Predicting 30-Day Hospital Readmission for Diabetes Patients using Multilayer Perceptron

Ti'jay Goudjerkan[1], Manoj Jayabalan[2]

School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia

Abstract—Hospital readmission is considered a key metric in order to assess health center performances. Indeed, readmissions involve different consequences such as the patient's health condition, hospital operational efficiency but also cost burden from a wider perspective. Prediction of 30-day readmission for diabetes patients is therefore of prime importance. The existing models are characterized by their limited prediction power, generalizability and pre-processing. For instance, the benchmarked LACE (Length of stay, Acuity of admission, Charlson comorbidity index and Emergency visits) index traded prediction performance against ease of use for the end user. As such, this study propose a comprehensive pre-processing framework in order to improve the model's performance while exploring and selecting a prominent feature for 30-day unplanned readmission among diabetes patients. In order to deal with readmission prediction, this study will also propose a Multilayer Perceptron (MLP) model on data collected from 130 US hospitals. More specifically, the pre-processing technique includes comprehensive data cleaning, data reduction, and transformation. Random Forest algorithm for feature selection and SMOTE algorithm for data balancing are some example of methods used in the proposed pre-processing framework. The proposed combination of data engineering and MLP abilities was found to outperform existing research when implemented and tested on health center data. The performance of the designed model was found, in this regard, particularly balanced across different metrics of interest with accuracy and Area under the Curve (AUC) of 95% and close to the optimal recall of 99%.

Keywords—Readmission; diabetes; multilayer perceptron; feature engineering

I. INTRODUCTION

Diabetes is one of the chronic non-communicable diseases that are on the rise with massive urbanization and a drastic change of lifestyle in many countries. It is expected to turn into the seventh most prevalent mortality factor by 2030 and millions of deaths could be prevented each year through better analytics [1]. Therefore, diabetes is on the health agenda of most developed and developing countries. Healthcare industry collects and process diabetes patient medical data in huge volume, diverse structure, and real-time flow of data [2]. With the rise of technology, both from the diagnosis and monitoring, storage and analysis, novel solutions are now available to better address challenges like non-invasive screening, tailor-made treatment, and hospital readmissions [3].

When assessing the quality of care delivered by a health center, readmission is the metric of choice. It measures the number of patients that need to come back to the hospital after their initial discharge. The readmission can be classified into three broad categories such as unavoidable, planned, and unplanned. The unavoidable readmission that is highly predictable mostly due to the nature of the pathology or patient's condition (i.e. cancer phase IV, metastasis). Secondly in the planned readmission which is directly prescribed by the healthcare professional to the patient (i.e. check-up, transfusion). Lastly, the unplanned is defined as readmission that shouldn't have happened given the practitioner's diagnosis and could have been avoided if proper care was given to the patient post-discharge [4]. Unavoidable and planned readmissions already are highly anticipated. However, predicting unplanned readmission is of prime interest due to its inherent uncertainty.

Unplanned readmission is the most useful type when evaluating the quality of care of a hospital as it highlights a practitioner's diagnosis or treatment error [4]. Beyond being a core indicator of the quality of care, unplanned readmissions also constitute a financial problem for nations [5], [6]. Therefore with a predictive model to assess unplanned readmission risk could optimize the quality of hospital services and state Medicare [7], [8].

According to [8], readmissions occurring after 30 days have less correlation with the quality of care from the health center and might be an impact due to external factors such as complications or patient's lifestyle. Numerous researches highlighted clear interest in 30-day unplanned readmission prediction models based on diabetes complications [7], [9]. However, predictability performance is quite low when dealing with unplanned readmission rates [4]. Moreover, several researchers have proposed a predictive model for readmission in healthcare for all types of diseases and only limited work are dedicated to diabetes. As different pathologies have different conditions and behaviors, prediction on specific pathology subset would highly benefit the prediction model's performance.

The purpose of this study is to propose a prediction model for 30-day unplanned readmission among diabetes patients in US hospitals. The analysis will be based on risk factors such as a patient's demographics, admission details, diagnosis, and medical data. In a broader sense, the goal of the study is to allow health centers to better anticipate and address unplanned readmissions while improving their quality of care and cost-efficiency.

II. RELATED WORK

Consequent efforts have been led so far to improve hospitals' readmission rate and predictability. However, due to

the limitation in term of data quality and volume, only a few models are found accurate and generalizable enough. Through initial research, it appeared that the LACE index (Length of stay, Acuity of admission, Charlson comorbidity index and Emergency visits) is so far the most preferred model of readmission prediction due to its ease use and implementation by a healthcare professional [10]–[13]. Despite its assets, this index does not pass the acceptable threshold achieving 0.56 to 0.63 c-statistics for unplanned short-term diabetics readmission prediction, hence urging the need for further research to improve the model's performance [4], [11]. To cope with the limitations of the previous approach this review will only list prediction models for unplanned readmission for diabetics limited to acceptable and above c-statistics. A total of 6 novel models were selected, ranging from 0.70 to 0.95 c-statistics. In that sense, this summarized list establishes an attempt to regroup the key novel models that might be generalizable and applicable to hospitals for diabetes patients' management.

Recent work using data from 130 US hospitals brought up state of the art performance across the different proposed models with accuracy and AUC up to 94% [14]. However, the work is unpublished and has not been verified by peers. Moreover, achieved the best performance with the Random Forest algorithm. Such high performance might be attributed to the exhaustive pre-processing of the data as well as data balancing with SMOTE algorithm.

A contemporary study led by [15] achieved comparable result with the use of a Recurrent Neural Network method achieving 0.80 c-statistics with 81.12% accuracy. The strength of this approach compare to the Collins' model is the use of an all-age larger database of 100,000 patients. This dataset also does not discriminate the length of previous continuous hospital enrolment that provides a higher level of predictability. However, the data used is old compared to the other models and there might be possible improvements in the factor selection process as 33 out of 56 variables were used for analysis. In [11], through Machine Learning algorithm achieved 0.70 – 079 c-statistics respectively for 30-70 and 0-30 age group on the same dataset with a sensitivity of 43.63% - 49.78% and specificity of 82.62% - 89.19%. Compare to the two previous models, the researcher applied a different algorithm to each defined population segments. The configuration of such algorithm is the ensemble average of "Extreme Gradient Boosted Trees Classifier with early Stopping, Nystroem Kernel SVM Classifier, Balance random forest classifier, and Gradient boosted greedy tress classifier with early stopping". Despite lacking accuracy, such segmentation approach might be used to improve the previous models.

Convolutional Neural Network presents deep learning as an efficient method for predicting hospital readmission of diabetic patients [16]. This model indeed achieves state of the art c-statistic performance of 95% and performs better than other machine learning models. Such model performance is attributed to both maximizations of the sample size and data engineering processes. As such, efficient feature selection, feature transformation and the use of SMOTE to address class imbalance inherent to medical data appear key to improve deep learning performances.

This review, far from being exhaustive, succeeded in extracting a limited number of optimum performing models among the current body of literature. Despite providing additional knowledge contribution to previous systematic reviews, the listed models are still limited in term of performance [9], [17]. As such, this review reveals a clear gap in term of model's classification performance, model generalizability and data pre-processing.

III. Methods

This section will highlight and rationalize the methods used in this study in order to achieve the objectives. While the selected dataset will be presented in section A particular attention will be given to the pre-processing stage in section B. In this section, details over cleaning, data reduction, transformation techniques but also pre-processing performance evaluation will be outlined (Section C). Namely, key pre-processing steps include feature Hotdeck imputation with Approximate Bayesian Bootstraps, ICD-9-DM clustering and feature selection using the Random Forest algorithm. Furthermore, the need for class balancing will be presented in section D and emphasize on Synthetic Minority Over-sampling Technique (SMOTE) algorithm. Particular attention will also be given to the modeling part in section E where Multilayer Perceptron (MLP) will be described.

A. Dataset

This study uses the Health Facts National Database (Cerner Corporation, Kansas City, MO), gathering extensive clinical records across hundreds of hospitals throughout the US [18]. The data subset used for analysis covers 10 years of diabetes patient encounter data (1999 – 2008) among 130 US hospitals with over 100,000 diabetes patient. Moreover, all the encounters used for analysis satisfy five key criteria:

- It is a hospital admission.

- The inpatient was classified as diabetic (at least one of three initial diagnoses included diabetes).

- The length of stay was comprised from 1 to 14 days.

- The inpatient underwent laboratory testing.

- The inpatient received medication during its stay.

B. Data Pre-processing

As real-world medical data are often noisy, a particular focus will be led on pre-processing task handling both missing data and inconsistencies but also by reducing the dataset and optimizing it for further model deployment [19]. While some pre-processing steps are based on an understanding of the data and background, this study will combine and implement the most relevant pre-processing identified in the body of literature [11], [14], [15]. As such, Hotdeck imputation with Approximate Bayesian Bootstrap (ABB), ICD-9-DM clustering and Random Forest feature selection constitutes some of the key pre-processing tasks.

In order to empower the dataset and improve the model performance, the importance of each input feature against the output variable will be assessed while non-important features will be excluded. Part of previous researches selected variables

based on medical expertise or based on the p-value of each variable extracted from linear regression. This study will consider another approach to feature selection using a random forest classification algorithm. Hence all relevant feature selection wrapper algorithm will be used to extract the variable importance measure for each feature with a random forest method.

The random forest was selected due to its speed of execution, and due to the fact, that it can be run without tuning of parameters while providing a numerical assessment of the important features. The selected method will perform a top-down search for important variables by comparing the original feature's with relevancy obtained at random, assessed applying their permutation and recursively eliminating unimportant variables to stabilize the testing.

The feature selection will follow the below process [20]:

1) Empower the information system by adding copies of all features (at least 5 shadow variables).

2) Shuffle the shadow attributes in order to minimize their correlation with the response.

3) Perform a random forest classification and collect the computed Z scores.

4) Identify the Maximum Z score in the Shadow Attributes (MZSA) and assign a hit to each variable scoring above the MZSA.

5) Perform a 2-sided equality test with the obtained MZSA for variables with undetermined relevancy.

6) Label all attributes with importance significantly below the obtained MZSA as "unimportant" and drop them from the information system.

7) Label all attributes with importance significantly above the obtained MZSA as "important".

8) Allow shadow variables are removed.

9) Iterate the same process until every attribute is assigned with a level of importance.

C. Pre-Processing Performance Benchmarking

Logistic Regression is one of the classification algorithms, which were used in assessing the performance of the pre-processing stages in predicting 30 days unplanned readmission among diabetes patients [4], [14]. Therefore, this study considers Logistic Regression to benchmark the pre-processing performance before building to the core model.

This algorithm is based on some strong assumptions including a binary target variable, no misclassified instances and clean from outliers [15]. The objective of the logistic regression model is to summarize data characteristic and obtain the optimal fitting model to define the relationship between the binary class target and the predictor variables. The model is optimized to generate the optimal coefficient for each variable to predict the logit transformation with the probability of the characteristic presence of interest based on the training set. Moreover, as opposed to the Linear Regression that selects parameters, which reduce the sum of squared errors, the logistic regression selects parameters which optimized the probability of perceiving the sample values.

D. Class Imbalance

Imbalanced data in a classification problem possess a significant challenge in quality of result obtained through the predictive models. This is defined by an uneven frequency distribution among each output class will lead to biases in the majority class. SMOTE is a commonly used technique to cope with this issue [21]. This method consists of artificially generating new records of the less represented class by employing the nearest neighbors of these observations. In parallel the majority class is under-sampled, contributing to well-balanced output classes.

While under-sampling would be a preferable method, the dataset is not large enough to allow further reduction that could significantly reduce the model performance. Hence, SMOTE can be viewed as an efficient way to overcome imbalanced data as it also increases the sample size [16].

E. Model

A Neural Network (NN) model is distinguished by an activation function, applied by interconnected processing nodes to convert the input into output. The initial layer of the NN collects the raw input then processes it and delivers the processed data to the hidden layers. These hidden layers process and deliver, in their turn, the processed information to the last layer, hence producing the final output. Moreover, a relevant cost function should be selected for optimal performance. Such a function has to learn how to provide the best solution to the classification problem.

Multilayer Perceptron (MLP) is the simplest form of NNs constituted of three layers. The initial layer is the input layer followed by a hidden layer and terminated by the output layer. Each layer can be composed of one or several neurons. Such a perceptron model collects multidimensional input and then processes it with an activation function and weighted summation. The training uses label data and learning methods that enhance the weights for the summation process. In order to achieve even greater performance and be able to deal with non-linear situations the model can be complexities by expanding the number of hidden layers, the number of neurons and the number of links between each layer. Such a model is called an MLP neural network.

IV. RESULTS AND DISCUSSIONS

The experiment was carried out using R programming. The data includes 101,766 rows and 49 columns (5,088,300 data points). Of all 50 columns, 37 are nominal, 13 are numeric. The output variable is the column labeled "readmitted" which is encoded a 3-class classifier including "<30 days", ">30 days", "Not readmitted". The full initial set of data also comprises 2 ID type variables, "Encounter ID" and "Patient Nbr".

A. Data Pre-Processing

1) Missing values: The first step in cleaning the data consist of handling missing values. Missing values refers to the absence, voluntary or not, of data in a record. While the initial step is to identify and encode missing values, the second step consists in addressing the missing values.

Each variable comprising missing values were independently analyzed, as the methods to be applied differs based on statistics but also best practices and industry knowledge. In this particular case, the missing values are encoded as "?", which is not a standard missing value format. Therefore, the first step in addressing missing values will be to encode them properly.

As a general rule, variables with 50% or more missing values should be dropped from the analysis. The variable medical specialty comprises 49% of missing observations. In term of proportion, the whole column should be dropped. However, based on background understanding and recommendation from previous researches such variable is of prime importance when predicting readmission [15]. Hence, the missing values were encoded as a new category labeled "Missing". Moreover, the social economic status of the patient is a critical factor in predicting readmissions; therefore variable such as "Payer code" should be preserved in the dataset. In addition, the listwise deletion was performed for variables with very few missing values as the dataset is large enough to maintain significant weight.

For the rest of the variables with low to average missing rate, imputation was conducted in order to maintain as much data as possible for further modeling [4], [15]. Indeed, imputation methods are crucial in order to minimize non-response bias and to generate efficient predictors. Many techniques can be applied for imputation such as mean imputation or KNN imputation. Considering the mixed nature of the variables, Hot-deck imputation appeared to be of ideal fit both in term of efficiency and accuracy. Such a method is recommended to eliminate non-answer bias in survey data [22]. The imputation is done by replacing missing values with observed values from similar observation. Hot-Deck imputation was applied, being suggested by the body of literature as a reliable and commonly used imputation technique for a similar type of data [23]. The method consists of selecting a donor for a recipient cell and the values of the donor are imputed for each missing observation. Moreover, the Approximate Bayesian Bootstrap (ABB) donor selection method was used for processing. Indeed, ABB provides benefits in adding relevant uncertainty when imputing missing values.

2) Inconsistencies: Data inconsistencies compromise data integrity and alter the performance of the algorithm. As a result, the second cleaning step resides in addressing such bias in data. Based on the body of literature this particular set of data has some specific inconsistent features to be addressed. For instance, several patients from the medical records had multiple admission and should not be treated as an independent encounter as it would bias other observations. In order to ensure a unique identifier for each patient, previous researchers suggested keeping only the first encounter when a patient had multiple record [14], [15]. Indeed, multiple encounter aggregation appeared not to be efficient while keeping the last encounter generated highly imbalanced data in term of output.

The dataset comprises records without any diagnosis ("diag_1", diag_2" and "diag_3" all missing). Previous literature advises on deleting observations meeting such condition as being a synonym of poor data quality [14].

Discharge Disposition refers to the person's location or status after admission in the healthcare center. As per supported by contemporary researchers, patients who died during their admission have no probability to be readmitted and should be hence excluded from the analysis [14], [15]. Therefore, all records with expired discharge were deleted. Moreover, patients discharged to hospice, referring to end of life care, were also omitted for the same reason.

Two variables, namely, "examide", "citoglipton" having the same observation ("No") for every record in the dataset. Such features will, as a result, be dropped from the analysis.

3) Data reduction: After having cleaned the data from missing values and other potential bias, it is important to optimize the feature and, mostly in this case, reduce the number of unique values for categorical variables. Hence clustering was performed to group similar observations into the same group (cluster). As this study focuses on improving the current classification models and not bringing up novel pre-processing techniques, this clustering step will mainly follow the scheme from existing literature [11], [14].

One of the most critical pre-processing steps for this set of data was to cluster the diagnosis codes, namely "diag_1", "diag_2" and "diag_3", from the ICD-9 -CM format into fewer comorbidity features. Indeed, each diagnosis variable comprises more than 700 individual ICD codes which make it unusable for further modeling and interpretation. Hence, the diagnoses codes were collapsed in nine categories including "Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Others".

Medical Specialty variable comprises more than 70 unique values and hence makes it extremely difficult to be used in models. As a result, medical specialty observations will be clustered based on a semantic term including the correction of typos. For example, all categories related to surgery will be clustered as "Surgery" (i.e. Surgeon, Surgery-General, Surgical Specialty).

Similarly, to diagnosis codes or medical specialty, admission source of the admission type variable could be further clustered. For example, trauma, urgent care, and emergency were merged as an emergency.

Among categorical variables, several contained excessive amount categories and hence had to be clustered to improve data quality. The initial medical specialty 73 categories were clustered down to 8 categories applying both a semantic and frequency-based approach. The initial diagnosis (1-3) 700+ categories were clustered down to 9 categories based on the ICD-9 categories and medical expertise.
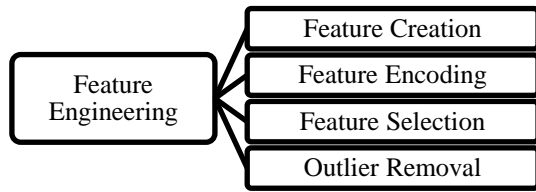
Fig. 1.   Feature Engineering Process.

4) Feature engineering: Several feature engineering steps were taken in order to perform both feature creation, encoding, feature selection, and data scaling. Indeed, while some feature engineering steps are based on the data and business understanding others such as variable encoding taken into account the requirements of future algorithms to be applied. In this study, the feature engineering step will be subdivided into 4 categories, refer Fig. 1.

a) Feature Creation: The design of three additional features which are "Service Utilization", "Count of Medication Change", and "Count of Medication Used" to enrich and empower the original dataset. While researchers used this approach for feature reduction by dropping individual variable after feature creation. However, this study maintained all variables and perform a feature selection algorithm to statistically assess the importance of each variable. Indeed, some individual grouped variable such as insulin might have a strong predictability weight and should, therefore, be preliminarily evaluated before any drop is performed.

The initial dataset contains three key features measuring the utilization of the hospital facilities by the patient such as "number_outpatient", "number_emergency" and "number inpatient" variables which measure how much the medical services were utilized by the given patient over the last year. "Service utilization" variable will be engineered by the sum of all above features.

The initial set of data includes 23 medication related features, each associated with 4 classes, namely "No", "Steady", "Up" and "Down". Such categories aim at assessing whether a change of medication occurred during the patient's admission. Several studies highlighted medication change as an influential factors towards readmission [14], [16], [24]. Hence, a new feature label "medication_change" will be engineered by counting all changes in medication for all records.

A higher total number of drugs might be an indicator of the condition's severity or care intensity of the patient. As a result, a new variable labeled "Count of Medication Used" will be engineered by summing all drug used during the hospital stay.

b) Feature Encoding: As neural networks will be applied in the later model stage, features will need to be encoded as numeric. The encoding process is discussed below.

Reduce Output Class to Binary: The objective of this study is to predict whether a patient will be readmitted or not within the next 30-days after discharge. Therefore, the scope of the study is limited to discrimination between <30 readmissions and no readmission. However, the dataset a hand comprises 3 classes for the "readmitted" output, including readmission occurring below 30 days (11.2%), above 30 days (34.9%) and non-readmission (53.9%). Dropping readmission occurring after 30 days would, therefore, result in a loss of more than one-third of all observations. Similar to other researchers, readmissions occurring after 30 days will be considered as non-readmission.

Encode Age as Discrete: The studied dataset provides the "Age" feature as 10-year groups. While some researchers decided to keep it encode as categorical nominal variables [4], [15], this study will convert it into numerical type. Indeed, with a categorical variable, the effect of increasing age on the output is less perceivable. Hence, following the method applied in recent literature, this study will consider the average age of each category at the midpoint of each age category [14]. As for an example, age group 10-20 will be converted to 15. As per this method, the age feature will be converted to numerical type.

Encode other variables: The studied dataset encoded most of the variable in string format including race, gender, medication change, and all the medication used features. Hence medication change will be re-encoded into 0 and 1 values respectively for "No" (no change) and "Ch" (change). Moreover, all medication used features will be simplified as "Change" and "No Change" and will be encoded as 0 and 1. A1C and Glucose serum test results will also be simplified into three categories, namely "normal", "abnormal" and "not tested". A similar process will be applied to remaining non-numerical variables.

c) Feature Selection: Several techniques were tested to perform feature selection. The initial test using logistic regression and regularization technique coefficient's p-values appeared to drop highly significant features both from a medical perspective but also as per highlighted by the body of literature. Hence, the random forest feature selection was performed. The variable importance was then computed during 60 iterations. A total of 25 predictors were hence selected after 60 iterations, see Table 1.

TABLE I.   SET OF SELECTED FEATURES

Selected Features (after 60 iterations)		
"race"	"medical_specialty"	"diag_1"
"age"	"num_lab_procedures"	"diag_2"
"admission_type_id"	"num_procedures"	"diag_3"
"discharge_disposition_id"	"num_medications"	"number_diagnoses"
"admission_source_id"	"number_outpatient"	"max_glu_serum"
"time_in_hospital"	"number_emergency"	"A1Cresult"
"payer_code"	"number_inpatient"	"insulin"
"change"	"diabetesMed"	"service_utilisation"
"count_medication_change"	-	-

Two of the engineered variables (count of medication change and service utilization) were selected as important variables by the feature selection algorithm. Hence it demonstrates the efficiency of such variable creation, empowering the dataset with highly predictive additional features.

d) Outlier Removal: The preliminary exploration of the numerical variables in the section highlighted that most of them were highly skewed with high kurtosis. As a rule of thumb, positively skewed variables were addressed with log(x+1) transformation while negatively skewed featured were addressed with a cube root method. Log(x+1) will be used instead of log(x) in order to retain the 0 in the data and avoid any missing value issue.

After having log transformed the data to ensure normal distribution, numerical data will then be standardized. This method appeared to lead to high accuracy and better model fit. Moreover, the data would be close to normal distribution. The outliers can now be treated using the coverage rule for normal distribution using standard deviation as a discrimination threshold. This study will follow the method used by previous researchers consisting of deleting any data outside of a 3 standard deviation range, corresponding to 99.7% of all values [14].

B. Class Balancing

The output variable "readmitted" appeared to be relatively imbalanced with the below distribution, refer to Table 2. Such distribution should be addressed as it may alter the generalizability of the model.

The SMOTE algorithm was found particularly efficient, generating a target class distribution close to 50%. The distribution after and before SMOTE is summarized in Table 3.

C. Pre-Processing Assessment

The pre-processing stage inherent performance was benchmarked by applying a logistic regression model. The comparison of the AUC of this proposed pre-processed data against the AUC of existing study pre-processing approach on the logistic regression model was performed. The AUC obtained on the proposed pre-processed dataset achieved AUC metrics of 62.2%. The current pre-processing techniques provide more added value in term of model performance than the most recent study achieving 61% [14].

On the other hand, the model's AUC metric falls short against [4] model scoring 67%. However, this previously mentioned model didn't include any data balancing in its pre-processing, and the trained and tested data is hence highly imbalanced. Therefore, such a difference in performance can be explained by the target class imbalance. As the model developed by [14] also use SMOTE to balance the data, it appears as a more reliable source of comparison. As a result, the proposed pre-processing appears to allow better model performance than the previous studies.

TABLE II. OUTPUT CLASS DISTRIBUTION

Class	Frequency
<30	11.2%
>30	34.9%
NO	53.9%

TABLE III. DATA BALANCING OUTPUT

Output Class	Observations before SMOTE	Observations after SMOTE
0	56146 (91.58%)	30990 (46.15%)
1	5165 (8.42%)	36155 (53.85%)
Total	61311	67145

D. Multilayer Perceptron

The proposed MLP model is constituted of one input layer, one hidden layer with uniform initialization and one output layer. Input layers are activated using PRelu function while the output layer is using the sigmoid activation function. Dropout (regularization) of 0.15 will be added after each input layers in order to limit overfitting and hence boost the model's performance. The model is compiled with a Mean Square Error (MSE) loss function and with Adam optimization function using accuracy as a metric. Batch size of 500 is applied across 600 epochs (iterations). 512 hidden units for each input layers will be utilized.

The model was assessed based on an 80:20 train-test ratio using both accuracy, precision, recall, and AUC as performance metrics. Machine learning algorithms being stochastic, the same algorithm is subject to randomness and might give slightly different results at each training. As a result, the proposed MLP model will be run and assessed five consecutive times in order to provide more reliable performance estimation, refer to Table 4.

The proposed model achieved a high score on all evaluated metrics, with optimal performance on the recall metrics (99%). The model also appears to be particularly balanced with high accuracy and AUC of 95%. Finally, the model performs the least in term of precision, but still achieves high performance of 93% on average.

TABLE IV. PROPOSED MLP MODEL PERFORMANCE ON TEST SET

Metric	Test 1	Test 2	Test 3	Test 4	Test 5	Average
Accuracy	96%	95%	95%	95%	95%	95%
Precision	95%	93%	93%	93%	93%	93%
Recall	98%	99%	99%	99%	99%	99%
AUC	96%	95%	95%	95%	95%	95%

E. Model Benchmarking

The proposed model will now be benchmarked with recent best performing models from the body of literature. The Convolutional Neural Network (CNN) proposed by [16], Random Forest Gini Algorithm from [14] and RNN developed by [15] will be used for this comparison, see Table 5.

The proposed model consistently outperforms the other models on both accuracy (1 point of increase) and recall metrics (9 points of increase). The AUC equals the one obtained by [16] with CNN and outperforms the other models. Despite, achieving similar AUC metrics (95%), the CNN model, however, falls short in term of accuracy (92%) while precision and recall metrics weren't available for comparison. While the proposed model achieves inferior performance, in term of precision compared to [15] Random Forest model (98%), the proposed model appears to outperform its counterpart on accuracy (95% vs. 94%), recall (99% vs. 90%) and AUC (95% vs. 94%).

The proposed MLP brought out state of the art performance on this particular hospital readmission problematic and dataset. The model efficiency was found particularly balanced across the key performance metrics all achieving above 93%. The main contribution of this model in term of performance is toward recall which achieves close to an optimal score of 99% and outperforms previous studies by 9 points. Such improvement is the main highlight of this specific problem as recall is of prime importance when dealing with medical data.

Such advances toward better hospital readmission classification can be attributed to the comprehensive data pre-processing of the noisy medical data addressing both missing values, inconsistencies, outliers, feature selection, and class imbalance challenges. The utilization of advanced parameters such as PReLU activation function, Adam optimizer or dropout to limit overfitting is also responsible for this enhancement. Moreover, the developed MLP model might also have faster training and use less computational power than some other models such as convolutional neural networks, and a recurrent neural network.

This particular research was also only oriented toward data from US hospitals and may not be generalizable to other countries or settings. The current data was comprehensive but not exhaustive as other important influential factors could have been captured. Classification quality is indeed dependent on both the data volume available and its variety. For example, features like deprivation indexes and access to care, which was shown to account for 58% of readmission rates variation, could further strengthen the model capabilities [14]. Another health center specific information like the patient's distance from services and hospital readmission rates could also provide great enhancement [4]. From a pathology point of view, diabetes type and duration would be a highly determinant factor in hospital readmission.

TABLE V.   ODEL BENCHMARKING

Metric	Random Forest Gini	CNN	RNN	Proposed MLP
Accuracy	94	92	81	95
Precision	98	-	-	93
Recall	90	-	-	99
AUC	94	95	80	95

V.   CONCLUSION

30-day hospital readmission of diabetes patients is of prime importance for health centers and is found very stressful due to the current models limit in term of performance and generalizability. To cope with this challenge, this study implemented a comprehensive pre-processing framework in order to improve the initial data quality, hence empowering the model's efficiency. The suggested pre-processing framework included comprehensive data cleaning, data reduction and transformation aiming at better optimizing and selecting prominent features for 30-day unplanned readmission among diabetes patients. Random Forest algorithm for feature selection and SMOTE algorithm for data balancing are some examples of methods during pre-processing.

The proposed Multilayer Perceptron model combined with this feature engineering was found to outperform other machine learning algorithms in term of prediction quality. More specifically, the performance of the designed model was found robust, scalable and particularly balanced across different metrics of interest with accuracy and Area under the Curve (AUC) of 95% and close to the optimal recall of 99%.

The studied dataset provides an array of information both in term of administrative data, demographics and medical data about hospital readmissions of diabetes patients. However, various limitations should be acknowledged. The data at hand has a limited time range (1999-2008), the availability of information spanning across a wider period could improve significantly the performance of the models. Furthermore, a newer set of data would be preferable to have more realistic information about hospital readmission for diabetes patients in recent years.

In term of pre-processing, under-sampling would be preferable to achieve better quality if the amount of data at hand was larger. Finally, the black box problem is the key limitation of this study in term of modeling. In fact, while multilayer perceptron provides state of the art classification performances, its interpretation remains limited. As such, future research could be led in this direction by the implementation of hybrid models, for example, to harness the prediction quality of deep learning while providing a certain degree of interpretability.

REFERENCES

[1] World Health Organization, Global report on diabetes. World Health Organization, 2016.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big Data for Health," IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[3] N. Bhardwaj, B. Wodajo, A. Spano, S. Neal, and A. Coustasse, "The Impact of Big Data on Chronic Disease Management," Health Care Manag. (Frederick)., vol. 37, no. 1, pp. 90–98, 2018.

[4] K. Hempstalk and D. Mordaunt, "Improving 30-day readmission risk predictions using machine learning," in Health Informatics New Zealand (HiNZ) Conference 2016, 2016, no. December.

[5] C. Baechle and A. Agarwal, "A framework for the estimation and reduction of hospital readmission penalties using predictive analytics," J. Big Data, vol. 4, no. 1, p. 37, Dec. 2017.

[6] J. Albritton, T. Belnap, and L. Savitz, "The Effect of the Hospital Readmission Reduction Program on the Duration of Observation Stays: Using Regression Discontinuity to Estimate Causal Effects," eGEMs (Generating Evid. Methods to Improv. patient outcomes), vol. 5, no. 3, pp. 1–7, Dec. 2017.

[7] D. Kansagara et al., "Risk Prediction Models for Hospital Readmission," JAMA, vol. 306, no. 15, p. 1688, Oct. 2011.

[8] Medicare.gov, "30-day unplanned readmission and death measures," 2017.

[9] H. Zhou, P. R. Della, P. Roberts, L. Goh, and S. S. Dhaliwal, "Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review," BMJ Open, vol. 6, no. 6, p. e011060, Jun. 2016.

[10] S. Damery and G. Combes, "Evaluating the predictive strength of the LACE index in identifying patients at high risk of hospital readmission following an inpatient episode: A retrospective cohort study," BMJ Open, vol. 7, no. 7, 2017.

[11] D. Mingle, "Predicting Diabetic Readmission Rates: Moving Beyond Hba1c," Curr. Trends Biomed. Eng. Biosci., vol. 7, no. 3, Aug. 2017.

[12] L. L. Low, N. Liu, S. Wang, J. Thumboo, M. E. H. Ong, and K. H. Lee, "Predicting 30-Day Readmissions in an Asian Population: Building a Predictive Model by Incorporating Markers of Hospitalization Severity," PLoS One, vol. 11, no. 12, p. e0167413, Dec. 2016.

[13] L. L. Low et al., "Predicting 30-Day Readmissions: Performance of the LACE Index Compared with a Regression Model among General Medicine Patients in Singapore," Biomed Res. Int., vol. 2015, pp. 1–6, 2015.

[14] C.-Y. Lin, H. S. Singh, R. Kar, and U. Raza, "What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients?," Berkeley, 2018.

[15] C. Chopra, S. Sinha, S. Jaroli, A. Shukla, and S. Maheshwari, "Recurrent Neural Networks with Non-Sequential Data to Predict Hospital Readmission of Diabetic Patients," in ICCBB 2017 Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, 2017, pp. 18–23.

[16] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, "Predicting Hospital Readmission among Diabetics using Deep Learning," Procedia Comput. Sci., vol. 141, no. November, pp. 484–489, 2018.

[17] P. Zhao and I. Yoo, "A Systematic Review of Highly Generalizable Risk Factors for Unplanned 30-Day All-Cause Hospital Readmissions," J. Heal. Med. Informatics, vol. 08, no. 04, 2017.

[18] B. Strack et al., "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," Biomed Res. Int., vol. 2014, pp. 1–11, 2014.

[19] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India," Int. J. Diabetes Dev. Ctries., vol. 36, no. 4, pp. 469–476, Dec. 2016.

[20] M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," vol. 36, no. 11, 2010.

[21] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," in 2017 International Conference on Engineering Technology and Technopreneurship, ICE2T 2017, 2017, vol. 2017–Janua.

[22] SAS, "The SURVEYIMPUTE Procedure," support.sas.com, 2018.

[23] D. Rithy, "Simulation of Imputation Effects Under Different Assumptions," in SAS Global Forum 2016, 2016.

[24] N. J. Wei, D. J. Wexler, D. M. Nathan, and R. W. Grant, "Intensification of diabetes medication and risk for 30-day readmission," Diabet. Med., vol. 30, no. 2, pp. e56–e62, Feb. 2013.