Student Gross Enrolment Ratio Forecasting: A Comparative Study Using Statistical Method and Machine Learning

Jamal Hussain, David Rosangliana*, and Vanlalruata

***Abstract*—Educational data mining has advanced substantially within the past decade. These mining strategies lay out a plan for increasing overall academic enrollment. An increase in student enrolment, in general, would enhance academic performance. Therefore, the student enrollment pattern demands great attention, as it is a vital performance indicator of academic sustainability. In this paper, student enrolment data is pre-processed to obtain the gross enrolment ratio (GER). GER analysis and forecasting were performed using the state of art models Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM). The purpose of this study is to analyze and compare student GER (time series data) using ARIMA (statistical methods) and LSTM (machine learning approach), forecast GER using a better method, and propose corrective measures for increasing student enrolment. The comparison results confirmed that LSTM out-performs ARIMA by an average of 0.1322% and 5.6% in both Root Mean Square Error (RMSE) and Accuracy. The predicted GER using LSTM for the academic year 2035 is 34.23% which is far lower than 50% which is targeted by Govt. of India. An in-depth analysis of student enrolment and GER in higher education in Mizoram was done, and corrective measures were proposed for enhancing GER.**

***Index Terms*—Machine learning, statistical method, GER prediction, forecasting model**

I. Introduction

Analysis and forecasting of student enrolment trends utilizing machine learning and statistical method play an important role in drawing up a plan to further improve higher education. This can help in coordinating institutional resources more efficiently and effectively. Any new course approval inclined to the rate of growth in GER can enhance the overall education policy. This state of art machine learning and statistical methodologies were explored and implemented in order to acquire a solid understanding of the educational forecasting framework. The data utilized in the study were acquired from the yearly reports of AISHE (https://aishe.gov.in/aishe/home) and Department of Economics and Statistics, Govt. of Mizoram (https://des.mizoram.gov.in/). Less access to higher education is a fundamental factor for the poor intake of higher education. Through New Education Policy (NEP) 2020, government of India wants to raise GER to 50% by 2035. However, during the last 10 years GER have only increased by an average of 10%. To achieve this 50% GER by 2035 a new strategic approach with data driven decisions is crucial.

Manuscript received October 27, 2022; revised November 18, 2022; accepted December 7, 2022.

Jamal Hussain, David Rosangliana, and Vanlalruata are with Department of Mathematics and Computer Science, Mizoram University, Aizawl, India.

*Correspondence: drosangliana@gmail.com (D.R.)

In India, children between the ages of 5 and 9 obtain basic education at a GER of about 99% (Class 1–Class V). The GER, however, considerably drops in the intermediate and higher secondary levels. The GER of Mizoram is at a mediocre level of 26.1% for higher education (AISHE Report 2019–2020). This level has raised concern on how to improve the GER considering internal and external factors.

Higher education institutions across the globe are increasingly interested in assessing their enrolment pattern so as to identify the components that could maximize their enrolment numbers. Student enrolment in higher education impact by food insecurity reports lack of time and inadequate money are the biggest barrier [1]. Enrolment projection is an important indicator for an academic decision making, education management and financial management [2]. This forecast depends on various factors such as the student's family wealth, parent's education level, fees, student assistance levels, and student's academic success. Interestingly, a study in [3] also indicates that when the family income increases the fees for education becomes less into consideration. Enrolment is also indirectly proportional to the unemployment rate. When the rate of unemployment increases, student enrolment decreases [4]. Prediction is a fundamental yet complicated component of analysing time series data. Enrolment data prediction [5] using different machine learning model for GER till academic year of 2020 shows the prediction result for 2019-2020 to be 26.94 which in actual is 26.1. However, this GER prediction implements only machine learning model, to extend the study domain of prediction and forecasting. This paper study focuses on comparing both ARIMA (statistical method) and LSTM (machine learning model). The main objective of this study can be summarized as comparing ARIMA and LSTM performance accuracy for forecasting GER data. Selecting more accurate model based on performance accuracy for predicting the known dataset, which is nothing but the pass enrolment dataset collected for each academic year. Analyze and forecasting future GER (2020 to 2035) using the selected model. Proposing corrective measures to increase GER of higher education.

The rest of the paper is organized as: Section II, Literature Review. Section III, presents the methodology used in this study. Section IV, consists of results and discussion. Lastly, Section V, draws conclusions from the study.

II. Literature Review

Although several studies [6] have been carried out using machine learning models for prediction and forecasting in academic data, this paper compares machine learning and statistical methods. This comparative study is essential to eliminate ambiguity regarding the accuracy performance

between the two. LSTM model from machine learning and ARIMA from statistical method were selected for performance comparison. Based on the pass enrolment prediction accuracy, either a machine learning or statistical model will be selected for future forecasting. This forecasted GER value plays a crucial role as it is the primary indicator of academic performance. Our study proposed corrective measures to increase GER in higher education based on this forecasted value. Recent related literature based on prediction are as follows:

Yağci and Çevik [7] reported research of variables impacting on student performance in Turkish student, where the investigations were done in Turkish high schools. It was established that anxiousness had a considerably detrimental influence on a performance of the students under study case.

Yang *et al*. [8] attempted forecasting of student performance from an online course. The statistical data used in this study is generated by analysing both the video viewing habits and their exams performance. By merging PCA with MLR, the accuracy of the MLR algorithm was enhanced from 71% to 81%.

Menaei-Bidgoli *et al.* [9] used Michigan state university data by applying data mining algorithm to gather new information about the enrolment of student in physics course. In the concluding the study recommended genetic algorithm (GA) upgraded the prediction model with an accuracy improvement of 10%.

Belachew and Gobena [10] at university of Wolkite, aims to predict student performance using machine learning models. The data collection includes GPA of all the final year students including their grades performances for all the previous semester starting from 1st year. A best prediction accuracy of 93.8% was achieved using Naïve Bayes approach.

Shanthini *et al*. [11] studied the use of decision tree (DT) classifiers with respect to algorithms, such as AdaBoost, Bagging and Grading to create distinct decision trees. In total, 401 samples from undergraduates were recorded. The suggested technique attained an accuracy of 97.5%.

Okubo *et al.* [12] studied neural network approach for students' performance prediction and draw a conclusion by stating that Recurrent Neural Network (RNN) is an excellent technique as compare to multiple regression for student grades performance prediction. Kyushu University data were used in this study.

Wang *et al.* [13] provided an aspect to the character of students in learning programming tasks. Two goals of student behaviours were investigated by implementing both the LSTM and RNN architecture. The studies were utilized to observe the students' activities over time, identify the at-risk youngsters, find knowledge gaps of the pupils, and offer teachers an early warning so they could provide further help.

Arindam and Joydeep [14] studied student performance using RNN. The data were gathered from Kaggle, and the model was evaluated with an ANN and DNN They rated the RNN having the greatest performance, with an accuracy of 85.5%.

Krauss *et al.* [15] have implemented multiple forecasting model including deep learning, support vector machine and simple neural network. These machine learning architectures were studied in details by comparing each of the output accuracy. They concluded that, deep learning is the most robust and accurate among the network they take into consideration.

Lee and Yoo [16] studied a RNN to analyse the performance of stock returns. The stock performance history was gathered as a data input. The variation in the inner layer of RNN in corresponding to the stock return were analysed in details. The results show that RNN prediction with respect to stock price depends on the parameters taken into consideration for real world use.

Sin and Muthu [17] analyzed big data using machine learning techniques to evaluate student performance prediction, course recommendation for the next semester and student behaviour analysis.

Lu *et al.* [18] made an early prediction on the performance of final year student at University of Northern Taiwan. Principal component regression is implemented by taking variable such as homework performance, quiz, online course viewing behaviour and tuition attendances after class.

Rechkoski *et al.* [19] predicted the student performance for the next academic year using collaborative filtering technique for the year between 2011 to 2016 at Macedonia institutions. This technique is based on probabilistic matrix factorization and Bayesian probabilistic models.

Bydžovská [20] conducted research from Masaryk University students' records using collaborative filtering technique to predict the performance of the student enrolled in first year before the exam. The pass academic records are used as the input datasets. This study shows that the finding is as efficient as using machine learning models.

Polyzou and Karypis [21] used Mninesota University historical data for a period of 12 and half year to predict student academic performance. The study used low range matrix factorization and dispersed linear technique and conclude a data with a specific course generate better accuracy for grade prediction.

Nuankaew, P *et al.* [22] developed a model to predict the success cluster rates of educational technologists in Thailand during the academic year 2015 to 2017. The dataset uses in this study consist of 98 students. The prediction model accuracy is claim to be 98.37 %.

Arjaria and Roy [23] used machine learning of tagging online learning materials automatically by identifying the subject. Two architectures known as KNN and Back propagation where implemented. The results show 84% and 93% accuracy. IEEE LOM 9.0 meta specification were used in the study.

III. Methodology

The methodology used in this study began with data pre-processing, in which year-by-year historical enrolment data were converted into their equivalent GER. This GER dataset was trained using ARIMA and LSTM. The trained model was tested against the known dataset using 5-fold validation. Accuracy and RMSE were calculated for each fold. In the model selection stage, prediction analyses are performed where LSTM performance is proven to be superior. Thereafter, LSTM is selected for forecasting GER data. This GER data is then analysed further by comparing it to the current enrolment trend and suggesting insight

improvements to boost GER.

IV. Proposed Model and Implementation

The overall proposed framework is illustrated as Fig. 1 below.

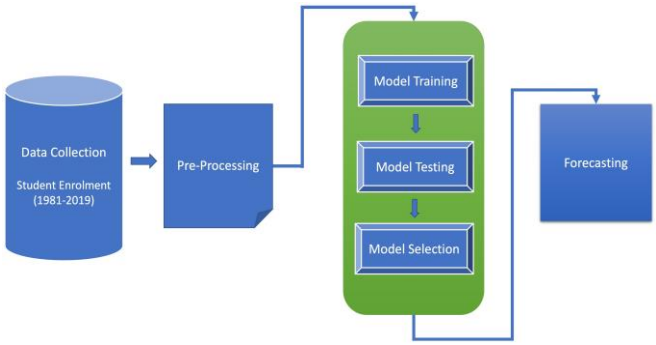

Fig. 1. Overview of the proposed framework.

The initial stage of the proposed model started with data collection, in which the education enrolment data was gathered and analysed. The pre-processing stage deals with data transformation, where year-wise enrolment data is converted to its corresponding GER. The ARIMA and LSTM were trained using the known datasets (GER) in the model training phase. The model testing phase uses the known GER data to test against the prediction accuracy by calculating the accuracy and RMSE. The model selection stage compares and selects the model based on the accuracy results obtained from the previous step to predict the unknown GER (2020-2050). The forecasting stage performs the actual GER forecast using the selected model.

A. Data Collection

The student enrolment data utilized in this study is acquired from two sources. The enrolment data during the years 1981 to 2010 is collected from Department of Economic and Statistic, Govt. of Mizoram and the year 2011-2019 is from AISHE issued by the Ministry of Education, Govt. of India. In total 39 years student enrolment data were analysed in this study. The datasets are divided into 5 folds, each consisting of 8 years student enrolments from 1st to 4th folds and 7 years student enrolment for the final 5th fold as in Table I. Training and testing were performed in each fold by taking 70% on each fold enrolment dataset and 30% for testing the models.

B. Data Pre-processing

Data pre-processing is a vital stage in statistical analysis and machine learning, since the arrangement of the data strongly affects the capacity of our model to learn. Intelligent algorithms will not only be enough to generate valuable insights from an inadequate data. Pre-processing stage not only assure the readiness of the data but might also increase the performance of models. In this paper, data transformation, is employed in the data pre-processing stage by converting the year wise student enrolment data to its corresponding GER, which is a major indicator for enrolment status. GER represents the number of students enrolled given as a percentage of the population between the age range of 18-23 years. Therefore, GER indicates a ratio of student enrolment based on the eligible population which makes the enrolment and population directly proportional to each other. Fig. 2 represents data before pre-processing which is the student enrolment data, while Fig. 3 and Fig. 4 are the transformed dataset which is GER.

Interestingly, while the number of female enrolments is lesser as compared with male enrolment for all the years taken into consideration (1981-2019) as in Fig. 2. But after pre-processing and converting to GER, we can find that for several years such as 1991 to 1996 and 1998 to 2004 the female enrolment is higher than male enrolment as in Fig. 3.

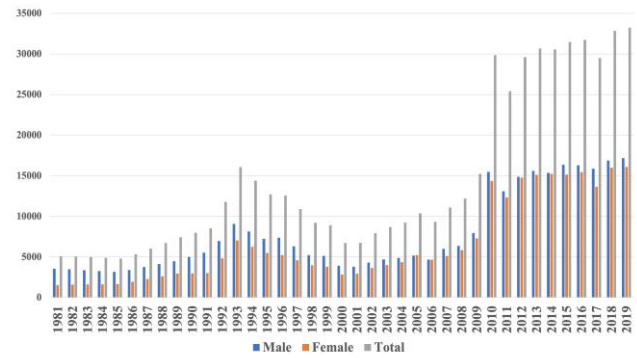

Fig. 2. Student enrolment 1981-2019.

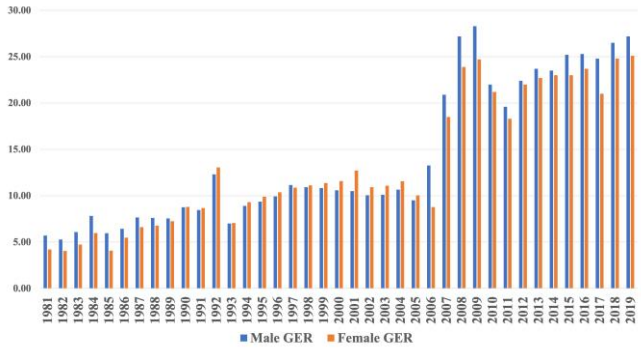

Fig. 3. Male and female GER 1981-2019.

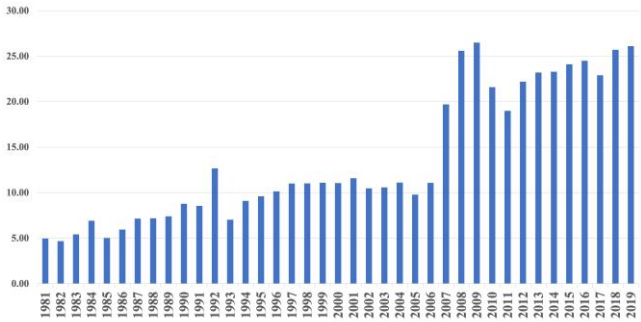

Fig. 4. Total GER 1981-2019.

C. Model Training and Testing

In our study we compare two state of art forecasting model known as ARIMA and LSTM. Both these models have gain favour in recent studies both in literature and application.

1) Training and testing

The GER generated by the pre-processing stage is an input to ARIMA, and LSTM for both training and testing. The input datasets were divided into both training and testing as 70% and 30% respectively. Training phase is

validated using 5-fold cross validation as this phase is the primary indication for model selection. The model selected in the stage will be used in the testing phase for forecasting GER for the year 2020 to 2035. This training and testing stage is critical as it is the primary determining factor for the how accurate our model can forecast the future GER. Learning weight, bias, epoch for LSTM model are the parameters which varies during the training stage. Likewise, the auto regression and moving average functions plays a critical role in ARIMA for determining the model accuracy.

2) Assessment metrics

The "loss" is a number that is generated by machine learning. Loss can be defined as a penalty for a poor guess or incorrect prediction. In other words, when the prediction is 100 % accurate then loss value will be 0. Therefore, to reduce the value of loss, two variables named weight and bias of the network are adjusted in every epoch. In addition to loss, researchers most often utilize the RMSE to evaluate the prediction accuracy. RMSE is generally used to calculate the variation between the projected value to the known actual value. The formula is as given in Eq. (1).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (1)$$

where N is the total number of observations which in our case is the number of years of student enrolment. $y_i$is the actual value of the GER for the known year; while, $\hat{y}_i$ is the predicted value of GER for the unknown year. To improve the prediction accuracy changes over each epoch is calculated using Eq. (2). Often variation in epoch can tune the model to increase the performance while overfitting is to be taken into consideration. This change is used by the weight and bias of the network during the training phase.

$$Change\% = \frac{Predicted\ Value - Origianal\ Value}{Original\ Value} \times 100 \quad (2)$$

3) Accuracy

The accuracy is measures by evaluate the capability of our prediction model by analysing the proportion of properly projected students' enrolment. Accuracy is determined from the confusion matrix. This representation can clearly depict the overall model performance during testing phase. Considering the following Fig. 5, the accuracy can be determined:

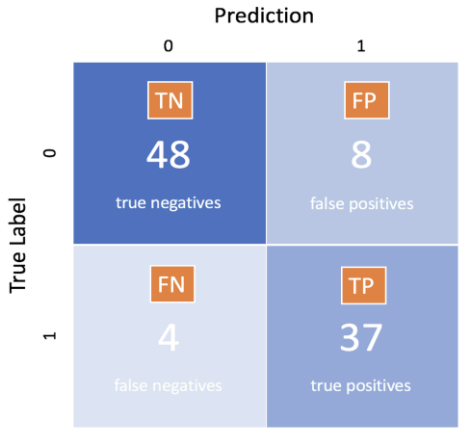

Fig. 5. Confusion matrix.

Confusion Matrix indicate the accuracy measurement for the model performance. True Negative (TN) and True Positive (TP) are correct prediction while False Positive (FP) and False Negative (FN) are incorrect prediction. The prediction accuracy increases as the value of TN and TP increases.

Hence, accuracy can be calculated using Eq. (3):

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+TP} \quad (3)$$

4) Predictive Mean Square Error (PMSE)

PMSE is applied to evaluate the quality of fit from our recommended models by measuring the variability in predicting accurate values. We seek to estimate how near our forecasted GER is to the real GER for the years prior to 2020. PMSE may be derived from Eq. (4)

$$\text{PMSE} = \frac{1}{M}\sum_{i=1}^{M}(o_i^a - o_i^p)^2 \quad (4)$$

where $O^a \in \{Decimal\ Number\}$ is the actual outcome and $O^p \in \{Decimal\ Number\}$ is the predicted outcome. The improved PMSE is derived from a nearly diagonal matrix of the generated confusion matrix. Unlike accuracy measurement, the smaller the value of PMSE, the better the model. PMSE equal to 0 means the prediction model is flawless.

5) Arima model

ARIMA is a statistical method used extensively to forecast and analyse the time series data. The concept of Autoregressive and Moving Average is combined by integrating it. The auto-correlation and partial auto-correlation are the underlying concept of ARIMA where AR is based on the concept of partial auto-correlation while MA is based on the concept of auto-correlation [24]. AR part of the ARIMA is based on a dependent connection between the observed data and the delay data [25]. The AR of p-order can be defined as the following Eq. (5).

$$X_t = \sum_{j=1}^{p} \emptyset_j X_{t-j} + \varepsilon_t \quad (5)$$

MA: Moving Average. A model which utilizes the dependency between an observation and a residual error of a moving average model applied to delayed data. The moving average process of the q-order or MA (q) is defined as:

$$X_t = \varepsilon_t - \sum_{j=1}^{p} \theta_j \varepsilon_{t-j} \quad (6)$$

Generally, the ARIMA model is represented as ARIMA (p,d,q), where p characterizes the order of the autoregressive process, d defines the order of the stationary data and q indicates the order of the moving average process. The ARIMA model can be mathematically expressed by:

$$(1-B)^d y_t = \frac{\theta(B)}{\emptyset(B)}\varepsilon_t \quad (7)$$

The terms AR and MA can be described as Eq. (8).

$$\emptyset(B) = 1 - \emptyset_1 B^1 - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$$

$$\theta(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_p B^p \quad (8)$$

Listing 1: ARIMA implementation

```
Inputs      : GER (Time series data)
Outputs     : Accuracy and RMSE for the forecasting data

1.  nsize ← length (GER) * 0.60
2.  training ← GER [0...nsize]
3.  testing ← GER [size... length (nsize)]
4.  historyData ← training
5.  predictionsData ← empty
6.  for each p in the range (lenght (historyData)) do
7.      model ← ARIMA (historyData, order=(4, 1, 0))
8.      modelFit ← model.fitting()
9.      forecasting ←  modelFit.forecast(predictionsData)
10.     predictions.append(forecasting)
13. end for
14. MSE = Find_MeanSquareError(historyData, forecasting)
15. RMSE = Find_sqrt (MSE)
16. ConfMatrix = PlotConfusionMatrix(historyData,
    forecasting)
17. Accuracy = ConfMatrix
18. Return RMSE
19. Return Accuracy
```

6) LSTM model

Learning big dataset with large variety of dependent variable were difficulty to be model accurately using the predecessor network known as RNN. Therefore, an extension version of RNN was created which can solve the vanishing gradient problems. Large datasets are often accompanied with larger dependency, to manage this dependency LSTM was introduced. LSTM are built with memory type structure which can remember the previous state and take decision based on the information from past dataset [26]. In other words, this extension version of RNN have the ability to learn sequential data by retaining information of all the relevant previous stages. Therefore, the memory introduce in LSTM makes it superior as compare with the predecessor RNN. In LSTM the choice of whether to retain or delete the information about the previous stages are controlled by a cell known as "gates". These gates perform by analysing the weight value assigned during the training phase of dataset. If the associated weighted value is lower than threshold the previous phase information is deleted, if not then the information is retained.

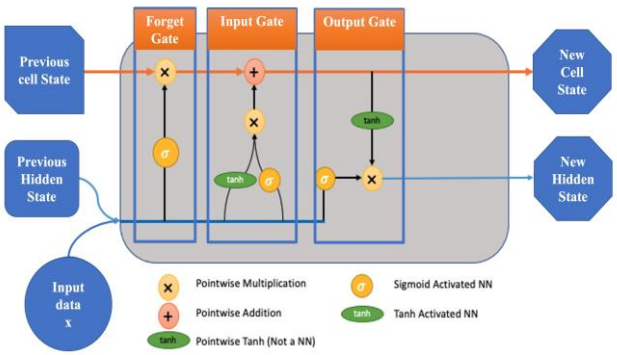

Fig. 6. LSTM Gates representation.

Generally, an LSTM architecture is build using three gates known as: Input, Forget, and Output gates. The forget gate is responsible preserving or deleting the information, while the input gate determine which information will be inserted into the LSTM memory. Finally, the output gate check in each epoch, weather the cell value makes any significant contribution to the model output. The overall gates can be depicted as Fig. 6.

The LSTM gate representation shows a graphical diagram of how the input data is processed. The unique feature of this model is how the previous input state contributes to the current input state in each gate. Input data is combined with the previous cell state with functions such as pointwise multiplication, pointwise addition, pointwise Tanh, sigmoid activation, and Tanh activation. Each of these functions is graphically represented in different colors. The line chart also gives the data flow along the LSTM path. This line chart is again classified into forget, input, and output gates. Details of these gates are as follows:

a) Forget gate: The first steps in LSTM known as the forget gates it decides which state of cell is important in considering both the hidden state and the input data. To make this gate functional both the new input x_t and previous input cell state h_{t-1} are trained using sigmoid function. The output of this always between the interval of [0 and 1]. When the value is closer to 0, the information is considered not relevant and closer to 1 means retaining more information. In other words when the output value f_t is 0 the network tends to forget and when the output is 1, the information is retained. This output of Forget gate can be represented as following Eq. (9):

$$f_t = \sigma(Wf_h[h_{t-1}], Wf_x[x_t], b_f) \quad (9)$$

where b_f is a constant with fixed value throughout the epoch and is also often known as bias value.

b) Input gate: This gate is responsible for adding new information to the cell state by considering the value of the given previous hidden cell and the new data which is input. The input gate is a combination of two function known as sigmoid Activation and $tanh$Activation. The sigmoid will make a decision on which value have to be change in cell. As sigmoid output a value of [0, −1], where 1 means allow to change or update all data whereas 0 means not allow to change or update. The "$tanh$" layer on the other hand output a value between [−1, 1], where it represents a candidate value from the cell state which will be added to a new memory cell of LSTM. The input gate output can be mathematically represented as following Eq. (10) and Eq. (11) as follows.

$$i_t = \sigma(Wi_h[h_{t-1}], Wi_x[x_t], b_i) \quad (10)$$

$$\tilde{c}_t = tanh(Wc_h[h_{t-1}], Wc_x[x_t], b_c) \quad (11)$$

Output of Eq. (10) i_t pass through a sigmoid function and tell us the old value is to be updated or not. While Eq. (11) output $\tilde{c}_t$ gives a list of vectors to the newly candidate list that will be inserted into a new memory cell. These two gate (forget and input) work in synchronized manner and update the new memory cell state accordingly.

c) Output gate: The output gate also known as the final gate is responsible for deciding which will be the new

hidden state. The input to this gate is a combination of previous hidden state, updated cell state and the newly input data.

A non-linear function known as **tanh** function is applied which generate a value between −1 and 1. This output is then multiplied to the output of the sigmoid layer. Eq. (12) and Eq. (13) give a mathematical representation as follows:

$$o_t = \sigma(I[h_{t-1}], W0_x[x_t], b_0) \quad (12)$$

$$h_t = o_t * \tanh(ct) \quad (13)$$

where output value LSTM is given by o_t , and the representation value between −1 and 1 is given by h_t.

The implementation of LSTM algorithm is described in Listing 2. Initially data were separated into 30% and 70% for both training and testing separately. A function call fit LSTM is used for training the algorithm with an input of training dataset, number of epoch and number of neurons. Line 8 and 9 set the model parameters by using RMS Prop as optimizer, loss is computed in MSE. Line 10 to 13 accomplishes training the actual LSTM model. In Line 12, the algorithm the following iteration is initiated by resetting the internal training stage example epoch. Line 14 is used to estimate the next step in (look ahead estimation one single) (look ahead estimation one single) Line 19 to 27 performs the prediction for the known dataset. Line 28 to 32 provides our model performance such as RMSE, Accuracy and PMSE.

Listing 2: LSTM implementation

```
Inputs : GER (Time series data)
Outputs    : RMSE, Accuracy and PMSE
1.    sizen ← length (Ger Dataset) * 0.70
2.    train ← GER [0 to size]
3.    test ← GER [size+1 to length (Ger Dataset)]
4.    set the random.seed(10)
5.    X ← training (70% of dataset)
6.    y ← training – X (30% of dataset)
7.    model = Use Sequential()
8.    model.layer .add(LSTM (neurons))
9.    model.layer.compile (loss=MSE, optimizer= RMSProp)
10.   for each in (X) do
11.         model.layer.fit(X, y)
12.         model.resume()
13.   end for
      return model
14.   ForcastGER ← model.predict(y)
      return ForcastGER
15.   MSE ← Mean Square Error(expected, ForcastGER)
16.  RMSE ← sqrt (MSE))
17.   Accuracy ← Confusion matrix
18.   Return RMSE and Accuracy.
```

D. Model Selection

The model selection compares both the statistics and machine learning models in terms of prediction accuracy. The compare model was developed during the training phase on the known time series data. This selection is performed on the testing phase, where prediction is performed against the truth data also known as the known datasets. Assessment Metrics such as RMSE, Accuracy and PMSE were utilized to examine the model.

In our model selection K-Fold cross validation technique is utilized. The total dataset is separated into 5 folds for the year between 1981-2019 as indicated in Table I below. The 2nd column indicates the starting year, 3rd column gives the ending year and the last column give the no of year with the fold sub-sets.

TABLE I: TRAINING AND TEST DATA DISTRIBUTION FOR 5 FOLDS

Fold No.	Start Year	End Year	No. of year
F1	1981	1988	8
F2	1989	1996	8
F3	1997	2004	8
F4	2005	2012	8
F5	2013	2019	7

The dataset for training and testing is scuffled so that the input and output are completely random. Thereafter the dataset is split into 5 folds as in Table I. The numbers of years covered in each fold are identical from F1 to F4, however due to the data distribution variation, F5 contains 7 years. The main purpose of scuffling and splitting the dataset is to achieve randomized dataset. This also enable to make the trained model more robust to predict unseen datasets. Selecting an accurate model after randomizing using K-fold during training generalized the model without overfitting the training phase. The overall process of model selection is depicted in Fig. 7 below.

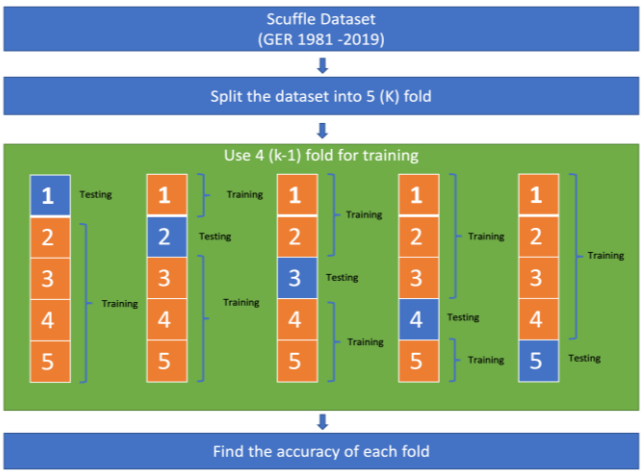

Fig 7. Model selection using 5-Fold.

Model selection using a 5-Fold graphical representation shows how much data were used for training and testing in each data fold. In Fold −1 out of [1,2,3,4,5] datasets, the first dataset (1) is used for training while the remaining dataset [2,3,4,5]. Likewise, the training dataset is randomized from FOLD 2 to FOLD 5 by taking a section of 2,3,4, and 5, respectively, as represented in Fig. 7. Data scuffling using K-FOLD is a common practice in training models as it helps generalize the model.

TABLE II: ARIMA PERFORMANCE

Fold No.	RMSE	Accuracy %
1	0.478	87
2	0.323	86
3	0.451	82
4	0.359	87
5	0.212	91
Average	0.3646	86.6

TABLE III: LSTM PERFORMANCE

Fold No.	RMSE	Accuracy %
1	0.321	91
2	0.218	83
3	0.193	94
4	0.231	97
5	0.199	96
Average	0.2324	92.2

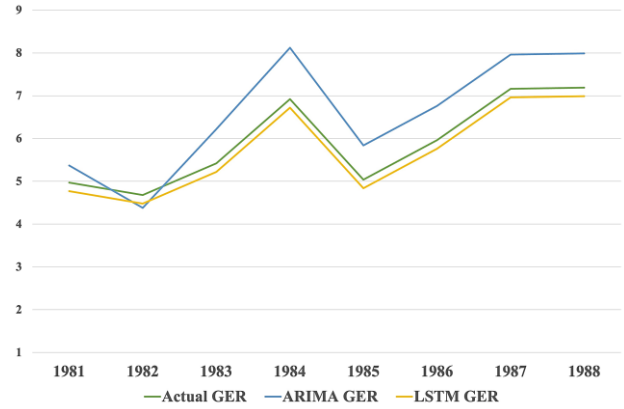

Fig. 8. Fold 1 ARIMA and LSTM prediction:
ARIMA and LSTM prediction against the actual GER for the enrolment year 1981-1988.

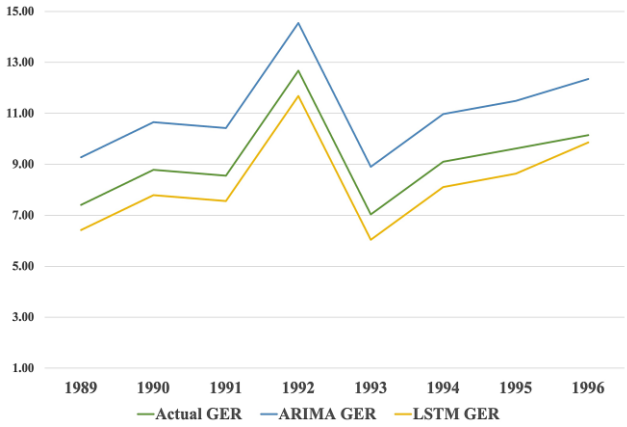

Fig. 9. Fold 2 ARIMA and LSTM prediction:
ARIMA and LSTM prediction against the actual GER for the enrolment year 1989 – 1996.

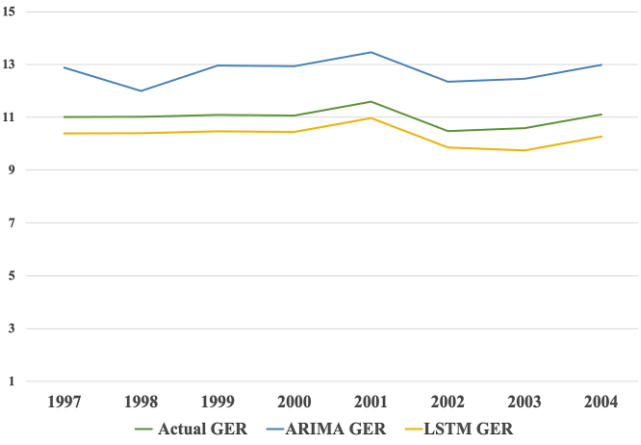

Fig. 10. Fold 3 ARIMA and LSTM prediction:
ARIMA and LSTM prediction against the actual GER for the enrolment year 1997 – 2004.

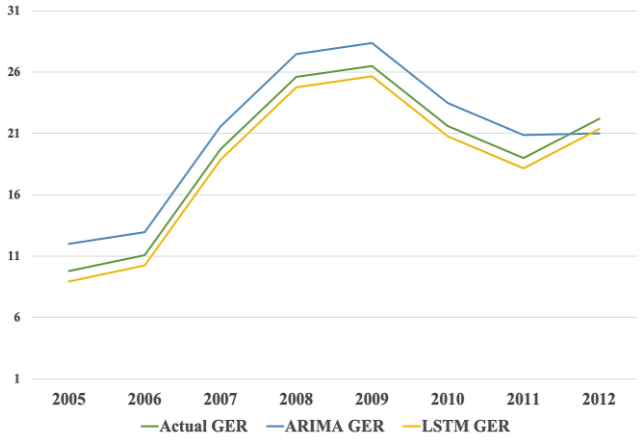

Fig. 11. Fold 4 ARIMA and LSTM prediction:
ARIMA and LSTM prediction against the actual GER for the enrolment year 2005 – 2012.

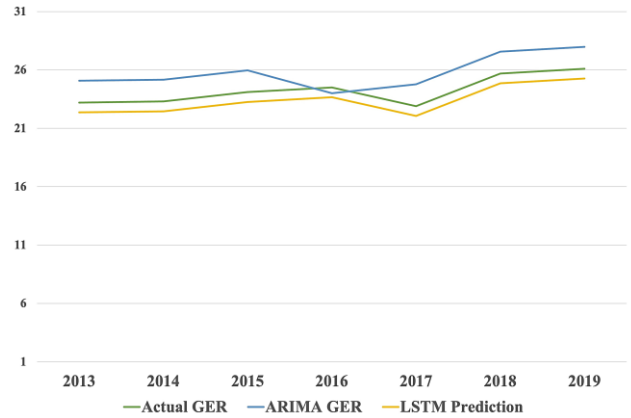

Fig. 12. Fold 5 ARIMA and LSTM prediction:
ARIMA and LSTM prediction against the actual GER for the enrolment year 2013 – 2019.

Performance of ARIMA and LSTM model are compared in Table II and III. Average performance metrics for both RMSE and Accuracy is higher in LSTM model by 0.1322 and 5.6 percentage respectively. Likewise, Fig. 8 to Fig. 12 also clearly represents the prediction value in LSTM is much closer to the actual known value. So, LSTM is selected for performing GER forecasting.

E. Forecasting

In this section, the actual forecasting of GER is performed using the selected LSTM model, as the previous section (model selection) has confirmed that LSTM is superior in prediction accuracy. A new LSTM model is built using 100% of the known GER data (1981 to 2019). In other words, all the available dataset is used for training the LSTM model. This approach is practically feasible as the comparison data in the model selection stage has suggested that LSTM-based models beat ARIMA-based models by a substantial margin. Therefore, the LSTM model is selected for performing the GER Forecasting. The Forecasted GER is given in Fig. 13.

Fig. 13 give the actual forecasted GER value using the selected LSTM model. This forecasted value gives a clear picture that by following the current trends and taking 38 years historical enrolment data, there will be an incremental increase in GER every year.

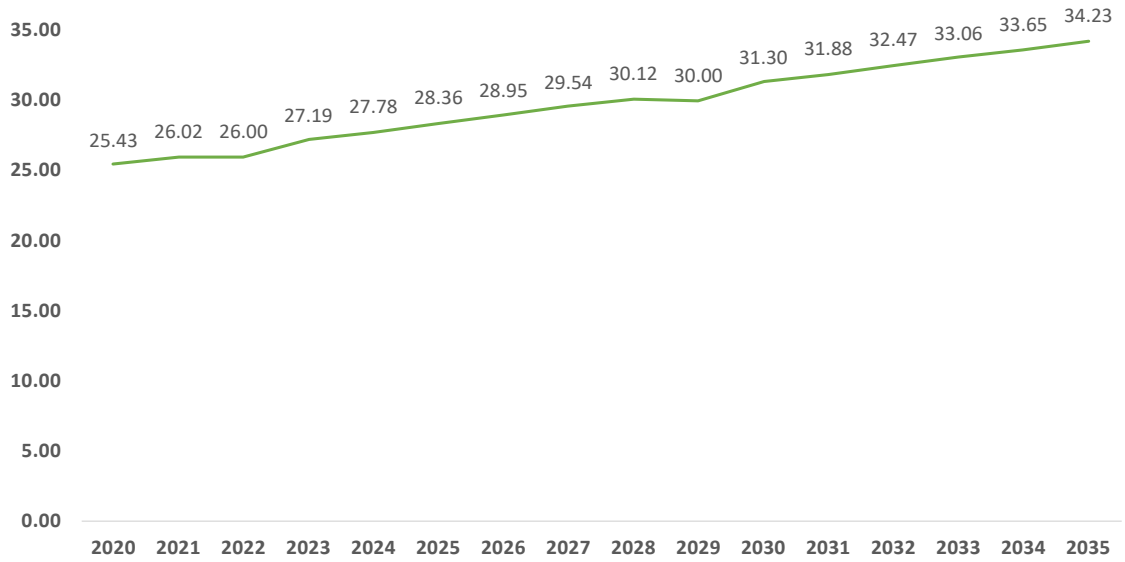

Fig. 13. GER forecasting using LSTM (2020 - 2035).

V. Results and Discussion

The study's findings indicate that LSTM is superior to ARIMA in forecasting time series data. A comparison of Tables II and III reveals that the LSTM-based algorithm predicts more accurately than ARIMA by 5.6% and that RMSE reduces in LSTM by 0.1322. Furthermore, GER prediction performance against the actual known datasets from Fig. 8 to Fig. 12 demonstrates that LSTM is significantly closer to the actual GER value, which favors LSTM as compared to ARIMA. These experimental findings direct the research to select the LSTM model for forecasting future GER data. The GER forecasted value (2020−2035) using the selected LSTM is plotted in Fig. 13. In this forecasting, GER increment is observed 14 times, and decrement is observed 2 times (2022 and 2029). This decrement in GER indicates that the number of student enrolment for these years is expected to be lower than the previous year. The average GER variation during these forecasting periods is 0.59. The forecasted GER mean and standard deviation are 31.73 and 12.22, respectively.

Improvement in accuracy and RMSE were not observed when increasing the number of epochs during the training phase. The exceptional performance observed through LSTM-based methods is due to the "iterative" optimization technique utilized in these approaches to obtain the best outcomes. The NEP 2020 aims to increase the GER to 50 % by 2035. The forecasting GER using our selected model (LSTM) for 2035 is 34.23 % which is lower than the NEP target by 15.8%. Considering the latest enrolment data in Table IV, the number of student enrolment from UG to PG drop significantly by 83.3 %. This is clear indication in shortage of seats and limited number of institutions for PG courses. The policymaker may consider increasing the number of seats and institution to achieved NEP 2020 target.

TABLE IV: Enrolment At Various Level in Mizoram (2019-2020)

District	Ph.D	M.Phil	PG	UG	PG Diploma	Diploma	Certificate	Integrated	Total
Aizawl	867	152	4221	19745	203	984	463	73	26708
Champhai				913		92			1005
Kolasib				544		87			631
Lawngtlai				1183		82			1265
Lunglei				1944		517			2461
Mamit				181		57			238
Saiha				359		64			423
Serchhip				427		78			505
Mizoram	867	152	4221	25296	203	1961	463	73	33236

VI. Conclusion

This research paper conducts a comparative analysis of two state-of-the art forecasting techniques: ARIMA and LSTM. In this process, time series data were pre-processed by transforming the yearly enrolment dataset into the equivalent GER. This GER data is then split into 5 folds, so as to train each fold separately. The results of each fold were examined and compared to the known dataset. The model selection phase chose the LSTM for future GER forecasting. This research finding provides policymakers with insight into future academic enrolment. Future research may be carried out by examining the status and reasons impacting student dropout at different levels of education since it has a direct and indirect influence on higher education (GER).

Conflict of Interest

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Prof. Jamal Hussain conducted the research, Vanlalruata analyse the data and David Rosangliana wrote the paper; all author has approved the final version.

REFERENCES

[1] K. M. Broton, K. E. Weaver, and M. Mai, "Hunger in higher education: Experiences and correlates of food insecurity among Wisconsin undergraduates from low-income families," *Social Sciences*, vol. 7, no. 10, p. 179, 2018.

[2] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, 032019, 2020, November.

[3] E. A. Hanushek and L. Woessmann. (2020). The economic impacts of learning losses. [Online]. Available: https://www.oecd.org/education/The-economic-impacts-of-coronavirus-covid-19-learning-losses.pdf

[4] D. Contini, F. Cugnata, and A. Scagni, "Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy," *Higher Education*, vol. 75, no. 5, pp. 785-808, 2018.

[5] J. Hussain, D. Rosangliana, and Vanlalruata. "Gross enrolment ratio prediction using artificial neural network," *International Journal of Recent Technology and Engineering*, vol. 8, no. 5, pp. 5006-5011, 2020

[6] H. Zhang, H. Nguyen, D. A. Vu, X. N. Bui, and B. Pradhan, "Forecasting monthly copper price: A comparative study of various machine learning-based methods," *Resources Policy*, vol. 73, p. 102189, 2021.

[7] A. Yağci and M. Çevik, "Prediction of academic achievements of vocational and technical high school (VTS) students in science courses through artificial neural networks (comparison of Turkey and Malaysia)," *Education and Information Technologies*, vol. 24, no. 5, pp. 2741-2761, 2019.

[8] S. J. Yang, O. H. Lu, A. Y. Huang, J. C. Huang, H. Ogata, and A. J. Lin, "Predicting students' academic performance using multiple linear regression and principal component analysis," *Journal of Information Processing*, vol. 26, pp. 170-176, 2018.

[9] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: An application of data mining methods with an educational web-based system," in *Proc. 33rd Annual Frontiers in Education*, vol. 1, pp. T2A-13, 2003, IEEE.

[10] E. B. Belachew and F. A. Gobena, "Student performance prediction model using machine learning approach: The case of Wolkite university," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 2, pp. 46-50, 2017.

[11] A. Shanthini, G. Vinodhini, and R. M. Chandrasekaran, "Predicting students' academic performance in the university using meta decision tree classifiers," *J. Comput. Sci.*, vol. 14, no. 5, pp. 654-662, 2018.

[12] F. Okubo, T. Yamashita, A. Shimada, and H. Ogata, "A neural network approach for students' performance prediction," in *Proc. the Seventh International Learning Analytics & Knowledge Conference*, 2017, March, pp. 598-599.

[13] L. Wang, L. Sy, L. Liu, and C. Piech, "Learning to represent student knowledge on programming exercises using deep learning," in *Proc. the 10th International Conference on Educational Data Mining*, 2017.

[14] M. Arindam and M. Joydeep, "An approach to predict a student's academic performance using recurrent neural network (RNN)," *International Journal of Computer Application*, vol. 181, no. 6, July 2018.

[15] C. Krauss, X. A. Do, and N. Huck, "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500," FAU Discussion Papers in Economics 03/2016, Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics, 2016 .

[16] S. I. Lee and S. J. Yoo, "Threshold-based portfolio: The role of the threshold and its applications," *The Journal of Supercomputing*, vol. 76, no. 10, pp. 8040-8057, 2020.

K. Sin and L. Muthu, "Application of big data in education DATA mining and learning analytics — A literature review," *ICTACT J. Soft Comput.*, vol. 5, pp. 1035–1049, 2015.

[17] O. H. Lu, A. Y. Huang, J. C. Huang, A. J. Lin, H. Ogata, and S. J. Yang, "Applying learning analytics for the early prediction of students' academic performance in blended learning," *Educ. Technol. Soc*., vol. 21, pp. 220–232, 2018.

[18] L. Rechkoski, V. V. Ajanovski, and M. Mihova, "Evaluation of grade prediction using model-based collaborative filtering methods," in *Proc. the 2018 IEEE Global Engineering Education Conference (EDUCON)*, Tenerife, Spain, 17–20 April 2018, pp. 1096–1103.

[19] H. Bydžovská, "Are collaborative filtering methods suitable for student performance prediction?" in *Proc. the Progress in Artificial Intelligence - 17th Portuguese Conference on Artificial Intelligence (EPIA)*, Coimbra, Portugal, 8–11 September 2015, pp. 425–430.

[20] A. Polyzou and G. Karypis, "Grade prediction with models specific to students and courses," *Int. J. Data Sci. Anal*., vol. 2, pp. 159–171, 2016.

[21] P. Nuankaew, T. Sittiwong, and W. S. Nuankaew, "Characterization clustering of educational technologists achievement in higher education using machine learning analysis," *International Journal of Information and Education Technology*, vol. 12, no. 9, 2022.

[22] S. K. Arjaria and D. Roy, "Learning content management using machine learning," *International Journal of Information and Education Technology*, vol. 2, no. 5, p. 472, 2012.

[23] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparative analysis of forecasting financial time series using arima, lstm, and bilstm," arXiv preprint arXiv:1911.09512, 2019.

[24] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," in *Proc. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, December, pp. 1394-1401, IEEE.

[25] H. Bousnguar, L. Najdi, and A. Battou, "Forecasting approaches in a higher education setting," *Education and Information Technologies*, vol. 27, no. 2, pp. 1993-2011, 2022.

Jamal Hussain is a professor in the Dept. of Mathematics and Computer Science, Mizoram University. His specialization is mathematical modelling and computer simulation of agricultural, biological, ecological and environmental systems. artificial neural network and intrusion detection systems.

David Rosangliana is a research scholar in the Department of Mathematics and Computer Science, Mizoram University. He is currently working in Govt. Zirtiri Residential Science College, Aizawl and have a teaching experience of more than 14 years in Under Graduate level.

Vanlalruata is a research scholar in the Department of Mathematics and Computer Science, Mizoram University. His area of interest is computer vision and machine learning.