IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF INSULIN LEVEL OF DIABETES PATIENT USING MACHINE LEARNING APPROACHES

[1]Ketan Gupta, [2]Nasmin Jiwani
[1,2]Research Scholar
[1,2]Department of Information Technology
[1,2]University of The Cumberlands, Kentucky, USA

Abstract: In today's world, there are so many patients with Diabetes having varying insulin levels and blood glucose levels in the human body. So, there is a requirement to constantly monitor the blood glucose level and improve the patient's condition with an adjusted insulin dose. Before each meal, they must take a dose of insulin. Doctors must calculate insulin dosages for each patient based on previous dose data and regular sugar levels. Our study uses a Machine Learning approach to develop the model, which uses an RNN (LSTM) and ANN algorithm to predict a patient's insulin chart. The algorithm was trained using the patient's 36-month chart, and the extended series of following insulin predictions are based on that data. The predictive model is used in this investigation.

Index Terms **– Machine Learning, LSTM, ANN, RNN, Diabetes**

I. Introduction

Diabetes is a metabolic ailment caused through hyperglycemia (high blood glucose level). Diabetes affects an estimated 3-4 percent of the global population (half of whom are undiagnosed), making it the most prevalent chronic ailments worldwide. Hyperglycemia is a condition caused by problems with insulin-secretion, action, or both. Long-term harm, dysfunction, and catastrophe of various organs, including the eyes, nerves, kidneys, heart, and blood vessels, have been related to diabetes' persistent hyperglycemia. This insufficiency causes pancreatic b-cell death, causing in insulin insufficiency and abnormalities that develop in endurance to insulin action and reaction mechanism. Insulin deficiency is caused by insufficient insulin secretion. This primary cause of hyperglycemia is improper insulin secretion and insulin action abnormalities. So, the significance of insulin dose is visible [1].

We took on the task of forecasting insulin charts for diabetic patients in this work. We collected 36 months of data (insulin chart) for a patient with code 33 (in case of regular insulin dose). Using the previous 36 months' data as training data, the next 36 months' charts are projected and compared to the actual data. Using data from a diabetic patient, we employed an RNN model to predict the insulin chart. Speech recognition, language modeling, translation, image captioning, and other challenges have been solved with RNN. RNN is primarily employed when there is a minor gap between the relevant information and the location where it is required. However, there are times when the extra context is required. Unfortunately, as the gap widens, RNNs lose their ability to learn to connect the dots. We solved this problem by employing a type of RNN known as LSTM (Long Short-Term Memory), that is meant to avoid the issue of long-term dependency. We trained 67 percent of our given dataset and tested 33 percent of it for RNN [2]. We employ RNN to solve the sequence of insulin levels of hospital patients one at a time. ANN is a prominent method for identifying unknown and hidden patterns in data that can be used to forecast insulin data. For the prediction of insulin levels, we use 80% train on data, and 20% predict on the train data.

II. Related work

In the study article [3], the authors published a comparative analogy on diabetes in-depth diagnosis using the Pimadiabetes complex dataset and showed that multidimensional neural networks (MNN) with Levenberg – Marquardt (LM) algorithms surpassed the additional neural network-based classifiers. The authors [4] executed tweet classification to predict different categories of abusive languages by analyzing machine learning and deep learning algorithms and conducted a comparative study to decide which algorithm accomplishes the best results in detecting abusive language. The feature vectors generated were based on two approaches, the bag-of-words method, and word embeddings. The outcomes showed that the deep learning algorithms based on the word embeddings approach performed well than other algorithms. Diabetes mellitus is a severe health condition that may cause several complications. It is worth researching how to anticipate and diagnose this illness using machine learning precisely. The authors [5] performed several experiments on different datasets of patients to predict Diabetes by using Decision Tree and RF algorithm in two different tools, WEKA and MATLAB. They concluded from the experiments that machine learning can anticipate or accurately predict Diabetes, but it is essential to perceive appropriate attributes, classifiers, and data mining methods.

In this paper [6], the authors have analyzed the results of classifying and clustering spam management of SMS using two discrete environments, RapidMiner and WEKA. They experimented using the similar dataset in alike environments, and the simulations gave the same results, and SVM was considered the best classifier in both environments. Machine Learning put-forward a versatile classification and regression mechanisms that can be put to use to address diagnosis problems in various medical fields. A comparative study of breast cancer prediction has used Artificial Neural Networks (ANN), k-Nearest Neighbors (KNN), and Bayesian Network Classifiers for comparison. Based on the analysis results, Artificial Neural Networks are better than KNN and Bayesian Classifiers in classification with 97.4% accuracy.[7] In this study [8], the writers introduced the RALE lung sound library support vector machine (SVM) and KNN machine learning algorithms to analyze respirational illnesses or chronic diseases using pulmonary acoustic signals. And demonstrated that the KNN classifier's generalization capacity is higher compared to SVM's. KNN's precision was calculated to be 98.26% relative to SVM's 92.19%. The authors in the paper [9] proved that, compared to other algorithms such as NB, C4.5, and Repeated Incremental Pruning to Produce Error Production (RIPPER), Support Vector Machine (SVM) gives 93%, the highest and apical precise value in their advanced spam filtering framework.

III. BACKGROUND

AI and ML are extremely trendy terms that appear to be used interchangeably. They are not the same thing, but the notion that they are can cause misunderstandings. Simply put, artificial intelligence is machines' ability to do tasks in a "smart" manner. Then again, AI uses artificial reasoning (AI) to empower frameworks to learn and improve without being customized [10].

Artificial Neural Networks (ANNs) Based Prediction: Artificial Neural Networks (ANNs) are biologically inspired computer simulations that perform specialized tasks, including grouping, classification, and pattern recognition. The strength of inter-neuron connections, known as synaptic weights, is learned and stored by a neural network (as shown in Figure 1).

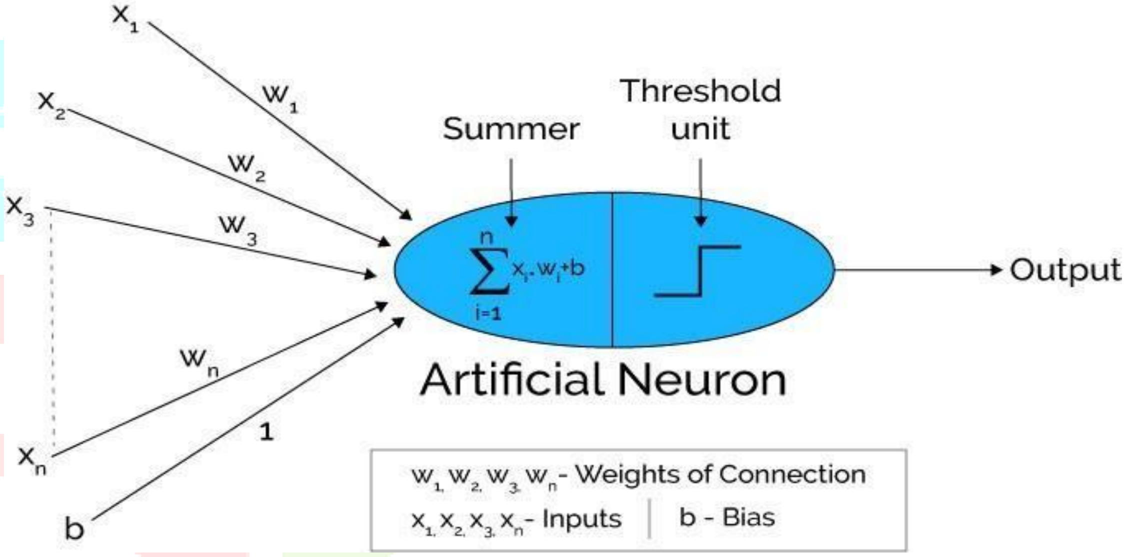

x_1
x_2
x_3
x_n
b
w_1
w_2
w_3
w_n
1
Summer
Threshold unit
$\sum_{i=1}^{n} x_i \cdot w_i + b$
Output
Artificial Neuron
w_1, w_2, w_3, w_n - Weights of Connection
x_1, x_2, x_3, x_n - Inputs
b - Bias

Figure 1: ANN Working Module

The loads doled out to each information are duplicated by it. The brain network's information to tackle an issue is called loads. The strength of the associations between neurons inside a brain network is much of the time meant by weight. Inside the figuring unit, the weighted sources of info are added to (artificial neuron). When the weighted all-out is zero, the inclination is utilized to make the outcome non-zero or increment the framework response. Predisposition's weight and information are constantly set to '1'.

Any number among 0 and limitlessness can be utilized as the total. Limit esteem is provided to restrict the response to the necessary worth. The aggregate is accommodated for this reason by the initiation work. The enactment work is the assortment of move capacities used to come by the ideal outcome. There are two sorts of enactment capacities: straight and non-direct.

Recurrent Neural Network based prediction

RNN is a brain network in which the past stage's result is utilized to contribute to the ongoing stage. The model efficiency feed-forward ANN models the sequence of varying lengths [11]. All sources of info and results in typical brain networks are free of each other; be that as it may, while foreseeing the following expression of an expression, the earlier words are required; thus, the earlier words should be recollected. Thus, RNN was created, which utilized a Hidden Layer to conquer the issue. The most significant and fundamental part of RNN is the Hidden state, which recalls explicit data about a grouping.

LSTM: Sepp Hochreiter and Jurgen Schmidhuber created LSTM in the late 1990s compared to alternative RNNs, hidden Markov models, and other sequence learning algorithms in several applications, which is generally insensitive to gap length [12]. LSTM model is well suited for sequential datasets, including speech, high-end resolution videos, and time series, as it can clutch continuing dependencies [13]. Long-term reliance is a problem that LSTMs are specifically designed to prevent. Long-term memory is an instinct; they don't have to work hard to develop it.

In regular RNNs, this repeating module will have a simple construction, such as, a single tanh layer. LSTMs have a chain-like design; yet, the reworking module has an other one. There are four brain network layers rather than one, each of which collaborates remarkably.

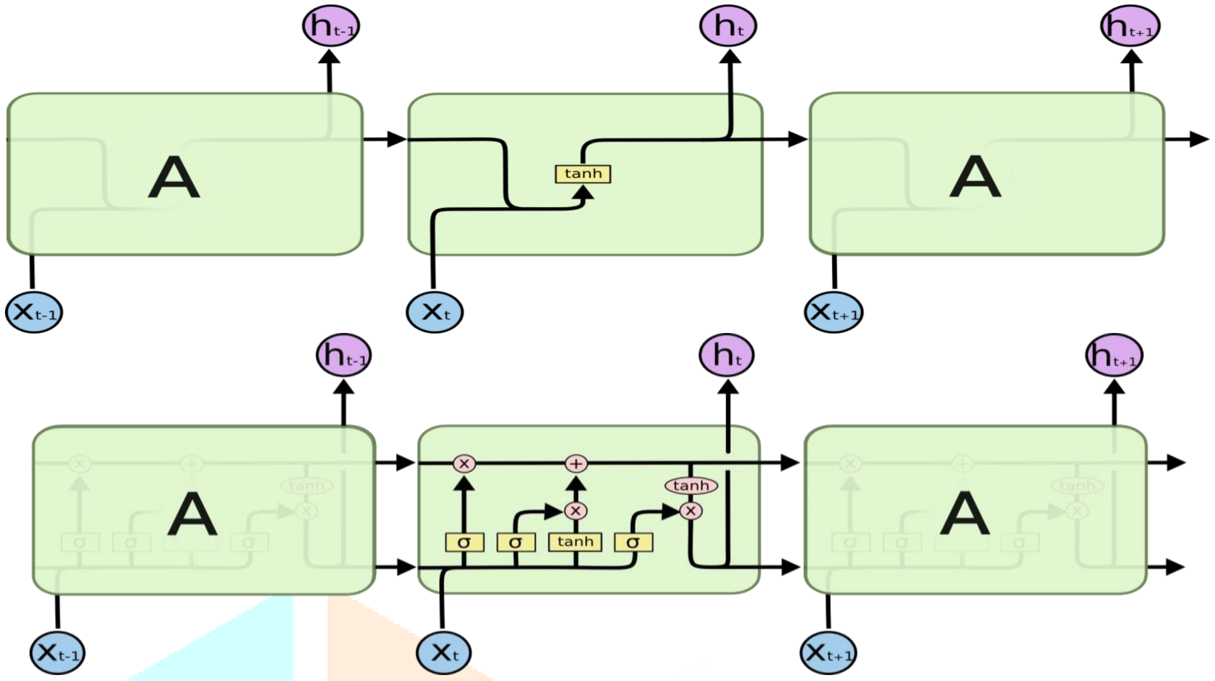


Figure 2(a) RNN with Single Layers **Figure 2(b) RNN with Four Layers**

This model is broken down into cells, each with its operations. Operation gates update the internal state variable of an LSTM as it is passed from one cell to the next.

Forget (Wf. [ht-1, xt] + bf) Gate ft = (Wf. [ht-1, xt] + bf)

This sigmoidal layer combines the result at time t-1 and the ongoing contribution at time t into a solitary tensor before applying a direct change and afterward sigmoid. Because of the sigmoid, the result of this entryway is somewhere in the range of 0 and 1. The inward state is expanded by this sum, which is the reason the entryway is known as a neglected door. The past inner state is neglected, assuming ft is 0, yet it is sent through unblemished on the off chance that ft is 1.

Input Gate it = Wi. [ht-1, xt] + bi)

That information door passes the old result and the new information using another sigmoid layer. This door yields a number somewhere in the range of 0 and 1. The worth of the information entryway increases the result of the up-and-comer layer.

Tanh = Ct (Wc. [ht-1, xt] + bc)

This layer returns a competitor vector to be added to the inner state in the wake of registering an exaggerated digression from the information and past result.

These standard updates the interior state: foot * Ct-1 + it * Ct = Ct

The neglected entryway duplicates the earlier state, adding to the small portion of the new applicant permitted by the resulting door.

Yield Gate

ht = Ot * tanh Ct = (Wo. [ht-1, xt] + bo)

This entryway controls the amount of the inward state conveyed to the result and works similarly to the others.

The organization will determine how much past result to keep, how much current contribution to store, and how much interior state to convey to the result because the three doors shown above have autonomous loads and predispositions.

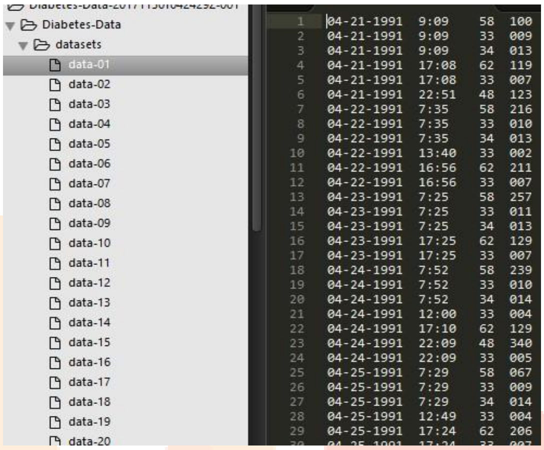
IV. DATASET DESCRIPTION

The dataset utilized in our exploration was taken from https://archive.ics.uci.edu/ml/datasets/diabetes [14], called UCI storehouse. There are two diabetes patient records: programmed electronic recording frameworks and paper records. Paper records have "coherent time" spaces, though the programmed gadget has an inbuilt clock that timestamps occasions (breakfast, lunch, and dinner). Each record in a diabetes document has four fields. (1) MM-DD-YYYY (2) Time in the arrangement XX: YY (3) Value (4) Code

The code field is decoded in an accompanying manner: Insulin Dose 33=Normal Insulin Dose 54 signifies NPH insulin segment, 35 means Ultralente insulin portion, 48 means undefined blood glucose estimation, 57 means unknown blood glucose estimation, etc.

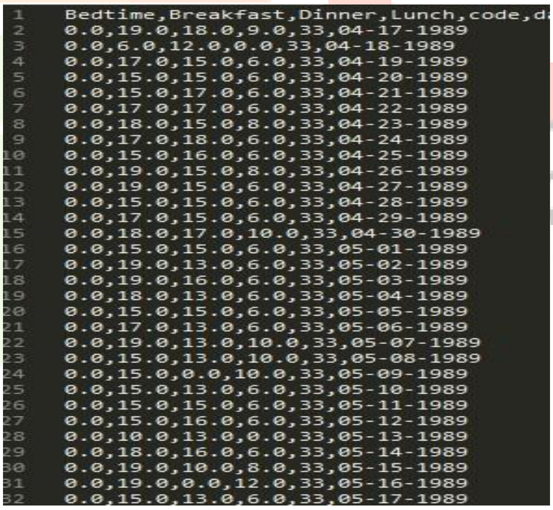
We gathered three years of information (insulin diagram) for a patient's code 33i.eReguler insulin portion. Expecting that this massive dataset fills in as preparing information, the accompanying one-month outline is gauged and contrasted with the truthful information afterward.

Original Data and Preprocessed Data (Code=33):



```
04-21-1991  9:09    58  100
04-21-1991  9:09    33  009
04-21-1991  9:09    34  013
04-21-1991  17:08   62  119
04-21-1991  17:08   33  007
04-21-1991  22:51   48  123
04-22-1991  7:35    58  216
04-22-1991  7:35    33  010
04-22-1991  7:35    34  013
04-22-1991  13:40   33  002
04-22-1991  16:56   62  211
04-22-1991  16:56   33  007
04-23-1991  7:25    58  257
04-23-1991  7:25    33  011
04-23-1991  7:25    34  013
04-23-1991  17:25   62  129
04-23-1991  17:25   33  007
04-24-1991  7:52    58  239
04-24-1991  7:52    33  010
04-24-1991  7:52    34  014
04-24-1991  12:00   33  004
04-24-1991  17:10   62  129
04-24-1991  22:09   48  340
04-24-1991  22:09   33  005
04-25-1991  7:29    58  067
04-25-1991  7:29    33  009
04-25-1991  7:29    34  014
04-25-1991  12:49   33  004
04-25-1991  17:24   62  206
```

Figure 3: The Original Dataset



```
Bedtime,Breakfast,Dinner,Lunch,code,d
0.0,19.0,18.0,9.0,33,04-17-1989
0.0,6.0,12.0,0.0,33,04-18-1989
0.0,17.0,15.0,6.0,33,04-19-1989
0.0,15.0,15.0,6.0,33,04-20-1989
0.0,15.0,17.0,6.0,33,04-21-1989
0.0,17.0,17.0,6.0,33,04-22-1989
0.0,18.0,15.0,8.0,33,04-23-1989
0.0,17.0,18.0,6.0,33,04-24-1989
0.0,15.0,16.0,6.0,33,04-25-1989
0.0,19.0,15.0,8.0,33,04-26-1989
0.0,19.0,15.0,6.0,33,04-27-1989
0.0,15.0,15.0,6.0,33,04-28-1989
0.0,17.0,15.0,6.0,33,04-29-1989
0.0,18.0,17.0,10.0,33,04-30-1989
0.0,15.0,15.0,6.0,33,05-01-1989
0.0,19.0,13.0,6.0,33,05-02-1989
0.0,19.0,16.0,6.0,33,05-03-1989
0.0,18.0,13.0,6.0,33,05-04-1989
0.0,15.0,15.0,6.0,33,05-05-1989
0.0,17.0,13.0,6.0,33,05-06-1989
0.0,19.0,13.0,10.0,33,05-07-1989
0.0,15.0,13.0,10.0,33,05-08-1989
0.0,15.0,0.0,10.0,33,05-09-1989
0.0,15.0,13.0,6.0,33,05-10-1989
0.0,15.0,15.0,6.0,33,05-11-1989
0.0,15.0,16.0,6.0,33,05-12-1989
0.0,10.0,13.0,0.0,33,05-13-1989
0.0,18.0,16.0,6.0,33,05-14-1989
0.0,19.0,10.0,8.0,33,05-15-1989
0.0,19.0,0.0,12.0,33,05-16-1989
0.0,15.0,13.0,6.0,33,05-17-1989
```

Figure 4: Processed Data Sample

V. METHODOLOGY

In the large diabetes dataset, it is difficult to predict all insulin doses from data. We here predict only insulin dose code=33 for the prediction. Among 36 months, data are trained here for predicting. We try to find which algorithm best predicts the sequence of Diabetes Data.

Our works for predicting insulin are as follows:

Take our raw diabetes data
Preprocess the raw data and find the data which contains 33 codes.
Apply RNN LSTM, ANN algorithm for predicting the following state sequence of insulin level.

We first take the raw Diabetes dataset, which contains 70000 data from various diabetes patients. We only take code=33 which contains the insulin dose, where we want to measure the next insulin level of long sequence data.

For this thinking, we choose any algorithm that predicts 36 months of breakfast to lunch or dinner with a consuming process time. Here we found many algorithms that predict the following observable sequence for fifteen or 20 days. Although these algorithms don't predict 2000 days of data at a time, we selected RNN long short-term memory for solving these algorithms. To prove this, we use the best prediction algorithm neural network. We compare two algorithms to know which one gives the best prediction. Finally, we use the java machine learning tool WEKA to find the best possible frequent insulin items set, and mining some association rules with accuracy takes 90 % accuracy. We will analyze RNN RMSE and Ann RMSE and what accuracy gives ANN for each dataset. We will analyze the epochs and training data for ANN and Train data score and test data score for RNN LSTM.

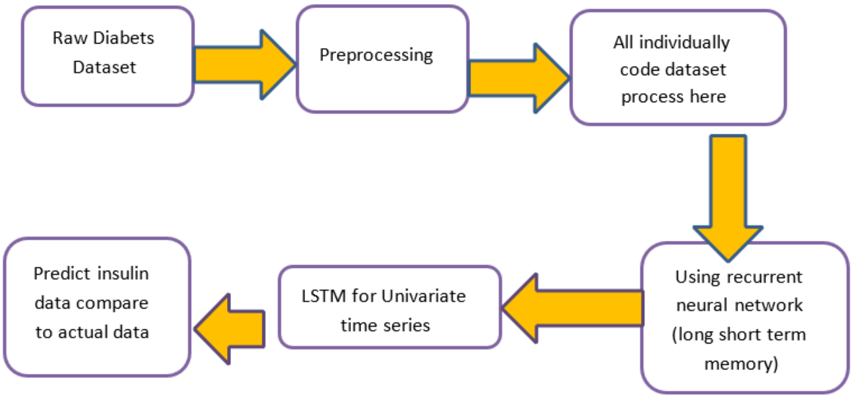

Raw Diabets Dataset
Preprocessing
All individually code dataset process here
Using recurrent neural network (long short term memory)
LSTM for Univariate time series
Predict insulin data compare to actual data

Figure 5: Working Methodology

Results and Discussion

After preprocessing data, we implement RNN LSTM for predicting the following state sequence of insulin data. RNN takes data as a sequence. For this, it is easy for RNN to predict any data sequence. We have a large 70 dataset of text files containing 1000 data. After preprocessing, we got only 130 data for each file of code =33, which contains only regular insulin doses. We have four states breakfast, lunch, and dinner. We here predict each of each state's insulin value by RNN. We measure 100 epochs for iteration to this algorithm. The performance of each prediction gives RNN a good test score RMSE, a bad test score RMSE, or an average test RMSE.

Breakfast Prediction

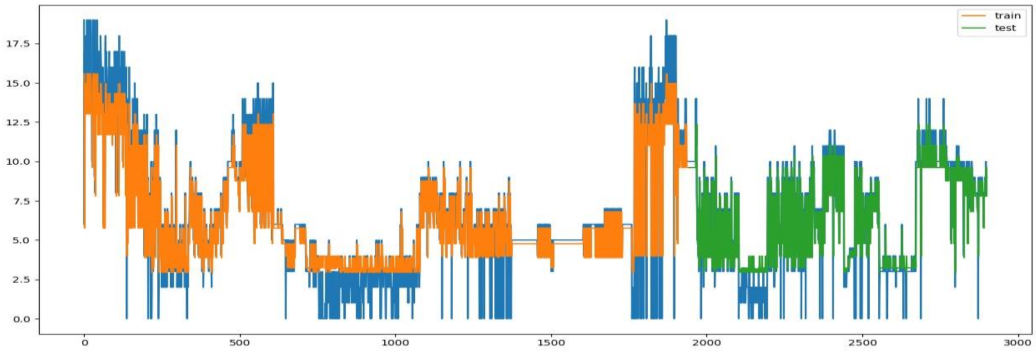


Figure 6: Breakfast Prediction (0-64) Dataset

In Figure 6, predictions are slightly better than in the previous plot. When it gets too much data, predictions are brilliant. Most of the time, predictions are slightly better. What is the benefit of RNN is that you don't need to calculate each of the states. It is the most significant advantage of RNN as it predicts everything, which gives sequence. This will be better if we give epochs 1000. RNN improves prediction in every state.

Lunch Prediction

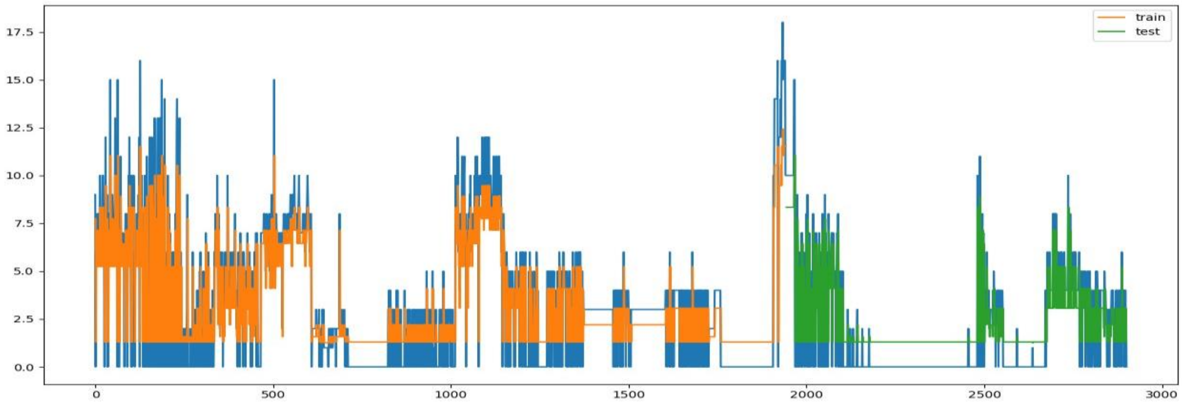


Figure 7: Lunch Prediction (0-64) Dataset

RNN predicts this well than the last three plots. Here is the dataset from (0-64) and 2905 data. RNN impressively predicts this extensive 2905-day data. Predicted data are plotted similarly to actual data. Moreover, we have got 90% accuracy in this graph (Figure 7). Loss functions are decreasing here. RNN LSTM proves that the observable sequence of data prediction is a master technique.

Dinner Prediction

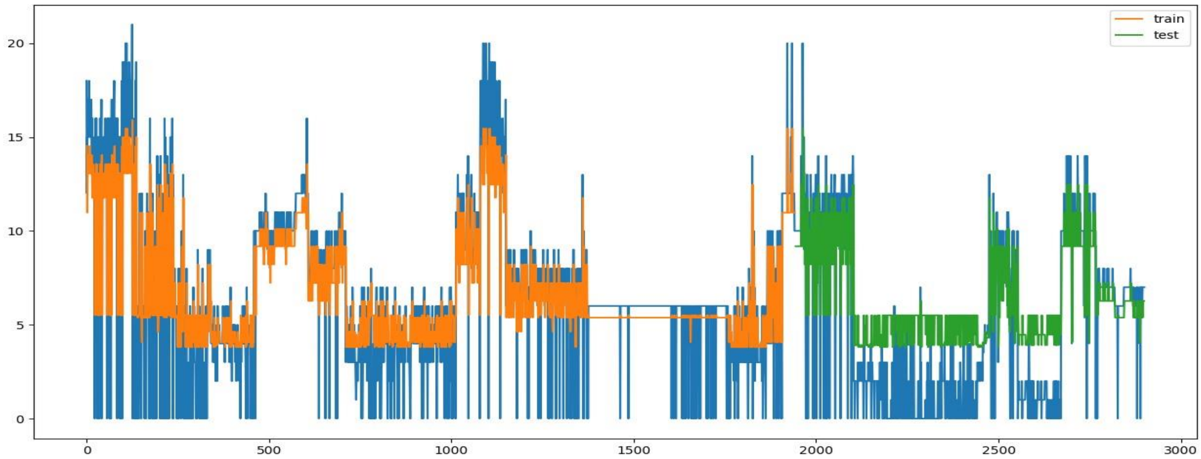


Figure 8: Dinner Prediction (0-64) Dataset

In Figure 8, we predicted 2905 data on the actual value. Prediction looks good to this dataset. This prediction improves because we take more data from the previous three plots. The loss might decrease when we take too much data and predict data would fit correctly on actual data.

Prediction Train Score and Test Score (RMSE)

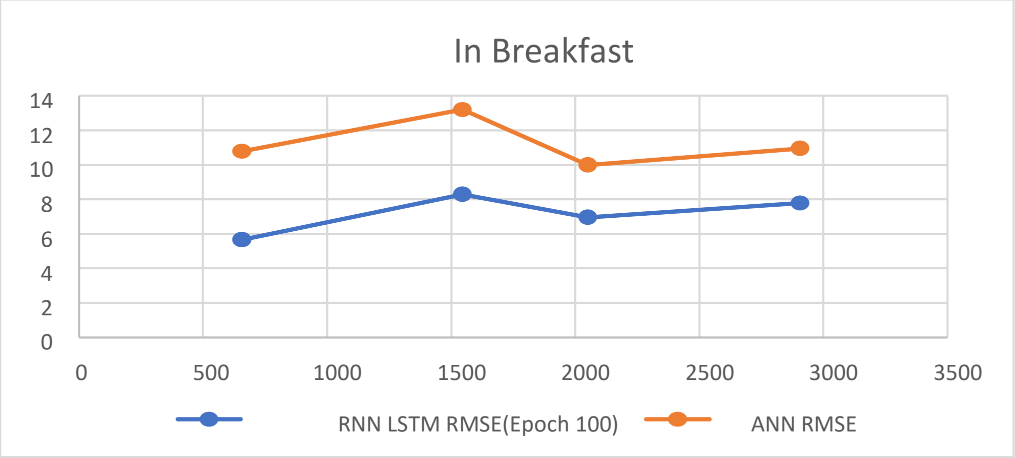


Figure 9: Breakfast Comparison

In Figure 9, RNN LSTM gives less RMSE with fewer epochs, whereas ANN gives too many errors with the same type. Here RNN predicts too well than the ANN. To solve this ANN, RNN uses three gates to overcome this neural network prediction.

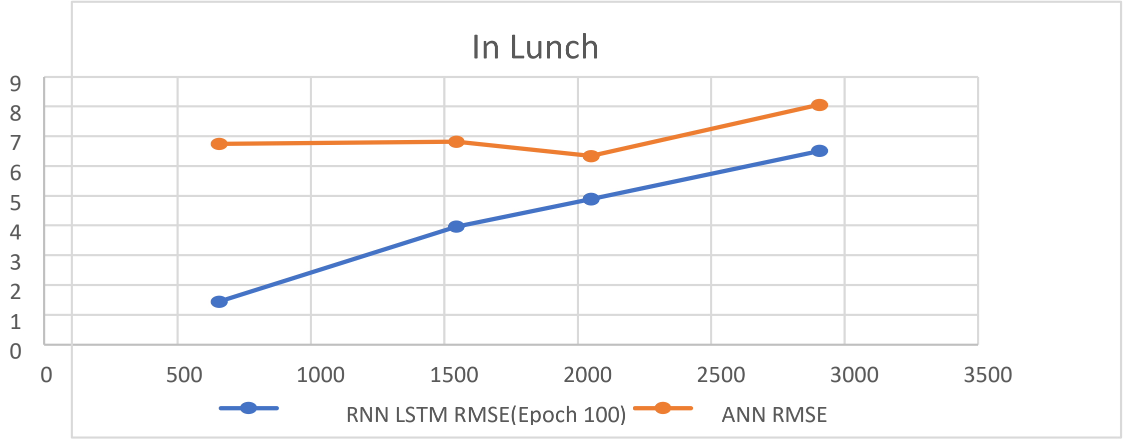


Figure 10: Lunch Comparison

Figure 10 RNN goes lower where ANN gives high errors with many epochs. Here RNN gives perfect prediction with a short epoch.

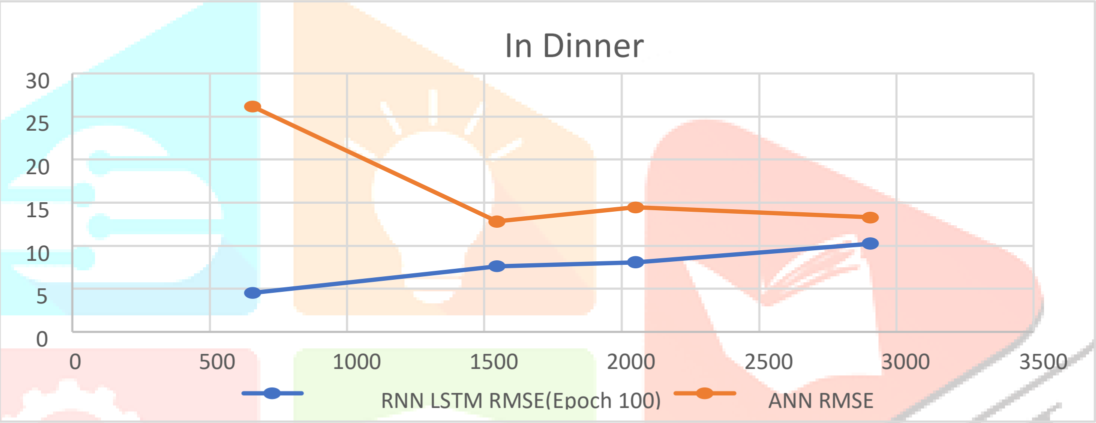


Figure 11: Dinner Comparison

In Figure 11 for dinner, RNN LSTM gives less error with tiny epoch where ANN gives too much error were prediction on insulin. Here RNN gives a better-predicted insulin sequence, whereas ANN gives less prediction than RNN.

VI. Conclusion & Future Work

We described a machine learning strategy for predicting insulin dose levels and reported experimental findings. We have shown the comparison between projected data and data in the graphs, and the results are pretty encouraging and dependable. All of the requirements for forecasting a lengthy sequence problem were met by our method. We trained our data in a time-consuming approach and achieved our goals. To summarise, training our system with a more extensive data set results in more error-free insulin prediction. We used RNN to predict the Regular Insulin dosage (code=33). We will work on various insulin doses in the future, such as NPH insulin dose (code=34) and Ultralente insulin dose (code=35).

VII. References

[1] A. M. Syed, M. U. Akram, T. Akram, M. Muzammal, S. Khalid, and M. A. Khan, "Fundus images-based detection and grading of macular edema using robust macula localization," *IEEE Access*, vol. 6, pp. 58784–58793, 2018.

[2] A. Cahn, A. Shoshan, T. Sagiv et al., "Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model," *Diabetes*, vol. 36, no. 2, pp. 1–8, 2020.

[3] Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, "A Machine Learning Approach",2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10- 12 December 2015 | Trivandrum.

[4] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.

[5] S. Islam Ayon, and M. Milon Islam, "Diabetes prediction: a deep learning approach," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 2, pp. 21–27, 2019.

[6] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, pp. 1–14, 2020.

[7] P. Suresh Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153, pp. 705-709.

[8] Ridam Pal, Dr. Jayanta Poray, and Mainak Sen, "Application of Machine Learning Algorithms on Diabetic Retinopathy", 2017 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology, May 19-20, 2017, India.

[9] Berina Alic, Lejla Gurbeta and Almir Badnjevic, "Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases", 2017 6th Mediterranean Conference on Embedded Computing (MECO), 11-15 JUNE 2017, BAR, MONTENEGRO.

[10] Dr. M. Renuka Devi and J. Maria Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730 © Research India Publications. http://www.ripublication.com

[11] Martinsson, John, Alexander Schliep, Bjorn Eliasson, Christian Meijner, Simon Persson, and Olof Mogren. "Automatic blood glucose prediction with confidence using recurrent neural networks." In *Khd@ ijcai*. 2018.

[12] Minyechil Alehegn and Rahul Joshi, "Analysis and prediction of diabetes diseases using machine learning algorithm": Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017

[13] T. E. Idriss, A. Idri, I. Abnane and Z. Bakkoury, "Predicting Blood Glucose using an LSTM Neural Network," *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2019, pp. 35-41, doi: 10.15439/2019F159.

[14] Michael Kahn, MD, PhD, Washington University, St. Louis, MO, https://archive.ics.uci.edu/ml/datasets/diabetes (accessed on 10/06/2022).