Prediction of diabetic patient readmission using machine learning

Juan Camilo Ramírez
Facultad de Ingeniería de Sistemas
Universidad Antonio Nariño
Bogotá, Colombia
juan.ramirez@uan.edu.co

David Herrera
Facultad de Ingeniería de Sistemas
Universidad Antonio Nariño
Bogotá, Colombia
herrera78@uan.edu.co

***Abstract*—Hospital readmissions pose additional costs and discomfort for the patient and their occurrences are indicative of deficient health service quality, hence efforts are generally made by medical professionals in order to prevent them. These endeavors are especially critical in the case of chronic conditions, such as diabetes. Recent developments in machine learning have been successful at predicting readmissions from the medical history of the diabetic patient. However, these approaches rely on a large number of clinical variables thereby requiring deep learning techniques. This article presents the application of simpler machine learning models achieving superior prediction performance while making computations more tractable.**

***Index Terms*—diabetes, hospital readmission, neural network, random forest, logistic regression**

I. Introduction

Hospital readmissions within 30 days are a health care quality metric, given their associated costs both to the patient and the clinical institution, and thus are one indicator of inefficiency in the healthcare system [1]–[3]. They are amply studied in a variety of medical conditions. However, they only recently have started to attract attention of researchers in the study of healthcare policies for diabetic patients [4]. Different machine learning approaches, including deep learning, have been attempted in order to predict a patient's risk of readmission based on their medical history with varying results [5]–[9]. The present investigation evaluates several machine learning models aimed at predicting readmission from clinical data recorded in previous visits by the diabetic patient. The techniques used include logistic regression, support vector machines, neural networks and random forests. The models are trained and evaluated over a publicly available dataset comprising patient data from a hospital network in the United States collected over the course of nearly ten years [5]. The performance of all the models tested is evaluated using several metrics, including F1 and ROC AUC (Area Under the Curve in Receiver Operating Characteristic analysis). Random forests is shown to outperform all the models under evaluation and to exhibit comparable or superior prediction rates than the other models trained over the same data and previously reported in the literature, including deep learning, while requiring significantly less computing power.

978-1-7281-1614-3/19/$31.00 ©2018 IEEE

II. Related work

Various investigations can be found in the literature seeking a reliable prediction of diabetic patient readmission using a variety of machine learning models and patient data sources. [5], for instance, use multivariable logistic regression in order to show that there is a decreased risk of 30-day readmission in diabetic inpatients who have their hemoglobin A1c (HbA1c) measured and that this association is true only for patients whose only primary diagnosis is diabetes. This model is trained over a preprocessed dataset derived from electronic health records, which has been made public on the UCI Machine Learning Repository and has subsequently been reused in related studies employing different prediction models and preprocessing methods, all evaluated with different measures. These metrics, derived from the model's exhibited number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), include accuracy (Equation 1), precision (Equation 2), recall (Equation 3), specificity (Equation 4), F1 (Equation 5) and the area under the ROC curve (ROC AUC score) obtained after plotting the model's recall (Equation 3) against the fall-out (Equation 6).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{F1} = \frac{2TP}{2TP + FP + FN} \tag{5}$$

$$\text{Fall-out} = 1 - \text{Specificity} = \frac{FP}{FP + TN} \tag{6}$$

Random forests models trained on the dataset compiled by [5] have exhibited good precision-recall scores (0.65) [10] whereas other classifiers fine-tuned through evolutionary algorithms (EA) have yielded good performance in terms of

accuracy, recall and specificity (0.97, 1.00, 0.97, respectively) [11]. Hybrid approaches that can capture more complex patterns between different features have been achieved combining Evolutionary Simulated Annealing method with sparse LOgistic Regression model of Lasso (ESALOR) improving accuracy, precision, recall and F1 (0.76, 0.77, 0.77, 0.86, respectively) over SVN and other conventional methods [9]. Various classifiers have shown varying performance metrics when patients in this dataset are grouped by age [7]. High ROC AUC (0.95) and F1 (0.92) scores have been achieved with deep learning models, including convolutional networks [8]. The same dataset has been used along with others in order to evaluate novel data mining and deep learning methods [12]–[14]. In other, related studies, risk factors for readmission have also been found by training other machine learning models on different clinical datasets collected in India and the United States [6], [15], [16]. In one of these, motivated by striking a balance between accuracy and interpretability, [16] proposes two novel methods: K-LRT, a likelihood ratio test-based method, and a Joint Clustering and Classification (JCC) method. Results are reported in AUC (0.7924) which fare better than conventional methods but does not surpass random forests (0.8453). However, their method allows for interpretation and identifying key features.

III. Methods

The dataset, originally compiled by [5] and publicly available in CSV format, is composed of electronic records spanning ten years (1999 through 2008) with various demographic and clinical variables per patient. The most salient aspects of the dataset can be summarised very briefly as follows: each row corresponds to a hospital visit by a patient and each patient may have more than one visit, *i.e.*, several rows may be associated to the same patient. Demographic information of the patient is stored as categorical variables, including `gender` and `race` as well as `age`, which appears as labels describing intervals measured in years (*e.g.*, `[0, 10)`, `[10, 20)`, `[20, 30)`, *etc.*). Columns `diag_1`, `diag_2`, and `diag_3` contain ICD9 codes indicating the diagnoses made during the visit. Each row includes also 24 features associated to different medications against diabetes, each one indicating if the drug, or a change in its dosage, was prescribed. The possible values for all these 24 columns are `'NO'` (not prescribed), `'Steady'` (no change in dosage), `'Up'` (increased dosage) and `'Down'` (decreased dosage). The class attribute indicates if the patient was readmitted after the visit and its possible values are `'NO'` (*i.e.*, no readmission), `'<30'` (*i.e.* readmission occurred within 30 days) `'>30'` (*i.e.*, readmission occurred after 30 days). Full details of the dataset, including detailed descriptions of the features mentioned earlier and others that have been omitted for brevity, can be found in the original study by [5].

Prior to training the models proposed in this paper, this dataset was preprocessed as follows: for all patients only the first visit was retained, *i.e.*, second and subsequent visits from the same patient were removed in order to ensure independence of the data. Eight columns were removed because most of the values in them were unknown or missing or because they do not pertain to the medical state of the patient (namely `encounter_id`, `patient_nbr`, `weight`, `admission_type_id`, `discharge_disposition_id`, `admission_source_id`, `payer_code` and `medical_specialty`). After this, numerical features were kept intact whereas categorical variables were mapped to numerical representations as follows: class labels `'NO'` and `'>30'` were merged into one, representing *'no 30-day readmission'* (encoded numerically as 0), while keeping the third class `'<30'` intact (encoded numerically as 1), representing *'30-day readmission.'* This way the problem is reduced to one of binary classification.

The remaining categorical columns were transformed as follows: ICD9 codes representing the diagnoses of the patient in each visit were grouped by their ICD9 Chapter,thus reducing the number of possible values in columns `diag_1`, `diag_2`, and `diag_3` to only 19. Thereafter, these three categorical attributes were transformed to numerical ones by replacing them with dummy variables. That is to say, `diag_1` was replaced with new, binary columns, one for each ICD9 Chapter code, in such a way that a value of 1 in one of these new columns indicates that a diagnosis of a disease or condition from this ICD9 category was made and recorded in the original `diag_1` whereas a value of 0 indicates otherwise. For instance, a value of 1 in dummy variable `diag_1_ICD9_9` would indicate that the patient's first diagnosis pertains to a condition in ICD9 Chapter 9, namely a disease of the digestive system, and a value of 0 would indicate otherwise. Variables `diag_2` and `diag_3` were replaced with dummy counterparts in the same manner. An analogous procedure was followed in order to replace other non-ordinal, categorical columns, such as `race` and `gender`, with dummy variables.

Ordinal, categorical values in column `age` were replaced directly with numerical values, with higher values reflecting higher age groups (*e.g.*, `[0, 10)` was encoded as 0, `[10, 20)` was encoded as 1, *etc.*). Finally, all 24 columns referring to medication prescriptions were converted to binary features by merging values `'Steady,'` `'Up,'` and `'Down'` into one single value representing `'Drug prescribed'` while value `'NO'` was kept intact to represent `'Drug not prescribed.'` Subsequently, each one of these 24 medication prescription features was replaced with a binary dummy variable, in a manner analogous to that described earlier for diagnosis attributes. After all these feature transformations the resulting dataset comprises 100 columns, including the class column.

In order to reduce the dimensionality of the data prior to training the prediction models, principal component analysis was conducted in order to reduce the number of features from 100 to 45 while preserving 98% of the variance. Thereafter, all features were normalised to a common scale, with unit variance and zero mean. After this, prediction models with logistic regression (LR), single layer perceptron (SLP), mul-

tilayer perceptron (MLP) and random forests (RF) were individually trained on the selected features using 10-fold cross-validation. Before doing this, the training data were balanced through oversampling since the original dataset was found to be highly unbalanced, with $63,417$ *'no 30-day readmission'* visits against only $6,152$ *'30-day readmission'* visits. Several performance metrics were calculated for each trained model, including ROC AUC and F1. Overfitting was prevented by the use of cross-validation during training and evaluation as well as by the application of oversampling only on the training data and not the testing data.

All the models were implemented in Python (`2.7.15`) using the library `scikit-learn` (`0.20`). The LR model was trained using the stochastic average gradient (SAG) solver implementation provided by `scikit-learn`. The rectified linear unit (ReLU) function was used as the activation function in the SLP and MLP models, with the latter having one 23-neuron hidden layer, whose size was chosen as a middle point between that of the input (45) and the output (1) layers. The RF model was trained with 100 trees in order to improve the estimates from the out-of-bag predictions without increasing the computational cost of training. During experimentation the models were observed to exhibit the best predictive and computational performance with these parameters and the results are presented in the following section of this article. Nevertheless, other parameter choices were also considered for each model during experimentation, *e.g.*, a larger hidden layer for MLP and logistic regression as an activation function for SLP, among others. These were omitted from this article for brevity.

IV. Results

The performance metrics obtained with each one of the models trained are listed in Table I. These metrics show that the least performing model was the single layer perceptron whereas the best prediction scores were achieved, by far, by the random forest model. These even exceed those reportedly obtained through deep learning techniques, listed in Section II, while requiring significantly less computing power. Notably, Table I also shows that none of the other models evaluated managed to obtain performance scores near those achieved by the random forest model, instead obtaining rather modest scores, mostly just over 0.5.

V. Conclusions

This research article presents a machine learning approach for the identification of diabetic patients at risk of requiring hospital readmissions. A reduction of these statistics should be expected to contribute towards the improvement of patients' well-being as well as towards a reduction in financial and reputational costs to healthcare institutions. This is particularly critical for patients of chronic conditions, such as diabetes. This creates the need for policies and strategies to reduce these statistics, especially methods to predict when a patient is at high risk of requiring readmission in the future. The best prediction rates reported in the recent literature have been achieved only through the use of deep learning on the patient clinical data collected by [5]. However, the present research article describes a novel method for efficiently managing the same dataset for the training of machine learning models, allowing better prediction rates without requiring deep learning. The best-performing model trained in this study, namely random forest, exceeds the prediction metrics reported by others in the recent literature using the same base dataset. This includes the precision-recall scores (0.65) reported by [10] as as well as the ROC AUC (0.95) and F1 (0.92) scores reported by [8] through the use of convolutional networks, while also exceeding or closely approximating the accuracy, recall and specificity (0.97, 1.00, 0.97, respectively) reported by [11]. To the best knowledge of the authors of the present study, no prediction models reported in the literature achieve the same performance metrics obtained by the RF model presented in this paper.

The main contributing factor to this result is the preprocessing of the patient data, which is achieved through a method not previously explored in the related literature using the same dataset. This method contemplates the simplification of the data in various ways, one being the reduction of several variables' domain with the intent of allowing the prediction models to generalise more easily. This includes the grouping of patients' diagnoses codes by their corresponding ICD9 Chapters, thus reducing the domain of these variables to only 19 possible values. This simplification was completed with application of a similar strategy on some of the other dataset's features, including the class attribute whose domain was adjusted from three to two possible values, thus transforming the original problem, *i.e.*, the prediction of a future readmission, into one of binary classification. This simplification of the problem, paired with the dimensionality reduction achieved through principal component analysis as well as the class balancing achieved through oversampling, results in a much more compact dataset from which generalisations can be learned by the prediction models presented. This, without requiring deep learning techniques and also achieving higher performance metrics than those reported previously in related literature using the same original dataset.

With the main exception of the domain reduction of some features described earlier, the data preprocessing procedure used in the present study and the methodology described in previous, related studies share some aspects, such as consideration of only first patient visits, feature selection and data balancing [8], [11]. As stated before, this simplification of the data is arguably a contributing factor to the predictive power of the models proposed. Nevertheless, interpretability of the prediction is still limited despite this simplification of the dataset, given the complexity of the random forest technique. Furthermore, the proposed data preprocessing clearly comes at the cost of reduced granularity. It can be reasonably hypothesised, though evidently not guaranteed, that these results could be further improved with larger datasets, *i.e.*, with higher numbers of patient visits with the same features. This investigation is focused on diabetes, however, this type of

Model	ROC AUC	F1	Precision	Recall	Accuracy
LR	0.5783	0.5550	0.5599	0.5515	0.5566
SLP	0.5229	0.5484	0.5129	0.5929	0.5147
MLP	0.6548	0.6164	0.6100	0.6095	0.6083
RF	0.9999	0.9974	0.9950	0.9999	0.9974

TABLE I

PERFORMANCE METRICS OF MODELS LOGISTIC REGRESSION (LR), SINGLE LAYER PERCEPTRON (SLP), MULTILAYER PERCEPTRON (MLP) AND RANDOM FORESTS (RF).

approach could be further evaluated in the context of other chronic conditions, such as heart disease.

REFERENCES

[1] W. Boulding, S. W. Glickman, M. P. Manary, K. A. Schulman, and R. Staelin, "Relationship between patient satisfaction with inpatient care and hospital readmission within 30 days," *The American Journal of Managed Care*, vol. 17, no. 1, pp. 41–48, 2011.

[2] R. N. Axon and M. V. Williams, "Hospital readmission as an accountability measure," *Jama*, vol. 305, no. 5, pp. 504–505, 2011.

[3] M. T. Kassin, R. M. Owen, S. D. Perez, I. Leeds, J. C. Cox, K. Schnier, V. Sadiraj, and J. F. Sweeney, "Risk factors for 30-day hospital readmission among general surgery patients," *Journal of the American College of Surgeons*, vol. 215, no. 3, pp. 322–330, 2012.

[4] D. J. Rubin, "Hospital readmission of patients with diabetes," *Current Diabetes Reports*, vol. 15, no. 4, p. 17, 2015.

[5] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records," *BioMed Research International*, vol. 2014, 2014.

[6] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Predictive risk modelling for early hospital readmission of patients with diabetes in India," *International Journal of Diabetes in Developing Countries*, vol. 36, no. 4, pp. 519–528, 2016.

[7] D. Mingle, "Predicting diabetic readmission rates: moving beyond Hba1c," *Current Trends in Biomedical Engineering & Biosciences*, vol. 7, no. 3, p. 555707, 2017.

[8] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, "Predicting hospital readmission among diabetics using deep learning," *Procedia Computer Science*, vol. 141, pp. 484–489, 2018.

[9] S. Tutun, S. Khanmohammadi, L. He, and C.-A. Chou, "A meta-heuristic LASSO Model for diabetic readmission prediction," in *Proceedings of the 2016 Industrial and Systems Engineering Research Conference (ISERC)*, 2016.

[10] M. S. Bhuvan, A. Kumar, A. Zafar, and V. Kishore, "Identifying diabetic patients with high risk of readmission," *arXiv preprint arXiv:1602.04257*, 2016.

[11] A. Choudhury, D. Greene, and M. Christopher, "Evaluating patient readmission risk: a predictive analytics approach," *arXiv preprint arXiv:1812.11028*, 2018.

[12] Ö. M. Soysal, "Association rule mining with mostly associated sequential patterns," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2582–2592, 2015.

[13] H. Kaschel, V. Rocco, and G. Reinao, "An open algorithm for systematic evaluation of readmission predictors on diabetic patients from data warehouses," in *2018 IEEE International Conference on Automation/XXIII Congress of the Chilean Association of Automatic Control*. IEEE, 2018, pp. 1–6.

[14] E. Chou, T. Nguyen, J. Beal, A. Haque, and L. Fei-Fei, "A fully private pipeline for deep learning on electronic health records," *arXiv preprint arXiv:1811.09951*, 2018.

[15] C. King, S. Atwood, M. Lozada, A. K. Nelson, C. Brown, S. Sabo, C. Curley, O. Muskett, E. J. Orav, and S. Shin, "Identifying risk factors for 30-day readmission events among American Indian patients with diabetes in the Four Corners region of the southwest from 2009 to 2016," *PLoS ONE*, vol. 13, no. 8, p. e0195476, 2018.

[16] T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams, and I. C. Paschalidis, "Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach," *Proceedings of the IEEE*, vol. 106, no. 4, pp. 690–707, 2018.