ISSN: 2096-3246
Volume 54, Issue 02, December, 2022

HYBRID CNN-SVM BASED PREDICTIVE HOSPITAL READMISSION MODEL FOR DIABETES PATIENTS

G. G. Rajput[1], Ashvini Alashetty[2]
[1,2]Department of Computer Science, Karnataka State Akkamahadevi Women's University, Vijayapura, Karnataka-586108
[1]ggrajput@yahoo.co.in, [2]ashwinialashetty@gmail.com

***Abstract*—**Today's healthcare systems are challenged with a number of changes, including an ageing population, a growing dependence on technology, and rising patient demands. These changes have produced healthcare systems that are more patient-centered and value-driven. The difficulty of expanding health care while keeping costs down is a significant obstacle for many organizations. Access to electronic health data has risen considerably due to advancements in processing power and widespread adoption of electronic medical records (EMRs). Because there are so many data to process, typical programs lack the processing ability to handle massive data sets properly. The combined Convolutional Neural Network (CNN) and Support Vector Machine (SVM) is an intelligent methodology used for prediction and has shown promising results. In this article, the hinge loss function for hybridizing CNN and SVM is described. In the preprocessing steps, feature engineering and the addition of missing variables are all handled. Using a CNN-SVM model, the major outcome of this paper is an improvement in the accuracy of readmission prediction as 78%.
***Index Terms*—**Machine learning, Support Vector Machine, Pre-processing, Diabetes, Hospital readmission

INTRODUCTION

Diabetes is a condition that lasts a lifetime and is now ranked among the top 10 causes of mortality worldwide. By 2045, there will be 700 million diabetics, a 51.18 percent increase from 2019. By 2035, diabetes is estimated to be the sixth greatest cause of mortality. In 2000, the Ministry of Global Health's figures indicated that diabetes was not among the leading causes of mortality; nevertheless, by 2016, it had moved to seventh position [1]. Due to its high mortality rate, diabetes is a prominent priority on the health agendas of both industrialized and developing nations. The health care industry obtains and analyses large amounts of medical data for diabetic patients. With the introduction of technology in diagnosis, monitoring, storage, and analysis, more effective problem-solving options are now accessible [2]. Chronic disorder patients have a tough time recovering completely, and the difficulty increases when readmission occurs. In the United States, according to the Centers for Medicare and Medicaid Services (CMS), 76% of hospital admissions were avoidable [3]. Patient traits, illness features, and health care system variables are risk factors for hospital readmission, especially for the elderly [4]. The following is a summary of this work's key contributions:

1) An improved version of the original data set used in the analysis is using a variety of data preprocessing techniques. Missing value problems and feature engineering are two of these tactics.
2) Using a hybrid SVM-CNN model with hinge loss function, the accuracy of hospital readmission prediction based on the patient's medical data is investigated.

The paper structure is as follows: The second chapter reviews algorithms for predicting hospital readmissions. Section 3 explains the proposed methodology and clinical report dataset. Section 4 analyses experimental outcomes. Section 5's model conclusions.

LITERATURE REVIEW

In recent years, a lot of study articles in healthcare have been published regarding predicting hospital readmission, as this can save a patient's life. The authors [5] demonstrate the increasing interest in the implementation of ML models. There was no noticeable difference in precision between regression and ML models. According to the authors [6], the majority of studies only reported on bias when questioned about performance metrics. The AUC for logistic regression was found to be between 0.70 and 0.80 based on statistical study, making it a popular option with average performance. When developing clinical treatments, regression techniques are always employed as the beginning point. However, they could not perform as well, which is a significant disadvantage. Utilizing numerous ML models in therapeutic situations is often smart.

SVM is a powerful classification learning approach [7] that was first introduced in 1995. In [8] built a model for predicting hospital readmissions using the C5.0 tree and SVM to account for database-specific variables. Here are the study's findings: The true positive rates (precise values) of C5.0 forecasts are indicative of the accuracy of SVM predictions. In general, the accuracy of the kit falls anywhere between 81% and 85%. Patients were divided into three groups (30, [30–70], and >70) in [9], which presented several classifications. Each group's model was individually developed using one or more machine learning (ML) approaches or a mix of ML methods.

The accuracy of the developed models was 84% for groups under the age of 30, 78.5 % for groups between the ages of 30 and 70, and 68.5% for groups above the age of 70. The proposed research in [10] used data preparation techniques such as normalization to predict how often individual patients will be readmitted. Comparing the performance of many ML approaches to that of an RNN-based model, we discovered that the RNN-based model was much more accurate, especially when dealing with non-sequential data. With the ROC Curve of 0.61 and an accuracy of 69.53 percent for the 2-layer basic neural network, as well as 0.80 and 81.12 percent for the 2-layer recurrent neural network, the proposed work is effective in both instances. CNN of deep learning has been applied to this problem [13] in order to accurately predict which diabetic patients will require hospital readmission.

With a 92% success rate, this model exceeds the bulk of machine learning (ML) models. This model, on the other hand, is dependent on data engineering techniques and expanding the sample size. Boosting deep learning's efficacy requires feature engineering, SMOTE to address class imbalance in clinical data, and normalization. The work proposes an SVM-CNN hybrid model for predicting the accuracy of hospital readmissions utilizing vast amounts of clinical data. This proposed model makes use of a data set with a number of problematic properties. CNN-SVM hybrid model to predict hospital readmission, hoping it would be better than previous attempts.

PROPOSED METHOD

We use a CNN model coupled SVM -based classifiers for readmission predictions. The planned research would use a 1999-2008 US diabetes dataset from 130 hospitals. This information was utilized to predict the chance of a 30-day readmission for a diabetic patient. Figure 1 displays the recommended model in its entirety. In the following sections, we will demonstrate the process in detail.

A Data pre-processing

There were 50 attributes in the original dataset's containing 101766 records, for a total of 5,088,300 data points [14]. Of the 50 attributes, 37 are categorical and 13 are numerical values. The attributes labeled as "readmitted" with an encoded "30 days" is the output variable, whereas the column labeled "Not readmitted" with an encoded ">30 days" is the other option. This data did not have the best structure for the intelligent model being presented, thus it might benefit from various pre-processing procedures. Two of the most prominent proposed pre-processing techniques are those for handling missing values and developing new features. Each will go into great depth.

B Missing value management

Data cleansing begins with the recovery of lost information. Missing values are information that is absent from a record. Identifying the missing values is the first step. The second phase addresses the missing values.

- Dropping columns with several missing values
- Eliminating characteristics with a few missing values
- Among 101766 hospitalized patients, 60832 were re-admitted and 39355 died not just from diabetes but also from several other connected conditions.
- We can see that payer code, medical specialty, and weight had more than fifty percent of their data missing, therefore we deleted them.
- The Race feature also has some missing values.

C Feature engineering

What we mean by 'Feature Engineering' is a technique whereby many features work together (feature creation, encoding, and data scaling). An individual's in- and out-patient visits, as well as their hospitalization history, are just some of the patient-related characteristics that may be accessed through the database. There are a total of 23 characteristics for 23 medicines that are unique to each drug in the database. Changing diabetics to a different medication regimen upon admission has been linked to decreased rates of readmission, according to the literature [15].

From this vantage point, the frequency with which individual patients required a change in their medication was determined, and updates were made public. We use a unique encoding scheme, mapping characteristics like "No" to 0 and "steady" to 1 or "up" to 1 or "down" to 1. For the treatment of diabetes, patients may get up to six different medications. Thanks to encoded values, we were able to develop a brand-new functionality. This was done to reduce complexity and explore a possible correlation between the number of alterations and the number of drugs. Reclassifying

inpatient care by consolidating classifications of admission and discharge types and admission origins.

The encoding of certain variables: Features such as "medication change" can be toggled between "No" and "Ch" or "no change" and "changed" and "Yes" and "no change" [17]. Assume the patient as one year old if their age range is 0–10 years; two years old if their age range is 10–20 years; and so on [18]. {'age': {'[0-10)': 1, '[10-20)': 2, '[20-30)': 3, '[30-40)': 4, '[40-50)': 5, '[50-60)': 6, '[60-70)': 7, '[70-80)': 8, '[80-90)': 9, '[90-100)': 10}}.

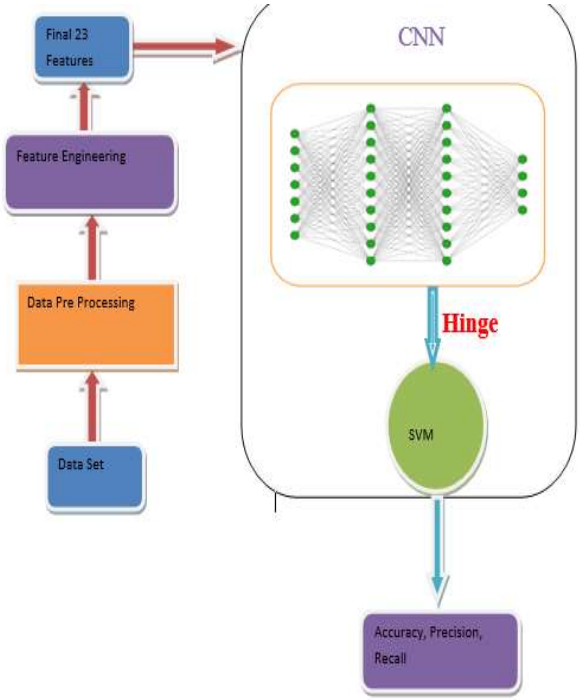

Fig. 1. Full proposed CNN-SVM model

In this case, the quantitative characteristics with numerical values are discrete and limited in scope. There is no way to address outliers in this dataset because quantitative and qualitative discrete data are both possible. Due to the sensitive nature of the information included in diag 1,diag 2, and diag 3, we must remove them. They don't add anything to the treatment's efficiency. We take in both quantitatively discrete and qualitatively categorical input variables, and our output variable is also a categorical. Chi-square tests are used to determine whether or not there is a relationship between two sets of input data. Following the elimination of extraneous details, 23 characteristics emerged. It aided in further research.

Prediction Requires AGE

Prediction requires admission type id.

Prediction requires discharge disposition id.

Prediction Requires admission source id

Prediction depends on time in hospital

Prediction needs num procedures.

Prediction needs num medications.

Prediction Requires number inpatient

Prediction Requires number diagnoses

Prediction depends on race AfricanAmerican

Prediction depends on race Caucasian

Prediction needs race Hispanic

Prediction needs gender female

Predictions need male gender.

max glu serum >300 for Prediction

max glu serum None for Prediction

Prediction Requires max glu serum Norm

A1Result >7 is important for prediction.

A1Result None is important for prediction.

Prediction requires change Ch.

Prediction Requires change No

dummyCat helps predict

3.3 Novel Hybrid CNN-SVM model

Utilizing the "hinge" loss function and their related hyper parameters, the proposed method combines the advantages of SVM and CNN. To prevent over-fitting and offer more accurate classification results, we replace CNNs' typical softmax classifier with a Support Vector Machine (SVM) classifier and back-propagate the network using its hinge loss function. We have used gridsearchCV to fine-tune the model, and the findings indicate that parameter values have a significant effect on the model's output.

The hinge loss is frequently employed as the goal function in SVM. For example, given a training dataset mi ∈ XP, i ∈{1,2,⋯,N}, yi ∈{−1,+1}, Let's pretend we have N samples, each of which has P dimensions; xw is the predicted value of a linearly separable SVM; Y is the proper classification category; and w is a learnable parameter of the classifier. In such case, the amount of its hinge loss may be calculated as in equation 1

$$Li = \max (0,1-miwTyi) \quad (1)$$

When miwTyi > 1, the loss value for the objective function of a binary classification issue is 0,

indicating that the loss function prediction is correct. Table 1 displays the proposed model's parameters.

TABLE I
MODEL PARAMETERS OF CNN_SVM CLASSIFIER

Classifier	**Learning rate**	**Optimizer**	**Loss function**	**Activation function**	**Batch size, Epoch**	**Accuracy**
CNN - SVM	0.01	Adam	Binary cross entropy	sigmoid	[15, 10]	76.2%
	0.001	Adam	Hinge	sigmoid	[10,10]	78.02%

The learning rate has a major impact on the speed and dependability of network training. Convergence may be sluggish if the learning rate is low, and it may fail if it is high. Increasing the learning rate might potentially accelerate the network's training, whilst reducing it could facilitate the process. For learning, rates of 0.01 and 0.001 were utilized. Python script shows that the syntax of mentioning the learning rate, optimizer, metrics for learning rate 0.01 and 0.001

```
adam = Adam(lr = 0.01)
model.compile(loss = ‘binaryclass entropy', optimizer = adam, metrics = ['accuracy'])
adam = Adam(lr = 0.001)
model.compile(loss = ‘hinge', optimizer = adam, metrics = ['accuracy'])
```

The optimizer recalculates and revises network parameters for model training and output. Adam, SGD, and RMSProp are well-known optimizers. Adam combines the benefits of Momentum and RMSProp by dynamically adjusting the learning rate of each parameter, utilizing first- and second-order moment estimate of the gradient. Compared to SGD and RMSProp, Adam's ability to restrict each learning rate to a certain range leads in less choppy parameters. When it comes to training a network, the activation function makes a big difference in both its learning capacity and its pace. The sigmoid, Tanh, and ReLU activation functions are the industry standards. Using the sigmoid activation function, the dense 2 module ensures that the values transmitted from the top layers are compressed to the interval [0, 1]. The popular ReLU activation function, in contrast to sigmoid and Tanh, can reduce gradient disappearance and hasten network convergence. For this reason, we implement ReLU as the fundamental activation function in our network model.

EXPERIMENTAL RESULTS

In this subsection, we verify that readmission to the hospital can be predicted using CNN-SVM based approaches. Accuracy, recall, and precision are some of the commonly used measures upon

which these findings are based.

A Performance evaluation metrics

The efficacy of an intelligently based model may be measured in terms of its accuracy, precision, and recall. A true positive (TP) is the outcome when the model properly predicts the positive category, and a true negative (TR) is the outcome when the model correctly predicts the negative category (TN). It is called a false positive (FP) when the model wrongly predicts a positive category, and a false negative (FN) when the model incorrectly predicts a negative category (FN). Precision and Recall are two measures of our model's efficacy that may be derived from its accuracy.

The efficiency of a rating system may be expressed as

Accuracy = TP + TN / TP + TN + FP + FN (2).

As such, it constitutes an element of the predictions that our model successfully made.

Precision is the fraction of correct positive identifications and is calculated as follows:

Precision=TP/TP+FP (3).

The proportion of true positives that are correctly detected is known as "recall," and recall is represented by the equation

$Recall=TP/TP+FN$ (4)

This hybrid CNN SVM model's accuracy is summarized in Table II.

TABLE II
ACCURACY METRICS OF PROPOSED CNN-SVM MODEL

	Precision	**Recall**	F1-score
0	0.94	**0.64**	0.76
1	0.69	**0.95**	0.80
Accuracy	-	**-**	0.78
Macro Avg	0.81	**0.80**	0.78
Weighted Avg	0.83	**0.78**	0.78

The ability of a model to find all the relevant cases within a data set. Mathematically, we define recall as the number of true positives divided by the number of true positives plus the number of false negatives. Precision: The ability of a classification model to identify only the relevant data points.

TABLE III

COMPARISON OF DIFFERENT MODELS WITH PROPOSED MODEL

	Algorithm	**Dataset**	**Total number of attributes**	**Accuracy%**
Sarwar et al[16]	Six ML algorithms are used to predict disease	PIDD data set	50	SVM and KNN 72%
Sneha et al [17]	RF, SVM, NB, DT, KNN	USA diabetes dataset	50	Navie bays algorithm 72.3%
Deepti et al [18]	NB, SVM	PIDD data set	50	Navie Bays algorithm 76.3%
Proposed work	**CNN-SVM**	**USA diabetes dataset**	**23**	**78%**

Table 3 compares the different datasets used to estimate the accuracy of diabetes patients' hospital readmission. In our innovative hybrid method, the number of attributes was decreased to 23, and an attained an accuracy of 78%.

CONCLUSION

In this work, we present a CNN-SVM-based intelligent model for predicting hospital readmission across a clinical data set, with the help of certain data pre-processing. Some feature engineering techniques were used for feature selection and the resolution of missing value problems as part of this pre-processing. In order to enhance the accuracy of readmission predictions, we combine support vector machines (SVMs) and convolutional neural networks (CNNs), specifically by swapping out the softmax classifiers typically used in CNNs for SVMs and training the network model with hinge loss function back-propagation. Without any data normalization, our CNN-SVM based model exhibited an overall accuracy of 78%. We compared the proposed model to state-of-the-art models that presented data using various pre-processing techniques. All values in the data set are preserved in the proposed model without any information loss. In this way, the training model was able to extract additional characteristics from the input data, hence improving the readmission prediction model's accuracy. Using certain clinical data, the proposed model can assist the healthcare industry in predicting hospital readmission for diabetic patients. That will have an immediate impact

on healthcare expenditures, along with the effectiveness and standing of the hospital.

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] 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.

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

[5] Mahmoudi, E.; Kamdar, N.; Kim, N.; Gonzales, G.; Singh, K.; Waljee, A.K.: Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ 369, 958 (2020) 15.

[6] Markazi-Moghaddam, N. Fathi, M. Ramezankhani, A.: Risk prediction models for intensive care unit readmission: a systematic review of methodology and applicability. Aust. Crit. Care 33(4), 367–374 (2020)

[7] Cheng, W.; Zhu, W.: Predicting 30-day hospital readmission for diabetics based on spark. In 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC), pp. 125–129 (2019)

[8] Ramirez, J.C. Herrera, D. "Prediction of diabetic patient readmission using machine learning" IEEE Colombian Conf. Appl. Comput. Intell 2019, 1–4 (2019)

[9] Glans, M.; Kragh Ekstam, A.; Jakobsson, U.; Bondesson, A.; Midlov, P.: "Risk factors for hospital readmission in older adults within 30 days of discharge: a comparative retrospective study". BMC Geriatr. 20(1), 467 (2020)

[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] 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.

[16] A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," *ICAC 2018 - 2018 24th IEEE Int. Conf. Autom. Comput. Improv. Product. through Autom. Comput.*, no. September, pp. 1–6, 2018.

[17] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, vol. 6, no. 1, 2019.

[18] Deepti Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018.