Prediction Of Medicine Consumption Using Machine Learning

Assistant Professor C R Manjunath, Shiful Islam Rony, Shuvo Nandi, Rajib Chandra Ghosh, Salman Pasha

Computer Science and Engineering Department, Jain University

Ramanagara District,Karnataka 562112,India

Abstract: The provision of pharmaceutical drugs in quantities accurate to consumption is an crucial point in the pharmaceutical industry and storage of medicines, as the production of large quantities of unnecessary drugs cause to a longer storage of drugs which is worthless. In the meantime most of the medicines have a short storage life.. A recurrent neural network with Long-Short Term Memory LSTM has been used by deep learning. The proposed methodology is founded on the seasonal number of prescription required quantities with the number of quarters as indicators. The purpose of this research is to forecast the drugs amount needed for one year. The proposed method is evaluated using two types of evaluation. The first one is based on MSE and the visualization of the actual data and forecasted data. The proposed method has reached a low value of MSE and the visualization graph is semi-identical, whereas the second evaluation method compares the result of the proposed method with traditional forecasting method. Multiple linear regression is a traditional prediction method used with the data set, whose results are relatively good and promising compared to the results of the traditional method.

Keywords : Drugs consumption forecasting, DNN, LSTM, Recurrent long-short term memory-deep learning based drug analysis and forecasting, RNN.

I. INTRODUCTION

Forecasting the drugs quantity is a method used to calculate the required amount of products with the intention to purchase. However, the quantification method includes assessing both the quantities which is required for a specific item and the monetary ways required to purchase the item. Accurate drug identification requisites lots of information, including the Essential Medicines List (EML), an approximate consumption, total prescriptions, minimal and maximal inventory levels, inventory replication, and epidemiological information. The quantities of drugs which is required should always be chosen in light of the resources and information available. If there is unavailivility of authentic information on previous consumption or morbidity, the use of the drug can be anticipated from data from other facilities, regions or countries [1]. This study shows real-world data, which has been processed to be accurate for the forecasting process. Forecasting future observations is one of the most significant tasks of time series analysis[2]. The univariate time series is a sequence of observations of a single random variable at different moments over a set of uniform time span. In fact, in forecasting the time series, the magnitude of the error increases over time, as uncertainty increases with expectations[3] Therefore, a compatible technique for the prognostic process should be selected in accordance with the rate of the available data. There are many ways that can be applied in the prognostic process. In the proposed paper, the deep learning technology is used for the quest of forecasting. Deep neural networks have been thoroughly studied in affiliated fields and have had some exceptional effects on solutions to a comprehensive range of problems. RNN models, for example, have become more popular in natural NLP research. A special case of RNN in particular, namely the Long Short-Term Memory (LSTM)[4] which has provided a distinct solution to many forecasting problems.

II. RELATED WORK

This section sums up the most recent and remarkable studies engaged with this problem. The authors worked in [1] methods of time series forecasting to predict the future consumption. The paper emphasizes on finding the most important indicators on the total quantity of the drug consumption. Consequently, the authors find 27 indicators for the qualitative and 5 for the quantitative. Monthly use indicators for the drug, seasonal changes, and data distribution were assessed. These above mentioned indicators were the most vital ones that influence the total sum of medicines annually required. In [5] the researchers put forward a Smart Sales Prediction Analysis to foretell the requirement of the Pharmaceutical distribution companies with a system that depend on real-time application, the forecasting time series method (ARIMA) and neural network. The hybrid method represents a real-time data mining methodology with a combination of a predetermined factor and a new factor. The outcome of the proposed method pointed out that it is a real-time approximation forecasting method. Evermore with this research scope, Oscar Chang and et al, have worked in a deep neural network in [6] to predict the weekly data of pharmaceutical products. In this paper, an auto-encoder is used as data retraction for the outward neural network. This recent information is entered with retractions into the third shallow network for the purpose of predicting its own week The system can produce a good and stable weekly forecast. The researchers held in [7] a new methodology for pre-evaluating the history of release and evolutionary behavior in a homogeneous category of pharmaceutical drugs. The methodology is based on dynamic meta-analysis ,mentioning the problem of predicting new products of drugs. The properties of drug spread conduct in a clear ranitidine class, whereas the meta-analysis is based on the dynamic assessment study of expantion parameters and the observation of drug sales information. This method is used to foretell the behavior of new drug labeling. The authors employed in [8] methods of time series forecasting to predict the future consumption and purchase of the drug RAPILYSIN LYPDINJ 2X1.16G/VIAL (RL). They used Auto-Regressive Moving Average (ARMA) and Auto-Regressive Integrated Moving Average (ARIMA). Using a very short time series of drug consumption, the MSE is sued for determining the accuracy of the forecasting method.

III. TIME SERIES DATA AND FORECASTING PROCES

Time series data is recording of observations of a specific subject over multiple time periods. The time series can be explained as a set of observable and equally spaced data recorded at regular intervals. The time series analysis involves methods for analyzing data for the purpose of extract compatibility and meaning from statistics and give a statement for the data. In fact, the time series forecasting process is the use of a model to predict future values founded on ones that have been observed subsequently .As the methods connected to this process, forecasting only rely on past values and presumes that the factors that have influenced the past and the present will continue to invaded the future. In the Meantime, if the time series of the future values can be assumed from their previous values, the time series is inevitable (deterministic). If the future of a time series can be partly determined by previous values, it means that the time series is stochastic or random[8]. Linear models, such as the (ARMA) and (ARIMA), are both successful linear methodologies, but their predictive power is constrained by linear behavior. As a result, this will not always be acceptable, especially for application in predicting nonlinear time series. Accordingly, many predictive techniques have been flourished over the past few decades. when it is compared to other methods ,each technique has some advantages and disadvantages too . There are certain factors that have arisen in the last century which led to the advancement of time-series forecasting. On the other hand, advancement in computer science resulted in the use of more complex algorithms. Whereas, on the other hand, the advancement of machine learning techniques, such as neural networks, has led to greatdevelopment in time series. In fact, neural networks are a very familiar technique in assuming the time series after being verified as an effective way of predicting the handling of nonlinear input and output variables, with the capability to estimate any function in specific circumstances. In other words, if the data are affected by nonlinear behavior, linear methods are incapable to model nonlinear latent relationships. Recently, neural networks used in the prediction of time series in particular have magnificently improved in terms of its performance. The popularity of neural networks is because of their high capability to emulate a wide range of basic nonlinear behaviors and to cover a wide range of fields. Therefore, many studies have used neural networks to predict time series and assimilate them with traditional methods. However, the results were not well-defined for the purpose of adopting a particular methodology[9]. In fact, neural networks have developed rapidly in deep learning, which is enumerated to be a revolution in machine learning and used in all fields such as speech recognition, natural language processing, prediction, classification, image and video processing.

III. DEEP LEARNING TECHNIQUE

Deep learning is the evolving form of traditional neural network technology. It can be explained as a neural network that is build on many layers[11]. The great advancement of computer science has aided to build the deep learning of multilayered neural networks. Therefore, the discovery of nonlinear patterns is one of the most significant deep learning capabilities[12]. Deep learning has been used broadly in speech, signal, image, video, and text analysis. The most recent performance improvement was found to be more than 30%. The power of deep learning remains in the automatic learning feature of voluminous data sets [13]. It facilitates the extraction of properties and allows for optimal learning processes[14] [10]. However, the main algorithms of deep learning used in many machine learning applications are convolutional neural network, deep belief network, deep neural network, and recurrent neural network[12]. Recurrent neural networks are a class of interconnected models that have an internal cell or short-term memory because of recurrent feedback connections and are competent for dealing with series problems such as speech classification, prediction, and generation[15]. The use of recurrent neural networks gave promising results in machine learning tasks. In contrast to the results presented through neural networks with feedforward, the network can deal with sequence data in varying lengths by relying on a repeated hidden cell that is activated each time based on the one the one that preceded [16]. In the proposed paper, the RNN (LSTM) is used to predict of the amount of drug brands used based on its past information

STM (Long-Short Term Memory)

RNN network proved very successful in various applications such as speech recognition, translation and image captions. The reason for the success of this network lies in the use of LSTMs: long-term memory, a special type of recurrent neural networks is LSTM. Figure (1) shows the LSTM cell diagram.

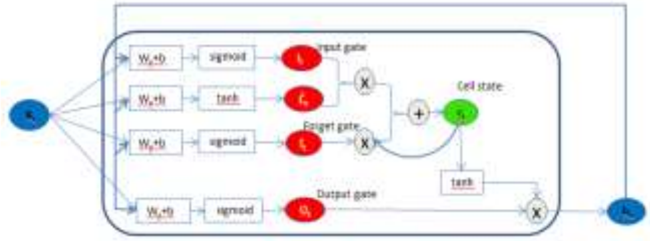


Figure 1: LSTM cell diagram

The LSTM unit differs from the traditional recurrent unit, maintaining the current memory Rn at time t. The input at time t is xt, ht−1, ct−1, and the output is ht, ct .These can be updated by means of the following equations [RNN-write]:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_t) \quad (2)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$
$$g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g) \quad (4)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$
$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

Where the logistic sigmoid function and ⊙ is is the operator denoted to the element-wise vector product. At each time step t, there are

- Input gate i_t
- forget gate f_t,
- output gate o_t,
- memory cell c_t,
- hidden unit h_t. h_0

and c_0 can be initialized to 0 and the parameters of the LSTM are W,U, b[LSM]. Figure (2) shows the RNN with LSTM diagram.

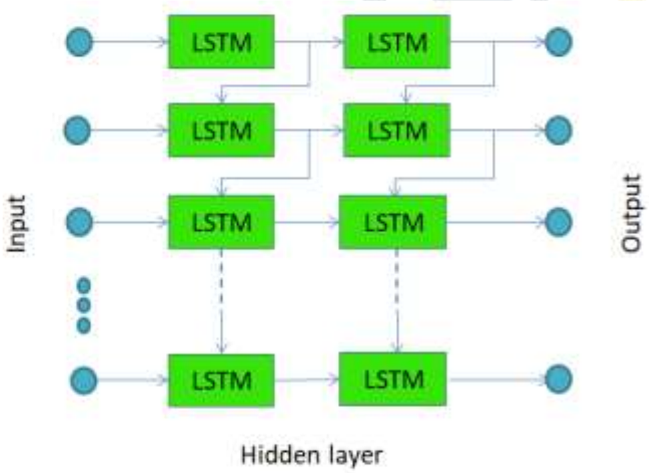
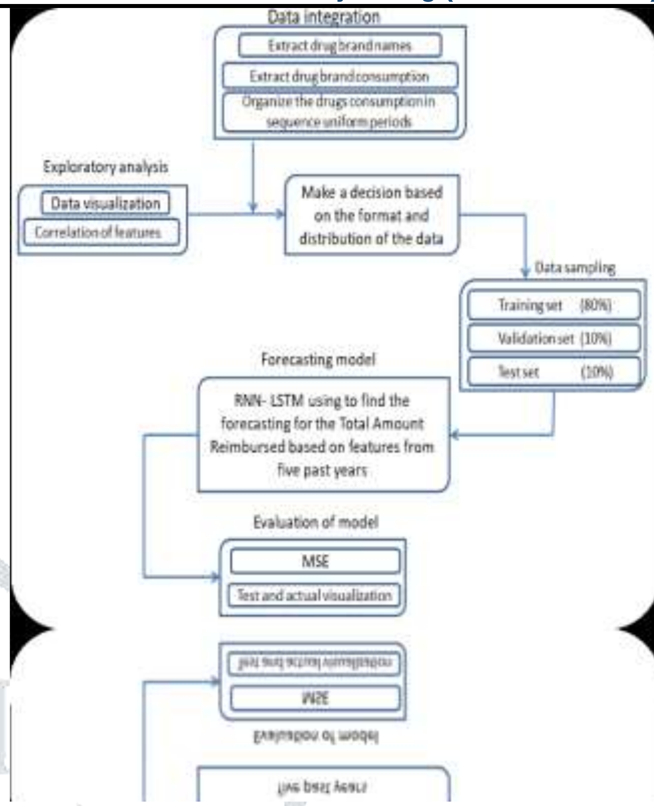


Figure 2: RNN with LSTM

IV. THE PROPOSED METHODOLOGY

This section explains the methodology steps in detail: As shown in Figure (3), the methodology steps consists of data integration, exploratory analysis, data sampling, forecasting model and the evaluation of model.



*5.1. **Data integration***: For the purpose of predicting the future need for drugs, the utilization data for these drugs should be available in advance. The proposed research was treated with a real-world data collection representing the use of drugs in the United States over a period of five years. This step included three sub-steps for configuring data:

Extract drug brand name: A collection of brand names for specific drugs has been extracted from the website specializing in drug and brands used in USA" www.drug.com". 135 drug brand names are extracted.

Extract drug brands consumption: At this point, the real world data collection, which represents the consumption of medicines in the United States, has been loaded for the past five years. The consumption restrictions for a group of drugs extracted at a previous point are extract from this dataset.

Organized the drug consumption in sequence uniform period: The real-world data set is unordered but contains important and valuable data, which makes it an essential step. It is known that the analysis of time series requires past data at fixed time intervals. In this step, the number of prescriptions and quantity required for each drug brand is collected over the quarter. In other words, the extracted data represents the total number of prescriptions and quantity required for each drug over four quarters of the year and for the previous five years. This is the most significant step in the search as the results of the forecasting process depend heavily on the organizer of the data. Otherwise, the process of prediction will be unsuccessful.

Data Normalization: this process is used to put the extracted data of drugs in a range of (0-1). Equation (7) is used to normalize the data set.

$$normalized = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (7)$$

5.2. ***Data exploratory analysis:*** This step is used to understand the nature of the data used in the research, and includes the following steps

Data visualization: Data is visualized at this stage. Figure (4) depicts the data of the required quantities over six years. It is possible to observe that the data does not have a specific format, as it is unstable. For some drugs, the total consumption has decreased, whereas others have increased over the years. Some drugs started to appear gradually over the years, while others disappeared. This seasonal pattern seems to be weak and unstable for drugs.

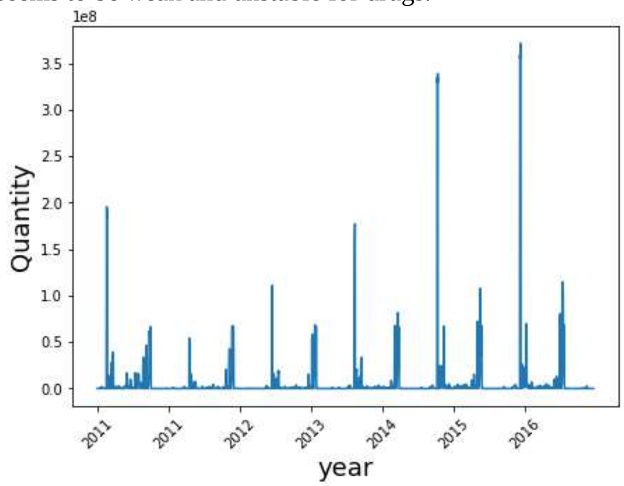


Figure 4: Quantity Required for six year from (2011-2016)

Figure (5) shows the total number of prescription for each drug. The data is organized into the total number of prescriptions: one each quarter, four quarters for each year. The data is integrated for six year from (2011-2016).

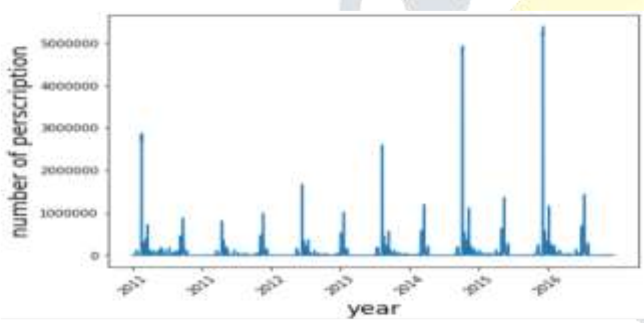


Figure 5: number of prescription for six year from (2001-2016)

Features correlation: With the aim of analyzing available data and after its visualization, the correlation between the number of prescriptions and the quantity required of the drug is calculated, eventually resulting in a correlation between them up to more than 90%, as Table (1) shows. This means that when the number of prescriptions increases, the required amount of the drug increases accordingly, and vice versa.

Table 1:Correlation of quantity required & number of prescription

	quantity required	number of prescriptions
quantity required	1.00	0.96
number of prescriptions	0.96	1.00

5.3. ***Data sampling:*** The training process is one of the most important stages in machine training. In order for the training to be successful, the data should be separated into the training set, the validation set, and the test set. The dataset obtained from the data integration step is as follows (135 drugs; the data for each drug was arranged over four quarters per year for a period of six years (2011-2016)). The training set is (80%), (10%) for the validation set, and (10%) for the test set. The data for the first five years (2011-2015) represent the deep network inputs that form the time series, whereas the data for the last year (2016) represents the desired value for the deep network, as the (LSTM) is a supervised network.

5.4. Forecasting model, Figure (6) shows the deep LSTM network designed for drugs forecasting in the future.

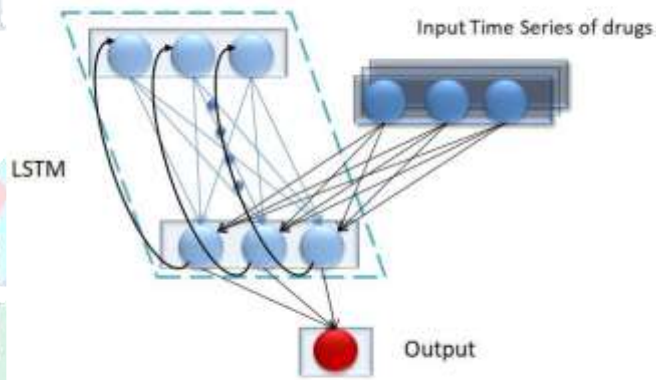


Figure 6:deep neural network (LSTM)

Deep drugs forecasting descriptive algorithm

Let:
t: t=0 to T where T represent the time series (5 previous years).
N: represent the number of LSTM blocks.
M: represent number of inputs (number of drugs is 135).
X_t: represent the input vector of x at time t.
H_{t-1}: previous cell output.
C_{t-1}: previous cell memory.
H_t: current cell output.
C_t: current cell memory.
the following weights for an LSTM layer:
• Input weights: W_i , W_f , $W_o \in R$ N×M
• Recurrent weights: U_i , U_f , $U_o \in R$ N×N
• Bias weights: b_i , b_f , $b_o \in R$ N
The vector formula for LSTM compute as

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$g_t = \tanh\left(W_g x_t + U_g h_{t-1} + b_g\right) \quad (4)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$
$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

Where σ is logistic sigmoid $(\sigma(x) = \frac{1}{1+e^{-x}})$

And $g(x) = h(x) = \tanh(x)$

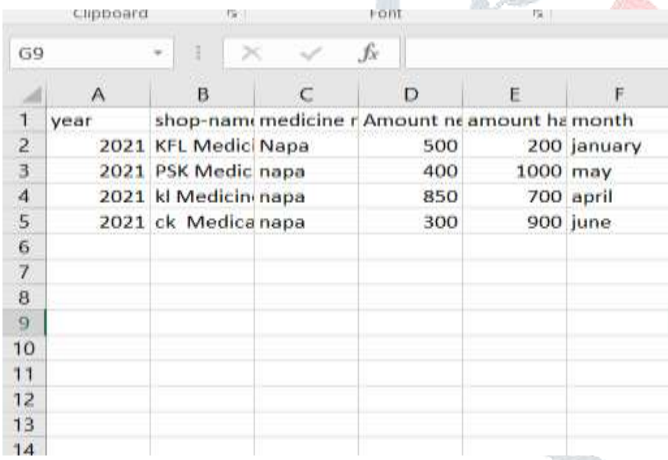
VI. EVALUATION OF FORECASTING MODEL

When constructing any model, this model should be evaluated. This is accomplished through:
Mean squared error one of the most famous error measures is used to evaluated the accuracy of model.

$$MSE = \sum_i \frac{(d_i - y_i)^2}{n} \qquad (8)$$

VII. RESULTS AND DISCUSSION

In this figure (1), We show the data of year, Shop-name, medicine needed, medicine have and the month in csv formet.



	A	B	C	D	E	F
1	year	shop-nam	medicine r	Amount ne	amount ha	month
2	2021	KFL Medici	Napa	500	200	january
3	2021	PSK Medic	napa	400	1000	may
4	2021	kl Medicin	napa	850	700	april
5	2021	ck Medica	napa	300	900	june

Figure 1: Sample of data

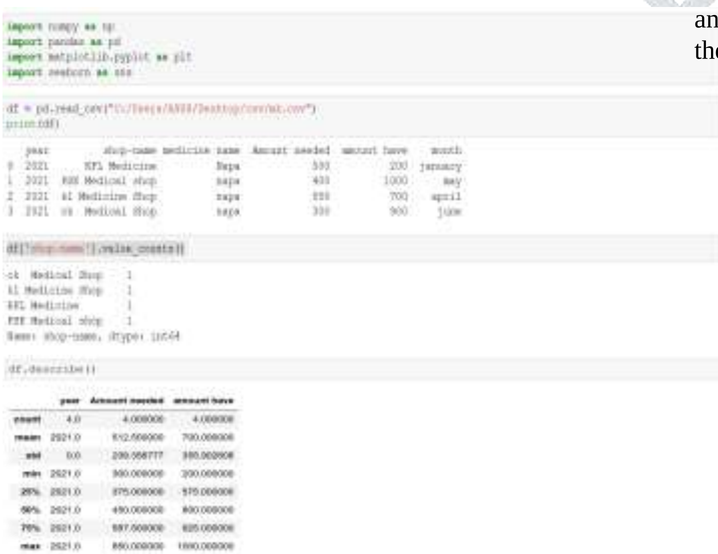


Figure 2: Sample of extracted data

- By using pandas DataFrames, we manipulate and visualize time series data.
- By using numpy library, we visualize the data in n-dimensional array.
- By using Matplotlib comprehensive library, we create static,animated and interactive visualization.
- By using seaborn visualization library, we visualize the based on matplotlib.
- By using describe () method, we calculate some statistical data like percentile, mean and std of the numerical values.

In figure (3), we show the different month of amount have and amount needed. The blue line is the amount needed and the orange line is amount have.

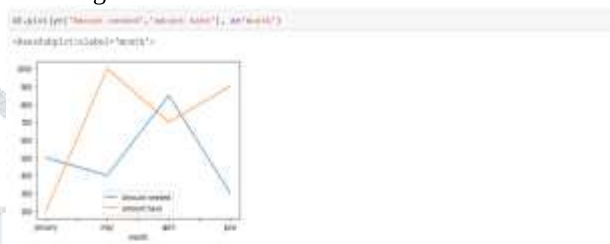


Figure 3: Number of amount have & amount needed in different month

In figure (4), we show the different shop-name of amount have and amount needed. The blue line is the amount needed and the orange line is amount have.

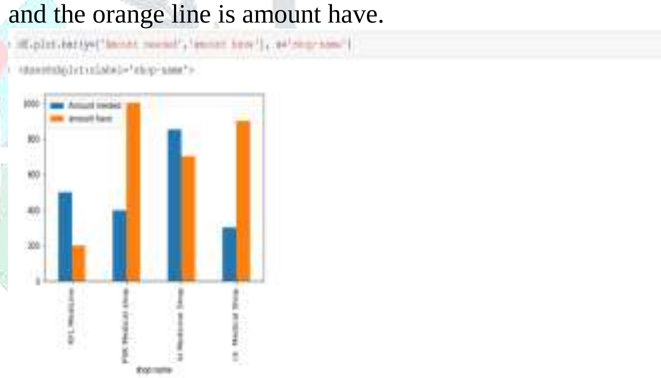


Figure 4: Number of amount have & amount needed in different shop-name

In figure (5), we show the medicine name of amount have and amount needed. The blue line is the amount needed and the orange line is amount have.

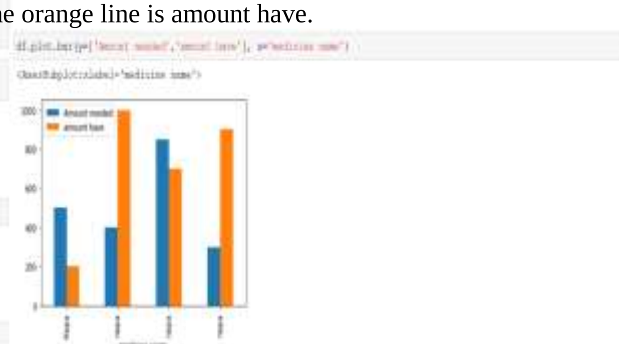


Figure 5: Number of amount have & amount needed in medicine

In figure (6), we show the amount needed and amount have in individual amount. In figure (6), we showing that if any shop have equal or more than 500 medicine. If they have they will only show the True and if they don't have they will show as a False(dttype:bool). And they will also show the shop name of available or not available as True or False.

Figure 6: Number of amount have & amount needed in individual amount

IX. CONCLUSION

The proposed paper presented an automated system from the beginning of its work to its end. The system consists of five steps, each of which has achieved the desired goal clearly. First, the data was extracted from real-world data. The most important step at this stage is the collection of data in a format suitable for time series. The duration of the test was six years, five of which represented the input data and the last year represented the forecasted data. In the second stage, the data is clarified by means of visualization. The importance of this stage is to identify the distribution of the data and understand its behavior. Third, the separation of data into training data and test data. Fourth, a deep network of LSTM type is designed, which gave good results. The research reached the lowest error value after testing more than the design of the deep network. Finally, for the purpose of displaying network performance, the result of deep network and real-world data are compared through visualization. The system has provided integrated as well as promising results for the purpose of predicting the future needs of medicines in order to control the quantities of medicines produced according to the supply and demand theory.

REFERENCES

II. [1] M. J. Iqbal, M. I. Geer, and P. A. Dar, "Evaluation of Medicines Forecasting and Quantification Practices in Various Evaluation of Medicines Forecasting and Quantification Practices in Various Public Sector Hospitals Using Indicator Based Assessment Tool," J. Appl. Pharm. Sci., vol. 7 (12), no. December 2017, pp. 072–076, 2018.

III. [2] P.-A. Cornillon, W. Imam, and E. Matzner-LZber, "Forecasting time series using principal component analysis with respect to instrumental variables Forecasting time series using principal component analysis with respect to instrumental variables," Comput. Stat. Data Anal., vol. 52, no. July, pp. 1269 – 1280, 2008.

IV. [3] I. A. Gheyas and L. S. Smith, "A Neural Network Approach to Time Series Forecasting," Proc. World Congr. Eng., vol. II, no. 1, pp. 1–5, 2009.

V. [4] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks," SIGIR,

VI. [5] K. N. Mahajan, "Business Intelligent Smart Sales Prediction Analysis for Pharmaceutical Distribution and Proposed Generic Model," vol. 8, no. 3, pp. 407–412, 2017.

VII. [6] Y. Tech, "A Deep Learning Algorithm to Forecast Sales of Pharmaceutical Products," no. September, 2017.

VIII. [7] R. Guseo et al., "Pre-launch forecasting of a pharmaceutical drug," Int. J. Pharm. Healthc. Mark., vol. 11, no. 4, pp. 412–438, 2017.

IX. [8] A. Papana, D. Folinas, and A. Fotiadis, "Forecasting the consumption and the purchase of a drug," Int. Conf. SUPPLY Chain. Funct., vol. 2, 2016.

X. [9] N. K. Zadeh, M. M. Sepehri, and H. Farvaresh, "Intelligent Sales Prediction for Pharmaceutical Distribution Companies : A Data Mining Based Approach," Hindawi Publ. Corp., vol. 2014, 2014.

XI. [10] T. Pham, T. Tran, and D. Phung, "Predicting healthcare trajectories from medical records : A deep learning approach Predicting healthcare trajectories from medical records : A deep learning approach," no. October, 2017.

XII. [11] E. AL-Shamery and A. AL-haq, "An Optimized Feed Forward Neural Network for Reducing Error Based Stoch Market Prediction," J. Eng. Appl. Sci., vol. 13, no. Special Issue 5, pp. 4616–4621, 2018.

XIII. [12] F. Jiang et al., "Artificial intelligence in healthcare : past , present and future," stroke Vasc. Neurol., vol. first publ, 2017.

XIV. [13] E. AL-Shamery and A. AL-haq, "Enhancing Prediction of NASDAQ Stock Market Based on Technical Indicators," J. Eng. Appl. Sci., vol. 13, no. Special Issue 5, pp. 4630–4636, 2018.

XV. [14] H. Khalid Obayes, N. Al – A'araji, and E. AL-Shamery, "Deep Neural Network for Enhancing Drug-Utilization Clustering," Int. J. Eng. Technol., vol. 8, pp. 290–298, 2019.

XVI. [15] J. Koutn and K. Greff, "A Clockwork RNN," arXiv, vol. 1402.3511v, 2014.

XVII. *[16] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Aaron, 2016.*

.