pHealth 2020
B. Blobel et al. (Eds.)

doi:10.3233/SHTI200628

Identification of Diabetes Risk Factors in Chronic Cardiovascular Patients

Oleg METSKER[a, 1], Kirill MAGOEV[a], Stanislav YANISHEVSKIY[a], Alexey YAKOVLEV[a], Georgy KOPANITSA[b] and Nadezhda Zvartau[a]
[a] *Almazov National Medical Research Centre, Saint Petersburg, Russia*
[b] *ITMO University, Saint Petersburg, Russia*

Abstract. Specific predictive models for diabetes polyneuropathy based on screening methods, for example Nerve conduction studies (NCS, can reach up to AUC 65.8 – 84.7 % for the conditional diagnosis of DPN in primary care. Prediction methods that utilize data from personal health records deal with large non-specific datasets with different prediction methods. Li et al. utilized 30 independent variables, which allowed to implement a model with AUC = 0.8863 for a Multilayer perceptron (MLP). Linear regression (LR) based methods produced up to AUC = 0.8 %. This way, modern data mining and computational methods can be effectively adopted in clinical medicine to derive models that use patient-specific information to predict the development of diabetic polyneuropathy, however, there still is a space to improve the efficiency of the predictive models. The goal of this study is the implementation of machine learning methods for early risk identification of diabetes polyneuropathy based on structured electronic medical records. It was demonstrated that the machine learning methods allow to achieve up to 0.7982 precision, 0.8152 recall, 0.8064 f1-score, 0.8261 accuracy, and 0.8988 AUC using the neural network classifier.

Keywords. Polyneuropathy, machine learning, neural networks, risk management

Introduction

Every third patient with diabetes suffers from diabetic polyneuropathy (DPN) [1]. This complication is a severe problem for a chronic patient because it can be accompanied by neuropathic pain and lead to a decrease in the quality of life. Neuropathic pain significantly disrupts sleep, negatively affects daily activity and life satisfaction. The earlier the diagnosis is found, the easier the disease can be managed. Several studies on the prediction of diabetes mellitus that utilize Linear regression (LR) [2,3] developed predictive models based on records with up to 967 independent variables available from large electronic health record archives resulted in up to 0.80% Area Under The Curve Receiver Operating Characteristics (AUC-ROC) [2] and enabling to formulate the most likely trajectory of comorbidities [3]. Application of the ensemble models approach based on the naïve Bayes, Linear regression (LR), Instance-Based Learner, support vector machine (SVM) [4], artificial neural networks (ANN) [5], decision trees [6], and random forest (RF) help to increase the accuracy of diabetes prediction with the accuracy of up to 74% [7,8].

[1] Corresponding Author. Oleg Metsker, Almazov National Medical Research Centre, Saint-Petersburg, Russia; E-mail: olegmetsker@gmail.com

Specific predictive models for diabetes polyneuropathy based on screening methods, for example Nerve conduction studies (NCS) [9], can reach up to AUC 65.8 – 84.7 % for the conditional diagnosis of DPN in primary care. Prediction methods that utilize data from personal health records deal with large non-specific datasets with different prediction methods. Li et al [10] utilized 30 independent variables, which allowed to implement a model with AUC = 0.8863 for a Multilayer perceptron (MLP). Linear regression (LR) based methods [2,11] produced up to AUC = 0.8 %.

This way, modern data mining and computational methods can be effectively adopted in clinical medicine to derive models that use patient-specific information to predict the development of diabetic polyneuropathy, however, there still is a space to improve the efficiency of the predictive models. The goal of this study is the implementation of machine learning methods for early risk identification of diabetes polyneuropathy based on structured electronic medical records.

1. Methods

The dataset contains a total number of 238590 laboratory records. Each record contains patient and episode identifiers, a timestamp, and a varying number of measured parameters, the total number of which is 31 (27 laboratory tests, retinopathy, nephropathy, age, and gender). The records cover the time period from 03-07-2009 to 22-08-2017 and include information about 5846 diabetic patients. Diagnosis served as the source of information about the target class values. We removed the patients with insufficient amount of data (< 6 parameters) and no data for the retinopathy and nephropathy from the dataset. We employed three approaches to preprocessing of the time series information:

- Replacement of each time series with its last value;
- Replacement of each time series with three of its characteristics: mean, length, and standard deviation;
- Replacement of each time series with its maximum, with the exception of hemoglobin time series, which were replaced with their minimums.

After that, we applied two strategies of dealing with missing values to ensure that all the patients have the same set of variables:

- Replacing the missing data with medians of the corresponding parameters;
- Deletion of parameters having too many missing values (>500 by default) and removal of all the patients having any missing values in the remaining parameters.

To identify subclasses within the class of patients with diabetes, the classification problem was solved. The scikit-learn library method of T-distributed Stochastic Neighbor Embedding (T-SNE) was used. All the parameters of the dataset were used for teaching the clustering model. The features were selected on the basis of correlation analysis of the target class. We used the scikit-learn library to calculate Pearson’s correlations of the features. The classification task for predicting the development of polyneuropathy was solved using the scikit-learn library. Laboratory data, patient age and gender were used as predictors. Each experiment ran in the setting of stratified 5-fold cross-validation (i.e., random 80% of patients were used for training and 20% for testing, target class ratios in the folds were preserved). For the performance assessment of SVM and DT classifiers, we ran it 100 times; and 100 x 5-fold cross-validation with total of 500 predictions. As an additional performance assessment score, we used the AUC of ROC. The AUC was calculated based on an average of 5 curves (one curve per fold in the setting of 5-fold cross-validation).

2. Results

After the patients having insufficient amount of data were removed from the dataset, it contained 5425 patients, 2342 (43.17%) of which did and 3083 (56.83%) did not develop polyneuropathy. Applying three ways to preprocess the time series and two approaches of handling missing data resulted in six datasets to apply our methods to. The T-SNE and DBSCAN algorithm was applied with parameters epsilon 0.4 and neighborhood threshold 35 to the patient vectors, containing creatinine and neutrophils values. Some examples of resulting clusters are presented in Figure 1. The initial cluster was removed from the picture from the picture. As a result of t- SNE clustering, 5 subclasses were identified (Table 1).

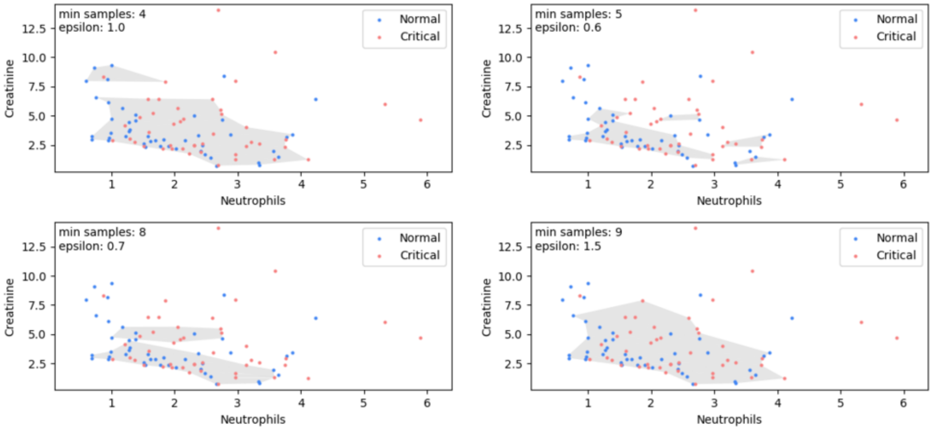
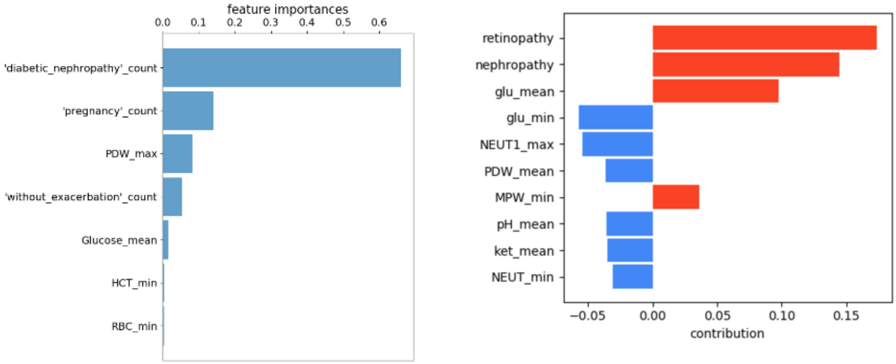


Figure 1. Outlier clustering example results

Table 1. Clusters description of T-SNE method.

Cluster No.	Count	Polyneuropahy %	Age, median	Age Standard deviation	Males, %	Females, %
#0	1877	59	65	9.79	0	100
#1	1628	43	60	13.74	100	0
#2	101	40	61	9.14	100	0
#3	930	22	31	6.53	0	100
#4	119	33	11	6.10	100	0
#5	111	23	39	20.29	99	1

There are two large clusters (#0 and #3) that are opposite to the percentage of patients with polyneuropathy, namely the smallest and the largest rate of polyneuropathy. Cluster # 0 has an increased rate of patients with polyneuropathy: 59 %. In cluster #3, on the contrary, the reduced number of patients with polyneuropathy is 22 %. However, these are both female clusters. In male clusters #1 and #2, the percentage is close to the average. Cluster #4 is slightly below average. In clusters 0, 1, and 2, patients are about the same age. The difference in cluster 1 is that there is an outlier on the left, and it is male in contrast to cluster 0. Cluster 3 patients are young men, younger than patients cluster 0, 1, and 2, and also male. Cluster 4 are young male patients, which explains their low percentage of polyneuropathy. Cluster 5 patients of different ages with a small portion of polyneuropathy are mostly men. It is evident that the female gender and the age of more than 50 years is a risk factor for the development of polyneuropathy in patients with diabetes mellitus. However, there is a particular group of young men (cluster 4) with the probability of polyneuropathy. As a result of the cluster analysis, several groups of patients were identified: younger men and women (clusters 4 and 3, the share of polyneuropathy is significantly lower), older men and women (clusters 1 and 2: men, 0: women, the share of polyneuropathy is significantly higher). The study of the peculiarities of clustering of the demographic and laboratory indices under study revealed the following regularities. Higher blood glucose (Glu_max) figures are associated with the development of polyneuropathy in older men. The higher the age, the less the antiatherogenic potential decreases (HDLs decrease), and the higher the rate of renal excretion (creatinine increases), both of which are associated with the development of polyneuropathy, and the detection of comorbid polyneuropathy, the lower the HDLs. This is a very important finding, since until recently, all attention has been paid only to the content of HDLs and triglycerides. Figure 2(a) shows the features' importance for the decision tree trained further. Figure 2(b) shows an example of a correlation matrix for selected parameters of the patient and polyneuropathy.



a) b)

Figure 2. Features importance for a decision tree (a). ANN interpretation (b)

The most significant features were those of neutrophils and glucose level in the blood and the urine. The correlation matrix shows the high correlation of MPW, RBC, HGB, and polyneuropathy. The results of the correlation analysis have an important clinical significance. A positive correlation of polyneuropathy presence with the patient's age was demonstrated. The mean platelet volume (MPV) has a positive correlation with glucoseuria and hyperglycemia. Previously, researchers proved a dependency between increased MPV and insufficient glycemic control [12]. Moreover, when observing patients receiving optimal treatment, evidence of a decrease in MPV following the normalization of glycaemia was demonstrated [13]. If we discuss the main features (indicators) affecting the development of polyneuropathy in diabetes, the presence of a decrease in the relative number of segmented neutrophils, which supports the postulate of the main contribution of monocyte cells in the process of inflammation development

[13], as well as signs of unsatisfactory glycemic control manifested in glucoseuria and hyperglycemia, become clinically important events. Performance measurements are presented in Table 2.

Table 2. Performance evaluation with comorbidities

	Precision	Recall	F1-score	Accuracy	AUC
SVM	**0.8328**	0.7221	0.7734	0.8120	0.8922
DT	0.7795	0.7864	0.7823	0.8054	0.8644
ANN	0.7982	**0.8152**	**0.8064**	**0.8261**	**0.8988**
LR	0.7934	0.8134	0.8031	0.8228	0.8926
Logistic	0.7961	0.8115	0.8036	0.8238	0.8941

3. Discussion and Conclusion

Clustering shows differences within patients with diabetes mellitus. Decision trees demonstrate pathophysiological mechanisms. It is possible to identify, under what conditions the current risks of polyneuropathy play the most significant role. It was demonstrated that inclusion of just two features, namely “nephropathy” and “retinopathy”, allows to increase the performance even further, achieving up to 0.7982 precision, 0.8152 recall, 0.8064 f1-score, 0.8261 accuracy, and 0.8988 AUC using the neural network classifier. Our results indicate the feasibility of creating a statistical model for predicting the development of diabetic polyneuropathy in patients with type 2 diabetes mellitus for prescribing or adjusting therapeutic strategies to reduce the risk of peripheral nerve damage, which should improve the quality of life of patients with diabetes.

Acknowledgements

The research was supported by Russian Science Foundation (project No. 17-15-01177). The work of Georgy Kopanitsa was financially supported by the Government of the Russian Federation through the ITMO fellowship and professorship program.

References

[1] Izenberg A, Perkins BA, Bril V. Diabetic Neuropathies. Semin Neurol 2015; 35: 424–430.

[2] Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data. 2015 Dec;3,4:277-87. doi: 10.1089/big.2015.0020.

[3] Oh W, Kim E, Castro MR, Caraballo PJ, Kumar V, Steinbach MS, Simon GJ. Type 2 Diabetes Mellitus Trajectories and Associated Risks. Big Data. 2016 Mar 1; 4,1: 25–30. doi: 10.1089/big.2015.0029.

[4] Zhang X. Support Vector Machines. In: Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer US, 2017: 1214–1220.

[5] Artificial Neural Networks. In: Encyclopedia of Machine Learning and Data Mining, 65–66. Boston, MA: Springer US; 2017.

[6] Fürnkranz J. Decision Tree. In: Encyclopedia of Machine Learning and Data Mining, 330–335. Boston, MA: Springer US; 2017.

[7] Bashir S, Qamar U, Khan FH. IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. J Biomed Inform. 2016 Feb;59:185-200. doi: 10.1016/j.jbi.2015.12.001.

[8] Ozcift A, Gulten A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput Methods Programs Biomed. 2011 Dec;104,3:443-51. doi: 10.1016/j.cmpb.2011.03.018.

[9] Olaleye D, Perkins BA, Bril V. Evaluation of three screening tests and a risk assessment model for diagnosing peripheral neuropathy in the diabetes clinic. Diabetes Res Clin Pract Nov 2001; 54,2: 115-128. https://doi.org/10.1016/S0168-8227(01)00278-9.

[10] Li CP, Zhi XY, Ma J, Cui Z, Zhu ZL, Zhang C, Hu LP. Performance comparison between logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. Chin Med J (Engl). 01 Mar 2012; 125,5:851-857.

[11] Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, De Cata P, Chiovato L, Bellazzi R. Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol. 2018 Mar;12,2:295-302. doi: 10.1177/1932296817706375.

[12] Coban E, Bostan F, Ozdogan M. The mean platelet volume in subjects with impaired fasting glucose. Platelets 2006; 17: 67–69.

[13] Demirtunc R, Duman D, Basar M, Bilgi M, Teomete M, Garip T. The relationship between glycemic control and platelet activity in type 2 diabetes mellitus. J Diabetes Complications. Mar-Apr 2009;23,2:89-94. doi: 10.1016/j.jdiacomp.2008.01.006.