Predicting Students' Performance Using an Attention-Based Long Short-Term Memory with Conditional Variation Autoencoders and Ensemble Model

Diana Frimpomaa, Abdul Salaam Gaddafi

Department of Computer Science, KNUST, Kumasi, Ghana

DOI: https://doi.org/10.51244/IJRSI.2025.120700007

Received: 19 June 2025; Accepted: 23 June 2025; Published: 28 July 2025

ABSTRACT

The study aims to develop an Attention-Based LSTM (Long Short-Term Memory) model with Conditional Variation Autoencoders (CVAE), and LIME (Local Interpretable Model-Agnostic Explanations), to accurately predict students' performance. The study employed data Processing and Augmentation using Conditional Variational Autoencoders (CVAE), an Attention-Based Long Short-Term Memory (LSTM) framework, and a final ensemble approach for refined predictions classifying students into three categories. The study utilized a primary dataset consisting of 500 students' records, the study establishes a performance benchmark. The study revealed that students who frequently raised their hands during lessons were identified by the model as more attentive and participatory, traits that strongly correlated with higher academic performance. It was evident that the developed LSTM with CVAE, and Ensemble model accurately predicts students' performance. The study concluded the Attention-Based LSTM outperformed the Deep Neural Network, K-Nearest Neighbor, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, and Artificial Neural Network which uses the same dataset in predicting student performance. The Attention-Based LSTM consistently outperforms these models, showcasing superior metrics in accuracy, precision, recall, and loss with an accuracy of 89.6%. The study recommended that the Attention-Based LSTM model's performance highlights how crucial it is for the educational industry to make use of cutting-edge analytical techniques. Also, when developing curricula, policymakers need to take the research on student involvement into account, including components that encourage active participation can result in better academic performance.

Keywords: Educational Data Analysis, Learning Analytics, Data Mining, Attention-Based LSTM, Conditional Variational Autoencoders

INTRODUCTION

In the digital era, educational institutions worldwide are generating vast and varied amounts of student data through platforms such as Learning Management Systems (LMS), examination databases, library usage logs, and records of student engagement [1]. This surge in both the volume and complexity of educational data has opened up new avenues for understanding and improving learning processes. Leveraging these datasets through Learning Analytics (LA), Data Mining (DM), and Machine Learning (ML) techniques has become a prominent strategy to predict student performance and identify at-risk learners that often exceed human capacity when handled manually. Accurate prediction of academic performance is increasingly critical, as student success in post-secondary education is not only a measure of institutional effectiveness but also a strong determinant of future employability and career development [2]. However, the expanding size and diversity of educational data have made the task of predicting student outcomes more challenging than ever [3].

The complex and multifactorial nature of academic success has made estimating students' performance based solely on observable traits or isolated data points insufficient. Educational Data Mining (EDM) has emerged as a powerful tool to uncover hidden patterns and relationships within student data, offering the potential to make more nuanced and reliable predictions. Among the earliest and most impactful applications of EDM is student

performance prediction, which enables timely interventions and informed decision-making to support learners' academic journeys. As Tsai and Gasevic point out, despite progress in performance prediction technologies, knowledge gaps remain, particularly regarding the optimal fusion of features and methods to maximize accuracy [4]. Existing literature tends to address student performance prediction in general terms, without adequately emphasizing the value of hybridized machine learning approaches [2] [5].

To better reflect real-world academic variance, this study employs a multi-class classification approach, segmenting student performance into three categories: Low, Middle, and High. While traditional predictive models such as decision trees, logistic regression, and support vector machines have shown utility, they often fall short in capturing complex temporal dependencies, hidden features, and latent structures inherent in educational data. Recognizing these limitations, this study proposes a hybrid deep learning framework that integrates an Attention-based Long Short-Term Memory (LSTM) network, Conditional Variational Autoencoders (CVAE), and an Ensemble Learning model. This combination aims to enhance predictive accuracy while also providing interpretability and robustness in modeling student learning trajectories

LITERATURE REVIEW

Educational Data Mining

In higher education settings, educational data mining (EDM) is used to extract meaningful data from various datasets to predict student achievement and improve performance. Mengash focuses on using data mining approaches to forecast academic success in college admissions [6]. The study analyzes data from computer science students in a Saudi governmental institution and predicts their first-year CGPAs using different data mining methodologies. The study emphasizes the importance of reliable admissions criteria and how pre-admission criteria can predict early university success. Kumar et al. explores the use of data mining to forecast student performance and identify crucial traits for prediction [7]. EDM research utilizes various data sources, such as online forums, databases, and standardized tests [7]. Additionally, Costa et al. compare the efficiency of mining educational data to predict students at risk of failing programming courses [8]. The results show the SVM algorithm as the most accurate predictor, suggesting that data mining can effectively identify students likely to struggle early on and prevent high failure rates.

Learning analytics

In the field of education, teachers and school administrators can access knowledge that can enhance teaching and learning through the use of various methods such as machine learning, human-computer interactions, data mining, natural language processing, and visualization [9]. Learning analytics, a powerful application of deep learning involves analyzing large and complex sets of educational data, including student achievement data, learning management system data, and social media data, using advanced algorithms to identify patterns, trends, and connections that can inform educational decisions. Learning analytics focuses on practice, while educational data mining focuses on approach. The use of learning analytics is explored in predicting the performance of students in Massive Open Online Courses (MOOCs) [9]. The study analyzes data from two separate MOOCs, including demographic information, prior knowledge and expertise of students in the subject area, their engagement with course content and activities, and their performance on evaluations. The findings suggest that several variables play a significant role in the effectiveness of MOOC performance prediction models. These variables include the level of student engagement with course content and activities, their prior knowledge and expertise in the subject, and their performance on early evaluations. The study also highlights that different predictive models may work better for different types of students or courses.

Namoun and Alshanqiti explored the use of various learning analytics and educational data mining models for predicting student performance [10]. Namoun and Alshanqiti found that decision trees, neural networks, logistic regression, and clustering are commonly used techniques [10]. Factors such as data quality, feature selection, and sample size can affect the accuracy of these models. To ensure equitable learning outcomes and bridge the digital divide, it is important to develop widely applicable models. Despite the challenges, the benefits of using educational data mining and learning analytics make the investment worthwhile.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

Concept of Deep Learning in Education

Researchers began using artificial neural networks (ANNs) in the early 1990s to simulate student learning and predict educational outcomes. An early study by Dedecker used an ANN to predict student grades in a beginner's psychology class [11]. The model achieved an accuracy of approximately 80% by training on demographic and pre-course performance data. Since then, deep learning has been widely applied in various educational applications, including personalized learning, mining educational data, and intelligent tutoring systems. The advent of deep learning techniques in the mid-2000s, capable of processing large and complex datasets like images and audio, marked a significant milestone in the field [11].. With the rise in popularity of massively open online courses (MOOCs) and other online learning platforms, there has been an explosion of data on student performance and learning habits. Deep learning models have been employed to analyze this data and develop prediction models for student success. For example, Cheng et al. (2016) utilized deep learning to forecast the rate of completion of MOOCs based on students' interactions with course material. The use of artificial neural networks, convolutional neural network models, recurrent neural networks, deep belief networks, and autoencoders are a few popular deep learning methods.

Artificial Neural Networks (ANN)

An essential part of deep learning is artificial neural networks (ANNs). ANNs are algorithms that draw inspiration from the composition and operation of organic brain neurons [12]. ANNs are made up of layers of interconnected nodes or neurons. Each neuron gets information from other neurons in the layer underneath it, processes the information, and then generates a result that is sent to the layer beneath it. The network's ultimate output is represented by the result of the final layer. In the course of training, ANNs modify the weights of the relationships between neurons. Backpropagation is a technique that involves incremental weight adjustments to reduce the discrepancy between the network's expected and actual outputs. Numerous tasks, like classification, regression, and grouping, can be carried out using ANNs. In the context of education, ANNs have been used for a range of applications, such as predicting students' performance, recommending personalized learning resources, and identifying at-risk students [13] [6]. ANNs have shown promising results in predicting students' performance, and their ability to handle large and complex datasets makes them suitable for analyzing educational data [13].

Convolutional Neural Networks (CNN)

For applications requiring image identification, convolutional neural networks (CNNs), a form of deep learning neural network, are highly effective [12]. They were first introduced in the 1980s and have since been extensively developed and improved upon. Convolutional, pooling, and fully linked layers are some of the layers that make up CNNs. The convolutional layer applies a collection of learnable filters to the image to extract features from the input image. The output is then down-sampled by the pooling layer, decreasing the input's dimensionality and making it easier to analyze. The fully linked layer categorizes the incoming image using the output of the preceding layers. The capacity of CNNs to learn spatial feature hierarchies is one of their main advantages [14]. They can detect basic features like edges and lines in the network's lower layers before using those features to discern more intricate patterns and forms in the network's higher layers. This system of hierarchy-to-feature learning makes CNNs particularly well-suited for image recognition tasks.

Deep Belief Networks

Several layers of restricted Boltzmann machines (RBMs) are used to create deep belief networks (DBNs), a specific type of deep neural network. DBNs employ an unsupervised learning algorithm that enables the model to learn a probability distribution of the input data. They have diverse applications, such as computer vision [15], speech recognition, and natural language processing [16]. DBNs are trained using a technique called greedy layer-wise training, where each layer is trained as an RBM. The hierarchical representations learned by DBNs allow them to recognize intricate patterns and correlations, making them well-suited for tasks like speech and image recognition. In education, DBNs have been utilized to predict student performance and provide personalized instruction recommendations [17]. However, training DBNs can be challenging due to their complex structure and data requirements. Interpreting the learned representations and understanding how the model generates predictions can also be difficult [16].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

Autoencoders

Autoencoders are a type of neural network that can compress data into a lower-dimensional representation and then restore it to its original dimensions [18]. This process is done by the encoder and decoder, which handle the compression and decompression respectively. Autoencoders have various applications, such as image compression, feature extraction, and anomaly detection [19]. In education, autoencoders have been used for analyzing and classifying student writing and speech patterns [20]. They can also extract features from educational data like student performance data and text data [21]. Autoencoders have been helpful in curriculum sequencing, determining the best sequence of instructional materials for individual students based on their learning needs [1]. One significant advantage of autoencoders in education is the ability to identify patterns in complex datasets and gain insights into students' learning processes [22]. By analyzing and extracting meaningful features from educational data, autoencoders can help educators personalize learning experiences and interventions to support academic success [23]. Deep learning has made significant contributions to the education sector. Intelligent Tutoring Systems (ITS) use deep learning techniques to provide individualized training to students [24]. These systems adjust to the student's learning style, rate of learning, and knowledge level, simulating the function of a human tutor

MATERIAL AND METHODS

Stage 1: Data Processing and Augmentation Using CVAE

The dataset was prepared, features were engineered, and CVAE was used to enrich the data in this stage to establish the groundwork for the proposed model.

Data Collection:

Utilizing data from an educational learning management system, Kalboard 360 (https://www.kaggle.com), the proposed Attention-Based LSTM model incorporates 16 comprehensive features from a dataset containing 500 student records. In addition to numerical characteristics like the number of times a student raised their hand, visited educational resources, viewed announcements, and participated in discussion groups, these characteristics also include nominal characteristics like gender, nationality, place of birth, stage, grade, section ID, topic, parent responsibility, and semester. Along with student absence days, the dataset also contains parental involvement factors, including parent happiness and parent answering. A summary of the dataset is found in Table 1.

Table 1: Summary of students' dataset

Name	Data Type	Distinct Values
Gender	Nominal	2
Nationality	Nominal	14
Place of Birth	Nominal	14
Stages	Nominal	3
Grades	Nominal	12
Section ID	Nominal	3
Topic	Nominal	12
Parent Responsible	Nominal	2
Semester	Nominal	2
Raised hand	Numeric	0-100
Visited Resource	Numeric	0-100
Viewing Announcement	Numeric	0-100
Discussion Group	Numeric	0-100
Parent Answering	Nominal	2
Parent Satisfaction	Nominal	2
Student Absent day	Nominal	2

The data gathered was carefully examined to find and correct any errors, inconsistencies, missing numbers, or outliers. The process of cleaning data entails several steps, including data validation, the elimination of duplicate records, the imputed filling in of missing information, and the detection and treatment of outliers. Due to 20 missing variables in the dataset, there were 480 clean records, comprising 305 men and 175 women after data cleaning.

Data Transformation:

To make several nominal attributes appropriate for model training, data transformation was conducted on the attributes. Binary data, encoded as '0' and '1', were created using nominal qualities such as a semester, gender, relation, parent answering a survey, parent-school satisfaction, and student absence days. For computational compatibility, additional nominal attributes such as place of birth, nationality, stage, topic, section, and section were translated into numerical data types. The prediction assignment in this study involved grouping students into three unique performance categories: Low, Middle, and High. Based on the interval values presented in Table 2, these classes are established. The classifications of "Low," "Middle," and "High" for students range from 0-69, 70-89, and 90-100 respectively. To speed up model training, the trials were run on a high-performance computing platform with GPUs. Scikit-learn was used for baseline models and performance assessment, and the PyTorch deep learning framework was used for model construction.

Table 2: Dataset Classes

Interval-Values	Class
0-69	Low
70-89	Middle
90-100	High

Data Augmentation Using Conditional Variational Autoencoders (CVAE):

Conditional modeling is added by CVAEs, a variation of Variational Autoencoders (VAEs). By doing this, more data were added to the dataset while maintaining conditional dependencies. As demonstrated in Figure 1, CVAEs address situations where you want to produce or encode data conditionally dependent on extra information.

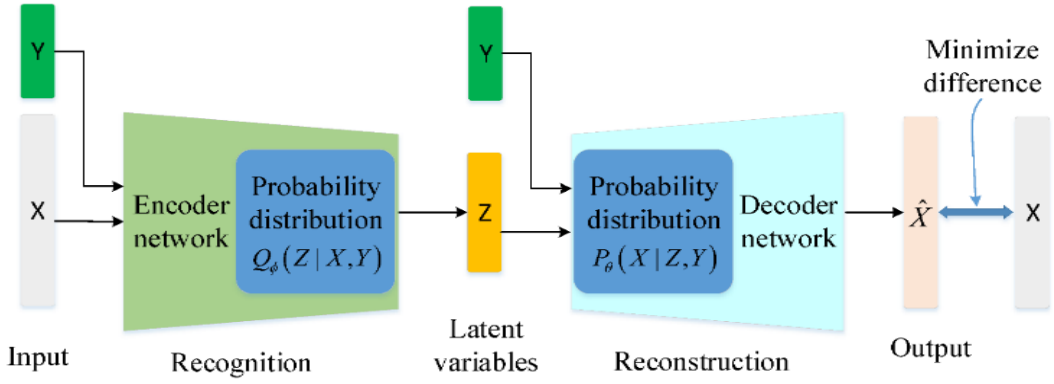

Fig. 1: Conditional Variational Autoencoders (CVAE)

Maximizing the Conditional Evidence Lower Bound (CELBO), a conditional form of the ELBO (Evidence Lower Bound), is the goal of training a CVAE, as illustrated in Equation 1:

$$\log P(X|C) \geq \text{CELBO} = \mathbb{E}_{Q(Z|X,Y)}\left[\log \frac{P(X,Z|C)}{Q(Z|X,C)}\right] \ldots\ldots\ldots (1)$$

where $P(X \mid C)$ represents the conditional probability of generating data X given the information C. $P(X, Z \mid C)$ represents the joint distribution of data X and latent variable Z conditioned on C.

$Q(Z \mid X, C)$ is the conditional encoder that approximates the posterior $P(Z \mid X, C)$.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

To increase the CELBO, the conditional encoder and decoder networks' parameters were optimized during training. This entails regularizing the conditional encoder to be near the conditional prior and minimizing the reconstruction error (negative log-likelihood).

Stage 2: Predictive Model using Attention-Based LSTM

A strong recurrent neural network RNN with a focus on modeling sequential data is the LSTM. It is renowned for having an excellent capacity for preserving data through lengthy sequences and capturing long-term dependencies. It is the perfect instrument for predicting pupils' academic performance based on their previous academic performance because of its special features. To store and retrieve pertinent data while processing input sequences as shown in Figure 2, the LSTM design contains memory cells and gates that regulate information flow. It includes the following essential elements:

Cell State (C_t): The horizontal line that passes through the top of the LSTM cell represents the cell state. It functions as a conveyor belt, transporting data from earlier time steps. The LSTM's long-term memory can be considered to be the cell state.

Hidden State (h_t): The output of the LSTM cell for the current time step t constitutes the hidden state. It acts as the short-term memory and holds data pertinent to the input being received at the moment and previous hidden states.

Three Gates:

Forget Gate (f_t): The information from the cell state that should be ignored or forgotten is decided by this gate. It outputs a value between 0 and 1 for each component of the cell state by taking input from the previous hidden state (h_(t-1)) and the current input (x_t). 0 indicates "completely forget," whereas 1 means "completely remember."

Input Gate (i_t): What additional information should be added to the cell state is decided by the input gate. Like the forget gate, it accepts input from x_t and h_t. A candidate cell state (C_t^i) is generated based on the present input.

Output Gate (o_t): The output gate determines the new hidden state h_t and also what part of the cell state to reveal as the output. It takes input from x_t and h_{t-1} and combines them with the candidate cell state (C_t^i) to produce the updated hidden state h_t.

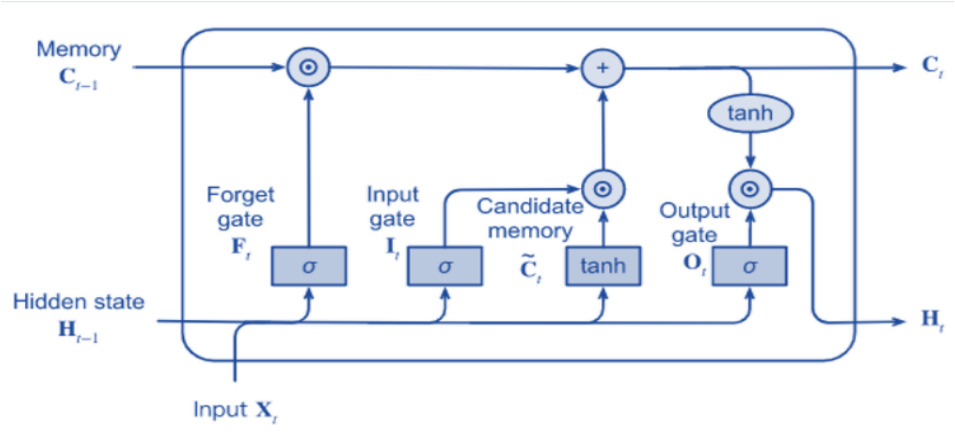

Fig 2: Long Short-Term Memory (LSTM) Architecture

Transformer Attention Mechanism:

The Transformer's attention mechanism uses self-attention layers to determine attention weights for each feature by taking into account how important that feature is to other aspects in the sequence. This gives the model the ability to evaluate the importance of different attributes while taking into account their links and dependencies.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

Three computations are involved in the self-attention process: queries, keys, and values. Queries are features that other features need to take care of. Values stand in for the data about the features being attended to, whereas keys stand in for the features themselves. The similarity between queries and keys is used to calculate the attention weights, which are then applied to the data to produce the attended representation.

Due to its adaptability for handling big values in the dot-product computation, Scaled Dot-Product Attention is used in this study to provide more stable and dependable attention weights to identify the features affecting student performance.

The mathematical formulation for Scaled Dot-Product Attention is shown in Equation 2:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QKT}{\sqrt{dk}}\right)V \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where Q represents the query matrix, with each row representing a query vector, K represents the keys matrix, with each row representing a key vector, V represents the values matrix, with each row representing a value vector, $\sqrt{d_k}$ is the scaling factor to mitigate the impact of large values in the dot-product computation. The dot-product between Q and ensuresK^Tmeasures the similarity between queries and keys. The attention weights are obtained using the softmax function, which ensures that they add up to 1 and reflect the relative significance of various features.

Multi-Head Attention was used at this point as an extension of the conventional attention process. The model can capture many facets of context and information by using multi-head attention, which enables it to concentrate on diverse portions of the input stream simultaneously. Essentially, it divides the attention mechanism into several heads, each with its own set of teachable parameters. To extract various types of information from the incoming data, these heads operate in tandem. The output of the final attention is created by linearly combining the findings from these heads, as seen in Figure 3. Equation 3 below gives the basic mathematical expression for multi-head attention:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head\_1},\text{head\_2},\ldots,\text{head\_h}) \times W\_o \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where head_i: The i-th attention head, and W_o: Weight matrix for the output

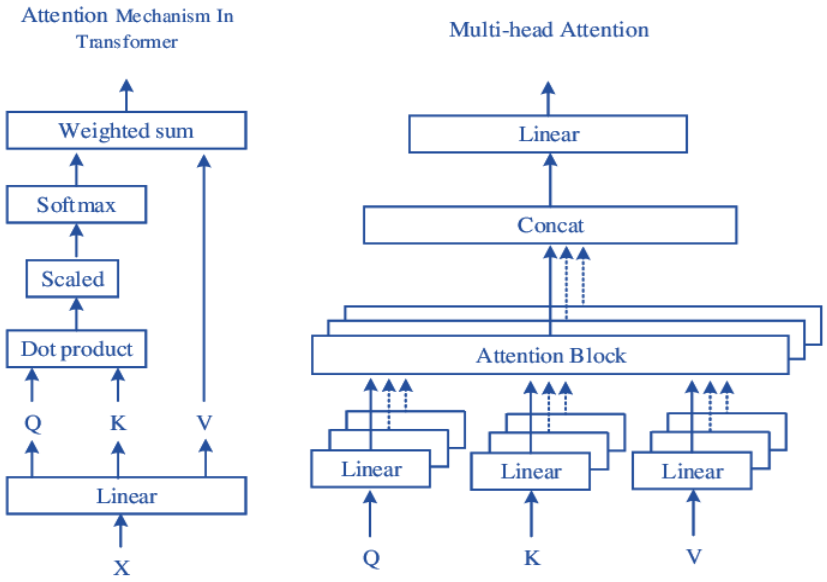

Fig 3: Multi-Head Attention Architecture

Stage 3: Ensemble of Models for Final Prediction

During this phase, ensemble modeling techniques which include variations of the Attention-Based LSTM were adopted, each with unique initializations or hyperparameters to combine the predictions of various models. The

suggested model's overall prediction accuracy and dependability were improved by this ensemble technique. By combining models that have been trained differently or exhibit variations in architecture, the aim is to reduce the risk of overfitting to specific patterns in the data.

Variations of the Attention-Based LSTM make up the basic models that were selected for the ensemble. With unique initializations or hyperparameters, each base model represents a different instantiation of the LSTM architecture. These changes allow the ensemble to identify various dependencies and trends in the student performance data.

Weight Assignment: The key step was to give the base models the proper weights. The weights controlled how much the predictions from each model contributed to the overall ensemble forecast. The model's performance on a validation set was frequently used to determine how much weight to assign. Underperforming models were given lower weights, while models that performed better in terms of generalization or prediction accuracy were given greater weights.

Ensemble Architecture: Simple combinations like majority voting or weighted averaging may be used in the ensemble's construction, as well as more complex strategies like stacking. Stacking, for instance, might incorporate a meta-learner that figures out how to best combine the predictions from the foundation models.

Integration of All Stages: Full Model Architecture

The three steps were integrated cohesively into the overall model architecture for predicting student performance, as illustrated in Figure 4. Each stage serves a particular function to improve the predictive model's accuracy, resilience, and interpretability. This extensive architecture is shown in its overall perspective here:

Stage 1: Data Processing and Augmentation using CVAE:

Data processing begins with the collection, cleaning, and integration of relevant student performance data.

Feature engineering techniques were applied to select, transform, and construct meaningful features.

Conditional Variational Autoencoders (CVAE) were utilized to augment the dataset, generating additional data points for improved model training.

Stage 2: Predictive Model (4-layered Stacked LSTM with Multi-head Attention):

This stage encompasses the heart of the predictive modeling efforts.

The architecture consists of a stacked LSTM with four layers.

Multi-head Attention mechanisms were inserted between the first and second, second and third, and third and fourth LSTM layers.

These attention mechanisms allow the model to capture both short-term and long-term dependencies within the student performance data while attending to crucial features.

Stage 3: Ensemble of Models for Final Prediction:

Ensemble techniques were employed to combine the predictions of multiple base models.

Diverse variations of the Attention-Based LSTM, each with unique strengths, were selected for the ensemble.

Weighted averaging was utilized to combine the predictions, with weights optimized to maximize predictive accuracy.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

Interpretability Enhancements:

Throughout the model, interpretability enhancements were integrated, with LIME (Local Interpretable Model-agnostic Explanations) being a notable component.

LIME generates explanations for individual predictions, offering transparent insights into the factors contributing to a student's predicted performance

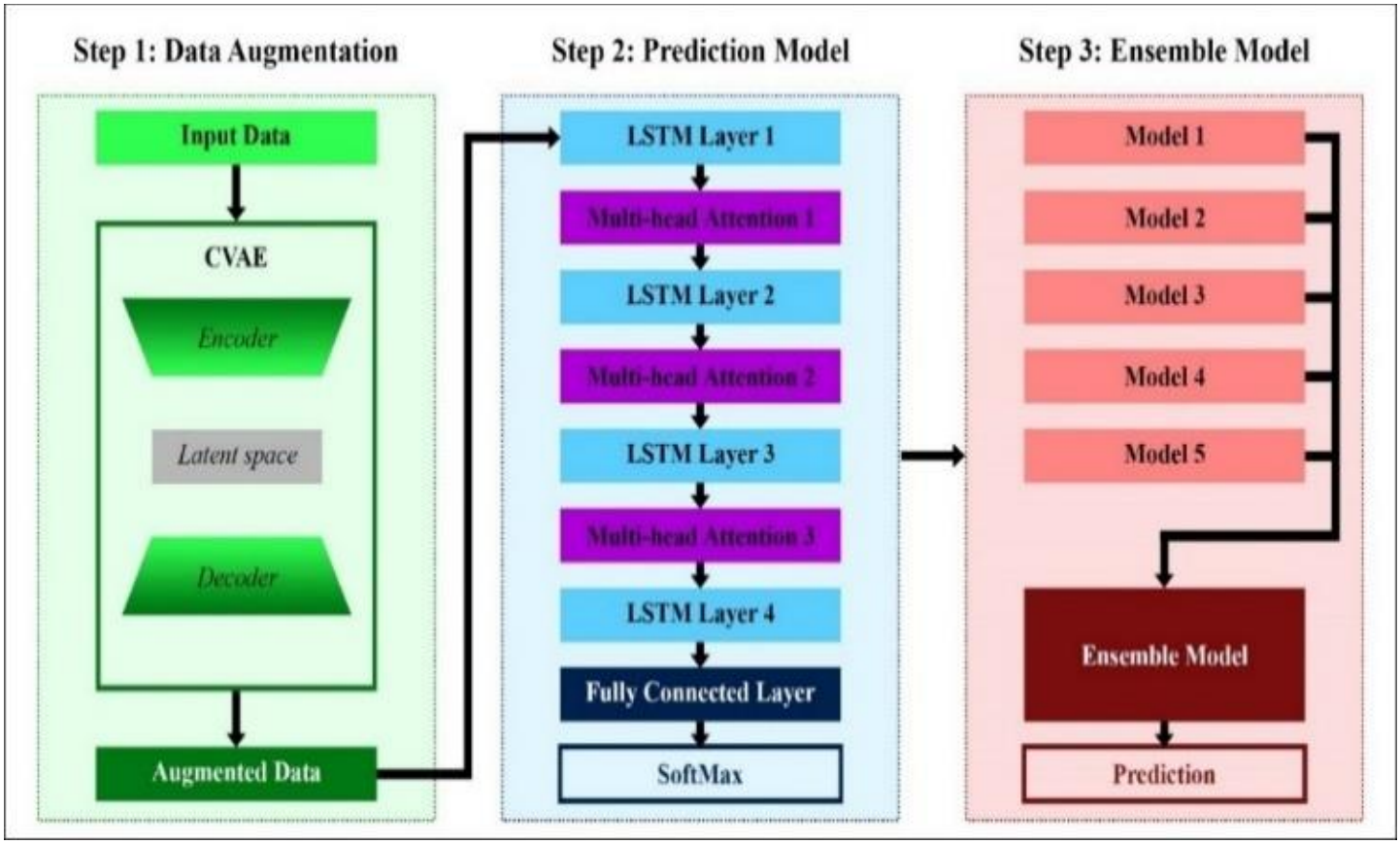

Fig 4: Proposed Model Architecture

Interpreting Factors Affecting Student's Performance Using LIME

A well-known interpretability technique called LIME (Local Interpretable Model-Agnostic Explanations) provides local explanations for certain predictions, providing insight into the model's decision-making process and aiding in a deeper understanding of the factors influencing students' performance. LIME clarifies the model's predictions by training an interpretable "surrogate" model on locally altered examples of the original data. These perturbed cases were produced by sampling and altering the feature values near the desired prediction. The surrogate model was then trained to approximate the black-box model using the behavior of the black-box model close to the prediction.

Model Evaluation Metrics

Metrics like accuracy, precision, recall, and loss were used to evaluate how well the suggested model performs. The model, Deep Neural Network, by Bendangnuksung and Prabu (2018), which utilized the same dataset, was contrasted with the Attention-Based LSTM model. The suggested model was again compared with six baseline models: K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and Artificial Neural Network (ANN) to measure the effectiveness of the proposed model.

Data Analysis

The analysis of the data were presented in tables, histograms, bar charts, box plots, and correlation maps

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

RESULTS AND DISCUSSION

Dataset Analysis

Presented in Figure 5 shows the distribution of the numerical attributes using histograms. Raised Hands feature exhibits a bimodal distribution, with a significant proportion of students either raising their hands infrequently (0–10 times) or very frequently (70–100 times). This suggested a polarization in classroom participation behaviour. Similarly, the visited resources attribute also displays a bimodal pattern, with peak observed in the 0-10 and 80–100 ranges, indicating that students tend to either access learning resources very rarely or very frequently. For announcements view variable, the distribution is right-skewed, with noticeable peak between 0 and 10. This suggested that a large number of students seldom view announcement. Minor peaks observe between 10 and 50 indicate some variability but overall, the distribution remain relatively uniform beyond the initial range. These findings highlight considerable variation in student engagement behaviors across different aspects of the learning platform, particularly in classroom interaction, resource utilization, and responsiveness to announcements.

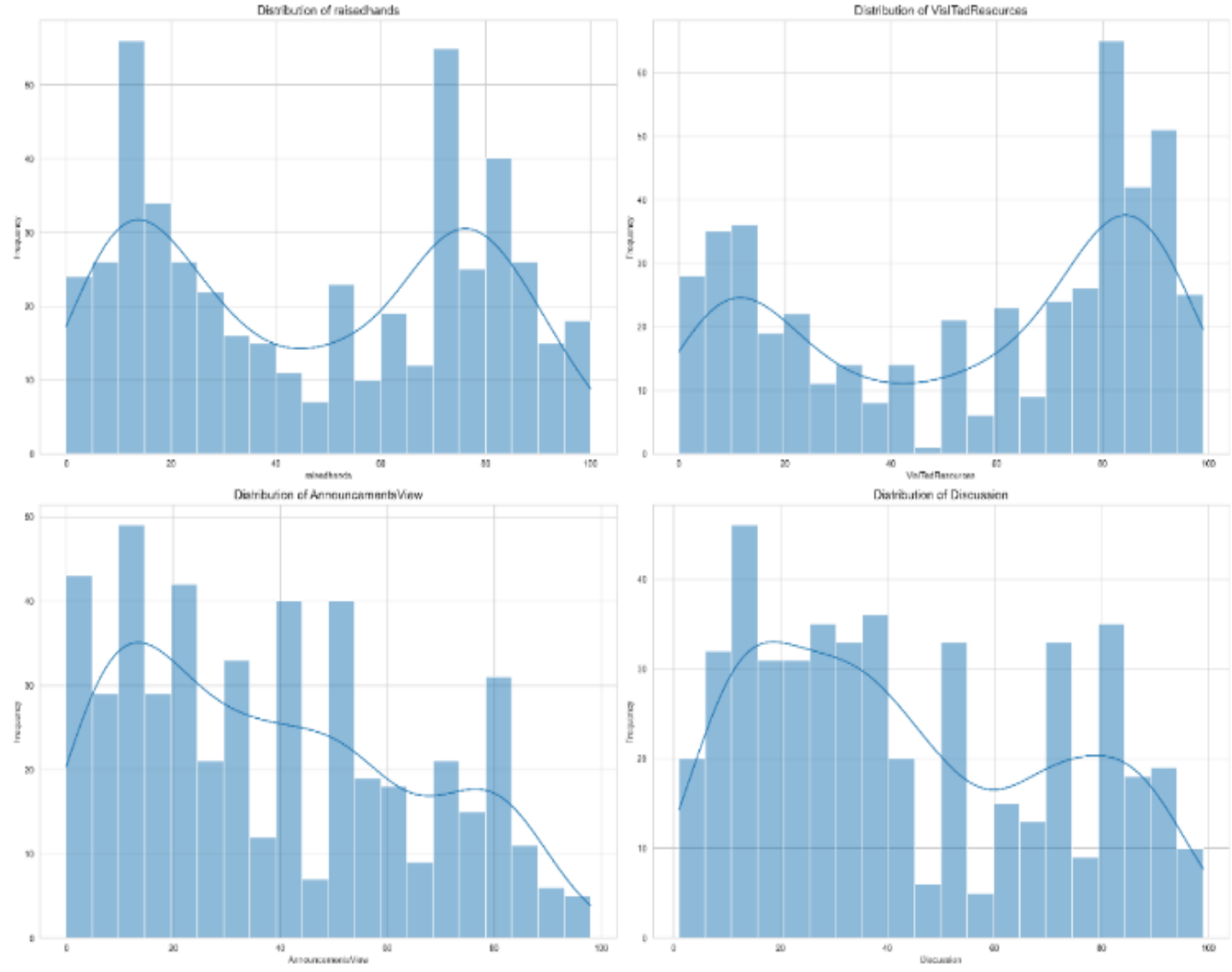

Fig 5: Distribution of the Numerical Attributes using Histograms

The bar graphs in Figure 6 offer valuable insight into the distribution of several categorical variables within the dataset. The data indicates that male students outnumber female students, suggesting a gender imbalance within the sample. Nationality-wise, although students from a variety of countries are included, the majority are from "KW," followed by "Jordan" and "Palestine," pointing to a largely regional representation within the student body. Parental engagement is also reflected in the data, with most parents having responded to the survey, indicating a notable level of involvement in their children's education. Similarly, parental satisfaction with the

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

school appears generally positive, as a larger number of parents rated their experience as "Good" rather than "Bad."

Student attendance records reveal that a significant proportion of learners had fewer than seven days of absence, implying good attendance habits among the majority. The data shows a relatively balanced spread of students performance. However, the Middle (M) class holds the highest number of students. This suggests that most students perform at an average level, with fewer students falling into the lower or higher performance brackets. These observations collectively provide a foundational understanding of the student demographic and behavioral patterns, which are critical in contextualizing performance prediction outcomes in the broader scope of educational research.

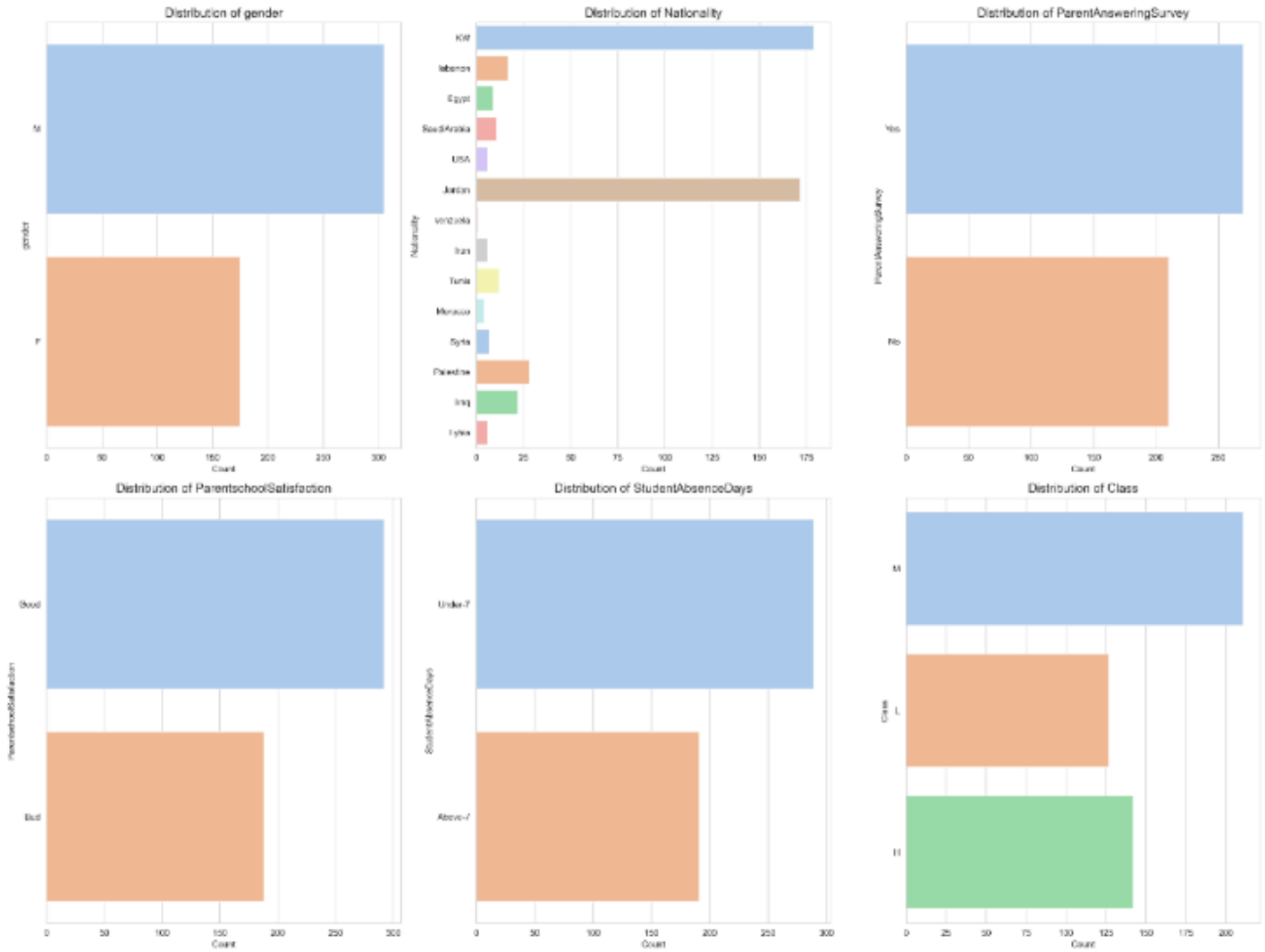

Fig 6: Distribution of Categorical Attributes

The box plots below in Figure 7 provide a visual representation of the distribution and variability of each numerical attribute, including potential outliers. Raised Hands: Most students raised their hands between approximately 15 and 75 times, with a median of around 50 times. There are some outliers at the lower end of the distribution.

Visited Resources: The interquartile range (IQR) is between approximately 20 and 84, with a median of around 65. There are several outliers on the lower end, indicating that some students visited resources very few times.

Announcements View: The distribution is somewhat right-skewed with a median of around 33. The IQR is between 14 and 58, and there are a few outliers at the lower end. The distribution is quite widespread with a median around 39. The IQR is between 20 and 70, and there are outliers on both ends, indicating varied participation in discussions,

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

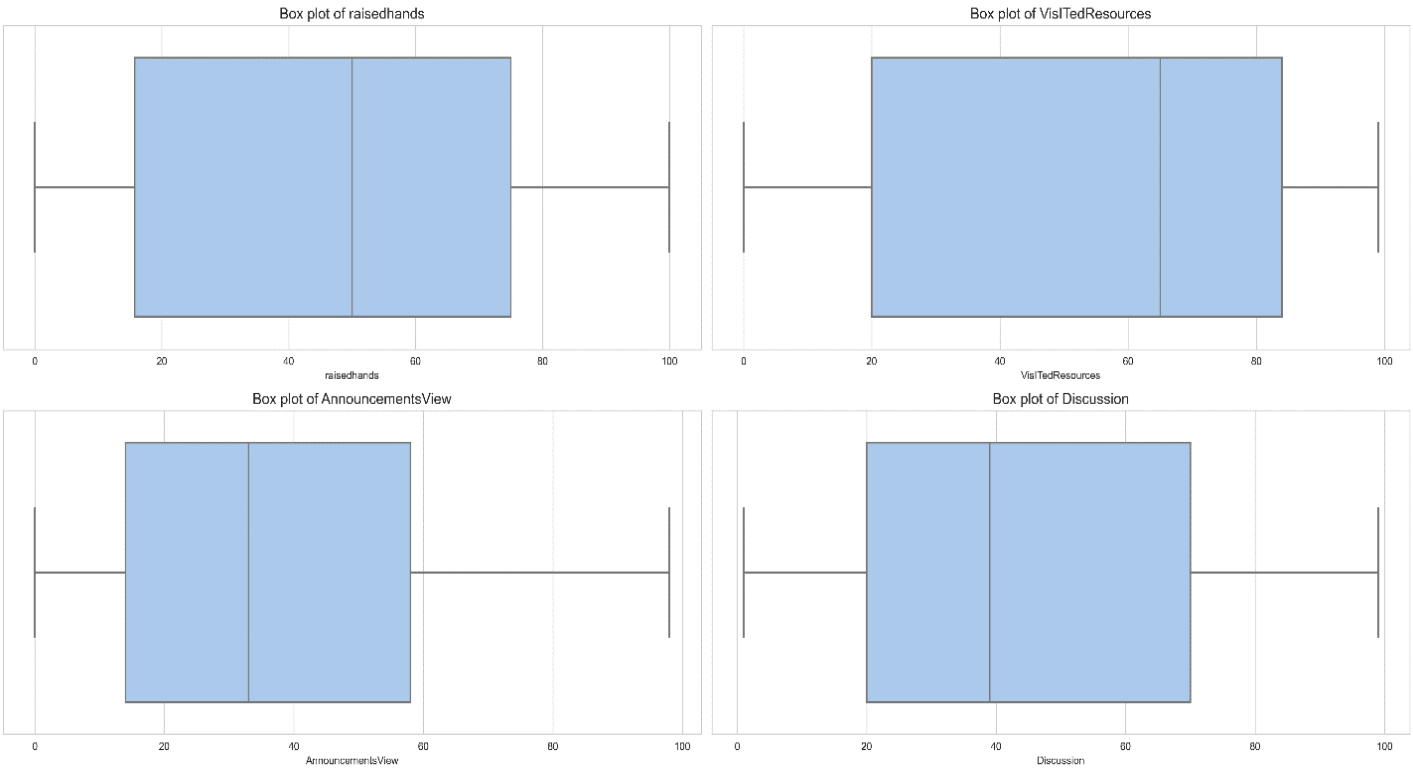

Fig 7: Visual representation of the distribution and variability of each numerical attribute, including potential outliers

The correlation heatmap presented in Figure 8 illustrates the pairwise relationships among the numerical attributes in the dataset. A strong positive correlation (r = 0.69) is observed between the number of times students raised their hands and the number of resources they visited A moderate positive correlation (r = 0.38) exists between raised hands and the number of announcements viewed. Similarly, the relationship between visited resources and announcement views also shows a moderate positive correlation (r = 0.61), indicating that students who explore resources are also inclined to keep up with class communications. Interestingly, the attribute related to discussion participation exhibits relatively low correlations with other variables, implying that the frequency of engagement in discussions does not strongly align with other behavioral indicators such as resource usage or announcement views.

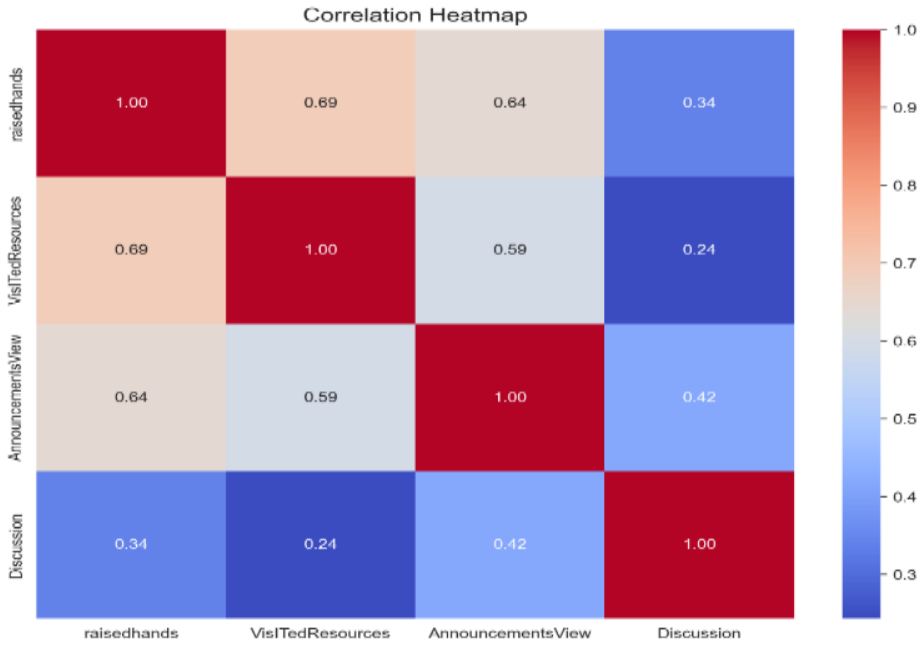

Fig 8: Pairwise Correlation Between the Numerical Attributes

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

Using LSTM, CVAE, and Ensemble model to predicts students' performance.

Three variants of an LSTM model are examined and integrated with an attention mechanism as displayed in Table 4. The study demonstrates the importance of data preprocessing steps such as attribute selection, cleaning, transformation, and reduction to enhance the quality of the data for training. The use of Conditional Variational Autoencoders (CVAE) in data augmentation was found to enhance model generalization.

Table 4: Evaluation of Three variants of LSTM with Attention Mechanism

Metrics	LSTM+ Attention	LSTM+ Attention (Augmentation)	LSTM+ Attention (Augmentation+ Ensemble)
Train Loss	0.6461	0.6466	0.3
Train Accuracy	0.8009	0.8018	0.95
Train Precision	0.8713	0.8702	0.96
Train Recall	0.8714	0.8696	0.96
Test Loss	0.7495	0.749	0.3
Test Accuracy	0.6956	0.852	0.896
Test Precision	0.8478	0.8478	0.96
Test Recall	0.85	0.8491	0.96

As shown in Table 4, the baseline LSTM+Attention model recorded a training loss of 0.6461, suggesting room for further optimization. Introducing data augmentation alone did not significantly reduce this value, with the training loss remaining nearly comparable. However, the incorporation of both augmentation and ensemble techniques led to a substantial improvement, reducing the training loss to 0.300. This sharp decline highlights the ensemble method's effectiveness in enhancing the model's fit to the training data. Similarly, training accuracy improved across the different models. The base model achieved an accuracy of 0.8009, while the augmented model showed a slight increase to 0.8018, a modest improvement likely attributable to the expanded training data.

The most notable gain was observed in the LSTM+Attention model with both augmentation and ensemble methods, which achieved an impressive training accuracy of 0.950, indicating highly effective learning of the training patterns. In terms of training precision and recall, the baseline model exhibited balanced values of 0.8713 and 0.8714, respectively. These values decreased marginally with the augmented model but improved significantly in the combined model, both reaching 0.960, reflecting a strong ability to predict the positive class during training.

When assessing generalization, the test loss results mirror the training outcomes. The baseline model recorded a test loss of 0.7495, suggesting weaker generalization to unseen data. The augmented version offered a marginal improvement with a test loss of 0.7490. However, the augmented and ensembled model demonstrated remarkable generalization with a significantly lower test loss of 0.300, indicating greater predictive accuracy on the test set. Test accuracy further substantiates this trend. The base model achieved a test accuracy of 0.6956, while the augmented model improved this metric to 0.852.

The augmented and ensembled version showed the strongest test performance, achieving a test accuracy of 0.896. Although slightly lower than the augmented-only version in this specific metric, it significantly outperforms the baseline model and provides a more balanced performance overall. Precision and recall on the test set also followed a similar pattern. Both the baseline and augmented models maintained similar values, approximately 0.8478 for precision and 0.8500 for recall. These figures rose substantially in the ensemble-augmented model, with both precision and recall reaching 0.960, indicating a high capacity to detect true positives and avoid false negatives in new data.

Fig 9 further visualizes the distribution of student performance classes. The dataset appears relatively balanced, with the majority of students categorized as "M" (Medium performance). Approximately 138 students fall into the "H" (High) category, while the remaining are classified as "L" (Low). This balanced distribution supports the robustness of model training and evaluation, as no single class dominates the dataset.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

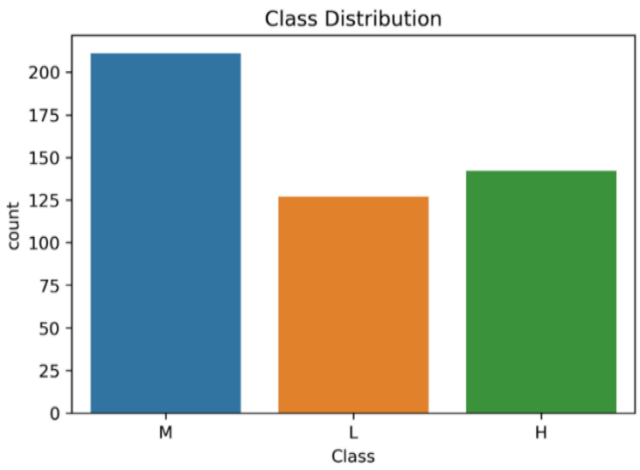

Fig 9: Distribution of Class

LIME providing insights into factors that contribute to students performance

The LIME analysis presented in Figure 10 provides valuable insights into the key features influencing student performance predictions.

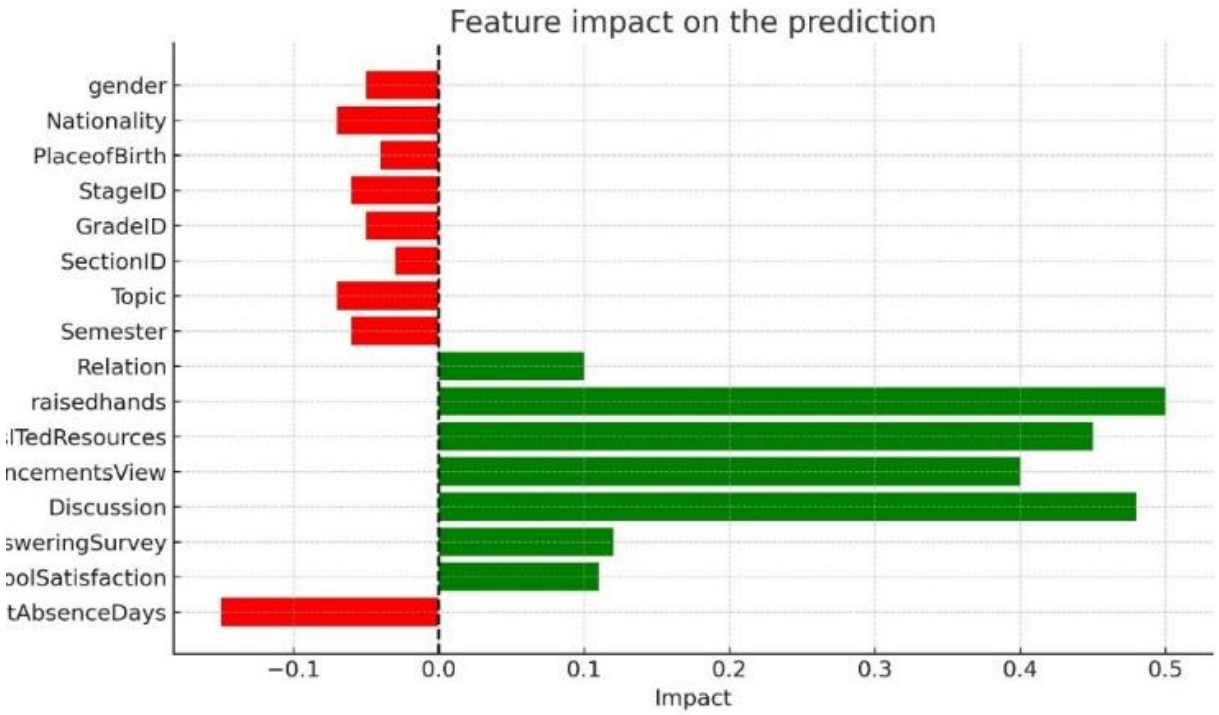

Fig 10: LIME interpretation feature impact on the prediction

As presented in Fig 10, the frequency with which a student raises their hand during class was identified as one of the most influential features. This behavior is indicative of active participation and attentiveness**,** suggesting that students who engage directly in classroom activities tend to perform better academically. The positive correlation between raised hands and academic performance aligns with pedagogical research that emphasizes the value of interactive learning environments. The number of learning resources accessed by a student also plays a significant role in performance prediction. Students who frequently utilize course materials and supplementary resources demonstrate a stronger commitment to learning and knowledge acquisition. This consistent engagement with educational content often translates to improved comprehension and higher academic achievement. It was also evident that the frequency at which students view announcements reflects their level of academic awareness and self-regulation. Those who regularly check updates and notifications are more likely to stay informed about important deadlines, tasks, and changes in the course structure.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VII July 2025

The contributions to class discussions whether in-person or online are also influential factor. Active involvement in discussions fosters critical thinking, deeper understanding, and knowledge retention. It also enables peer-to-peer learning, where students benefit from diverse perspectives and collaborative problem-solving. The LIME interpretations provide a transparent, interpretable view of the model's decision-making process, highlighting key behavioral traits that drive student success. These findings reinforce the importance of fostering student engagement across various dimensions as part of an effective learning strategy.

Accuracy of proposed predictive model as compared to existing prediction methods

In comparison to the proposed model, Attention-Based LSTM with Augmentation and Ensemble models, Table 5 shows the accuracy of each baseline model on the goal of predicting student performance:

Table 5: Performance of Baseline Models as Compared to Attention-Based LSTM

Classifier	Accuracy (%)
K-Nearest Neighbor (KNN)	73.5
Decision Tree (DT)	82.2
Support Vector Machine (SVM)	83.3
Naïve Bayes (NB)	80.0
Random Forest (RF)	85.4
Artificial Neural Network (ANN)	80.0
Deep Neural Network	84.3
Proposed Model	**89.6**

The results show that the Random Forest (RF) model, with an accuracy of 85.4%, performs better than the other baseline models. Due to its ensemble structure, RF can effectively capture complicated correlations in the data. The Support Vector Machine (SVM), which has an accuracy of 83.3%, also performs well. These reference models help analyze the suggested deep learning model, which has a high accuracy rate of 89.6%. The DNN model in the research exhibits a remarkable accuracy of 84.3%, while the LSTM+Attention (Augmentation+Ensemble) model has a superior accuracy of 89.6%, highlighting the efficacy of combining attention mechanisms, data augmentation, and ensemble methods in improving model performance. The model serves as a light for the promising potential of incorporating attention mechanisms, data augmentation, and ensemble approaches in moving forward in the field of educational data mining to create a prediction system that exemplifies accuracy and reliability

DISCUSSION

From the study, a bimodal distribution emerged in several key behavioral indicators. Notably, students who frequently raised their hands during lessons were identified by the model as more attentive and participatory, traits that strongly correlated with higher academic performance. This suggests that active classroom involvement remains a reliable marker of student success. Similarly, the frequency with which students visited online resources proved to be a strong indicator of their academic commitment. Regular access to educational materials was associated with improved grades, reinforcing the idea that resource utilization supports deeper learning. The analysis also showed that students who frequently viewed academic announcements were more likely to stay informed and engaged, contributing positively to their performance. In addition, student participation in in-person or virtual discussions was found to enhance understanding and knowledge retention, supporting more favorable academic outcomes. These findings are consistent with earlier research by Mengash and Namoun and Alshanqiti, both of whom emphasized the link between behavioral engagement and academic success in digital learning environments [6] [10].

From a model performance perspective, the Attention-Based LSTM outperformed baseline models across various evaluation metrics including accuracy, precision, and recall. Its ability to effectively capture patterns in temporal data and highlight critical features suggests it is well-suited for the complexities of educational data analysis. This strong performance supports its potential as a transformative tool in educational data mining. The results also align with those of Tsiakmaki et al., whose Deep Neural Network, applied to the same dataset,

similarly demonstrated robust predictive capabilities [14]. Perhaps most significantly, the Attention-Based LSTM model opens the door to personalized learning paths. By accurately predicting student performance across a range of behavioral features, the model can help educators identify specific needs and tailor interventions accordingly. Its ability to detect bimodal patterns in student engagement means it can serve both high-performing and at-risk learners, offering targeted support based on individual behavior profiles

CONCLUSION AND RECOMMENDATIONS

Based on a variety of academic and behavioural characteristics, the model fared better than baseline models at accurately predicting students' success. LIME provides information on the variables influencing students' expected success. It was observed that elements including conversation, resource visits, raised hands, and announcement views had a bigger effect on students' performance. The study concluded that the Attention-Based LSTM outperformed the Deep Neural Network, K-Nearest Neighbor, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, and Artificial Neural Network which uses the same dataset in predicting student performance. The Attention-Based LSTM model's performance highlights how crucial it is for the educational industry to make use of cutting-edge analytical techniques. For greater insights, educators and legislators must be urged to switch from conventional models to more advanced ones. Also, when developing curricula, policymakers need to take the research on student involvement into account. Including components that encourage active participation can result in better academic performance. Educational institutions should dedicate resources towards providing stakeholders, educators, administrators, and students alike, with training on the implications and practical uses of these models.

ACKNOWLEDGMENT

We acknowledge the support received from selected institutions for the dataset.

REFERENCES

1. Gaftandzhieva, S., Docheva, M., & Doneva, R. (2021). A Comprehensive Approach To Learning Analytics In Bulgarian School Education. Education And Information Technologies, 26(1), 145–163.
2. Alyahyan, E., & Dustegor, D. (2020). Predicting Academic Success In Higher Education Literature Review And Best Practices. International Journal Of Educational Technology In Higher Education, 17. Https://Doi.Org/10.1186/S41239-020-0177-7.
3. Nawang, H., Makhtar, M., & Shamsudin, S. N. W. (2018). Classification Model And Analysis On Students' Performance. Journal Of Fundamental And Applied Sciences, 9, 869. Https://Doi.Org/10.4314/Jfas.V9i6s.
4. Tsai, Y. S., & Gasevic, D. (2017). Learning Analytics In Higher Education-Challenges And Policies: A Review Of Eight Learning Analytics Policies. 233–242. Https://Doi.Org/10.1145/3027385.3027400.
5. A Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review On Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414–422. Https://Doi.Org/Https://Doi.Org/10.1016/J.Procs.2015.12.157.
6. Mengash, H. A. (2020). Using Data Mining Techniques To Predict Student Performance To Support Decision Making In University Admission Systems. IEEE Access, 8, 55462–55470. Https://Doi.Org/10.1109/ACCESS.2020.2981905.
7. Kumar, M., Singh, A. J., & Handa, D. (2017). Literature Survey On Student's Performance Prediction In Education Using Data Mining Techniques. International Journal Of Education And Management Engineering, 7(6), 40–49. Https://Doi.Org/10.5815/Ijeme.2017.06.05.
8. Costa, E. B., Fonseca, B., Santana, M. A., De Araújo, F. F., & Rego, J. (2017). Evaluating The Effectiveness Of Educational Data Mining Techniques For Early Prediction Of Students' Academic Failure In Introductory Programming Courses. Computers In Human Behavior, 73, 247-25.
9. De Marcos-Ortega, L., Garcia-Cabot, A., Garcia-Lopez, E., Ramirez-Velarde, R., Teixeira, A. M., & Martínez-Herráiz, J. J. (2020). Gamifying Massive Online Courses: Effects On The Social Networks And Course Completion Rates. Applied Sciences, 10(20), 7065.
10. Namoun, A., & Alshanqiti, A. (2021). Predicting Student Performance Using Data Mining And Learning Analytics Techniques: A Systematic Literature Review. In Applied Sciences (Switzerland), 11 (1): 1–

28).
11. A Dedecker, A. P., Goethals, P. L., Gabriels, W., & De Pauw, N. (2004). Optimization Of Artificial Neural Network (ANN) Model Design For Prediction Of Macroinvertebrates In The Zwalm River Basin (Flanders, Belgium). Ecological Modelling, 174(1-2), 161-173.
12. Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H. & Wang, C. (2018). Machine Learning And Deep Learning Methods For Cybersecurity. IEEE Access, 6, 35365-35381
13. Altabrawee, H., Abdul, O., Ali, J., & Ajmi, Q. (2019). Predicting Students' Performance Using Machine Learning Techniques. In Pure And Applied Sciences, 27:1-15.
14. Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Transfer Learning From Deep Neural Networks For Predicting Student Performance. Applied Sciences (Switzerland), 10(6).
15. Lv, Z., & Qiao, L. (2020). Deep Belief Network And Linear Perceptron Based Cognitive Computing For Collaborative Robots. Applied Soft Computing Journal, 92:1-21.
16. Nguyen, G. N., Viet, N. H. Le, Elhoseny, M., Shankar, K., Gupta, B. B., & El-Latif, A. A. A. (2021). Secure Blockchain Enabled Cyber–Physical Systems In Healthcare Using Deep Belief Network With Resnet Model. Journal Of Parallel And Distributed Computing, 153, 150–160.
17. Sokkhey, P., & Okazaki, T. (2020). Hybrid Machine Learning Algorithms For Predicting Academic Performance. International Journal Of Advanced Science Applications, 11(1), 32-41.
18. Deng, J., Xu, X., Zhang, Z., Frühholz, S., & Schuller, B. (2017). Semisupervised Autoencoders For Speech Emotion Recognition. IEEE/ACM Transactions On Audio, Speech, And Language Processing, 26(1), 31-43
19. Silhan, T., Oehmcke, S., & Kramer, O. (2019). Evolution Of Stacked Autoencoders. In 2019 IEEE Congress On Evolutionary Computation (CEC) (Pp. 823-830). IEEE.
20. Solorio-Ramírez, J. L., Saldana-Perez, M., Lytras, M. D., Moreno-Ibarra, M. A., & Yáñez-Márquez, C. (2021). Brain Hemorrhage Classification In CT Scan Images Using Minimalist Machine Learning. Diagnostics, 11(8), 1449
21. Chen, Y., Li, Y., Narayan, R., Subramanian, A., & Xie, X. (2016). Gene Expression Inference With Deep Learning. Bioinformatics, 32(12), 1832-1839.
22. Zafar, M. H., Khan, N. M., Mansoor, M., Mirza, A. F., Moosavi, S. K. R., & Sanfilippo, F. (2022). Adaptive ML-Based Technique For Renewable Energy System Power Forecasting In Hybrid PV-Wind Farms Power Conversion Systems. Energy Conversion And Management, 258, 115564
23. Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment Of Machine Learning Performance For Decision Support In Venture Capital Investments. Ieee Access, 7, 124233-124243.
24. Vanlehn, K. (2011). The Relative Effectiveness Of Human Tutoring, Intelligent Tutoring Systems, And Other Tutoring Systems. Educational Psychologist, 46(4), 197-221.