Received: March 9, 2021. Revised: April 26, 2021.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Enhanced Predictive Modelling for 30-Day Readmission Diabetes Patients Based on Data Normalization Analysis

Eman H. Zaky[1]* **Mona M. Soliman[2]** **A. K. Elkholy[1]** **Neveen I. Ghali[3]**

[1]*Faculty of Science, Al-Azhar University (Girls branch), Cairo, Egypt*
[2]*Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt*
[3]*Faculty of Computers and Information Technology, Future University in Egypt, Cairo, Egypt*
* Corresponding author's Email: eman.hanfy@azhar.edu.eg

Abstract: In the field of health care, one of the most important problems is predicting the possibility of hospital readmission due to its important role in caring for patients with chronic diseases such as diabetes. Such predictions affect the health care costs and the hospital's efficiency and reputation. In this paper, an intelligent-based model is developed to predict the reintroduction of the patient into the hospital. This model is based on using some Machine Learning (ML) algorithms such as Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Also, it proposes the use of a Deep Learning (DL) based network such as a convolutional neural network (CNN). Both ML and DL are used as classifiers to predict hospital readmission. The main problem is the input noisy data to these classifiers. These noisy data reduce the accuracy of the readmission prediction model. Sequential pre-processing steps are proposed to get over such a problem. These pre-processing steps provide solutions to missing values, feature engineering, and normalization problems. The main contribution of this work is improving readmission prediction rate by solving the data normalization problem. Two types of data normalization (e.g. z-score and min-max normalization) are applied, results show there is a difference in accuracy, z-score normalization is better than min-max normalization when comparing ML methods and DL models, CNN is the best with an accuracy of 0.894% in case of z-score normalization. Moreover, the model performance is improved with an accuracy of 0.924% when non-normalized data is used as input to the model. The proposed Non-normalization technique successes in providing superior results compared to some previous techniques which are displayed data by using Ensemble, Normalization, and Ensemble by age group techniques.

Keywords: Machine learning (ML), Deep learning (DL), Feature engineering, Pre-processing, Diabetes, Hospital readmission, Normalization - Non-normalization techniques.

1. Introduction

Diabetes is one of the most frequent non-communicable chronic diseases, which is expected to turn into the fifth most prevalent factor of mortality by 2035, as the statistics of the Ministry of Global Health in 2000 proved that diabetes was not among the prevalent mortality factors, but in 2016 statistics, it became the seventh most common factor [1]. Deaths are so prevalent that diabetes is a priority in the health agenda for developed and developing countries.

The health care sector collects and processes medical data for diabetes patients in huge quantities varied size, structure, and real-time data flow with the advent of technology, both from diagnosis, monitoring, storage, and analysis, new solutions are now available for the best facing challenges [2].

For patients with chronic diseases, complete recovery from the disease is difficult and the difficulty increases when it leads to subsequent readmission. In the United States, it was reported through the Centers for Medicare and Medicaid Services (CMS) that 76% of patients who were admitted to the hospital could have been avoided [3]. Some studies have shown that the risk factors for

readmission to the hospital, especially for the elderly are several, patient characteristics, disease characteristics, and health care system factors that predict readmission [4].

Other studies have also used clinical data collected from patients or patient hospital records [5]. Hospital discharge data were also used to identify factors associated with hospital readmission. Research has shown that reliance on these factors for determining readmission is better than random guessing [6]. In this research, ML and DL methods will be implemented to predict readmission into the hospital.

ML are algorithms that allow machines to learn through mathematical representations that enable machines to imitate the way humans learn.

DL is a subset of ML, it can be expressed as the new development of ML, it is a mechanical algorithm that mimics human notices, and it is the closest technology to how humans learn. The architecture of DL methods uses a neural network architecture, it refers to the heeding layers found in these neural networks.

In this paper, an intelligent model was proposed to perform the task of predicting hospital readmissions in the extensive clinical records, intelligent model is based on using either traditional ML algorithms or the DL model. Both methods are used as a binary classifier with only two outputs (0 as not readmitted and 1 as readmitted). The data set used in this proposed model has many features with different kinds of problems.

The main contributions of this work can be summarized as follow:

1) An enhanced version of the input data-set is introduced using some data pre-processing methods. These methods include: dealing with missing values problems, feature engineering, and data normalization.

2) An investigation analysis is performed on the prediction accuracy for hospital readmission based on patient's medical records. Such analysis shows the effect of using data normalization and Non-normalization techniques on readmission prediction accuracy.

3) An intelligent model is proposed to perform the task of predicting hospital readmissions using a modified version of clinical records. This model based on using traditional ML methods and DL model.

4) Compare intelligent methods in two cases normalization and Non-normalization techniques.

5) To display the efficiency of the DL method with Non-normalization techniques.

6) A comparative study between the proposed model and other state-of-the art models is introduced to ensure the superiority of our proposed model.

This paper is organizing as follows: section 2 provides a literature view of different hospital readmission prediction models. The proposed model with full description of the clinical report dataset is introduced in section 3. The experimental results are reported and discussed in section 4. Section 5 provides the conclusions of this proposed model.

2. Literature review

In the field of health care, some research articles have been developed during the past years in terms of anticipating readmission to the hospital as it is the critical application of health care to preserve the individual's life.

In [7], a hierarchical logistic regression model was developed for 567,850 patient records. It takes the IRF risk modification model into consideration of patient demographics, hospital diagnosis, procedure codes, and function in IRF admission, comorbidities, and past hospital use. It also took into account the number of days of IRF discharged through re-admission. The result was as follows: The 30-day average re-admission rate for IRFs was 12.4 ± 3.5, and the risk standard readmission rate was 13.1 ± 0.8. The statistic for our risk adjustment model was 70%.

In [8], the prediction model for hospital readmission was developed taking into account the unique features of the database learned by using the C5.0 tree as the primary classifier and (SVM) as a secondary classifier. The results from this study are as follows: SVM predictions are characterized by accurate values (true positive rates) of C5.0 predictions. The overall accuracy of the kit ranges from 81% to 85%.

In [9], several different classifications were proposed, as the study divided patients into 3 groups according to Age [<30, between [30 - 70], > 70]. A separate model was built for each group using some ML algorithms or combining them such as (random forest, different types of gradient enhanced trees, and SVM). The results from developed group models are: Group < 30 with accuracy 84%, group between [30 - 70] with accuracy 78.5%, and finally Group > 70 with accuracy 68.5%.

In [10], the study proposed was based on a model that predicted the number of patients who would be readmitted to the hospital by using pre-processing for data like normalization. Then compared some of the ML algorithms with Recurrent Neural Network (RNN) based model, RNN performance is very accurate compared to machine learning, especially on

non-sequential data. Hence, it can be used in health care to target high-risk patients, reduce the readmission rate, and provide the best health care. This study a chief in the case of Simple Neural Network (2-layer) [Area Under ROC Curve=0.61, accuracy 69.53%], in case of Recurrent Neural Network (2-Layer) [Area Under ROC Curve=0.80, accuracy 81.12%].

In [11], a model consisting of 5 models selected from 15 models was used, which were variants of logistic regression, decision trees, neural networks, and naive Bayes enhanced [12]. These models were selected after analyzing their accuracy, the performance of this model was on an unbalanced dataset 63.5% accuracy.

In [13], CNN of deep learning has been used in the problem of readmission of diabetic patients to hospital as an effective prediction method. This model achieves 92% performance and is better than other ML models. Whereas, this model relies on sample size enlargement and data engineering processes.

As such, the use of feature engineering, SMOTE to address the class imbalance inherent in clinical data, and the use of normalization are key to improving deep learning performance.

In [14], multivariate logistic regression was used with a ROC of 72.0%, including the DERRI proposal (The risks of early readmission for diabetes Contents), of the 43 suggested features, 13 statistically significant were selected and DERRI trained on them, first proposed that the HBA1C level had little to do with the readmission for 30 days. According to his research, low socioeconomic status, Ethnic / Ethnic Minority, common Burden of illness, public insurance, emergency or urgent admission, and previous last hospitalization history are some of the important factors responsible for 30-day patient readmission.

In conclusion, the state-of-the-art models depended on several methods which are displayed data by using Ensemble, Normalization, and Ensemble by age group techniques. As the data that we deal with is very more diversity, variety, and are closer to each other. Normalization leads, perhaps, to a loss of information, so normalization destroys diversity and thus the strength of recognition without normalization. The proposed model shows a predominance of Non-normalization technique in terms of accuracy measures.

3. The proposed model

An intelligent model based on normalization and Non-normalization techniques is proposed to develop a choice between two classes (0 as not readmitted and 1 as readmitted). We utilize the use DL model (e.g. CNN and RNN model) with three ML based classifiers for the prediction of readmission, the used classifiers are (KNN), (LR), and (SVM). This proposed study is performed on a dataset represent 10 years (1999-2008) of clinical care at 130 hospitals across the United States and is provided by the Center for Clinical and Translational Research at Virginia Commonwealth University. This data was used to predict the probability of readmission within the next 30 days for a patient with diabetes. The full proposed model is shown in Figure 1. It will be illustrated in full detail in the following sections.

3.1 Dataset description

The dataset contains 50 features [15], the information is extracted from the database with the following standard:

1. The case must be registered in the hospital.
2. In the database, only a diabetic is considered.
3. The duration of nursing stays in the hospital within the period from 1-14 days.
4. Lab tests were performed during the encounter.
5. Medicines were provided during the encounter.

In the end, 101.766 matches were identified that met all of the previous five conditions. It was used for further analysis. The full list of features and descriptions is set out in Table 1.

3.2 Data pre-processing

The original dataset can be described as 101766 records with 50 features which means 5,088,300 data points. For all 50 features, 37 are nominal, 13 are numeric. The output variable is the column labeled “readmitted” which is encoded “<30 days”, and “Not readmitted” with encoding “>30 days”. Data in this form was not ideally suited for the proposed intelligent model so some pre-processing methods should be running on it. The main three proposed pre-processing methods are: dealing with missing values, feature engineering, and normalization. Each one will explain in full detail.

3.2.1. Dealing with missing values

The first step in cleaning up data is processing lost values. The meaning of missing values indicates that the absence, voluntary or not, of data in a record. The first step is to identify missing values. The second step is addressing the missing values in three cases.

- Dropping columns with a large number of missing.
- Dropping attributes with a small number of missing

Table 1. Full list of features and description in the initial dataset [9]

Feature Name	Description and Values
Race	Values: African American, Asian, Caucasian, Hispanic, and Other
Gender	Values: female, male, unknown/invalid
Age	Grouped in 10-year intervals: [0-10), [10-20), …, [90,100)
Weight	Weight in pounds
Admission type	Integer corresponding to 9 distinct values
Discharge disposition	Integer identifier corresponding to 29 distinct values
Admission source	Integer identifier corresponding to 21 distinct values
Time in hospital	Integer number of days between admission and discharge
Payer code	Integer identifier corresponding to 23 distinct values
Medical specialty	Integer identifier of a specialty of the admitting physician, corresponding to 894 distinct values
Number of lab procedures	Number of lab tests performed during the encounter
Number of procedures	Number of lab test performed during the encounter
Number of medications	Number of distinct generic names administered during the encounter
Number of outpatient visits	Number of outpatient visits of the patient in the year preceding the encounter
Number of emergency visits	Number of emergency visits of the patient in the year preceding the encounter
Number of inpatient visits	Number of inpatient visits of the patient in the year preceding the encounter
Diagnosis 1	The primary diagnosis
Diagnosis 2	Secondary diagnosis
Diagnosis 3	Additional secondary diagnosis
Number of diagnoses	Number of diagnoses entered to the system
Glucose serum test result	Indicates the range of the result or if the test was not taken
A1c test	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7”, if the results were greater Change of medications, Indicates if there was a change in diabetic medications (could be dosage or generic name). Values: “change” and “no change”
Change of medications	Indicates if there was a change in diabetic medications (could be dosage or generic name). Values: “change” and “no change”
Diabetes medications	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
24 features for medications	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide,
Readmitted within 30-days	Days to inpatient readmission. Values: “1” if the patient was readmitted in less than 30 days and “0” for no record of readmission.

values.

● Some variables (e.g. drugs named citoglipton and discriminatory information for predicting readmission so decided to drop these two variables.

3.2.2. Feature engineering

Feature Engineering meaning several features perform (feature creation, encoding, and data scaling). While some feature engineering is based on the data and business understanding others [17].

Feature creation: There are several variables for patients in the database, such as the number of inpatients, emergency room visits, and outpatient visits for a particular patient. We added these three to create a new variable called service utilization.

service_utilization = number_outpatient + number_emergency + number_inpatient

Feature encoding: In the database, the number of drug-specific variables reaches 23 features for 23 drugs. There is research showing that changing drugs for diabetics upon admission is associated with lower rates of readmission [16].

From this standpoint, the number of drug changes for each patient was calculated, and new features were announced. The reason for this was to simplify and discover a relationship with the number of changes, regardless of which drug was changed.

Re-encoding admission type, discharge type, and admission source into fewer categories [9].

Encoding some variables: For example, “medication change” feature from (no change) “No” to 0 and (changed) “Ch” to 1, “gender” feature from “male” to 1 and “female” to 0 and” diabetesMed” feature from “Yes” to 1 and “No” to 0.

The A1C test and the results of the serum glucose test refer to the normal, abnormal, and untested categories [17].

If the patient's age category is 0-10 years, then assume the age = 1 year, the patient's age category is 10 - 20 years, then assume the age = 2 years, and so on [18].

As the subject of the study is whether or not the patient will be readmitted to the hospital within 30 days. This feature contains < 30, > 30, and the no readmission category. Dual classification is use, and combined readmission after 30 days and non-readmission into one category: replace (>30 with 0), replace (< 30 with 1) and replace (NO with 0).

Feature scaling (Normalization of dataset): Many ML algorithms aim to find trends in the data by comparing the features of the data. However, the problem is when the features are at very different levels. Normalization brings data on a common scale [19], [20]. This work focuses on two types of normalizations.

Z-Score normalization: the features are scaled in a way that they end up having properties of a standard normal distribution with a mean equal to zero and a standard deviation of one to get these coefficients.

To scale all values on data sets to Z using the following equation

Calculate $$Z = \frac{x+\mu}{\sigma_\mu} \quad (1)$$

subject to:

$$\mu = \frac{\sum_i^n x_i}{N} \quad (2)$$

$$\sigma_\mu = \sqrt{\frac{\sum_i^n (x_i-\mu)^2}{N}} \quad (3)$$

$$\sigma_\mu = \sqrt{\frac{\sum (x_1-\mu)^2 + (x_2-\mu)^2 + \ldots\ldots + (x_n-\mu)^2}{N}} \quad (4)$$

Eq. (1) represents the scale value (z) for each value of data sets (attribute).

Eq. (2) represents the mean average (μ) of the data set based on the data points or observation and the total number of data points in the data set. Where x_i is data points or observation, N is the total number of data points in the data set.

Eqs. (3) and (4) represent the standard deviation of the population based on the population mean, data points, and the number of data points in the population.

Min-Max normalization: the data is scaled in such a way that the values usually range between [0, 1]. Min-Max Normalization Formula can be described as follow:

$$v' = \frac{(v-min(A))}{(max(A)-min(A))} \quad (5)$$

Eq. (5) can be used to transform a value v of a numeric attribute A to v' in the range [0, 1]. Where $min(A)$ and $max(A)$ are the minimum and maximum values of the attribute.

3.3 The proposed intelligent based model

ML based methods and DL based model, as the intelligent based model, is built in spyder Python 3.7 environment with processor intel(R) Core(TM), i5-2500 CPU @3.30 GH_z using the Scikit-learn [21], TensorFlow [22], and Keras [23]. For (ML) based methods, various algorithms are used to build classifiers for the prediction of readmission. The implemented classifiers are (KNN), (LR), and (SVM). (LR) is based on predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, and true or false [25]. (KNN) as a classifier is based on a similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories then stores all the available data and classifies a new data point based on the similarity. (SVM) use the idea of creating the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future [26].

In DL, CNN is the network that employs a mathematical operation called convolution and uses convolution in place of general matrix multiplication in at least one of their layers [27]. Where each layer applies a linear transformation followed by a non-linearity to the preceding layer. To calculate the output of the CNN let:

$$X \in R^{N\times D} \quad (6)$$

$$w^k \in R^{d_{k-1}\times d_k} \quad (7)$$

$$X_{k-1}w^k \in R^{N\times d_k} \quad (8)$$

Eq. (6) represents the input data, where each row of X is D-dimensional data and N is the number of

training examples.

Eq. (7) represents a matrix of a linear transformation applied to the output of layer k–1. Where $x_{k-1} \in R^{N \times d_{k-1}}$,

Eq. (8) represents the obtaining a d^k-dimensional at layer k. For example, each column of w^k could represent a convolution with some filter (as in CNN).

After that the architecture chive Fully connected neural networks where all the nodes, or neurons, in one layer is connected to the neurons in the next layer.

$$\psi_k(x) = max\{0, x\} \tag{9}$$

Eq. (9) can be represented ψ_k as a non-linear activation function (PRelu) this applied to each entry of $x_{k-1}w^k$.

To generate the kth layer of CNN as

$$x_k = \psi_k\,(X_{k-1}w^k) \tag{10}$$

Where x_k is the output of CNN

The intelligent model based on the enhanced version of data-set is developed. It provides an investigation study on the effect of using normalized or non-normalized data within traditional ML methods(e.g LR, KNN, SVM) or (CNN, RNN) models. The proposed model classify between two classes (0 as not readmitted and 1 as readmitted).

CNN model is applied with one input layer, three hidden layers with uniform initialization, and one output layer. Softmax activation function was chosen for the output layer, while PRelu activation function was chosen for input layers. The selected optimization algorithm was Adam. Added Dropout with rate=0.1 after hidden layers to limit overfitting and hence DL model.

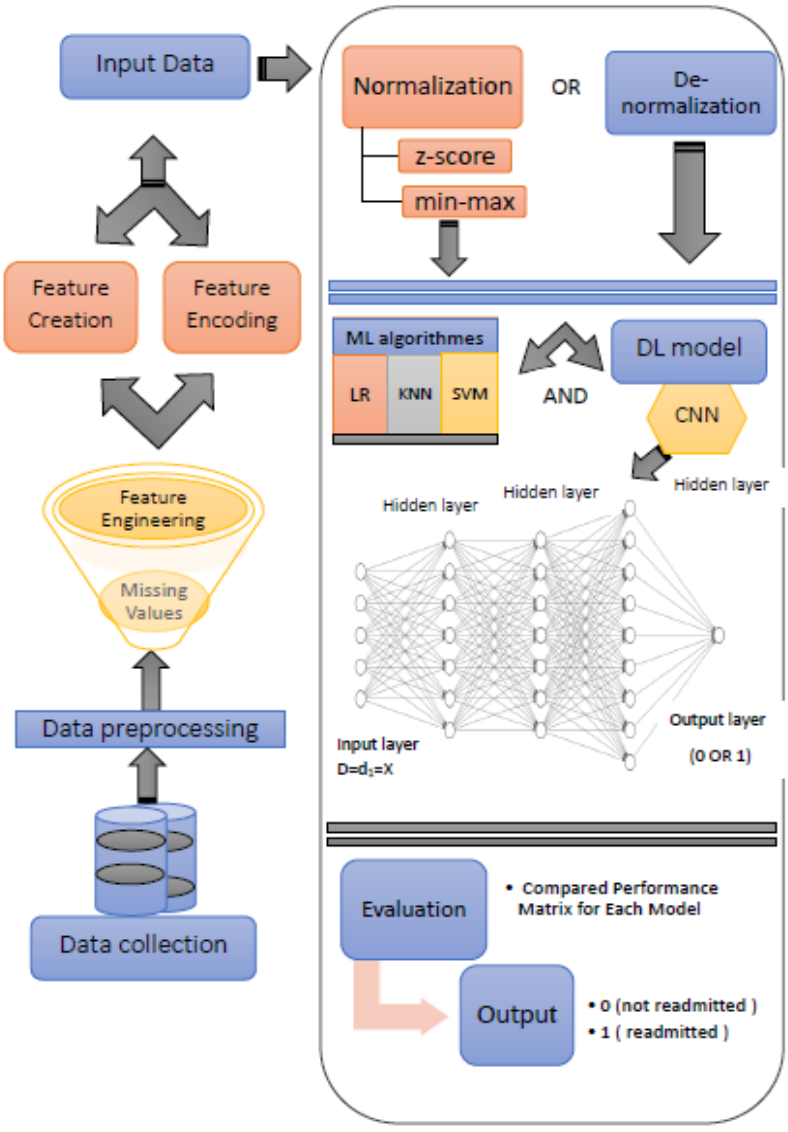


Figure. 1 Proposed model for prediction of Hospital readmission in diabetes data set

A comparison for predict readmission is done over the pre-processed data using ML algorithms (LR, KNN, SVM) and DL (CNN) models over normalization and Non-normalization techniques.

The main contribution of this work is resolving such problems using data pre-processing methods. Data pre-processing is an essential step for noisy data to enhance classification accuracy. pre-processing data are started by solving the missing values problem using data cleaning through different steps. Followed by some feature engineering to collect the most important and discriminate features. Finally, the normalization process is applied to bring data on a common scale. Mainly focus on providing a detailed analysis of the effect of using normalized data or non-normalized data on the readmission prediction accuracy.

4. Experimental results

This section validates the efficiency of using ML and DL based methods to predict hospital readmission. Two experiments are designed and report associated results. The first one use z-score and min-max normalization techniques with both ML classifiers and DL based network. The other experiment is done using DL and ML methods without normalization. The results are compared among all experiments. Such analysis is made to show the effect of normalization techniques on prediction accuracy. Also, the intelligent model which is based on Non-normalization technique is compared to some previous state of the art models that based on data Ensemble, Normalization, and Ensemble by age group techniques. All results are based on a set of standard evaluation metrics such as overall accuracy, recall, and precision.

4.1 Performance evaluation metrics

Accuracy, Precision, and Recall are used to indicate the performance of intelligent based model. When the model correctly predicts the positive category, in this case, the result is a true positive (TP), and likewise, when the model correctly predicts the negative category the negative result is true (TN). When the model incorrectly predicts the positive category, in this case, the result is a false positive (FP), and likewise, when the model incorrectly predicts the negative category the negative result is false (FN). From that accuracy can calculated, Precision, and Recall to indicate the performance of our model [11].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

Eq. (11) represents Accuracy in which is a measure of the rating model's performance. In other words, it is part of the predictions that our model got correctly.

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

Eq. (12) represents Precision, precision is a proportion of positive identifications that are actually correct.

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

Eq. (13) represents Recall, recall is the proportion of actual positives that are identified correctly.

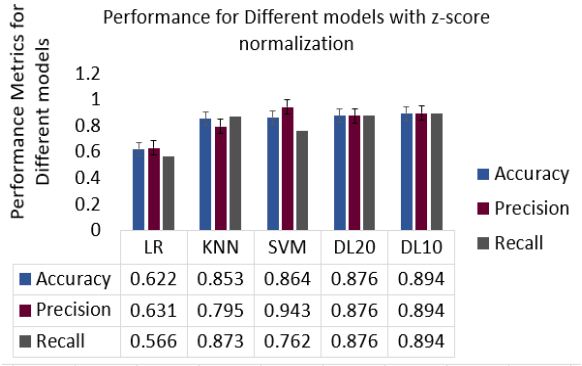
4.2 An intelligent model for readmission prediction using data normalization

In this experiment, both ML-based classifiers against DL based model are compared for hospital readmission in diabetes data set after applying z-score and min-max scaling as two normalization pre-processing steps. The following Table 2 compare the DL model in two cases (with test size 20 (DL(20)) and test size 10 (DL(10)) with ML algorithms (LR, KNN, SVM).

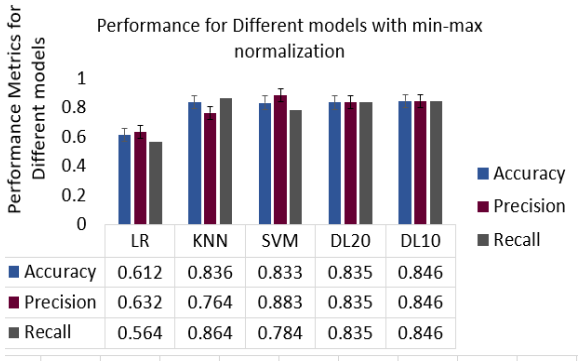
Fig. 2 shows a compare performance metrics for ML algorithms and the DL model when z-score normalization and min-max normalization are used with test size 10. As shown in Fig. 2 the performance of DL model is more accurate in predicting the use of ML in both cases, when z-score and min-max normalization are used. ML algorithms always need structured data, while DL networks rely on ANN

Table 2. Performance of DL and ML with two normalization pre-processing steps

z- score normalization (standardization)			
models	**Accuracy**	**Precision**	**Recall**
LR	0.622	0.631	0.566
KNN	0.853	0.795	0.873
SVM	0.864	0.943	0.762
DL(20) test	0.876	0.876	0.876
DL(10) test	0.894	0.894	0.894
min-max scaling (normalization)			
models	**Accuracy**	**Precision**	**Recall**
LR	0.612	0.632	0.564
KNN	0.836	0.764	0.864
SVM	0.833	0.883	0.784
DL(20) test	0.835	0.835	0.835
DL(10) test	0.846	0.846	0.846



(a)



(b)

Figure. 2 Compare performance for different models: (a) with z-score normalization and (b) with min-max normalization

(Artificial Neural Networks) layers. Therefore, the performance of DL was better as the data is not structured but it is multi-dimensional data. Also, ML algorithms need human intervention when actual output is not required. But DL does not need human intervention because the nested layers in neural networks place data through hierarchies of various concepts, which ultimately learn through their errors. However, we note that in DL, it is the quality of the outcome depends on the quality of the data.

4.3 An intelligent model for readmission prediction using data Non-normalization

In this experiment, applying different ML algorithms (LR, KNN, SVM) and DL models (CNN, RNN) on the dataset, with default parameters and without using any normalization are proposed, where we have more diversity and variety in our data. Normalization destroys diversity and scaling data in one range that leads to a loss of information. Hence test the power of recognition without normalization. The performance of the DL model at its best efficiency is compared at epochs = 1000 with some ML algorithms, as shown in the following table 3.

Fig. 3 shows the compare performance metrics for the DL model and ML algorithms As shown in

Table 3. Performance of DL, and ML without normalization and different epochs

Without normalization			
models	**Accuracy**	**Precision**	**Recall**
LR	0.642	0.624	0.544
KNN	0.872	0.822	0.951
SVM	0.886	0.985	0.773
RNN	0.837	0.842	0.862
DL(epochs=200)	0.902	0.902	0.902
DL(epochs=400)	0.913	0.913	0.913
DL(epochs=600)	0.915	0.915	0.915
DL(epochs=800)	0.912	0.912	0.912
DL(epochs=1000)	0.924	0.924	0.924

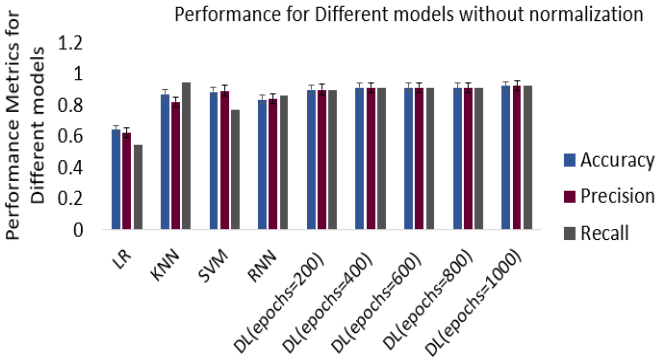


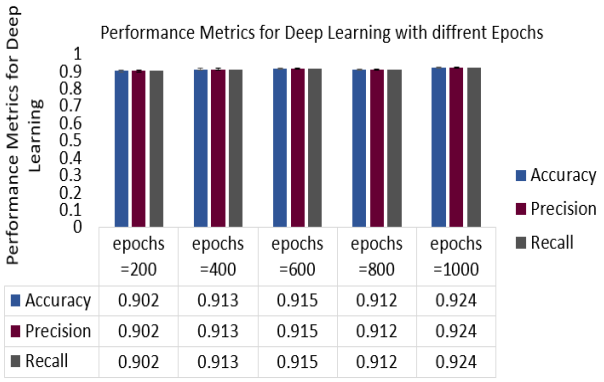
Figure. 3 Compare performance for different model using Non-normalization

Fig. 3 the performance of the DL model is more accurate in prediction than ML methods in case Non-normalization step is applied. A performance can saw improvement when increase the number of epochs. On the other hand, CNN takes constant inputs and gives a steady output that allows it to calculate results at a faster pace and more efficiently.

Applying the DL model on the dataset, with default parameters, without using any normalization and increase epochs to 400, 600, 800, 1000 the accuracy will be increased. Results are obtained in the following Table 3 in case of test size (10).

Fig. 4 shows the performance metrics for the DL model by increasing the number of epochs. As shown in Fig. 4 the performance of the DL model increasing by increasing the number of epochs without any overfitting until reaches standardized at epochs=1000.

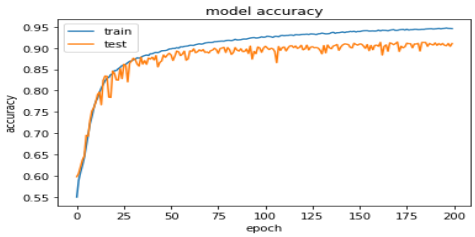
Figs. 5 and 6 show accuracy and loss curves at different numbers of epochs (e.g. 200,400,600,800, and 1000). As shown in these Figures, the performance of our model improves when increasing the number of epochs without any overfitting until it reaches a stability stage and this is the stage in which the model is not affected by any increase, and overfitting also occurs. This starts from 1200 epochs.



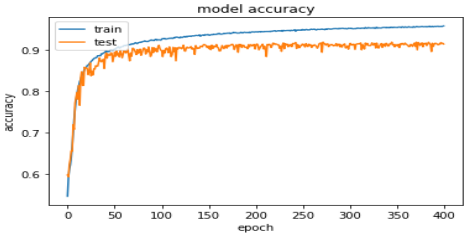
	epochs =200	epochs =400	epochs =600	epochs =800	epochs =1000
Accuracy	0.902	0.913	0.915	0.912	0.924
Precision	0.902	0.913	0.915	0.912	0.924
Recall	0.902	0.913	0.915	0.912	0.924

Figure. 4 Performance metrics for Deep learning with different epochs

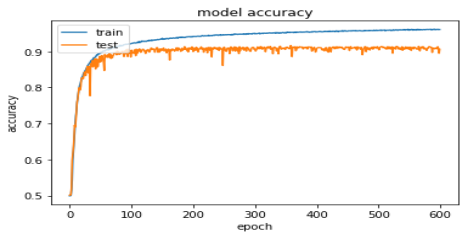
By increasing the number of epochs the model has a better chance to learn from data with higher accuracy provided that no overfitting occurs. The accuracy curves improved by increasing numbers of epochs. The figures indicate no change in accuracy for 400, 600, and 800 epochs. The model achieves highest performance with increasing accuracy at the number of epochs 1000 without overfitting. The accuracy is increased to be 0.924% compared to (0.905%, 0.913%, 0.914%, 0.913%) at (200, 400, 600, 800) epochs respectively. This means the model at this level is able to learn well data without normalization it.



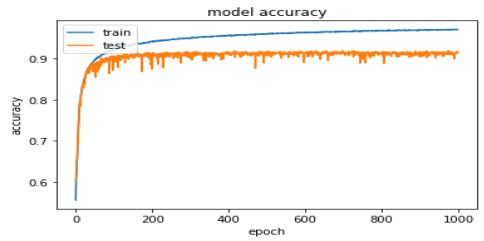
(a)



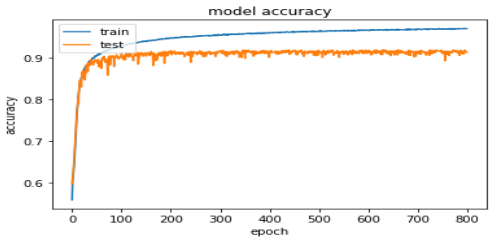
(b)



(c)

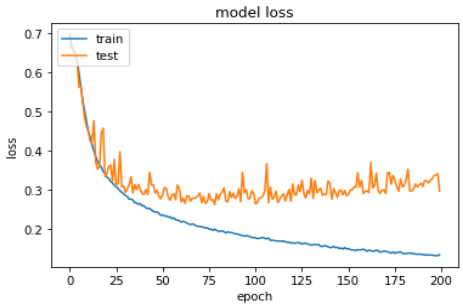


(d)

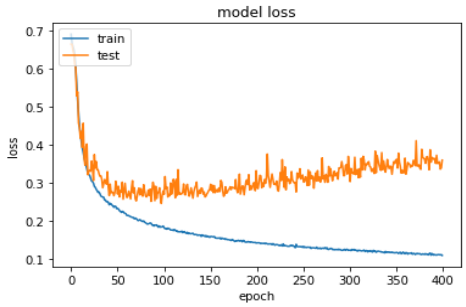


(e)

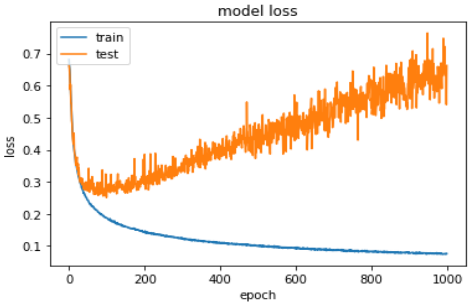
Figure. 5 Model accuracy without overfitting at : (a) epochs=200, (b) epochs=400, (c) epochs=600, (d) epochs=800, and (e) epochs=1000.



(a)



(b)



(c)

Figure. 6 Model loss curves without overfitting at: (a) epochs=200, (b) epochs=400, and (c) epochs=1000

4.4 Comparative analysis for ML and DL

In this section, a comparative analysis is applying between the performance of our model with normalization and Non-normalization techniques.

Fig. 7 shows a comparative analysis based on accuracy measures among different ML classifiers.

All classifiers are compared for z-score and min-max normalization. Also, consider the case for non-normalization of the input data-set.

Fig. 4 shows the performance metrics for (LR, KNN, and SVM) with z-score, min-max, and without normalization. It is observed that the best normalization.

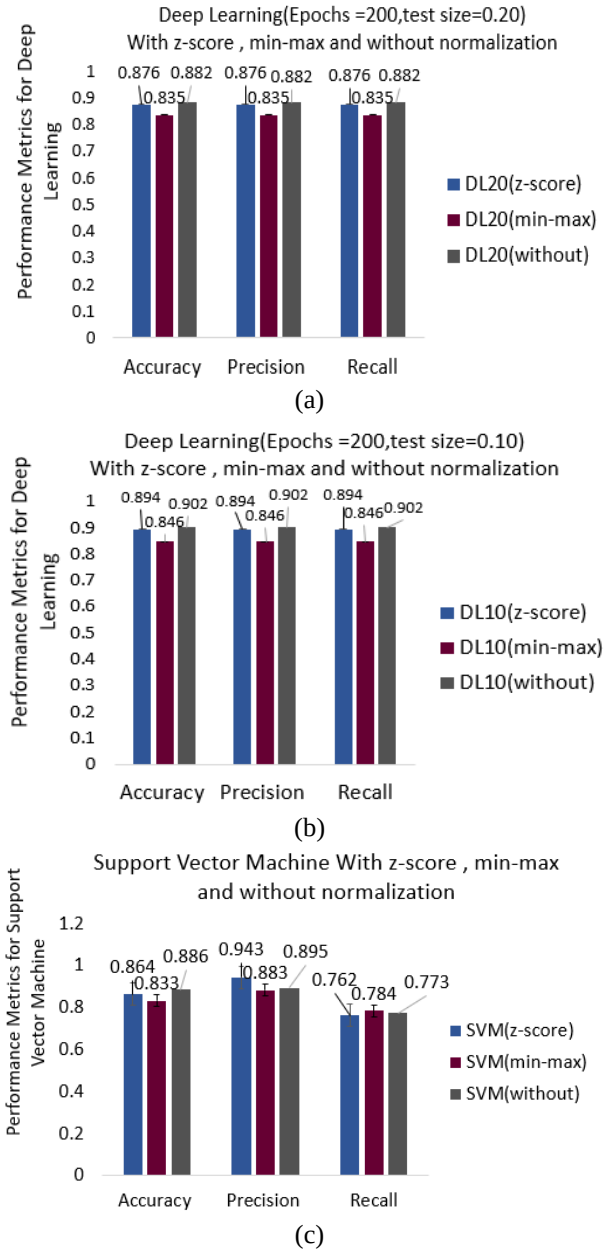


Figure. 7 Performance metrics with z-score, min-max, and without normalization: (a) LR, (b) KNN classifier, and (c) SVM

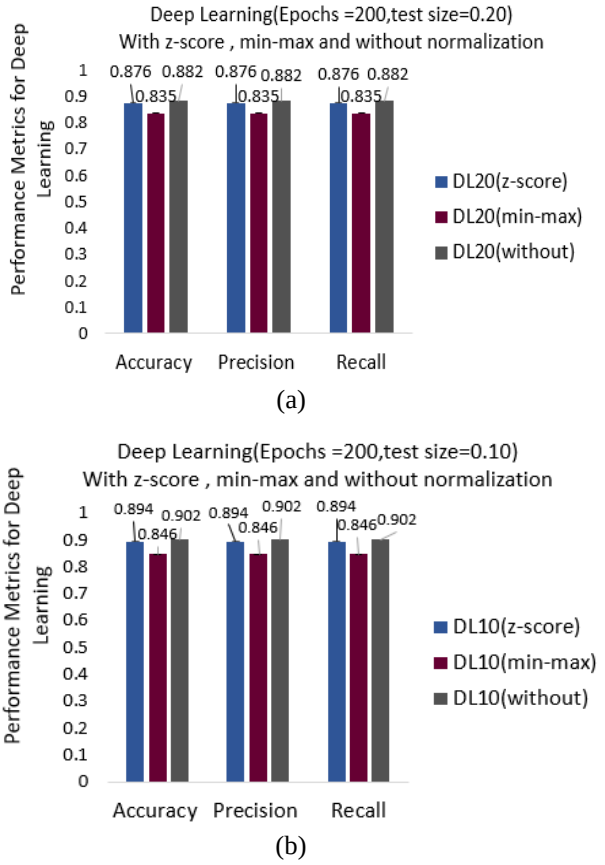


Figure. 8 Performance metrics with z-score, min-max, and without normalization: (a) for DL at test size 20 and (b) DL at test size 10

It is observed that the best performance is achieved when using the data set without using any normalization.

Applying the DL model on the dataset, with default parameters and without using any normalization. The results are obtained in two cases test size (20 & 10).

Fig. 8 shows the performance measures for DL when we use z-score normalization, min-max, and without normalization at test sizes 20 & 10. As shown in Fig. 8 the best performance is achieved when we use the data set without using any normalization at test size 10.

4.5 Comparative analysis against state-of-the-art models

In this section, a comparison between the proposed intelligent model with highest accuracy measures (Non-normalization based model) and other state of the art models illustrated before in related work section is discussed. Table 4 provides a complete analysis of such a comparison. It compares our non-normalization-based model with other complete analysis of such a comparison. It compares

Table 4. Comparison results

Model	LR	KNN	SVM	Simple Neural Network	RNN	CNN	Computing environment	Method
	Readmission Prediction Accuracy							
The proposed intelligent model	**0.642%**	**0.872%**	**0.886%**	**0.873%**	**0.837%**	**0.924%**	Personal computer system (anaconda3)	ML, RNN and CNN with Non-normalization Technique
H. N. Pham, et al [11] 2019	0.635%	-	0.2946%	0.7999%	-	-	Personal computer system	ML with Ensemble Technique
A. Hammoudeh, et al [13]2018	-	-	-	-	-	0.92%	Personal computer system	CNN with Normalization Technique
Chopra, Chahes, et al [10] 2017	0.6291%	-	0.6488%	0.6953%	0.8112%	-	Personal computer system	ML and RNN with Normalization Technique
D. Mingle [9] 2017	Machine Learning Methods Hitting ~ 0.84% accuracy			-	-	-	Personal computer system	ML with Ensemble Technique by Age Group

our non-normalization-based model with other models reported in [9-11, 13]. Each one of these models used a different ML and DL model with different pre-processing methods (e.g. [9] used ML with Ensemble Technique by Age Group, [10] used ML and RNN with Normalization, [11] used ML with Ensemble Technique, and finally [13] used CNN with normalization technique. The proposed intelligent model used non-normalized data achieves accuracy for ML algorithms as follow: LR=0.642%, KNN=0.872%, SVM=0.886%), Simple Neural Network =0.873%, and for reported accuracy for DL models as follow: RNN=0.837%, and CNN=0.924%. The advantage of Non-normalization technique is the ability to keep all features. It allows the classifier model to get benefit of all features. As shown in table-4 all intelligent models based on non-normalization techniques are achieved performance better than all state-of-the-art models. All state-of-the-art model applies normalization method on the input data. This normalization imposes the data values to be in the range from [0:1] in the case of min-max or [-1:1] in the case of z-score. Such restriction on data values makes a loss in some features that the model needs to learn for better performance. The proposed non-normalized model saves all data values, allow the prediction model to discover all features, and allowing the training model to be fed with more data. Once the model uses more features in training, it is able to provide classification results with high accuracy values.

5. Conclusion

In this work, an intelligent model based on (ML and DL) algorithms were proposed to predict hospital readmission over a clinical data set, after applying some pre-processing on the input data. These pre-processing included solving missing values problems, feature selection using some feature engineering methods, data normalization using z-score and min-max, and Non-normalization technique. The performance of our model was tested under two different conditions.

In the first experiment, DL and ML performed higher when z-score was used compared to min-max normalization. When the normalization range is between -1 and 1, more accurate results was obtained. Our DL based model reported an overall accuracy of 0.894% and 0.846% using z-score and min-max normalization, respectively.

In the second experiment, DL and ML performance were higher for non-normalized data. Also, DL performance was the best without using any normalization. Our DL based model reported an

overall accuracy of 0.924% without using any data normalization.

The proposed model was compared to the state of the art models that displayed data based on different pre-processing methods such as : ML with Ensemble, CNN with Normalization, ML and RNN with Normalization, and ML with Ensemble by age group techniques. The comparison shows a predominance of the non-normalization technique with an overall accuracy 0.924%. The proposed non-normalization-based model succeed in saving all data values without any loss in information. It allowed the training model to discover more features from input data which enhance the accuracy of the readmission prediction model. The proposed model can be used to help the health care sector to predict hospital readmission for diabetes patients using some clinical data. That will have a direct effect on the health care costs and the hospital's efficiency and reputation.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Eman H. Zaky gives the idea of the system, Eman H. Zaky; software and designed the experiments, Eman H. Zaky, Neveen I. Ghali, and Mona M. Soliman; formal analysis, investigation, resources, data preparation, Eman H. Zaky; writing—original draft preparation, Neveen I. Ghali, and Mona M. Soliman; review and editing, Eman H. Zaky, Neveen I. Ghali, and Mona M. Soliman; supervised the study, analyzed the results, and verified the findings of the study.

References

[1] World Health Organization, https://www.who.int/ar/news-room/fact-sheets/detail/the-top-10-causes-of-death, [Online] [accessed: 5 - 2 - 2020].

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

[3] Medicare Payment Advisory Commission. *Report to the Congress: promoting greater efficiency in Medicare*. Medicare Payment Advisory Commission (MedPAC). 2007.

[4] G. F. Anderson, and EP. Steinberg, "Predicting hospital readmissions in the Medicare population", *Inquiry*, Vol. 22, No. 3, pp. 251-258, 1985.

[5] L. Boult, C. Boult, P. Pirie, and J. T. Pacala, "Test-retest reliability of a questionnaire that identifies elders at risk for hospital admission", *Journal of the American Geriatrics Society*, Vol. 42, No. 7, pp. 707-711, 1994.

[6] N. Allaudeen, J. L. Schnipper, E. J. Orav, R. M. Wachter, and A. R. Vidyarthi, "Inability of Providers to Predict Unplanned Readmissions", *Journal of General Internal Medicine*, Vol. 26, No. 7, pp. 771-76, 2011.

[7] L. C. Daras, M. J. Ingber, J. Carichner, D. Barch, A. Deutsch, L. M. Smith, A. Levitt, and J. Andress, "Evaluating Hospital Readmission Rates After Discharge From Inpatient Rehabilitation", *Arch Phys Med Rehabil*, Vol. 99, No. 6, pp. 1049-1059, 2018.

[8] L. Turgeman, and J. May, "A mixed-ensemble model for hospital read- mission", *Artificial intelligence in medicine*, Vol. 72, pp. 72-82, 2016.

[9] D. Mingle, "Predicting Diabetic Readmission Rates: Moving Beyond Hba1c", *Current Trends in Biomedical Engineering & Biosciences*, Vol. 7, No. 3, pp. 1-21, 2017.

[10] 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: *Proc. of International Conf. On Computational Biology and Bioinformatics*, Newark, NJ, USA, pp. 18-23, 2017.

[11] H. N. Pham, A. Chatterjee, B. Narasimhan, C. W. Lee, D. K. Jha, E. Y. F. Wong, and M. C. Chua, "Predicting hospital readmission patterns of diabetic patients using ensemble model and cluster analysis", In: *Proc. of International Conf. On System Science and Engineering (ICSSE)*, pp. 273-278, 2019.

[12] L. X. Li, and S. S. Abdul Rahman, Students, "learning style detection using tree augmented naive Bayes", *Royal Society open science*, Vol. 5, No. 7, 2018.

[13] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, "Predicting Hospital Readmission among Diabetics using Deep Learning", *Procedia Computer Science*, Vol. 141, No. November, pp. 484-489, 2018.

[14] D. J. Rubin, "Correction to: hospital readmission of patients with diabetes",*Current diabetes reports*, Vol. 18, No. 4, pp. 1-9, 2018.

[15] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, J. K. Cios, and N. J. Clore, "Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records", *BioMed Research International*, pp. 1-11, 2014.

[16] T. Goudjerkan, and M. Jayabalan. “Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron”, *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 2, pp. 268-275, 2019.

[17] 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.

[18] K. Hempstalk, and D. Mordaunt, “Improving 30-day readmission risk predictions using machine learning”, In: *Proc. of International Conf. On Health Informatics*, New Zealand (HiNZ), pp. 1-5, 2016.

[19] Codecademy, https://www.codecademy.com/articles/normalization#:%CB%9C:text=Min%2Dmax%20normalization%3A%20Guarantees%20all,with%20the%20exact%20same%20scale [Online] [accessed: 5 - 6 - 2019].

[20] S. Raschka, “About feature scaling and normalization and the effect of standardization for machine learning algorithms”, *Polar Political Legal Anthropology Rev*, pp. 67-89, 2014.

[21] Machine Learning in Python, http://scikit-learn.org [Online] [accessed: 8 - 12 - 2019].

[22] Machine Learning in Python, https://www.tensorflow.org [Online] [accessed: 8 - 12 - 2019].

[23] Machine Learning in Python, https://keras.io [Online] [accessed: 8 - 12 - 2019].

[24] Machine Learning Crash Course, https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall [Online] [accessed: 7 - 5 - 2020].

[25] Jr. Hosmer, W. D., S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, Vol. 398, John Wiley & Sons, 2013.

[26] C. Cortes, and V. Vapnik, "Support-vector networks", *Springer, Machine learning*, Vol. 20, No. 3, pp. 273-297, 1995.

[27] Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, Vol. 1, No. 2, pp. 800, Cambridge, MIT press, 2016.