FORECASTING CONSUMER PRICE INDEX (CPI) USING DEEP LEARNING AND HYBRID ENSEMBLE TECHNIQUES

Submitted in partial fulfillment of the requirements
for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

Mohammed Adnan A (Reg.No - 39110630)
J Prince Immanuel (Reg.No 39110802)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119

APRIL - 2023

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with —A‖ grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Mohammed Adnan A (39110630) and J Prince Immanuel (39110802)** who carried out the Project Phase-II entitled **“Forecasting consumer price index (CPI) values using deep learning and hybrid ensemble technique”** under my supervision from December 2022 to April 2023.

Internal Guide

Dr. Roobini M.S. M.E., Ph.D

Head of the Department

Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva voce Examination held on 24.03.23

InternalExaminer **ExternalExaminer**

DECLARATION

I, **J Prince Immanuel (39110802),** hereby declare that the Project Phase-2 Report entitled **Forecasting consumer price index (CPI) values using deep learning and hybrid ensemble technique"** done by me under the guidance of **Dr. Roobini M.S.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

Prince

DATE: 24/04/2023
PLACE: Chennai **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph. D**, **Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.,** Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. Roobini M.S. M.E., Ph.D,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-2 project work.

I wish to express my thanks toall Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

In today's day and age, economic crises are all over the world due to high inflation. Inflation is a rise in price of the goods and services produced in a country. Prices rise, which means that one unit of money, buys fewer goods and services. This loss of purchasing power impacts the cost of living for the common public which ultimately leads to a deceleration in economic growth. Thus, it has a negative impact on the purchasing power of the people. Depending upon the selected set of goods and services used, multiple types of baskets of goods are calculated and tracked as price indexes to calculate inflation or deflation. One type of price index proposed in this project is Consumer Price Index (CPI) which examines the weighted average of prices of a basket of goods and services such as transportation, food and medical care. This paper proposes different deep learning time series models such as LSTM, BI-LSTM and hybrid ensemble learning to forecast the Indian consumer price index (CPI). These two single RNN models (LSTMS and BI-LSTMS) are compared with the hybrid ensemble learning model to see which gives better forecasting results for the consumer price index.

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Consumer price index (CPI) is the measure of change in the consumer prices of goods and services for a household over time. It is a metric to determine a country's inflation/deflation rate for the proper economic development and enhancing the livelihood of the people. Inflation is the result of an increase in consumer price index value referring to increased prices of goods and services. In contrast, a decrease in consumer price index value results in deflation. Although an escalation in prices results in a decline of purchasing power of the people, a small amount of inflation is necessary for sustainable economic growth of a country. This paper proposes the forecasting of CPI values using deep learning time series models.

The mid 1980s saw a period of advanced economic growth which is known as "great Moderation". This led to the increase the various domains in the economy especially in housing domains. This intern led to an economic bubble [1]. This bubble had a lift off in the year 2008 also known as the Global financial crisis of 2008 experienced a fall out of market prices and also the surprising fall of the housing domains. And also more recently the covid pandemic break out which had innumerable amendment in the economy [1][2]. This had a big impact to the market players. This led to the ability to forecast and learn more about the trends of the economic growth [2]. Due to the advisory growth in data science, the forecasting using Time Series has gained a lot of popularity in the financial world. The successful prediction of inflation which allows the shareholders to enter the market and exit the market at an appropriate time could return a lot of sanatives [3].

LSTMS a Time Series model which is gaining popularity today is widely used. But LSMTS are a product of RNN which is its parent [4]. However RNN is discarded due to the problem of vanishing gradient which came about with LSTMS and BI-LSTMS. LSTMS has solved the problem of RNN network which improves the memory and enables to work with larger data inputs [5],[6]. By taking the Root Mean Square

Error(RMSE) and Mean Absolute Error(MSE), it can be seen that the LSTMS show a better outcome as compared to the RNN model [7].

LSTMS works with the concept of a gate that stores a huge quantity of data. This helps the model to remember data and produce accurate results. These gate concepts already produce quality results but the problem is it does not produce back propagation whereas BI-LSTMS have extra training. In some cases, it is necessary to train the past and future values to predict the present values and this can be done using BI_LSTMS but in this case, LSMTS do not backpropagate and hence it might cause a training deficiency on the model. The object is to explore whether the extra set of training is required for the proposed dataset by which we can identify which deep learning model, either LSTMS or BI-LSTMS is better [8].

After that we compare both the deep Learning models with the proposed hybrid ensemble learning technique. Although these deep learning models are used by many it does have many issues in terms of accuracy. These hybrid models helps us to combine 2 to 3 models to give a better performance when compared to a novel model [9]. This paper has illustrated hybrid ensemble learning by combining LSTMS, ANN and using a boating algorithm such as adaboost (ANN-LSTM-ADA). Although ANN-LSTM-ADA beat the LSTMS model, it was just a shy away from having a better accuracy than BI-LSTM. but the main advantages was that the difference wasn't huge and when in terms of time complexity the hybrid model was much better and efficient and more user friendly in gpu power usage compared to BI-LSTM model.

CHAPTER 2

LITERATURE SURVEY

This problem statement has been extensively studied over the years by researchers in a bid to create a solution. Various solutions have been proposed in this area of research and some of their findings are summarized below.

The model proposed by Oren Barkhan [2] used Hierarchical Recurrent Neural Network model to predict disaggregated inflation components of the Consumer Price Index (CPI). The prediction is done on lower levels of CPI index which is highly volatile by using the higher levels of CPI index. This does not calculate CPI index as a whole but predicts the model using specific parameters of the CPI index.

Md. Arif Istiake Sunny [3] put forward different methods of RNN such as LSTMS and BI_LSTMS to forecast the stock price gains in the financial world and compares both the RNN models by using rmse values. The rmse values measured by tuning the following hyperparameters :

- Number of epochs
- Hidden Layers
- Dense Layers
- Units used in Dense Layers

From this paper it has been concluded that BI_LSTMS generated lower rmse values when compared to LSTMS.

The work of Sepp Hochreiter [5] explains how LSTMS solves the problem about storing information over extended time intervals. This paper also proposes backpropagation techniques and how to overcome insufficient decaying error backflow. LSTM also solves complex, artificial long time lag tasks that have never been solved by previous recurrent network algorithms.

Khaled A [4] has considered bidirectional LSTMS and stacked LStms in forecasting real world dataset for a stock market. These deep learning models are then benchmarked with shallow neural networks and simple forms of LSTMS model. This paper uses different metrics to measure the performance. Hyper tuning is done by alternating the number of hidden layers.

The short-term prediction was more accurately measured when compared to the long term by both stacked LSTMS and Bidirectional LSTMS. These models also outperformed the shallow neural networks and vanilla LSTMS.

Neda Tavakoli [6] suggests that the RNN-LSTMS model outperforms the ARIMA models on the basis of forecasting. It also compared BI-LSTMS with vanilla LSTMS and showed that additional training i.e. back propagation is required to display better forecasted results. Furthermore the additional number of hidden layers is examined to fine tune the parameters to improve the proposed model.

Md. Kowshera [7] brings about a suggestion combining two or more algorithms called hybrid deep learning models would give better results when compared to the single deep learning models. This paper illustrates how the integrated models of ANN-LSTMS and ANN-Bi-LSTMS are performing better than vanilla LSTMS and Bi-LSTMS.Overall it can be seen from this paper that ANN-Bi-LSTMS conveyed better results compared to all other models.

Shaolong Sun [8] model consists of a hybrid ensemble learning approach to predict the financial time series data. This ensemble technique is comprised of a boosting algorithm which is ada boost in combination with LSTM, ELM, SVR and MLP to enhance the parent model. The model is measured by metrics such as MAPE and is compared with different ML and DL models such as ARIMA, SVR, ELM and LSTM.

AdaBoost-LSTM model outperforms all other models and is best suited for nonlinearity and irregularity, such as exchange rates and stock indexes.

Shiqing Sang [10] suggests that existing methods either lack theoretical support or demand large integrated models. This brought about the rise in use of hybrid models to overcome these weaknesses and produce more efficient results. This paper proposes the model Ensembles of Gradient Boosting Recurrent Neural Network (EGB-RNN) which combines the gradient boosting ensemble framework with three types of recurrent neural network models, namely Minimal Gated Unit (MGU), Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM).

Different datasets are used on these hybrid models mentioned above which brought about varying results in each dataset. Moreover these hybrid models outperformed all the parent models proving that hybrid models can fix the gap which this vanilla model poses.

2.1 INFERENCES FROM LITREATURE SURVEY

[3] Data was more focused on a single stock market and was not generalized on different categories. This model is very specific and does not generalise properly on different sets of data.

[4] Although the short term predictions were highly accurate, the data which was considered was not of a real world application. The results would have been varying if the real time datasets were used.

[6] The use of Bi-Directional LSTMS can produce better forecasting results at the expense of less efficiency such as higher time complexities and a requirement of high usage of GPU power.

[7] Hybrid model being efficient in terms of predicting may not always be robust. A wide range of data is necessary to make these models more compatible to any appropriate data set.

[10] Lack of evidence of why extreme gradient boosting on RNN networks are susceptible to data. Moreover only boosting algorithms were considered ignoring bagging and stacking models. This leaves out the possibilities of attaining the best possible results.

2.2 OPEN PROBLEMS IN EXISTING SYSTEM

There are three standard problems when it comes to forecasting:

1. Better accuracy
2. Time Complexity
3. GPU Power usage

To measure accuracy we use metrics such as root mean squared error, mean absolute error etc. In order to forecast the future values a decent accuracy score for the metrics mentioned should be obtained

The time complexity refers to the total amount of time it takes to process each epoch. The lower the time complexity the better the efficiency of the model.

The processing power required by the gpu should be minimum for better user friendly options. Higher power usage higher the hardware requirement which leads to high capital spent on the GPUs.

There is a direct correlation between these three components. Higher accurate predicted value would result in less efficiency such as high Time Complexity and high GPU Power usage. Similarly lesser time required for the model to process will result in lower accuracies. Hence an equilibrium must be found between these 3 components.

No. of Epochs	LSTM RMSE	Time (min)	BI-LSTM	Time (min)
10	0.0011000	3	0.0007167	8
20	0.0007250	6	0.0006459	15
50	0.0004933	15	0.0004219	40
100	0.0004928	30	0.0004127	70
250	0.0031980	75	0.0003568	200

Fig 2.1 [3] indicating the varying time in training the models

CHAPTER 3

REQUIREMENT ANALYSIS

This project's main objective is to create a forecasting engine to forecast any consumer representative basket by determining the inflationary rates in the future. Essential requirements of the system and its design constraints are discussed below.

3.1 REQUIREMENTS

Data requirements

Consumer price index dataset is taken from www.india.gov.in website. It consists of two columns which are data and CPI values. The CPI values are record monthly starting from the year 2016 till 2021. This is a raw dataset and preprocessing is done to provide a clean data to get insights and display results respectively.

Performance requirements

The algorithm must be fast and efficient in processing the data, a hybrid ensemble learning algorithm to calculate the forecasted results. The deep learning algorithm must be accurate in forecasting future values.

Hardware Requirement for Implementation

- System architecture: Windows- 64-bit x86, 32-bit x86; MacOS- 64-bit x86; Linux- 64-bit x86, 64-bit aarch64 (AWS Graviton2 / arm64), 64-bit Power8/Power9, s390x (Linux on IBM Z & LinuxONE).

- Minimum 15 GB disk space to download and install.

- Minimum of 8gb RAM and 4gb VRAM for GPU

Software Requirement for Implementation

- Operating system: Windows 8 or newer, 64-bit mac OS 10.13+, or Linux.
- Installing Anaconda Individual edition 64-bit (PY 3.7)
- Use Jupyter notebook in Anaconda Navigator for running project Python notebook.
- Python Notebook also works in Google Colab, Kaggle Notebook editor.

3.2 DESIGN CONSTRAINTS:

Listed below are some of the models that have been designed while doing the project:

1. Building the LSTM model to showcase the accuracy of the predicted values
2. This LSTM model is then compared with the Bi-LSTM model
3. These models are compared on the basis of accuracy, time complexities and GPU power usage.
4. Our proposed Hybrid ensemble learning algorithm is then built and forecasted.
5. The Hybrid ensemble learning consists of ANN , LSTM and ada boost algorithms
6. This ANN-LSTM-ADA is compared with the other two models to overcome the disadvantages of these single deep learning models.

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

From the literature review, we are able to see the shortcomings of existing and proposed solutions to solve the problem at hand.

Our proposed system aims to overcome the shortcomings of the existing models as well as the ones discussed in the literature survey and prove as a viable solution that could be implemented.

In contrast to some of the approaches that do not consider the efficiency of time complexities and GPU power usage during the processing of the epochs. Our hybrid ensemble learning model takes into consideration the efficiency of the model as well.

We believe that attaining the equilibrium between the three important factors such as accuracy, Time Complexity and GPU power usage can be used to forecast the Consumer price index efficiently. We take into consideration of the dataset of consumer price index of India recorded monthly.

The proposed model can be broken down into two main modules which are:

- Proposed Hybrid ensemble learning model
- hybrid model vs LSTM vs Bi-LSTM

Our hybrid ensemble learning model consists of three deep learning models such ANN, LSTM and ada boost. These models are combines to produce the forecasting of the consumer price index which is better in accuracy and better in efficiency compared to the single deep learning models.

The architecture and working of the proposed model is discussed below.

4.1 SYSTEM ARCHITECTURE

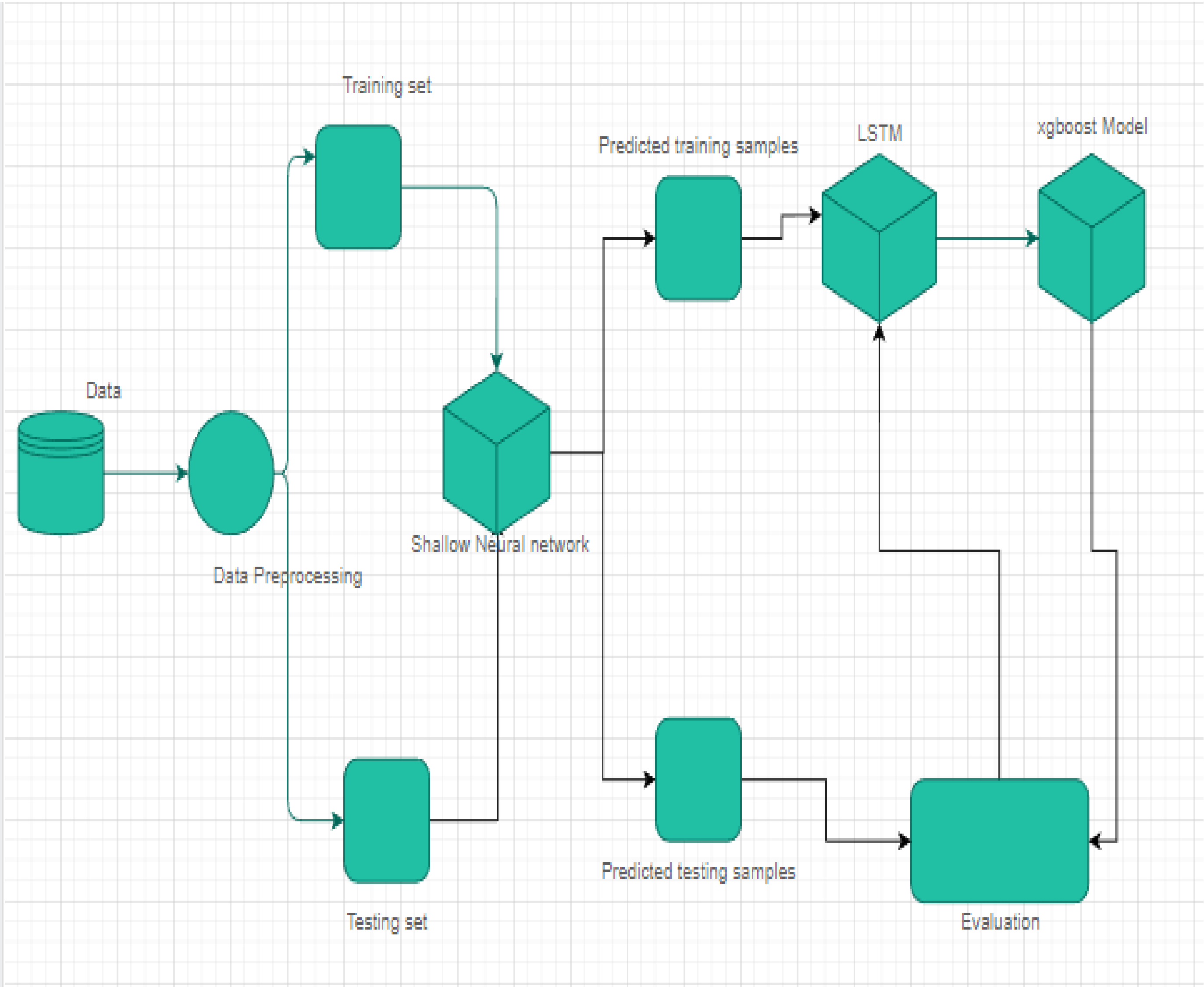

Fig 4.1 ANN-LSTM-ADA Model

4.2 MODELS IDENTIFIED:

Artificial Neural Networks:

Relevant pre-processing is done on the data and is split into the predictor and the target variables. Then the predictor variables are then fed into the ANN model with one hidden layer consisting of 50 units. The hidden layer had an activation function called relu instead of sigmoid and tanh function.

Sigmoid and tanh functions are not considered in this model because of the saturation issue. This issue arises when weights of hidden layers either activate or not i.e 1 or 0. It does not give a probabilistic value basis its decision on discrete values.

It uses a metric called mean squared error and optimizer adam to down scale the set of values.

Artificial neural networks are forecasting methods that are based on simple mathematical models of the brain. They allow complex nonlinear relationships between the response variable and its predictors.

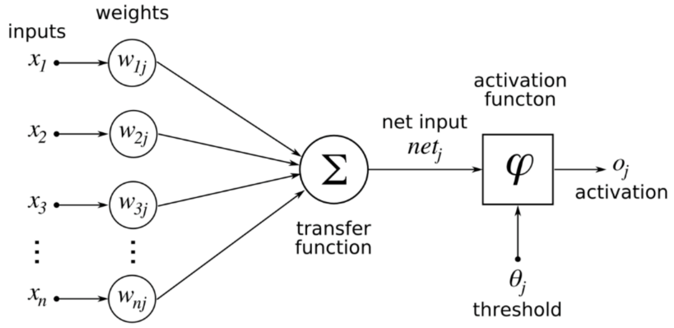

Fig 4.2 Shallow Network Model

Long Short Term Memory:

Deep learning model which is the LSTM is used to forecast the CPI index. This deep learning model is derived from its parent RNN. RNN is discarded due to the problem of vanishing gradient which brought about LSTM. LSTM is a neural Network. These neural networks are like black box. we feed in some inputs from one side and receive some outputs from the other side.

The output consisting of the down scaled values from the ANN model is fed into the LSTM algorithm with the timestamp of n = 60.

The LSTM model is built using three hidden layers each consisting of 50 units. The optimizer used is called adam optimizer and the metrics used was mean squared error. Adam optimizer is selected to achieve good time complexities.

LSTM Algorithms receives input with a time stamp of 60 from the Consumer price index dataset with the dimensions of the input being 760, 60, 1. The output dimensions are 760, 0. This input and output data is fitted in our model to forecast future values.

LSTM in our model runs on a epoch count of 25, 50, 75 and 100 with a batch size of 64. The epoch counts are chosen such that the equilibrium results can be achieved to reduce the processing power.

LSTM trains the data only once and does no back propagation which leads to good time complexities and saves a lot of processing power and hence being user friendly.

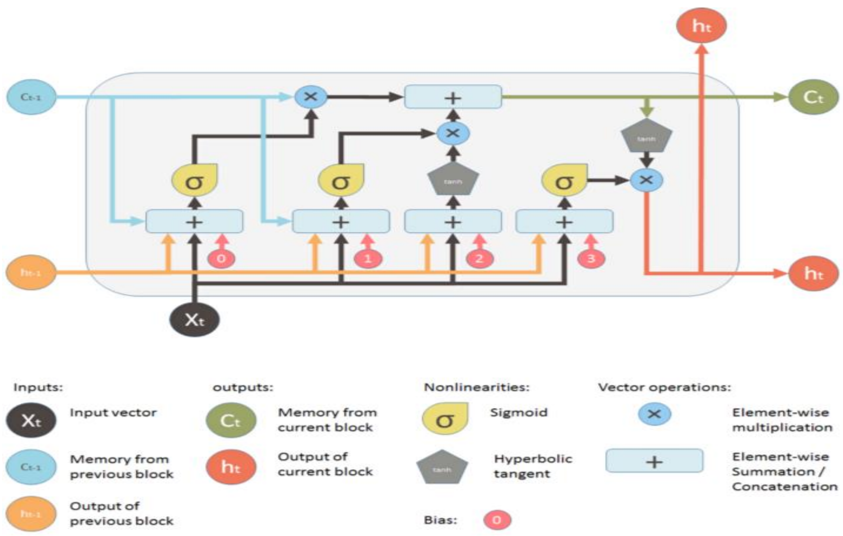

Fig 4.3 LSTM architecture

ADA Boost regressor:

After LSTM models are run and done processing, the output is fed into the ada boost algorithms as inputs in order to further increase the accuracy.

Ada Boost is a boosting algorithm consisting of ensembled number of tress which process individually and their resulting average is considered to be the output.

Hene implementing the boosting algorithm will enhance the accuracy of this hybrid model, without disturbing the time complexities and GPU power usage.

while the vanilla LSTMS use single output values, the boosting algorithm always takes the average from the set of values given to it, thus ensuring higher accuracy without compromising the efficiency.

LSTMS VS BiLSTM VS ANN-LSTM-ADAboost:

From the table it can be inferred that BI-LSTMS out performs the LSTMS and ANN-LSTM-ADAboost models but has higher time complexities and higher power usage.

Hybrid models turned out to give almost compelling results to Bi-LSTMS with shortened time complexities and minimal power usage achieving efficiency and accuracy.

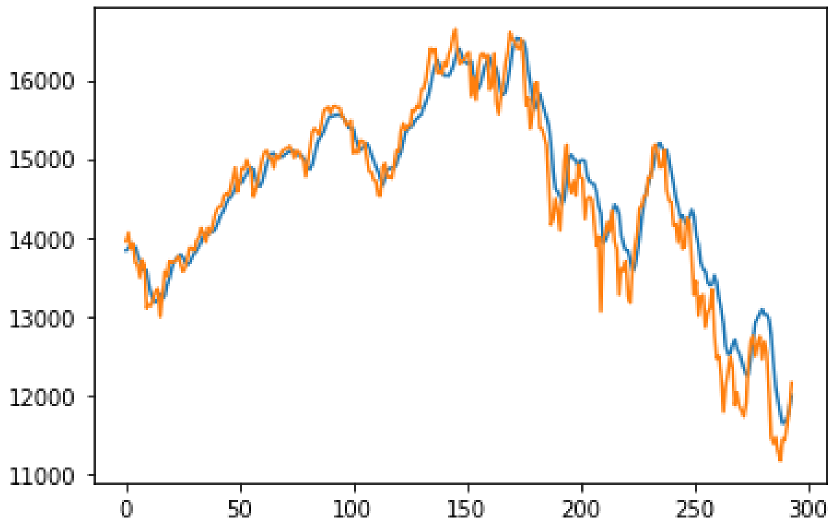

Fig 4.4 BILSTM model running on 50 epochs

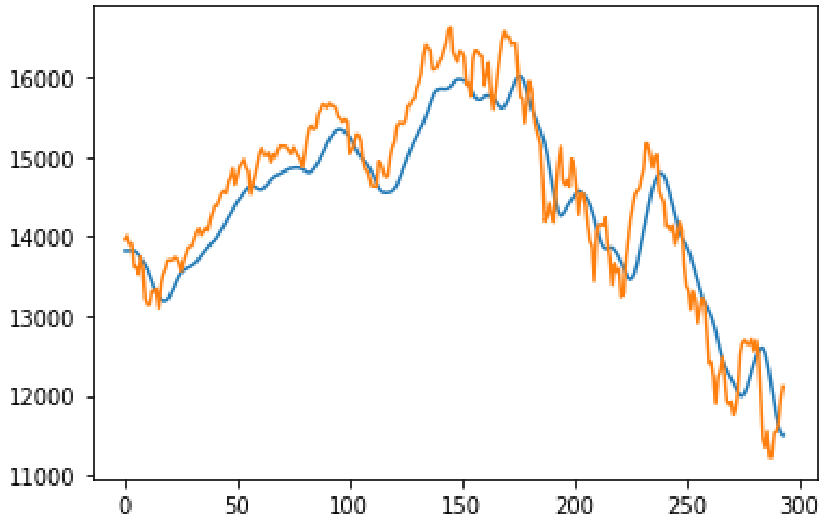

Fig 4.5 ANN-LSTM-ADA model running on 50 epochs

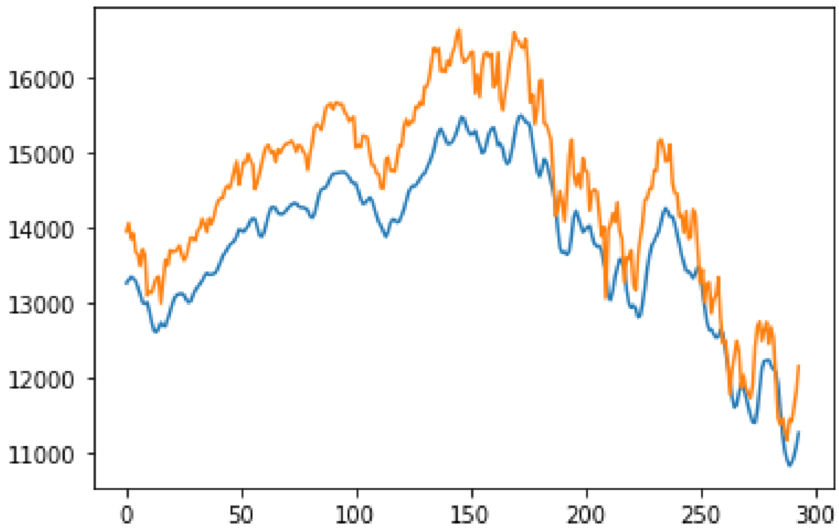

Fig 4.6 LSTM model running on 50 epochs

Model	No. of epochs	RMSE value	Time (min)
BILSTM	50	348	15.8
LSTM	50	783	8.3
ANN-LSTM-ADA	50	441	8

Fig 4.7 Comparison between LSTM, BILSTM and Hybrid model values

References:

[1] Manuel B. Aalbersa The Great Moderation, “the Great Excess and the global housing crisis”Department of Geography, KU Leuven/University of Leuven, Heverlee, Belgium Published online: 20 Jan 2015.

[2] Oren Barkana, Jonathan Benchimolb, Itamar Caspi ,Eliya Cohenc, Allon Hammer, Noam Koenigsteinc “Forecasting CPI Inflation Components with HierarchicalRecurrent Neural Networks” Ariel University, Bank of Israel, Tel-Aviv University, International Journal of Forecasting, forthcoming, published on 20 jun 2022.

[3] Md. Arif Istiake Sunny1, Mirza Mohd Shahriar Maswood1, Abdullah G. Alharbi2, Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model, Proceedings of NILES2020: 2nd Novel Intelligent and Leading Emerging Sciences Conference.

[4] Alex Sherstinsky, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network, 21 January 2020, journal homepage: www.elsevier.com/locate/physd

[5] Khaled A. Althelaya, El-Sayed M. El-Alfy, Salahadin Mohammed, Evaluation of Bidirectional LSTM for Short- and Long-Term Stock Market Prediction, International Conference on Information and Communication Systems (ICICS)

[6] Sepp Hochreiter, Jurgen Schmidhuber, LONG SHORT-TERM MEMORY, Neural Computation 1735-1780, 1997.

[7] Bibhuti Bhusan Ramakar Jha1, Anshuman Singh1, Deepak Kumar1, Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting, Accepted: 13 July 2019, Polish Academy of Sciences & Polish Academy of Sciences 2019.

[8] Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin The Performance of LSTM and BiLSTM in Forecasting Time Series, 2019 IEEE International Conference on Big Data (Big Data)

[9] Md. Kowshera, Anik Tahabilderb, Md. Zahidul Islam Sanjidc, Nusrat Jahan Prottashad Md. Shihab Uddine, Md Arman Hossainf, Md. Abdul Kader Jilanig LSTM-ANN & BiLSTM-ANN: Hybrid deep learning models for enhanced classification accuracy, 10th international young scientist conference on Computational Science.

[10] Shaolong Sun, Yunjie Wei1, Shouyang Wang, AdaBoost-LSTM Ensemble Learning for Financial Time Series Forecasting ICCS Camera Ready Version 2018.

[11] Shiqing Sang , Fangfang Qu, and Pengcheng Nie, Ensembles of Gradient Boosting Recurrent Neural Network for Time Series Data Prediction.

[12] Anusha Garlapati, Doredla Radha Krishna, Kavya Garlapati, Stock Price Prediction Using Facebook Prophet and Arima Models,2021 6th International Conference for Convergence in Technology (I2CT) Pune, India. Apr 02-04, 2021.

[13] DINH Thi Thu Huong, VU Van Truong, BUI Thu Lam Forecasting of Consumer Price Index using the ensemble learning model with multi-objective evolutionary algorithms:Preliminary results, 2015 International Conference on Advanced Technologies for Communications (ATC).

[14] Yanbing Lin, Hongyuan Luo , Deyun Wang , Haixiang Guo, and Kejun Zhu An Ensemble Model Based on Machine Learning Methods and Data Preprocessing for Short-TermElectric Load Forecasting Received: 14 July 2017; Accepted: 7 August 2017; Published: 11 August 2017.

[15] Fei Qian, Xianfu Chen, Stock Prediction Based on LSTM under Different Stability, 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics.