Predicting Hospital Readmission Patterns of Diabetic Patients using Ensemble Model and Cluster Analysis

Hung N. Pham
School of Information & Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
hungpn@soict.hust.edu.vn

Anurag Chatterjee
Institute of Systems Science
National University of Singapore
Singapore, Singapore
anurag.chatterjee@u.nus.edu

Balasubramanian Narasimhan
Institute of Systems Science
National University of Singapore
Singapore, Singapore
e0146758@u.nus.edu

Choon Wee Lee
Institute of Systems Science
National University of Singapore
Singapore, Singapore
e0146925@u.nus.edu

Diksha Kumari Jha
Institute of Systems Science
National University of Singapore
Singapore, Singapore
diksha.jha@u.nus.edu

Edric Yeng Fai Wong
Institute of Systems Science
National University of Singapore
Singapore, Singapore
e0267640@u.nus.edu

Stella Ellyanti
Institute of Systems Science
National University of Singapore
Singapore, Singapore
e0146509@u.nus.edu

Quang H. Nguyen
School of Information & Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
quangnh@soict.hust.edu.vn

Binh P. Nguyen
School of Mathematics and Statistics
Victoria University of Wellington
Wellington, New Zealand
binh.p.nguyen@vuw.ac.nz

Matthew C. H. Chua
Institute of Systems Science
National University of Singapore
Singapore, Singapore
mattchua@nus.edu.sg

***Abstract*—Diabetes is a chronic illness that affects around 425 million people globally in 2017, and this is predicted to increase to 629 million by the end of 2045. The ability to analyze and predict the readmission patterns of diabetic patients would allow the optimization of hospital resources and assessment of treatment effectiveness. This paper proposes an ensemble model to predict hospital readmission by choosing from a pool of 15 models, made up of variants of Logistic Regression, Decision Trees (DT), Neural Network (NN) and Augmented Naïve Bayes (NB) networks. The final ensemble model was assembled using the five best models, determined based on individual model accuracy and the Jaccard distance between them, to maximize overall accuracy and sensitivity. The final ensemble contained DT (CHAID), Tree Augmented Naïve Bayes network, DT (CHAID with boosting), Neural Network with bagging and DT (CART with boosting). Compared against existing predictive models, the proposed ensemble was able to achieve improved sensitivity at 56% while maintaining comparable accuracy at 63.5%. Cluster analysis after performing principal component analysis on the dataset revealed 4 distinct clusters of patients. Patients with a history of in-patient visits or if they received a high amount of treatment in their current visit were found more likely to be readmitted.**

***Index Terms*—diabetes prediction, ensemble model, decision trees, neural networks, naïve bayes networks, cluster analysis**

I. Introduction

Diabetes is a chronic condition associated with abnormally high levels of sugar (glucose) in the blood and can be classified into type 1, type 2, gestational and other specific types of diabetes by WHO [1]. Globally, in 2017 it was estimated that 425 million people had diabetes in 2017 – this is predicted to increase to 629 million by the end of 2045 [2]. In the United States (US), more than 30 million people had diabetes in 2017 [2]. As the quality of inpatient care is associated with early readmission [3], the ability to predict hospital readmission for diabetes patient would prove invaluable. Furthermore, this also allows intervention like inpatient diabetes education which is associated with less frequent hospital readmission [4] to be administered and tracked.

Various classification models had been evaluated and proposed. Duggal *et al.* [5] explored various classification models and concluded that random forest outperformed Naïve Bayes, Logistic Regression, Adaboost and Neural Network in predicting readmission. Chopra *et al.* [6] concluded that recurrent neural network was able to outperform logistic regression, support vector machine (SVM), decision tree, random forest and simple neural network in the prediction of diabetic readmission.

With regards to the dataset used in this paper, Strack *et al.* [7] had studied and presented their findings on HbA1c measurements on hospital readmission and concluded that it is a useful predictor of readmission. Munnangi and Chakraborty [8], using the same dataset, trained a model using

978-1-7281-0525-3/19/$31.00 ©2019 IEEE

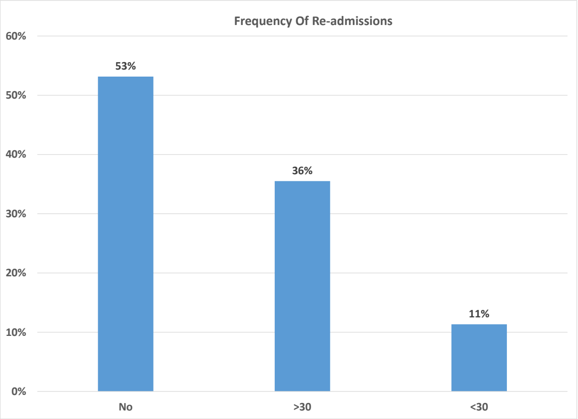


Fig. 1. Distribution of readmitted patients in original data.

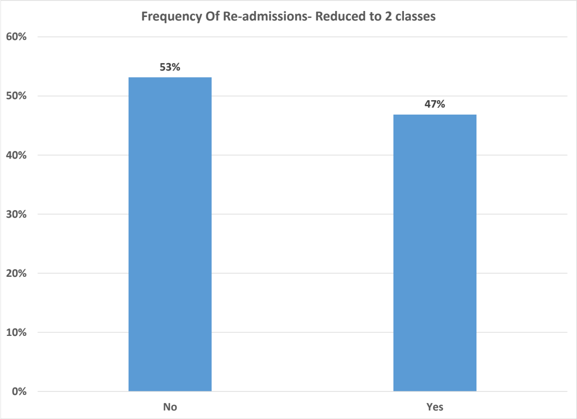


Fig. 2. Distribution of readmitted patients after the number of categories is reduced to 2.

SVM that was able to predict readmission with an accuracy of 63.3%.

While numerous papers on boosting the accuracy on diabetes prediction using ensembled models exist [9]–[11], there does not appear to be any studies that explores the use of cluster analysis to analyze the readmission patterns or use an ensemble model to predict diabetic readmission.

In this research, we analyze the readmission patterns using cluster analysis and optimize the prediction for hospital readmission of diabetic patients via the use of an ensemble model, which is derived from constituent models selected from a larger model pool.

II. Data Exploration and Preparation

A. Data

The dataset used in this study was collected from the UCI machine learning repository which represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks [12].

B. Data Exploration

The distribution of the target variable was analyzed, and it was seen that there were relatively very few patients who were readmitted within 30 days as shown in Figure 1. To ensure a more equal distribution of the target variables, the values for readmitted within 30 days and more than 30 days were merged to one field called YES, denoting that such patients were readmitted. The distribution of readmitted patients after the number of categories is reduced to 2 as shown in Figure 2.

Figure 3 presents that the proportion of patients who have high values of blood glucose test also have higher readmissions. This could be since these patients are chronic diabetes patients with high values of blood glucose.

C. Data Preparation

Extreme Values (Outliers) Extreme values are identified via plotting the data on a box-plot, where they are defined as data points that exceeds 150% the interquartile range above the upper quartile or below the lower quartile. It was observed that there were no significant outliers. However, it was observed that the distribution for the number of outpatient and emergency visits were highly skewed and required further transformation before modeling.

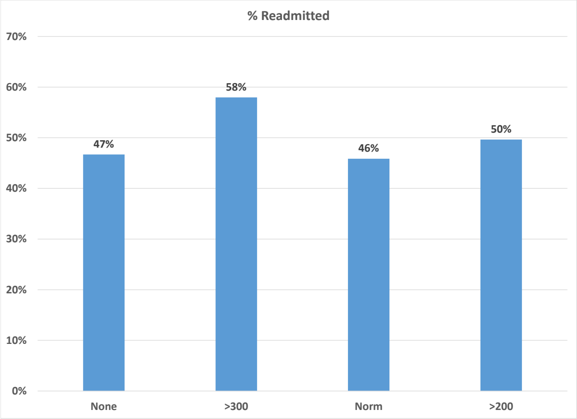


Fig. 3. Proportion of patients readmitted with result of blood glucose test

Missing Data The key focus in guiding the handling of missing data was to minimize the amount of bias introduced to the dataset, as this will negatively impact any models generated subsequently. Missing data, often denoted by "?" in the dataset, are summarized as below:

1) Race (2% missing). These data were merged with the category of "others".
2) Weight, payer code and medical speciality. These columns were removed due to large amount (> 50%) of missing data.

Irrelevant Data As the intent of the study was to predict readmission, it was deemed that patients who died should not be included. These samples (about 1.6%) was removed from the dataset.

After data preparation is completed, a sample size of

100,111 (original size: 101,766) remained.

Data Splitting The available data was split into 70% for training and 15% each for validation and testing.

III. Feature Engineering

To ensure the data is suitable for use in the various models, the following was applied to the dataset.

A. Dummy Coding

As some of the modeling approaches used do not handle categorical independent variables very well, some of the variables (e.g., race and gender) were dummy coded.

B. Combination of Categories

It was observed that some of the categorical variables contains numerous categories which makes dummy coding unrealistic (e.g., "Discharge Disposition" and "Admission Source" contains 29 and 21 distinct values, respectively). In these cases, the categories are merged based on domain knowledge and common sense. Some categories which made administrative sense, but little modelling sense were merged, for example, patients who "expired at home", "expired at medical facility" or "expired in an unknown place" were deemed to have "expired".

C. Creation of New Variables

In other cases, multiple variables appear can be collapsed into one. For example, the 24 features for medications, which are all type 2 diabetes medication, were collapsed into just 3 features – Diabetes Med (Increasing), Diabetes Med (Steady) and Diabetes Med (Decreasing).

D. Transformation of Skewed Variables

Amongst the variables in the dataset, it was observed that the data for "Number of outpatient visits" and "Number of emergency visits" was skewed. Hence, the common logarithm of the values was used instead.

The data dictionary for the original data can be referenced from the table in [13] while the created variables are listed as follows.

1) Age2: Categorization of age into 3 distinct groups based on trends observed by Beata Strack *et al.* [7].
2) Admission_type_id2: Admission type with similar categories merged.
3) Discharge_disposition_id2: Discharge Disposition with similar categories merged.
4) Admission_source_id2: Admission Source with similar categories merged.
5) Number_outpatient_log: Common logarithm value for the number of outpatient visits.
6) Number_emergency_log: Common logarithm value for number of emergency visits.
7) Diabetes Med (Up): Consolidates the number of "increases" from the list of 24 medications
8) Diabetes Med (Steady): Consolidates the number of "steady" from the list of 24 medications
9) Diabetes Med (Down): Consolidates the number of "decreases" from the list of 24 medications.
10) Diabetes (Change): Captures if there was a change in medication prescribed.
11) Readmitted2: Redefines the readmission criteria into a dichotomous one (i.e., yes/no readmission for diabetes patients).

IV. Methods

A. Performance Metrics

The basic performance parameters that the study considers are the model accuracy and sensitivity (i.e., recall). The basis of comparison will be based on the study by Munnangi and Chakraborty [8] where they trained a high performance SVM with linear kernel and were able to achieve an accuracy of 63.3%, sensitivity of 49.7% and specificity of 75.1%.

B. Predictive Modeling

Several models were trained on the training data using SPSS modeler. A pool of 15 models were kept aside which performed the best on the validation data. These were:

- Logistic Regression
- Variants of Decision Tree (CHAID, C4.5, CART, with boosting / bagging)
- Variants of Bayes Networks (Naïve Bayes, TAN)
- Variants of Neural Network (with boosting / bagging)

The best classification accuracy of 62.9% on the validation data set was obtained using a boosting of decision trees built using the CART technique. To improve the accuracy, a mechanism to ensemble the pool of models was developed. It has been shown that to get better classification accuracy it is better to ensemble many instead of all the models at hand [14] by selecting the ensemble from a library of models [15]. We used a novel approach that considers the accuracy and the correlation of the available pool of models as the criterion to select the models for the ensemble.

We used the Jaccard distance [16], which is a measure of dis-similarity for binary vectors to evaluate the correlation of the output predictions of any 2 models that would be considered in the ensemble. Theoretical [17], [18] and empirical research [19] have shown that a good ensemble consists of individual classifiers in the ensemble that are both accurate and have errors in distinct parts of the input space. To capture this metric as a single value we compute the product of the Jaccard distance and the model accuracy for the models in the candidate pool. Hence, the probability of selection for a model m given that a model, M is selected can be written as

$$P(m|M) = Accuracy(m) * JaccardDistance(m, M), \quad (1)$$

where the JaccardDistance is Jaccard distance of models m and M.

The resulting ensemble model when evaluated on the test partition showed better performance than all its constituent models and the considered baseline.

Figure 4 describes the work flow for training available predictive models on train and validation data set, then obtain

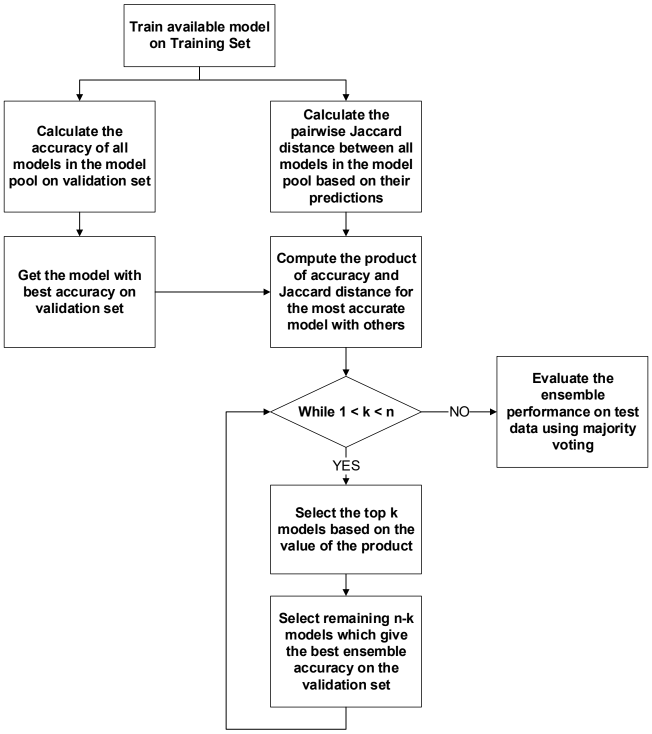

Train available model on Training Set
Calculate the accuracy of all models in the model pool on validation set
Calculate the pairwise Jaccard distance between all models in the model pool based on their predictions
Get the model with best accuracy on validation set
Compute the product of accuracy and Jaccard distance for the most accurate model with others
While 1 < k < n
NO
Evaluate the ensemble performance on test data using majority voting
YES
Select the top k models based on the value of the product
Select remaining n-k models which give the best ensemble accuracy on the validation set

Fig. 4. Training the ensemble classifier for ensemble size of n

the ensemble classifier and evaluate its performance on test data set.

C. Cluster Analysis using k-Means

Beside training an ensemble classifier from best predictive models, we also utilize an unsupervised learning method that is cluster analysis to analyze the patients that are part of the dataset. Cluster analysis is a machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.

The first analysis is performed using k-Means clustering, which aims to partition the 100K+ input samples into k clusters.

1) Features Selection and Preprocessing: The numeric columns in dataset including time_in_hospital, total_procedures, num_medications, number_outpatient_log, number_emergency_log and number_inpatient were considered to evaluate the patient groups. The total_procedures column was created as the sum of the num_lab_procedures and num_procedures columns. Totally, there are 7 numeric features selected.

Principal Component Analysis (PCA) was performed on these 7 features. Based on the Scree plot criterion, 3 principal components were chosen to be retained. Figure 5 shows a Scree plot of eigenvalue vs. number of components in PCA.

Based on the rotated factor loading, the interpretations of the first 3 components are as follows

PC1: The combination of time a user spends in the hospital, total amount of medications they are administered and the total number of procedures that are performed on them during the encountered. So, PC1 represents *the magnitude of treatment for the patient during the current encounter*.

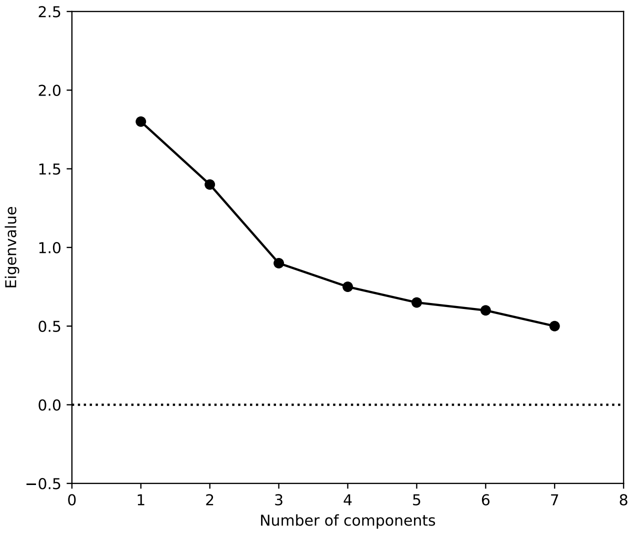
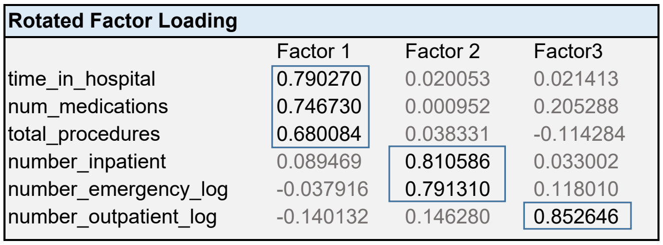


Fig. 5. Scree plot of Eigenvalue vs no. of components



Rotated Factor Loading

	Factor 1	Factor 2	Factor3
time_in_hospital	0.790270	0.020053	0.021413
num_medications	0.746730	0.000952	0.205288
total_procedures	0.680084	0.038331	-0.114284
number_inpatient	0.089469	0.810586	0.033002
number_emergency_log	-0.037916	0.791310	0.118010
number_outpatient_log	-0.140132	0.146280	0.852646

Fig. 6. Rotated Factor Loading matrix for the factors

PC2: The combination of number of emergency visits and number of inpatient visits for the patient in the year preceding the encounter. PC2 represents *the magnitude of prior expensive visits by the patient* [20].

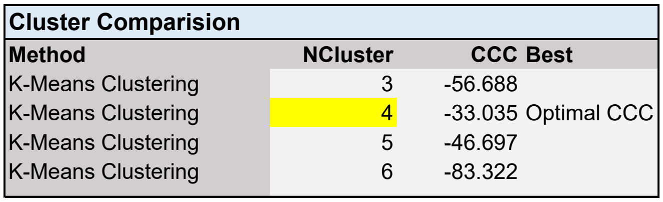
PC3: It is just the number of outpatient visits for the patient in the preceding year. It represents *the number of comparatively less expensive visits by the patient prior to the current encounter* [20].

Figure 6 presents rotated factor loading matrix for 3 factors which are chosen Principle Components (PC1, PC2, PC3).

2) Cluster Analysis: The k-Means clustering is performed on the 3 components that are the output of the principal component analysis. Cluster sizes from 3 through 6 were evaluated. The analysis of the clusters proceeded with 4 clusters, which were claimed to be the optimal number of clusters based on the Cubic Clustering Criterion (CCC) [21].

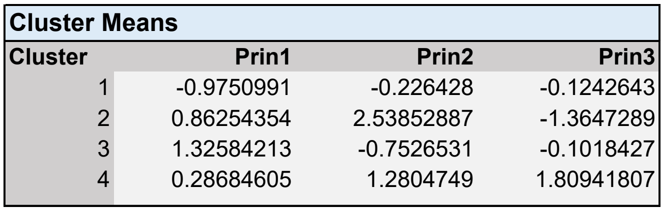
Figure 7 presents the cubic clustering criterion (CCC) values for 4 clusters. The cluster means for the 4 clusters are obtained as shown in Figure 8 and the below are the details of the clusters.

Cluster 1: These are the patients who had less prior encounters and they also received relatively less treatment during the current encounter.



Cluster Comparision		
Method	NCluster	CCC Best
K-Means Clustering	3	-56.688
K-Means Clustering	4	-33.035 Optimal CCC
K-Means Clustering	5	-46.697
K-Means Clustering	6	-83.322

Fig. 7. CCC values for different number of clusters



Cluster Means			
Cluster	Prin1	Prin2	Prin3
1	-0.9750991	-0.226428	-0.1242643
2	0.86254354	2.53852887	-1.3647289
3	1.32584213	-0.7526531	-0.1018427
4	0.28684605	1.2804749	1.80941807

Fig. 8. Cluster means for number of clusters is 4

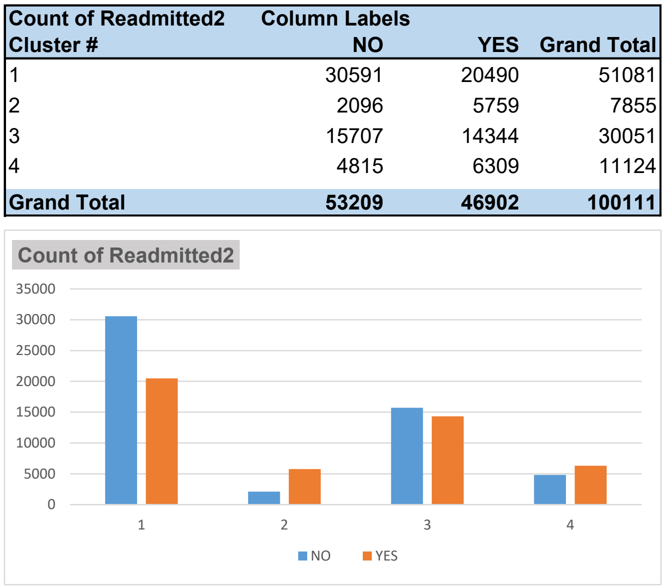
Cluster 2: These are patients who had multiple prior in-patient or emergency encounters and received relatively high treatment.

Cluster 3: These are the patients who received high treatment during their current encounter but have less prior encounters.

Cluster 4: These patients have had multiple encounters in the past, but they received relatively less treatment during their current encounter.

Figure 9 shows the distribution of the number of readmissions observed within the clusters. It can be seen that the maximum proportion of readmissions among the total number of patients in the cluster occurs for Cluster 2 at 73.3%.

Cluster 1 contains the most number of readmitted as well as non-readmitted patients with the non-readmitted ones in a majority as shown in Figure 9.



Count of Readmitted2	Column Labels		
Cluster #	NO	YES	Grand Total
1	30591	20490	51081
2	2096	5759	7855
3	15707	14344	30051
4	4815	6309	11124
Grand Total	53209	46902	100111

Fig. 9. Distribution of the number of readmissions observed in the clusters

Since Cluster 2 has patients who have a history of in-patient and emergency encounters (high values of PC2) and needed to receive relatively high treatment during the current encounter (value of PC1), the pattern continues for the current encounter. Indeed, the number of in patient encounters is the most important predictor for the target variable [8]. However, patients in Cluster 1 have less differentiating characteristics, so they contain the bulk of the patient population and most of these patients are not readmitted.

V. Results

Table I shows the confusion matrix for ensemble model predictions. Performance metrics including Accuracy, Specificity and Sensitivity were measured for each predictive model and compare with ensemble model. These performance metrics results were reported in Table II.

The classification accuracy obtained using the ensemble was 63.5% on the 15% test partition of the data which is better than the baseline.

TABLE I
Confusion matrix for ensemble model predictions

	Actual		
Predicted	Readmitted ?	Yes	No
	Yes	4036	2297
	No	3178	5699

TABLE II
Comparison of ensemble model with constituent models

	Accuracy	Specificity	Sensitivity
DT-Chaid	62.6%	74.9%	48.7%
NB-TAN	59.5%	64.0%	54.8%
DT-CHAID-Boosting	62.4%	68.6%	55.3%
NN-Bagging	62.8%	63.5%	61.9%
DT-CART-Boosting	62.9%	74.8%	49.3%
Ensemble	63.5%	71.3%	55.9%

Compared against the SVM model in the baseline, the ensemble has a comparable accuracy but better sensitivity. A higher sensitivity implies that the model has a higher true positive rate, i.e., it is better at predicting cases of re-admission than the baseline model.

VI. Discussion and Conclusion

We have analyzed the readmission patterns of diabetic patients in the USA using supervised learning (Predictive models) and unsupervised learning (Cluster Analysis using k-means). We perform the general data mining steps of data exploration, preparation, feature engineering to set the stage to predict whether a patient would be readmitted. We train several models for the predictive analysis. To exceed the accuracy of the best model that we train, we devise a novel ensemble strategy to perform the prediction using only a subset of the trained models. We consider the combination of the model accuracy and Jaccard distance between the models to come up with the ensemble model that does better than any of the constituent models and our baseline.

We also perform cluster analysis to analyze the characteristics of the patients who get readmitted. We found that patients who have history of prior in-patient visits or received particularly high treatment during the current encounter are more likely to be re-admitted.

A. Limitation and Future Work

The dataset on which the models have been trained are now almost 10 years old. As of the current date, we are not aware of any publicly available dataset that contains similar data for hospital readmissions but has released after the considered version. As lifestyle patterns and access to healthcare facilities change with time [22], the performance of our model on recent hospital readmission data might not be good enough. Further, the models have been developed focusing on the readmission of diabetic patients in the US, the performance of these models for other geographies or for other chronic illness patterns have not been evaluated, which are shown to be dependent on different factors [5], [23]. Also, the correlation among the models is only captured using one metric as part of this work.

There are many other model diversity measures that can also be explored as future work [24]. We evaluated a majority voting and a one vs. rest approach, with majority voting giving better results than the one vs. rest approach. However, there are other ensembling approaches that can be evaluated as future work [25]. Finally, we had observed that we get 4 distinct clusters of patients as the output of the cluster analysis. As a future work we could train different prediction models, e.g., [26]–[30], which could even be ensembles that would predict whether a patient is likely to be readmitted given that they belong to that cluster.

Acknowledgment

This research is funded by Hanoi University of Science and Technology (HUST) under grant number T2018-PC-016.

References

[1] Rydén L., Standl E., et al. Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary: The task force on diabetes and cardiovascular diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD). *European Heart Journal*, 28(1):88–136, 2007.

[2] International Diabetes Federation. IDF diabetes atlas, 8th edition. http://diabetesatlas.org/IDF_Diabetes_Atlas_8e_interactive_EN/, 2017. Brussels, Belgium.

[3] Carol M. Ashton, David H. Kuykendall, et al. The Association between the Quality of Inpatient Care and Early Readmission. *Annals of Internal Medicine*, 122(6):415–421, 03 1995.

[4] Sara J Healy, Dawn Black, et al. Inpatient diabetes education is associated with less frequent hospital readmission among patients with poor glycemic control. *Diabetes Care*, page DC_130108, 2013.

[5] Reena Duggal, Suren Shukla, et al. Predictive risk modelling for early hospital readmission of patients with diabetes in India. *International Journal of Diabetes in Developing Countries*, 36(4):519–528, 2016.

[6] Chahes Chopra, Shivam Sinha, et al. Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients. In *Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics*, pages 18–23. ACM, 2017.

[7] Beata Strack, Jonathan P DeShazo, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014, 2014.

[8] Hephzibah Munnangi and Goutam Chakraborty. Predicting readmission of diabetic patients using the high performance support vector machine algorithm of SAS® Enterprise Miner™.

[9] Manal Alghamdi, Mouaz Al-Mallah, et al. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PloS one*, 12(7):e0179805, 2017.

[10] Hui Chen, Chao Tan, Zan Lin, and Tong Wu. The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. *Computers in Biology and Medicine*, 50:70–75, 2014.

[11] Xun Wei, Fan Jiang, et al. An ensemble model for diabetes diagnosis in large-scale and imbalanced dataset. In *Proceedings of the Computing Frontiers Conference*, pages 71–78. ACM, 2017.

[12] Diabetes 130-US hospitals for years 1999-2008 data set. https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008, 2008.

[13] List of features and their descriptions in the initial dataset. https://www.hindawi.com/journals/bmri/2014/781670/tab1/.

[14] ZH Zhou, J Wu, and W Tang. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 137(1-2):239–263, 2002.

[15] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning*, page 18, 2004.

[16] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

[17] LK Hansen and P Salamon. Neural network ensembles. *IEEE Transactions on PAMI*, 12(10):993–1001, 1990.

[18] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, pages 231–238, 1995.

[19] Sherif Hashem. Optimal linear combinations of neural networks. *Neural Networks*, 10(4):599–614, 1997.

[20] Linda S Elting, Charles Lu, et al. Outcomes and cost of outpatient or inpatient management of 712 patients with febrile neutropenia. *Journal of Clinical Oncology*, 26(4):606–611, 2008.

[21] GW Milligan and MC Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[22] M Weinberger, EZ Oddone, and WG Henderson. Does increased access to primary care reduce hospital readmissions? *New England Journal of Medicine*, 334(22):1441–1447, 1996.

[23] Pedro Almagro, Bienvenido Barreiro, et al. Risk factors for hospital readmission in patients with chronic obstructive pulmonary disease. *Respiration*, 73(3):311–317, 2006.

[24] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[25] TG Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[26] Binh P. Nguyen, Wei-Liang Tay, and Chee-Kong Chui. Robust biometric recognition from palm depth images for gloved hands. *IEEE Transactions on Human-Machine Systems*, 45(6):799–804, Dec 2015.

[27] Binh P. Nguyen, Hans Heemskerk, Peter T. C. So, and Lisa Tucker-Kellogg. Superpixel-based segmentation of muscle fibers in multi-channel microscopy. *BMC Systems Biology*, 10(S5:124):39–50, 2016.

[28] Xuan Chen, Binh P. Nguyen, Chee-Kong Chui, and Sim-Heng Ong. Automated brain tumor segmentation using kernel dictionary learning and superpixel-level features. In *Proc. of the IEEE SMC 2016*, pages 2547–2552, Oct 2016.

[29] Xuan Chen, Binh P. Nguyen, Chee-Kong Chui, and Sim-Heng Ong. Reworking multilabel brain tumor segmentation – an automated framework using structured kernel sparse representation. *IEEE Systems, Man, and Cybernetics Magazine*, 3(2):18–22, Apr 2017.

[30] Xuan Chen, Binh P. Nguyen, Chee-Kong Chui, and Sim-Heng Ong. An automatic framework for multi-label brain tumor segmentation based on kernel sparse representation. *Acta Polytechnica Hungarica*, 14(1):25–43, Apr 2017.