variables decrease the model generalization and the overall accuracy of the classifier [28]. Furthermore, adding more variables to a model increases the overall complexity of the model. There were 50 features in the dataset, see Appendix 1, some were useless or did not have helpful information regarding our goal. They must be eliminated to improve the trained model and increase the accuracy [29]. This step aims to obtain a smaller volume of data while, at the same time, achieving analytical results that are similar to those of the original whole dataset and also reducing the complexity of models [30,31]. Our dataset came with variables that require selecting features with contributory factors to the target, known as high dimensionality data. Feature selection is crucial in model design due to its immense influence on model performance by reducing overfitting false alarm rates, minimizing training time, and improving accuracy [32]. In the present research, the medical file of each patient includes 49 features and one label (target/class), indicating whether the patient has been readmitted within 30 days of discharge. In the dataset, variables such as "weight" and "medical expertise" had 97 % and 53 % missing values, respectively; therefore, they cannot be used to build the model due to their missing values. Moreover, the only medication that differs among the patients is "the amount of insulin", whereas the rest is primarily identical among patients. Therefore, the latter does not affect the modeling, and only "Insulin" is selected as a variable. Also, some variables in our dataset do not affect the results. Hence, we must eliminate those variables, including "Encounter ID", "insurance number", and "payment code". Moreover, since "the initial diagnosis" is the primary diagnosis in this research, we only consider it (namely, Diag_1).

In order to make the forecast more realistic and accurate, we considered all significant attributes, unless those which are not helpful, regarding our goal. So, we obtained 20 factors and one variable related to the label of each sample, as shown in Table 1.

One of the main goals of this research is to select the features highly dependent on the target. The Chi-square analysis evaluates the relationship between each feature and the target, i.e., "patient readmission", and weights the features accordingly [33]. The selected 20 features, applying a Chi-square analysis, were scored, ranked, and considered as the essential factors affecting early readmission. This step ensures data conformity with ML algorithms and efficiency in model performance [32]. Spencer et al. [34] experimentally assessed the performance of models derived by machine learning techniques by using relevant features chosen by various feature-selection methods, including Chi-square testing, Relief, and symmetrical uncertainty to create distinctive feature sets. They concluded that the best model they created used a combination of Chi-square feature selection. In addition, Ray et al. [23] extracted the most essential features from the original dataset using Chi-squared testing to get better performance of their ML models. They believed unnecessary features would increase not only the time to train the model but also the danger of overfitting.

Since we had numerical and categorical features in the dataset, Chi-Square analysis was used for feature selection. We did not intend to reduce the dimensions of our problem so that we could get *more realistic results*. We seek the optimal feature combinations to improve the correct prediction of early readmission.

Table 1
Explanation of the selected features.

NO.	Features	Type	Values/Explanation
1	Race	Poly nominal	Oceania, Asia, America, Spain, and Others
2	Gender	Binominal	Male, Female
3	Age	Ordinal	[0–10], (10,20], …, (90–100]
4	Admition_type_id	Integer	1, 2, 3, …, 8 Admission Type
5	Discharge_disposition_id	Integer	1, 2, 3, …, 29 Clearance destination
6	Admission_source_id	Integer	1, 2, 3, …26 Admission source
7	Time_in_hospital	Integer	Number of days from the time of reception to release
8	Num_lab_procedures	Integer	Number of tests performed during the patient's hospitalization
9	Num_procedures	Integer	Number of procedures performed except the tests performed during the patient's hospitalization
10	Num_medications	Integer	Number of different medications administered during the patient's hospitalization
11	Number_outpatient	Integer	Number of outpatients before hospitalization during that year
12	Number_emergency	Integer	Number of emergency visits before hospitalization during that year
13	Number_inpatient	Integer	The number of visits led to hospitalization during that year
14	Diag_1	Integer	Three digits code Initial diagnosis
15	Number_diagoses	Integer	Number of diagnoses from the time of hospitalization
16	Max_glu_serum	Binominal	Yes, No Glucose serum test
17	A1Cresult	Real	A1C test result
18	Insulin	Integer	Amount of Insulin
19	Change	Binominal	Yes, No Change in medicine
20	DiabetesMed	Binominal	Yes, No Prescribing diabetes medications

2.4. Normalization

Normalizing data is a critical step in data mining that can eliminate the differences due to various aspects. This paper used the z-transformation to normalize the dataset. The Normalize operator in Rapid Miner software was used for this step. Fig. 2 presents the results of applying this method versus not applying it for various algorithms.

As shown in Fig. 2, normalizing the data increases the precision, accuracy, and f-measure values of SVM and NN algorithms while it does not affect the performance of RF.

2.5. Modeling

This research applies various classification algorithms, including SVM, RF, and NN, using Rapid Miner software. Moreover, hyperparameter tuning for the SVM is done by integrating a Genetic algorithm using Python software, which identifies the most appropriate "c" and "gamma" parameters. The research steps are shown in Fig. 1.

2.5.1. Support vector machine

SVM has become the most common prediction technique in data mining [31]. The nominal values of variables such as "Race", "Gender", "Age", "Max_glu_serum", "A1Cresult", "Insulin", "Change", and "DiabetesMed" were converted to numerical values. For example, for the variable "Race", which had five different nominal values, a unique 1 to 5 was assigned to each value. Moreover, using the normalization operator via the z-transformation method, the mean and variance of the data were set to zero and 1, respectively.

The critical parameters that significantly affect the efficiency of the SVM algorithm include the type of Kernel function, c, and gamma. The Kernel of this algorithm has various types, such as linear, polynomial, RBF (Radial Basis Function), and sigmoid, as shown in Table 2 [19]. The linear Kernel function is a particular case of the polynomial Kernel function. The linear Kernel function performs well in more straightforward problems, but the polynomial Kernel function can be more useful in complex problems [30]. The RBF Kernel function is another type of Kernel function that is usually used in problems where no prior knowledge of the type and nature of the data is available [35]. Also, the sigmoid Kernel function is a particular case of the RBF Kernel function [36]. The type of function is essential and depends on the type and nature of the problem. Therefore, one cannot introduce a function as the function suitable for SVM, but this choice can vary based on the conditions. Hence, all the above functions have been utilized in this research to select the best Kernel.

The parameter c is the penalty parameter. This parameter controls the balance between the smooth decision boundaries and the classification of training data. The parameter gamma is the Kernel factor for RBF and sigmoid functions. Larger gamma values mean the algorithm performs the fitting precisely according to the training datasets.

2.5.2. Genetic algorithm

In medical research for data mining, some optimization algorithms such as Genetic algorithms, neural networks, and decision trees have been used to improve the SVM algorithm's accuracy [10,19]. In this research, we use a combination of this algorithm with GA in Python software to estimate the best values of the parameters, namely c and gamma [19]. The GA chromosome is encoded in binary mode, and the parameters c and gamma are transformed into binary sequences, then they are connected in series as chromosome sequences [37]. To this end, the initial/default values of c and gamma parameters are imported to GA, and the best values for these parameters are determined using "Pop Crossover", "Mating Pool", "Fitness", and "Mutation" functions. Then, these values are exported to the SVM algorithm as optimal parameters, called the GA-SVM model, and the results are compared to those of other methods. The RBF Kernel function is used for the SVM algorithm due to its superior evaluation criteria compared to linear, polynomial, and

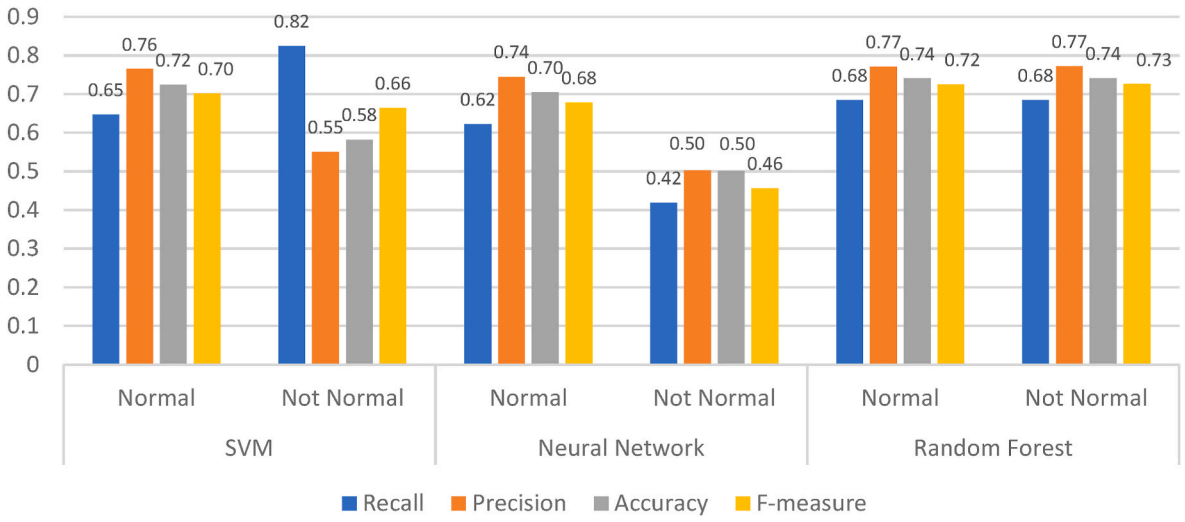

Fig. 2. Normalization results compared with not-normalization.

Table 2
The four different Kernel functions of the SVM algorithm.

Function	Kernel function
$k(x,y) = (x,y)$	Linear
$k(x,y) = ((x,y)+p)^q, q \in N$	Multi-nominal
$k(x,y) = \exp(-\gamma\|x-y\|^2), \gamma > 0$	RBF
$K(x,y) = \tanh(\gamma(x,y)+c), \gamma > 0, c > 0$	Sigmoid

sigmoid [38].

2.5.3. Neural network

In this study, similar to the research done by Bhuvan et al. [20], we considered three layers for our network, including a hidden layer consisting of 8 nodes, an input layer with 20 nodes, and an output layer with two nodes. These two nodes represent the readmission of patients within 30 days of discharge and their non-readmission. The other parameters of this algorithm remain as default [20]. Moreover, this algorithm required all data to be converted to numerical form, similar to the SVM algorithm. Hence, the same technique for converting nominal values to ratio ones has been used in this algorithm. In addition, similar to SVM, different features have been made independent of scale in this method via normalization and play identical roles in the classification.

2.5.4. Random forest

Random forest is considered a supervised learning algorithm. It might be used both for classification and regression problems [39]. Instead of searching for the most critical features during the division of a node, this algorithm seeks the best-specific features among a random set of features [22]. Compared with the decision tree (DT) algorithm, RF selects the observations randomly, decides on the creation features of several trees, and then uses the average of the results [40]. However, random forest most often prevents overfitting by creating a random sub-tree of features and creating smaller trees using this tree [41]. Then, it combines the random sub-trees.

2.6. Validation

The models' performance is evaluated by calculation of the confusion matrix. The performance of each classifier on positive and negative classes is determined by true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Next, each classifier is analyzed using the accuracy, recall, precision, and f-measure metrics [3].

Moreover, the K-fold cross-validation is done by splitting the samples set into K sample subsets and choosing K-1 sample sets to be the training set and the rest to be the test set. Repeating cross-validation K times, each subset is verified once. The results of the test are averaged as a single estimation. This validation technique lies in the repeating use of randomly generated sub-samples for training and validation, and each result is verified once. It prevents overfitting but is time consumption [37]. The most common value considered in related scientific texts for k equals 10 [17]. Using the 10-fold cross-validation, in which 90 % of data is used for model training whereas 10 % of data is used for testing, the accuracy of SVM training is chosen as the fitness function. Repeating the algorithm, the best training model is selected.

3. Results

3.1. Chi-square analysis results

The p-values of all 20 selected features were calculated as less than 0.05, confirming that they all are significant factors affecting the early readmission of diabetic patients. In Table 3, the factors are ordered according to their importance. The most important factor, according to this method, is the "Discharge_disposition_id", which has various states, such as a discharge to home or another section of the hospital, that contribute primarily to the readmission of diabetic patients. The "Admition_type_id" includes the admission of patients from ER or other sections, for instance, and the number of inpatients is adequate in the readmission of diabetic patients.

3.2. Results related to SVM

Fig. 3 shows the results of implementing the SVM algorithm with four different Kernel functions. As mentioned before, the two parameters affecting the performance of this algorithm, i.e., the c and gamma parameters, were manually changed from their default values to 1 and 0.1, respectively, so that better results could be obtained. As seen in the graph, the performance of the SVM algorithm with the RBF Kernel function produces the best results compared to the other three Kernel functions. Hence, we will use the RBF to compare the results of other algorithms to that of the SVM algorithm.

3.2.1. Hyper-parameter tuning results (GA-SVM)

As mentioned previously, to improve the results accuracy and the evaluation metrics of the SVM algorithm, the combination of this algorithm with the Genetic algorithm in Python software has been used with the following results, presented in Table 4.

Table 3
The ranked factors affecting the early readmission risk.

No.	Factor	Weight	No.	Factor	Weight
1	Discharge_disposition_id	806.975	11	Num_medications	55.268
2	Admition_type_id	313.436	12	Number_outpatient	55.216
3	Number_inpatient	258.695	13	Diag_1	46.668
4	Admission_source_id	166.961	14	Insulin	40.277
5	Max_glu_serum	142.192	15	Num_procedures	30.737
6	Num_lab_procedures	80.887	16	Number_emergency	23.528
7	Race	80.290	17	DiabetesMed	14.499
8	Number_diagoses	76.917	18	Change	8.002
9	Age	67.917	19	A1Cresult	1.258
10	Time_in_hospital	59.313	20	Gender	0.985

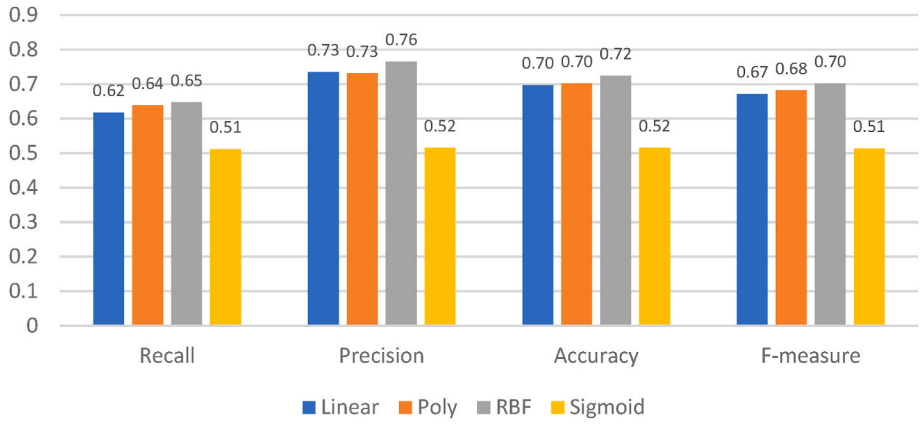

Fig. 3. Validation of SVM algorithm with different Kernel functions.

Table 4
The result of the GA-SVM algorithm.

Designed classifier	c and gamma (initial/ default values)	Recall	Precision	Accuracy	F-measure	c and gamma (tuned values)
GA-SVM	c = 19, γ = 0.6	0.67	0.76	0.74	0.71	c = 1.7203, γ = 3.5424

As shown in Fig. 4, tuning c and gamma for SVM using the Genetic algorithm GA-SVM increases the recall, accuracy, and f-measure.

3.3. Results related to NN

Table 5 displays the results of implementing the Multi-layer Perceptron Neural Network algorithm with training cycle, momentum, and learning rate parameters of 200, 0.01, and 0.9, respectively, with a hidden layer size 1 with eight nodes.

Table 5
The performance evaluation results of the Neural Network algorithm.

Algorithm	Recall	Precision	Accuracy	F-measure
Neural Network	0.6220	0.7447	0.7044	0.6778

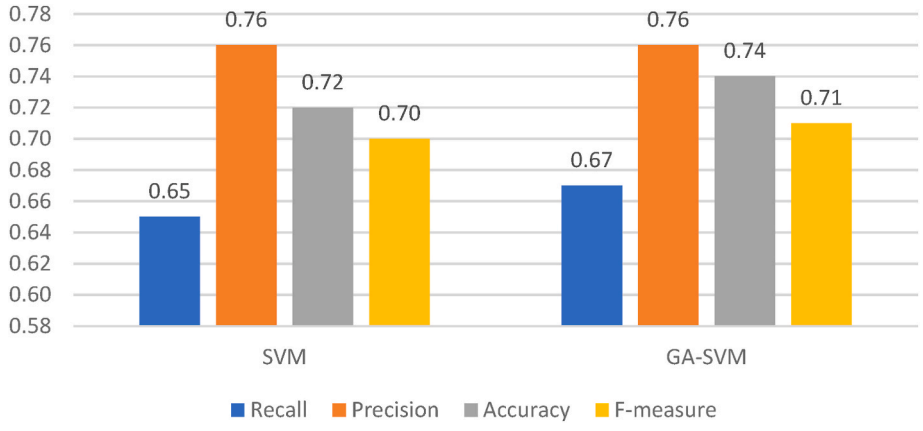

Fig. 4. Relative comparison of SVM with GA-SVM.

3.4. Results related to RF

Table 6 shows the results of implementing the random forest algorithm with three different validation benchmarks.

As shown in Table 6, the Gini index often produces better results than other benchmarks. To compare this algorithm to others, we use the Gini index with values of 150 for the number of trees and 12 for the maximum depth of the trees to result in a higher performance.

Several essential rules obtained from the RF algorithm are presented as follows.

I. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = female and admission source >12), THEN the patient will not be readmitted.
II. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = female and admission source <12 and time in hospital >3.5 and num diagnosis >4.5), THEN the patient will be re-hospitalized.
III. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = female and admission source <12 and time in hospital >3.5 and num diagnosis <4.5 and diabetesMed = Yes), THEN the patient will be re-hospitalized.
IV. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = female and admission source <12 and time in hospital <3.5 and admission type >4.5 and num medication >13.5), THEN the patient will be re-hospitalized.
V. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = female and admission source <12 and time in hospital <3.5 and admission type >4.5 and num medication <13.5), THEN the patient will not be re-hospitalized.
VI. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = female and admission source <12 and time in hospital <3.5 and admission type <4.5 and num medication <13.5), THEN the patient will be re-hospitalized.
VII. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = male and num lab procedures >41.5 and num diagnosis >4.5 and admission type >4.5), THEN the patient will not be re-hospitalized with the possibility of 87 %.
VIII. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = male and num lab procedures >41.5 and num diagnosis >4.5 and admission type <4.5), THEN the patient will be re-hospitalized.
IX. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = male and num lab procedures>41.5 and num diagnosis <4.5), THEN the patient will be re-hospitalized.
X. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = male and num lab procedures>41.5), THEN the patient will be re-hospitalized.
XI. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = male and num lab procedures >41.5 and num diagnosis <4.5), THEN the patient will be re-hospitalized.
XII. IF (num inpatient >0.5 and discharge disposition >15.5 and gender = female and admission source <12 and time in hospital >3.5 and num diagnosis <4.5 and diabetesMed = 0), THEN the patient will not be re-hospitalized.

Table 6
The performance evaluation result of the Random Forest algorithm with different benchmarks.

Benchmark	Recall	Precision	Accuracy	F-measure
Gain Ratio	0.4684	0.8772	0.7014	0.6107
Information Gain	0.6740	0.7708	0.7368	0.7191
Gini Index	0.6840	0.7710	0.7404	0.7249

3.5. Performance evaluation of the models

In this section, we compare the performances of the GA-SVM, SVM, RF, and NN algorithms to identify the best algorithm. As shown in Fig. 5, the RF algorithm has exhibited better performance concerning all the evaluation metrics compared to other algorithms. Therefore, this algorithm can be used to identify patients who are readmitted. The use of genetic algorithms to estimate the best value of SVM hyper-parameters has been effective and has raised the evaluation metrics. In general, the random forest algorithm performs better than others (Fig. 5), which means that this algorithm is more appropriate for predicting the readmission risk.

4. Discussion

In this research, 66221 medical records related to diabetic patients are available. Almost 11200 instances were readmitted in the hospital within 30 days of discharge, and about 55000 instances were never readmitted. To overcome the imbalanced data issues, we used balanced sampling with a sample size of 5000 for each class. As explained in subsection 2.2.2, we did not use “Synthetic Minority Oversampling TEchnique" to solve the class distribution imbalance problem because of the disadvantages of SMOTE, discussed earlier. After data cleansing, imputation of missing values, and normalization, during the feature selection phase, several not helpful features were eliminated regarding our goals. To confirm the selected features (also called factors), a Chi-square analysis was done, and the factors’ scores/weights were calculated. The ranked factors are presented in Table 3.

The classification was done by most employed data mining algorithms, including SVM, random forest, and neural networks. As shown in Table 2, four Kernel functions were used in the SVM algorithm to reach the best function and results. Also, three different benchmarks were used in the random forest algorithm to obtain better results. In addition, the SVM algorithm was combined with the Genetic algorithm in Python software. Through this integration, the hyper-parameter tuning for SVM was done by GA. As shown in Fig. 6, our experimental research has proved that SVM performance is be boosted via the Genetic algorithm approach. It is also apparent that the random forest algorithm has a higher evaluation criterion than the other algorithms, meaning it performs better in identifying the readmission probability of diabetic patients.

Compared to very few similar articles, for example, the research done by Cui et al. [19], our results are acceptable, as presented in Table 7.

The little difference in the results is because Cui et al. [19] have used the SMOTE to balance the classes, which has increased the evaluation metrics by creating false data, but this research has used actual data. Moreover, that research eliminated the data with missing values. However, in the present research, these values were replaced with the mode value of the corresponding feature, which is also essential. On the other hand, considering attributes of minimal contributions can lead to unreal improvement in performance. Nevertheless, regarding our goals, our population, and the characteristics of our samples, we considered all 20 significant features for modeling. However, as shown in Table 7, the evaluation metrics have increased in both research studies employing the strategy of hyper-parameter tuning with GA.

5. Conclusion

Most research on the readmission risk corresponds to heart disease, and only a few papers have investigated the readmission of diabetic patients. Another advantage of this research is that, unlike many others who use methods such as the SMOTE, which attempts to balance the classes by creating unreal data and hence increase the accuracy of these methods, this research attempts to resolve the problem of false data by using a balanced sampling method. Furthermore, applying the Chi-

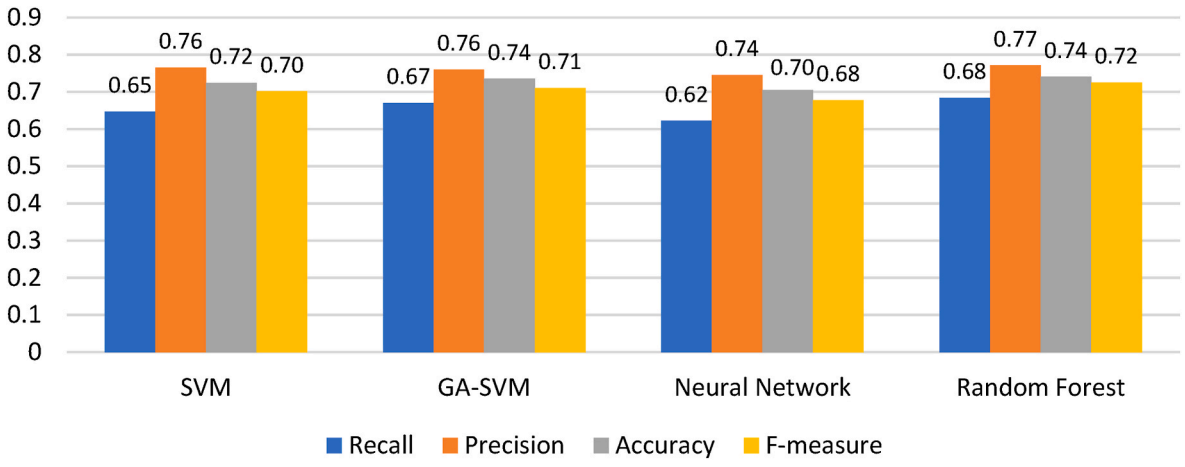

Fig. 5. Validation and comparison of the research models.

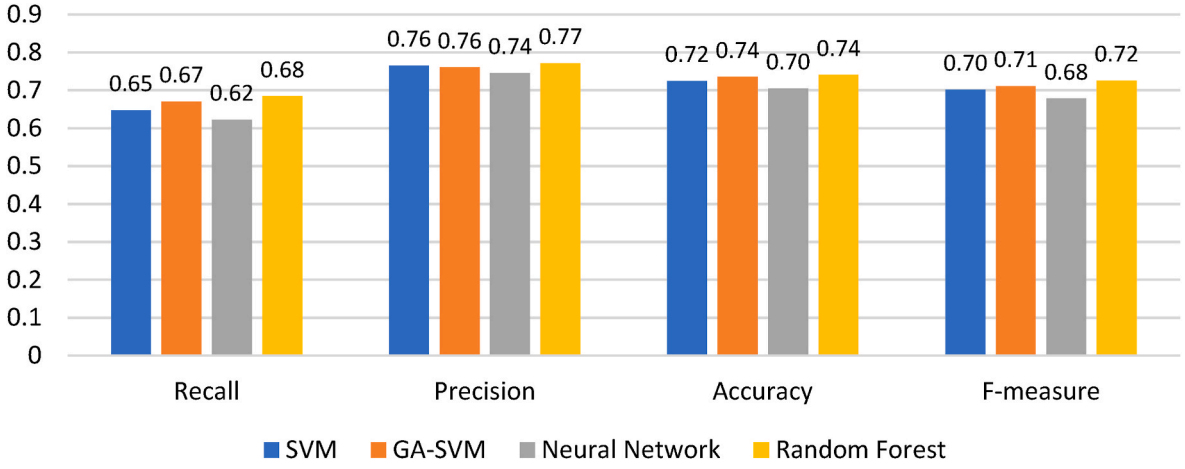

Fig. 6. Performance evaluation of the research algorithms.

Table 7
The comparison of our results with a similar study.

Algorithm	Recall	Accuracy	F-measure
SVM in this research	0.6468	0.7240	0.7009
GA-SVM in this research	0.6700	0.7352	0.7100
SVM in the research done by [19]	0.7213	0.7024	0.7046
GA-SVM in the research done by [19]	0.8289	0.8102	0.8198

square analysis, the selected features were scored, ranked, and confirmed as the critical factors affecting early readmission of diabetic patients.

Considering performance evolution metrics of f-measure, accuracy, precision, and recall, it was concluded that the combination of the SVM with the Genetic algorithm performs better than the SVM algorithm alone. However, the random forest algorithm is more accurate than other algorithms in general and performs better than other algorithms in determining the readmission status of the patients. Despite using different algorithms in this research, one can use other classification techniques or a combination in future research to increase the prediction accuracy. Furthermore, similar databases could be modeled with the designed models in this research, particularly with GA-SVM, and compare the results. GA approach might be used to further enhance the accuracy of other classification algorithms. Other meta-heuristic algorithms, such as particle swarm optimization (PSO) and Grey Wolf Optimizer (GWO), can also be used for this purpose. Future researchers might develop a hybrid feature selection technique, for example, based on PSO and a Chi-square analysis, to enhance the prediction accuracy. The knowledge gained in this research, if converted to a clinical decision support system, is a step in improving social health, preventing a waste of resources, and reducing hospital costs.

Funding

No funding was received to assist with the preparation of this manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) did not use ChatGPT or any other AI-assisted technologies at all.

Declaration of competing interest

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.

Data availability

Data will be made available on request.

Appendix 1. The list of the dataset variables

No.	Variable Name	Role	Type	Description
1	encounter_id	ID		Unique identifier of an encounter
2	patient_nbr	ID		Unique identifier of a patient
3	race	Feature	Categorical	Values: Caucasian, Asian, African American, Hispanic, and other
4	gender	Feature	Categorical	Values: male, female, and unknown/invalid
5	age	Feature	Categorical	Grouped in 10-year intervals: [0, 10), [10, 20), …, [90, 100)
6	weight	Feature	Categorical	Weight in pounds.
7	admission_type_id	Feature	Categorical	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
8	discharge_disposition_id	Feature	Categorical	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
9	admission_source_id	Feature	Categorical	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
10	time_in_hospital	Feature	Integer	Integer number of days between admission and discharge
11	payer_code	Feature	Categorical	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay
12	medical_specialty	Feature	Categorical	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
13	num_lab_procedures	Feature	Integer	Number of lab tests performed during the encounter
14	num_procedures	Feature	Integer	Number of procedures (other than lab tests) performed during the encounter
15	num_medications	Feature	Integer	Number of distinct generic names administered during the encounter
16	number_outpatient	Feature	Integer	Number of outpatient visits of the patient in the year preceding the encounter
17	number_emergency	Feature	Integer	Number of emergency visits of the patient in the year preceding the encounter
18	number_inpatient	Feature	Integer	Number of inpatient visits of the patient in the year preceding the encounter
19	diag_1	Feature	Categorical	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
20	diag_2	Feature	Categorical	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
21	diag_3	Feature	Categorical	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
22	number_diagnoses	Feature	Integer	Number of diagnoses entered into the system
23	max_glu_serum	Feature	Categorical	Indicates the range of the result or if the test was not taken. Values: >200, >300, normal, and none if not measured
24	A1Cresult	Feature	Categorical	Indicates the range of the result or if the test was not taken. Values: >8 if the result was greater than 8 %, >7 if the result was greater than 7 % but less than 8 %, normal if the result was less than 7 %, and none if not measured.
25	metformin	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
26	repaglinide	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
27	nateglinide	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
28	chlorpropamide	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
29	glimepiride	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
30	acetohexamide	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
31	glipizide	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
32	glyburide	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
33	tolbutamide	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
34	pioglitazone	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
35	rosiglitazone	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
36	acarbose	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
37	miglitol	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
38	troglitazone	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed

(*continued on next page*)

(*continued*)

No.	Variable Name	Role	Type	Description
39	tolazamide	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
40	examine	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
41	citoglipton	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
42	insulin	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
43	glyburide-metformin	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
44	glipizide-metformin	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
45	glimepiride-pioglitazone	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
46	metformin-rosiglitazone	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
47	metformin-pioglitazone	Feature	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
48	change	Feature	Categorical	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: change and no change
49	diabetesMed	Feature	Categorical	Indicates if there was any diabetic medication prescribed. Values: yes and no
50	readmitted	Target	Categorical	Days to inpatient readmission. Values: <30 if the patient was readmitted in less than 30 days, >30 if the patient was readmitted in more than 30 days, and No for no record of readmission.

References

[1] C. J. Ejiyi, Z. Qin, J. Amos, M.B. Ejiyi, A. Nnani, T. U. Ejiyi, V.K. Agbesi, C. Diokpo, C. Okpara, A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms, Healthcare Analytics 3 (2023), 100166, https://doi.org/10.1016/j.health.2023.100166. ISSN 2772-4425.

[2] Benjamin Sly, Anthony W. Russell, Clair Sullivan, Digital interventions to improve safety and quality of inpatient diabetes management: a systematic review, Int. J. Med. Inf. 157 (2022), 104596, https://doi.org/10.1016/j.ijmedinf.2021.104596. ISSN 1386-5056.

[3] Nitish Biswas, Khandaker Mohammad Mohi Uddin, Sarreha Tasmin Rikta, Samrat Kumar Dey, A comparative analysis of machine learning classifiers for stroke prediction: a predictive analytics approach, Healthcare Analytics 2 (2022), 100116, https://doi.org/10.1016/j.health.2022.100116. ISSN 2772-4425.

[4] R. Islam, A. Sultan, N. Tuhin, S.H. Saikat, M.R. Islam, Clinical decision support system for diabetic patients by predicting type 2 diabetes using machine learning algorithms, Journal of Healthcare Engineering (2023), https://doi.org/10.1155/2023/6992441. Article ID 6992441 |.

[5] Mengji Chen, Taj Malook, Ateeq Ur Rehman, Yar Muhammad, Mohammad Dahman Alshehri, Aamir Akbar, Muhammad Bilal, A. Muazzam, Khan, Blockchain-Enabled healthcare system for detection of diabetes, J. Inf. Secur. Appl. 58 (2021), 102771, https://doi.org/10.1016/j.jisa.2021.102771. ISSN 2214-2126.

[6] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, John N. Clore, Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records, BioMed Res. Int. 2014 (2014). Article ID 781670, 11 pages.

[7] in-Jung Kim, Robert H. Aseltine, Sara R. Tabtabai, Understanding the burden of 30-day readmission in patients with both primary and secondary diagnoses of heart failure: causes, timing, and impact of Co-morbidities, Am. J. Cardiol. (2023), https://doi.org/10.1016/j.amjcard.2023.09.086. ISSN 0002-9149.

[8] L.A. Hernández, L. Guilbert, E.M. Sepúlveda, F. Rodríguez, F. Peñuñuri, V. H. García, C. Zerrweck, Causes of revisional surgery, reoperations, and readmissions after bariatric surgery, Rev. Gastroenterol. México 88 (3) (2023) 232–237, https://doi.org/10.1016/j.rgmxen.2021.12.006. ISSN 2255-534X.

[9] A. Hammoudeh, G. Al-Naymat, I. Ghannam, N. Obied, Predicting hospital readmission among diabetics using deep learning, Procedia Comput. Sci. 141 (2018) 484–489.

[10] A. Sundararaman, S.V. Ramanathan, R. Thati, Novel approach to predict hospital readmissions using feature selection from unstructured data with class imbalance, Big data research 13 (2018) 65–75.

[11] Y. Xue, D. Klabjan, Y. Luo, Predicting ICU readmission using grouped physiological and medication trends, Artif. Intell. Med. 95 (2019) 27–37.

[12] Joan Masip, Francesc Formiga, Josep Comín-Colet, Xavier Corbella, Short term prognosis of heart failure after first hospital admission, Med. Clínica 154 (2) (2020) 37–44, https://doi.org/10.1016/j.medcle.2019.03.029. ISSN 2387-0206.

[13] S. Jiang, K.-S. Chin, G. Qu, K.L. Tsui, An integrated machine learning framework for hospital readmission prediction, Knowl. Base Syst. 146 (2018) 73–90.

[14] A. Garmendia, M. Graña, J.M. Lopez-Guede, S. Rios, Neural and statistical predictors for time to readmission in emergency departments: a case study, Neurocomputing 354 (2019) 3–9.

[15] C. Neto, et al., Different scenarios for the prediction of hospital readmission of diabetic patients, J. Med. Syst. 45 (1) (2021) 1–9.

[16] S. Alajmani, H. Elazhary, Hospital readmission prediction using machine learning techniques, Hospital 10 (4) (2019).

[17] Gupta Himanshu Gupta, Himanshu Singh, Anil Kumar, Texture and radiomics inspired data-driven cancerous lung nodules severity classification, Biomed. Signal Process Control 88 (Part A) (2023), 105543, https://doi.org/10.1016/j.bspc.2023.105543. ISSN 1746-8094.

[18] S.J. Im, et al., Hospital readmission prediction using discriminative patterns, in: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2020.

[19] S. Cui, et al., An improved support vector machine-based diabetic readmission prediction, Comput. Methods Progr. Biomed. 166 (2018) 123–135.

[20] M.S. Bhuvan, A. Kumar, A. Zafar, V. Kishore, Identifying Diabetic Patients with High Risk of Readmission, 2016 arXiv preprint arXiv:1602.04257.

[21] A.S. Fialho, et al., Data mining using clinical physiology at discharge to predict ICU readmissions, Expert Syst. Appl. 39 (18) (2012) 13158–13165.

[22] N. Kaieski, et al., Application of artificial intelligence methods in vital signs analysis of hospitalized patients: a systematic literature review, Appl. Soft Comput. 96 (2020), 106612.

[23] S. Ray, K. Alshouiliy, A. Roy, A. AlGhamdi, D.P. Agrawal, Chi-Squared Based Feature Selection for Stroke Prediction Using AzureML," 2020 Intermountain Engineering, Technology and Computing, (IETC), Orem, UT, USA, 2020, pp. 1–6, https://doi.org/10.1109/IETC47856.2020.9249117.

[24] Qiushi Liu, Yiguo Xue, Guangkun Li, Daohong Qiu, Weimeng Zhang, Zhuangzhuang Guo, Zhiqiang Li, Application of KM-SMOTE for rockburst intelligent prediction, Tunn. Undergr. Space Technol. 138 (2023), 105180, https://doi.org/10.1016/j.tust.2023.105180. ISSN 0886-7798.

[25] Joao Fonseca, Fernando Bacao, Geometric SMOTE for imbalanced datasets with nominal and continuous features, Expert Syst. Appl. 234 (2023), 121053, https://doi.org/10.1016/j.eswa.2023.121053. ISSN 0957-4174.

[26] Pengfei Sun, Zhiping Wang, Liyan Jia, Zhaohui Xu, Smote-kTLNN, A hybrid re-sampling method based on SMOTE and a two-layer nearest neighbor classifier, Expert Syst. Appl. 238 (2023), https://doi.org/10.1016/j.eswa.2023.121848. Part A, 2024, 121848, ISSN 0957-4174.

[27] Hongjiao Guan, Long Zhao, Xiangjun Dong, Chuan Chen, Extended natural neighborhood for SMOTE and its variants in imbalanced classification, Eng. Appl.

Artif. Intell. 124 (2023), 106570, https://doi.org/10.1016/j.engappai.2023.106570. ISSN 0952-1976.
[28] Atefeh Mansoori, Masoomeh Zeinalnezhad, Leila Nazarimanesh, Optimization of tree-based machine learning models to predict the length of hospital stay using genetic algorithm, Journal of Healthcare Engineering (2023), https://doi.org/10.1155/2023/9673395.
[29] Jianhua Dai, Weiyi Huang, Chucai Zhang, Jie Liu, Multi-label feature selection by strongly relevant label gain and label mutual aid, Pattern Recogn. 145 (2023), https://doi.org/10.1016/j.patcog.2023.109945, 2024, 109945, ISSN 0031-3203.
[30] M.E. Hossain, S. Uddin, A. Khan, Network analytics and machine learning for predictive risk modeling of cardiovascular disease in patients with type 2 diabetes, Expert Syst. Appl. 164 (2021), 113918.
[31] I. Kavakiotis, et al., Machine learning and data mining methods in diabetes research, Comput. Struct. Biotechnol. J. 15 (2017) 104–116.
[32] Love Allen Chijioke Ahakonye, Cosmas Ifeanyi Nwakanma, Jae-Min Lee, Dong-Seong Kim, SCADA intrusion detection scheme exploiting the fusion of modified decision tree and Chi-square feature selection, Internet of Things 21 (2023), 100676, https://doi.org/10.1016/j.iot.2022.100676. ISSN 2542-6605.
[33] M.M. Islam, M.J. Rahman, D.C. Roy, M. Maniruzzaman, Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach, Diabetes Metabol. Syndr.: Clin. Res. Rev. 14 (3) (2020) 217–219.
[34] R. Spencer, F. Thabtah, N. Abdelhamid, M. Thompson, Exploring feature selection and classification methods for predicting heart disease, DIGITAL HEALTH 6 (2020), https://doi.org/10.1177/2055207620914777.
[35] A. Onan, Mining opinions from instructor evaluation reviews: a deep learning approach, Comput. Appl. Eng. Educ. 28 (1) (2020) 117–138.
[36] L. Turgeman, J.H. May, A mixed-ensemble model for hospital readmission, Artif. Intell. Med. 72 (2016) 72–82.
[37] S. Yuanyuan, W. Yongming, G. Lili, M. Zhongsong, J. Shan, The comparison of optimizing SVM by GA and grid search. 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), 2017, pp. 354–360, https://doi.org/10.1109/ICEMI.2017.8265815. Yangzhou, China.
[38] G. Swapna, R. Vinayakumar, K. Soman, Diabetes detection using deep learning algorithms, ICT express 4 (4) (2018) 243–246.
[39] O. Ben-Assuli, J.R. Vest, Data mining techniques utilizing latent class models to evaluate emergency department revisits, J. Biomed. Inf. 101 (2020), 103341.
[40] M. Fratello, R. Tagliaferri, Decision trees and random forests, Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics 1 (2018) 3.
[41] X. Zhou, et al., Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree, Reliab. Eng. Syst. Saf. 200 (2020), 106931.