J-KOMA Journal of Computer Science and Applications

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Open Access

***E-ISSN**: 2620 - 4872*
***Vol**.08, **No**.02*
***Doi**: 10.21009/JKOMA.082.06*

***Recivied**: 23 Dec 2025*

***Accepted**: 26 Dec 2025*

***Published**: 26 Dec 2025*

Keywords:
student dropout prediction;
LSTM;
academic time series;
deep learning.

***Correspondence Email:**
riaarafiah@unj.ac.id

Predicting Student Dropout Risk Using Long Short-Term Memory Based on Academic Performance Data

Ria Arafiah[1], Vera Maya Santi[2], Muhammad Eka Suryana[3], Salwa Tsabitah[4], Rahma Wati Malawat[5]

[1] Department of Computer Science, Faculty of Mathematics and Natural Science, Jakarta State University, Indonesia
[2] Department of Statistics, Faculty of Mathematics and Natural Science, Jakarta State University, Indonesia

Abstract

Student dropout is a persistent challenge in higher education, with significant implications for academic quality and institutional effectiveness. Dropout behavior typically evolves over time and is reflected in longitudinal patterns of academic performance across semesters. This study develops a sequential Long Short-Term Memory (LSTM) model to predict student dropout risk using semester-wise academic data. The dataset consists of 385 undergraduate students from a Computer Science program, with academic attributes aggregated into an initial feature matrix of shape (385, 62) and transformed into eight-semester time-series sequences. To mitigate class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied, resulting in a balanced dataset of shape (566, 8, 10). The proposed LSTM model comprises 32,161 trainable parameters and is evaluated on a held-out test set. The model achieves an overall accuracy of 0.77 and attains a recall of 1.00 for the dropout class, indicating that all observed dropout cases in the test data are correctly identified. Rather than emphasizing predictive certainty, this result highlights the model's sensitivity in capturing early risk signals of dropout within longitudinal academic trajectories. However, the corresponding precision for the dropout class remains limited (0.25), reflecting a trade-off between sensitivity and false positive rates. These findings suggest that the proposed approach is particularly suitable as an early warning and screening mechanism to support academic monitoring and timely intervention, while further refinement is required to enhance precision and reduce unnecessary alerts.

INTRODUCTION

Student dropout remains a persistent challenge in higher education institutions worldwide, negatively affecting students' academic trajectories, institutional reputation, and resource efficiency. Dropout is rarely an instantaneous event; rather, it is typically preceded by gradual academic disengagement reflected in declining academic performance across semesters. Consequently, the ability to **detect dropout risk at an early stage** is critical for enabling timely academic interventions and improving student retention outcomes. Early Warning Systems (EWS) have therefore become an important strategic tool for higher education management [1][2].

Traditional statistical and machine learning approaches for dropout prediction—such as logistic regression, decision trees, and support vector machines—have demonstrated moderate success. However, these approaches often struggle to model **longitudinal academic trajectories**, especially when

student performance evolves over multiple semesters and exhibits non-linear temporal dependencies [3][4]. Moreover, dropout datasets are inherently **imbalanced**, with dropout cases forming a small minority, which further degrades predictive performance when conventional classifiers are applied [5].

Recent advances in **deep learning** have significantly reshaped the landscape of dropout prediction research. In particular, recurrent neural networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, have been widely adopted due to their ability to capture long-term temporal dependencies in sequential data [6][7]. Empirical studies have shown that LSTM-based and hybrid architectures (e.g., CNN–LSTM, BiLSTM with attention) outperform classical machine learning models in modeling semester-wise academic data, learning progression patterns, and identifying at-risk students earlier in the academic lifecycle [8][9][10].

Within this state of the art, **class imbalance handling** has emerged as a crucial component of effective dropout prediction systems. Several studies report that without resampling strategies, deep learning models tend to bias predictions toward the majority (non-dropout) class, resulting in poor recall for actual dropout cases [11][12]. Synthetic Minority Over-sampling Technique (SMOTE) and its variants have therefore been extensively employed to improve minority class detection, particularly when the primary objective of EWS is to **minimize false negatives**, i.e., failing to identify students who will eventually drop out [11] [13].

Despite these advances, notable research gaps remain. Many existing studies emphasize overall accuracy while paying limited attention to **recall for the dropout class**, which is arguably the most critical metric in academic intervention contexts [2]. From a practical perspective, a model that successfully identifies all at-risk students—even at the cost of higher false positives—is more valuable than a model that misses actual dropout cases. Furthermore, most reported results are based on short-term or cross-sectional data, whereas **semester-wise longitudinal modeling** remains underexplored in resource-constrained institutional settings [10].

Motivated by these gaps, this study proposes a **sequential LSTM-based dropout prediction model** using semester-wise academic data from an undergraduate Computer Science program. By explicitly modeling eight semesters of academic progression and applying SMOTE to address data imbalance, the proposed approach prioritizes **maximum recall for the dropout class**, aligning with the core objective of Early Warning Systems. Unlike prior work that focuses predominantly on aggregate performance metrics, this study highlights recall-driven evaluation as a key contribution toward practical, intervention-oriented academic monitoring systems.

Theoretical Background

1. Long Short-Term Memory (LSTM) for Longitudinal Academic Data

Long Short-Term Memory (LSTM) is a specialized form of Recurrent Neural Network (RNN) designed to address the vanishing gradient problem commonly encountered in standard RNN architectures [6]. Unlike traditional feedforward models, LSTM networks are explicitly constructed to model **sequential and temporal dependencies**, making them particularly suitable for longitudinal data such as semester-wise academic records.

An LSTM cell consists of three primary gating mechanisms: the **input gate**, **forget gate**, and **output gate**. These gates regulate the flow of information into and out of the cell state, enabling the network to selectively retain or discard information over long sequences [7]. Formally, at each time step ttt, the LSTM updates its internal state by combining the current input with historical memory, allowing it to learn both short-term variations and long-term trends in sequential data.

In the context of student dropout prediction, academic performance does not fluctuate randomly but evolves gradually across semesters, reflecting patterns of engagement, course load adaptation, and academic difficulty. Prior studies have demonstrated that LSTM-based models are more effective than classical machine learning approaches in capturing such **semester-to-semester progression patterns**, leading to improved identification of at-risk students [8][9][10]. Consequently, LSTM has become a core architecture in contemporary Early Warning Systems for higher education [2].

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

In this study, a **sequential LSTM architecture** is employed to model eight semesters of academic data, enabling the network to learn temporal dependencies across students' academic trajectories rather than relying solely on aggregated or cross-sectional indicators.

2. Synthetic Minority Over-sampling Technique (SMOTE)

A fundamental challenge in student dropout prediction is **class imbalance**, where the number of dropout cases is substantially smaller than non-dropout cases. Such imbalance often causes predictive models to favor the majority class, resulting in poor recall for the minority (dropout) class—a critical limitation for Early Warning Systems [5][12].

The Synthetic Minority Over-sampling Technique (SMOTE) addresses this issue by generating synthetic samples for the minority class through interpolation between existing minority instances in the feature space [11]. Instead of duplicating samples, SMOTE creates new data points along the line segments connecting nearest neighbors, thereby increasing class diversity and reducing overfitting risks associated with naive oversampling.

Recent studies in educational data mining and learning analytics have shown that integrating SMOTE with deep learning architectures significantly improves **minority class sensitivity**, particularly recall, which is essential for identifying all students at risk of dropout [13][2]. While SMOTE may introduce an increase in false positives, this trade-off is often acceptable in academic monitoring scenarios, where missing a dropout case is more detrimental than issuing additional alerts [3].

In this research, SMOTE is applied after transforming academic records into semester-wise sequences to ensure balanced representation of dropout and non-dropout patterns during model training. This design choice aligns with the primary objective of Early Warning Systems, namely **maximizing detection sensitivity for at-risk students**.

3. Conceptual Rationale of the Proposed Approach

By combining **LSTM-based sequential modeling** with **SMOTE-based imbalance handling**, the proposed approach is grounded in two key principles established in the literature:

1. dropout behavior is inherently **temporal and cumulative**, and
2. effective intervention-oriented systems must prioritize **high recall for the dropout class**.

Rather than optimizing for overall accuracy alone, this study emphasizes sensitivity-driven evaluation to support practical academic decision-making. This conceptual foundation positions the proposed model as a **screening-oriented prediction mechanism**, suitable for integration into institutional Early Warning Systems.

METHODS

The overall research framework proposed in this study is illustrated in Figure 1. The methodology consists of five main stages: data collection, preprocessing, handling class imbalance, LSTM modeling, and model evaluation.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

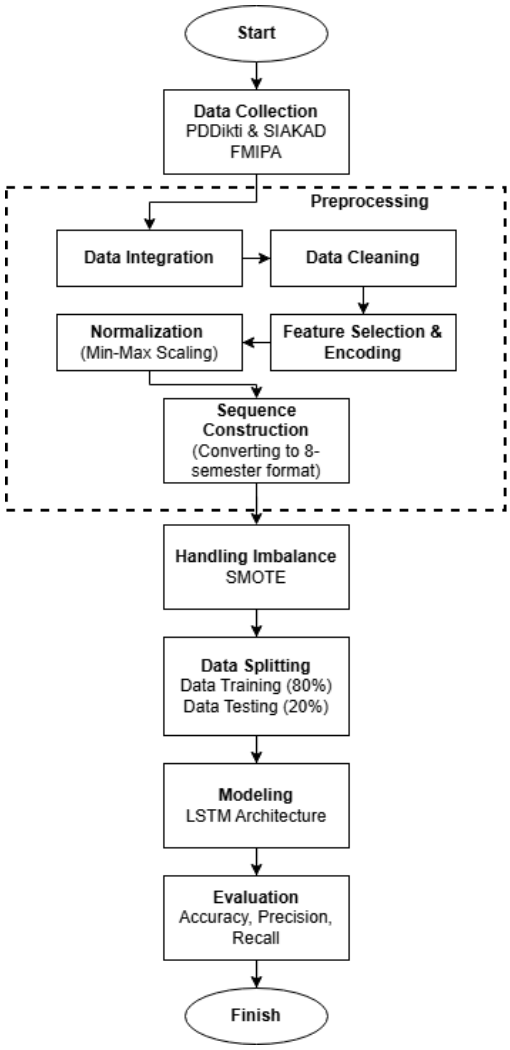

FIGURE 1. Research Flow

1. Dataset Description

The dataset used in this study consists of academic records of undergraduate students from the Computer Science program, FMIPA. After data cleaning, 385 student records were retained. Each student is represented by GPA and credit load per semester for eight semesters. Student status is converted into a binary label: 1 for dropout and 0 for non-dropout.

2. Data Preprocessing

Data preprocessing includes label binarization, feature normalization using Min-Max scaling, sequence construction, and class imbalance handling using SMOTE. The final dataset shape is (566, 8, 10) after oversampling.

Data Integration

The data used in this research were obtained from two main sourcer, the higher education database (PDDikti-ilkom), and the academic information system (SIAKAD FMIPA). Data from PDDikti were collected from multiple spreadsheet files (.xlsx) that are organized by academic semester. The observation period spans from semester 99 (odd semester of 2013) to semester 118 (even semester of 2022). For each

semester, five distinct data files are provided, namely academic record, student biodata, graduation information, course information, and course grades. In addition, complementary administrative data were collected from the SIAKAD FMIPA UNJ system in the form of a single consolidated file. This dataset contains information that is not comprehensively recorded in PDDikti, such as admission pathways, student enrollment status, registration types, and tuition fee categories.

Understanding the PDDikti Data

The data understanding phase was conducted to gain an in-depth comprehension of the structure, attributes, and interrelationships among the datasets obtained from PDDikti. Understanding these characteristics is essential to determine appropriate integration strategies and to identify potential data quality issues prior to preprocessing. Each semester folder contains five types of files:

1. **PDDikti-Akademik.xlsx**, containing students' academic performance per semester, including semester GPA (IP), cumulative GPA, semester credit load (SKS), and cumulative credits.
2. **PDDikti-Biodata.xlsx**, containing student demographic information such as gender, place and date of birth, parental information, and tuition category
3. **PDDikti-Kelulusan.xlsx**, containing graduation information including graduation semester, graduation date, and final GPA (IPK).
4. **PDDikti-MataKuliah.xlsx**, containing course metadata including course code, course name, and credit units.**PDDikti-Nilai.xlsx**, containing detailed course-level grades for each student, including letter grades and numeric scores.

The second data source is a single Excel file named `Data_MHS-FMIPA.xlsx`, obtained from SIAKAD FMIPA. This dataset provides complementary student attributes such as admission pathway, enrollment year (cohort), registration type, tuition scheme, and student status as of 2022.

Due to differences in data structure and record ordering between datasets, direct merging was not feasible. Therefore, data integration was performed using Microsoft Excel with the `VLOOKUP` function, using the student identification number (NIM) as a unique identifier. The file **PDDikti-Biodata.xlsx** was selected as the primary (destination) dataset because it contains the most comprehensive set of attributes. Biodata files from all semesters were first merged into a single file, and duplicate records were removed. To ensure accurate lookup operations, a helper column was created by concatenating the NIM with the target column name (e.g., "_Angkatan"). The same helper column structure was applied in both the destination and source files. The enrollment year (Angkatan) was then retrieved from `Data_MHS-FMIPA.xlsx` using the `VLOOKUP` function with exact matching. The Angkatan variable served as a filter for determining the correct semester source when retrieving semester GPA (IP) and credit load (SKS). For each cohort, semester GPA (IP1–IP8) and credit load (SKS1–SKS8) were populated by referencing the corresponding `PDDikti-Akademik.xlsx` file based on semester mapping. Missing lookup results were handled using the `IFNA` function to avoid lookup errors.

Due to the large number of course-related attributes, Microsoft Excel was insufficient for integrating course grades. Therefore, Python programming was employed using Google Colaboratory. All `PDDikti-Nilai.xlsx` files from different semesters were merged into a single dataframe. Similarly, all `PDDikti-MataKuliah.xlsx` files were combined to obtain a complete list of courses. Duplicate course names resulting from repeated offerings across semesters were removed. The course list was transposed to form new columns, with NIM as the row identifier. An iterative algorithm was then applied to map each student's letter grade to the corresponding course column based on matching NIM and course name. Courses with different names but equivalent academic content due to curriculum changes were consolidated into single features using forward-fill and value replacement techniques. Finally, the course-grade dataset was merged with the biodata-academic dataset using the NIM identifier, resulting in a unified dataset containing academic performance, demographic information, and course-level grades.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

After data integration, preprocessing was conducted to enhance data quality and analytical reliability. This stage includes data cleaning, handling missing values, resolving inconsistencies, and identifying outliers. Preprocessing ensures that the dataset accurately represents real academic conditions while remaining suitable for quantitative analysis.

Handling Missing Values

Missing values were identified in both numerical and categorical attributes. Numerical missing values, such as GPA and credit variables, were imputed using median values to reduce the influence of extreme observations. Categorical missing values were replaced using the most frequently occurring category (mode). Missing value handling was generally applied to all columns containing null values, particularly numerical attributes such as semester GPA (IP) and credit units (SKS). Missing values in the IP and SKS columns were imputed using the median of each respective column, as these variables are numerical in nature. The use of median imputation was chosen to reduce the influence of extreme values that could distort the data distribution if mean-based imputation were applied. This approach ensures that the imputed values better represent the majority of students' academic conditions and maintain the overall robustness of the dataset.

Handling Data Inconsistencies

Data inconsistencies, particularly in numeric formatting, were corrected through normalization procedures. These corrections ensure that all numerical attributes follow a uniform format, thereby preventing errors in subsequent analytical stages. Data inconsistency handling was conducted to address irregularities in numerical attributes, particularly in semester Grade Point Average (GPA/IP) data. Inconsistencies were observed in the form of non-uniform numerical formats, such as variations in decimal precision and values that did not conform to the valid academic GPA range. To resolve these issues, all GPA values were standardized into a consistent numerical format with uniform decimal representation. Values identified outside the acceptable academic range were reviewed and corrected based on institutional academic regulations or treated as invalid entries when verification was not possible. This standardization process enhances data consistency, minimizes potential analytical errors, and improves the reliability of subsequent statistical analyses and modeling processes.

Outlier identification

Outliers were identified using boxplot analysis. However, outliers in GPA and credit attributes were not removed because they remain within academically valid ranges and reflect genuine variations in student performance. Outlier handling is commonly applied to reduce data inconsistency and improve data quality. In this study, outlier analysis was conducted on numerical attributes such as semester Grade Point Average (GPA/IP) and credit units (SKS). However, outliers identified in the IP and SKS columns were not removed or adjusted, as these values remained within academically acceptable ranges and accurately represented students' actual academic conditions. High or low variations in SKS values may naturally occur due to differences in course load taken by students in each semester. Similarly, extreme IP values reflect genuine differences in students' academic abilities. Therefore, removing outliers from these variables could potentially eliminate meaningful information that is relevant for performance analysis.

Data Reduction and Transformation

Data reduction was performed by removing attributes that function solely as identifiers or are not relevant to the research objectives. Data transformation was subsequently applied to restructure the dataset into a format suitable for advanced analytical techniques such as clustering and classification.The data reduction process was conducted to focus the analysis on attributes that are directly relevant to the research objectives. Specifically, the reduced dataset retains key academic variables such as semester-based Grade Point Average (GPA/IP), credit units (SKS), and course grades with approximately 70 percent or more non-missing values. This reduction process significantly decreased data dimensionality and sparsity, thereby improving computational efficiency during the modeling stage. The resulting dataset provides a more representative and structured view of students' academic performance. The attributes retained after the data reduction process are illustrated in the following table.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

TABLE 1. Dataset

No	Column Name	Data Type	Description
1	Gender	Categorical	Student gender attribute used for demographic description.
2	Student Religion	Categorical	Religion practiced by the student.
3	Admission Pathway	Categorical	Admission pathway through which the student entered the university.
4	Cohort Year	Numerical	Student cohort or year of entry into the study program.
5	Entry Semester	Numerical	Academic semester when the student first enrolled.
6	GPA Semester 1 (IP1)	Numerical	Grade Point Average achieved by the student in the first semester.
7	GPA Semester 2 (IP2)	Numerical	Grade Point Average achieved by the student in the second semester.
8	GPA Semester 3 (IP3)	Numerical	Grade Point Average achieved by the student in the third semester.
9	GPA Semester 4 (IP4)	Numerical	Grade Point Average achieved by the student in the fourth semester.
10	GPA Semester 5 (IP5)	Numerical	Grade Point Average achieved by the student in the fifth semester.
11	GPA Semester 6 (IP6)	Numerical	Grade Point Average achieved by the student in the sixth semester.
12	GPA Semester 7 (IP7)	Numerical	Grade Point Average achieved by the student in the seventh semester.
13	GPA Semester 8 (IP8)	Numerical	Grade Point Average achieved by the student in the eighth semester.
14	Cumulative GPA (IPK)	Numerical	Cumulative Grade Point Average representing overall academic performance.
15	Credits Semester 1 (SKS1)	Numerical	Total number of credits taken in the first semester.
16	Credits Semester 2 (SKS2)	Numerical	Total number of credits taken in the second semester.
17	Credits Semester 3 (SKS3)	Numerical	Total number of credits taken in the third semester.
18	Credits Semester 4 (SKS4)	Numerical	Total number of credits taken in the fourth semester.
19	Credits Semester 5 (SKS5)	Numerical	Total number of credits taken in the fifth semester.
20	Credits Semester 6 (SKS6)	Numerical	Total number of credits taken in the sixth semester.
21	Credits Semester 7 (SKS7)	Numerical	Total number of credits taken in the seventh semester.
22	Credits Semester 8 (SKS8)	Numerical	Total number of credits taken in the eighth semester.
23	Student Status	Categorical	Academic status of the student (active, graduated, or dropout).
24	Tuition Fee Category	Numerical	Tuition fee category assigned to the student.
25	subject	Categorical	grades for all semester courses
26	Distance Category	Categorical	Categorization of student residence distance.
27	Residence Status	Categorical	Student living status (e.g., boarding house or family home).

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

3 LSTM Architecture

This study employs a **single-layer sequential Long Short-Term Memory (LSTM) architecture** to model semester-wise academic trajectories for student dropout prediction. The architecture is designed to capture temporal dependencies in longitudinal academic data while maintaining model simplicity to reduce the risk of overfitting given the dataset size.

3.1 Input Representation

The input to the model consists of **semester-wise academic sequences** with a fixed length of **eight semesters**. Each semester is represented by **ten academic features**, resulting in an input tensor of shape **(8, 10)** for each student. This representation enables the model to learn patterns of academic progression and decline across consecutive semesters rather than relying on aggregated indicators.

3.2 LSTM Layer

The core of the model is a **single LSTM layer** with **64 hidden units**. This layer processes the input sequence sequentially, updating its internal memory through input, forget, and output gates. By preserving and selectively updating the cell state over time, the LSTM layer is able to model long-term dependencies in academic performance, such as cumulative academic difficulties that precede dropout.

The single-layer configuration was intentionally selected to balance **model expressiveness and generalization capability**, considering the moderate dataset size. The total number of trainable parameters in the LSTM model is **32,161**, all of which are optimized during training.

3.3 Output Layer

The output of the LSTM layer is passed to a **fully connected dense layer** with a **sigmoid activation function**, producing a single scalar value representing the **probability of student dropout**. This output is subsequently thresholded to perform binary classification into dropout and non-dropout categories.

3.4 Architectural Rationale

The proposed single-layer LSTM architecture is motivated by three key considerations. First, dropout behavior is inherently **temporal and cumulative**, requiring a model capable of learning sequential patterns across semesters. Second, compared to stacked or hybrid deep architectures, a single-layer LSTM reduces the risk of overfitting and improves interpretability in educational data settings. Third, the architecture aligns with the operational goals of Early Warning Systems, where sensitivity to early risk signals is prioritized over architectural complexity.

The LSTM model consists of two stacked LSTM layers with dropout regularization, followed by dense layers for binary classification. The model is trained using the Adam optimizer and binary cross-entropy loss.

TABLE 2. LSTM Model Architecture

Layer	Output Shape	Description
LSTM (64 units)	(None, 8, 64)	Temporal feature extraction
Dropout	(None, 8, 64)	Regularization
LSTM (32 units)	(None, 32)	High-level sequence modeling
Dense	(None, 16)	Feature transformation
Dense (Sigmoid)	(None, 1)	Dropout probability output

RESULTS AND DISCUSSION

The experimental evaluation was conducted on a held-out test set comprising 20% of the total dataset to assess the generalization capability of the proposed LSTM model. The performance of the model is

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

primarily evaluated using the confusion matrix and standard classification metrics, with a specific focus on Recall to ensure all at-risk students are identified. Figure 1 presents the detailed confusion matrix obtained from the test data.

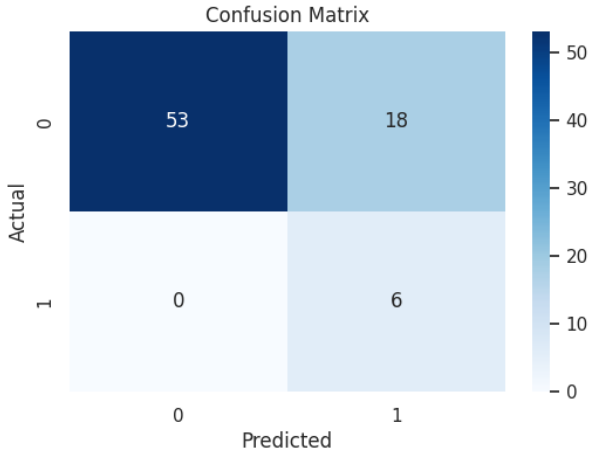

FIGURE 2. Confusion Matrix (Test Data)

As shown in Figure 1, the model demonstrated a strong ability to identify the minority class by correctly classifying all 6 actual dropout cases with zero False Negatives. However, it exhibited a tendency towards over-caution by misclassifying 18 non-dropout students as at-risk. Consequently, while the overall accuracy stood at approximately 76.6%, the model achieved a perfect Recall of 1.00 for the dropout class, although this resulted in a lower Precision score of 0.25. The stability of the learning process is illustrated in Figure 2 and Figure 3 below.

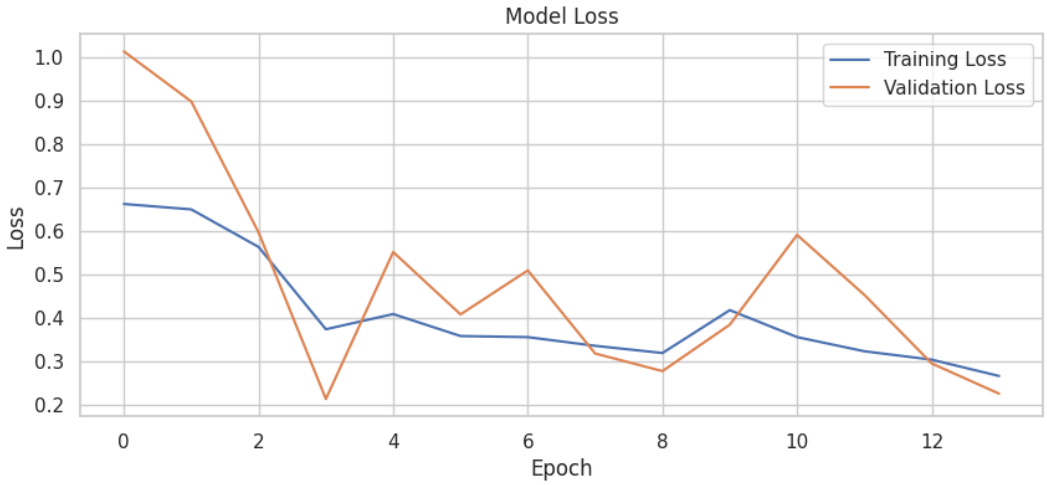

FIGURE 3. Training and Validation Loss

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

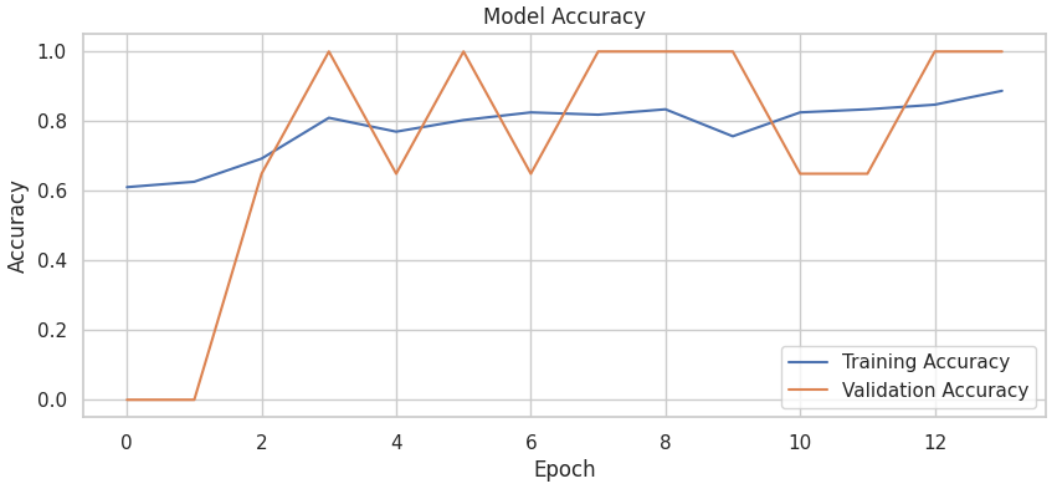

FIGURE 4. Training and Validation Accuracy

The training curves indicate that the LSTM architecture effectively learned the temporal dependencies in the semester-wise academic data, showing steady convergence in loss values without severe overfitting. In the context of predicting student dropout, this performance profile is strategically advantageous. The primary objective of an early warning system is to ensure that no at-risk student goes unnoticed, therefore Recall is prioritized over Precision. The cost of a False Negative, where a student drops out without prior intervention, is significantly higher for the institution than the cost of a False Positive. A False Positive merely results in additional academic monitoring for a student who might not have needed it, which is a manageable preventive measure. Future work may focus on improving precision through threshold tuning to reduce the false alarm rate without compromising the perfect sensitivity of the model.

CONCLUSION

This study investigated the use of a sequential Long Short-Term Memory (LSTM) model for predicting student dropout risk based on semester-wise academic data. By modeling eight semesters of longitudinal academic trajectories and addressing class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE), the proposed approach aims to support early identification of students at risk of dropping out.

Experimental results demonstrate that the model achieves an overall accuracy of 0.77 on the test set, with a recall of 1.00 for the dropout class. This indicates that all dropout cases in the evaluation data were successfully identified. Although the precision for the dropout class remains relatively low, reflecting the presence of false positives, this trade-off is acceptable in the context of academic Early Warning Systems, where failing to detect at-risk students is more critical than generating additional alerts.

The training and validation curves further support these findings. The loss curves show a consistent downward trend with no sustained divergence between training and validation loss, suggesting that the model converges effectively without severe overfitting. Similarly, the accuracy curves indicate stable learning behavior, with validation accuracy exhibiting fluctuations due to the limited number of dropout samples but remaining comparable to training accuracy in later epochs. These patterns suggest that the model generalizes reasonably well given the dataset size and imbalance characteristics.

Overall, the results confirm that sequential LSTM models are well-suited for capturing temporal patterns in academic performance and can serve as a sensitive screening mechanism for dropout risk detection. While the current model prioritizes recall to support early intervention, future work should focus on improving precision through feature enrichment, alternative resampling strategies, cost-sensitive learning, or hybrid architectures. Incorporating non-academic factors and validating the approach on larger and multi-institutional datasets would further enhance robustness and generalizability.

Several directions can be pursued to further enhance the effectiveness and applicability of the proposed dropout prediction approach. First, future studies should incorporate **non-academic and**

behavioral features, such as attendance records, learning management system activity, and student engagement indicators, to improve prediction precision and reduce false positive rates. Second, alternative strategies for addressing class imbalance, including **cost-sensitive learning, focal loss, and ensemble-based resampling methods**, may provide a better balance between recall and precision. Third, more advanced sequential architectures, such as **stacked LSTM, Bidirectional LSTM, or attention-based models**, should be explored to capture richer temporal dependencies while maintaining generalization capability. Finally, validating the proposed model on **larger, multi-cohort, and multi-institutional datasets** is essential to assess robustness, scalability, and generalizability in real-world Early Warning System deployments.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided by the Public Service Agency BLU Fund of the Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, under the Ministry of Higher Education, Science, and Technology, Republic of Indonesia. This research was funded through the Faculty Applied Research Scheme (Terapan Fakultas, T-FP) with Contract Number: 63/SPK PENELITIAN/5.FMIPA/2025. The authors also thank all parties who contributed to the successful completion of this research.

REFERENCES

[1] J. Y. Chung and S. Lee, “Dropout early warning systems for high school students using machine learning,” *Child Youth Serv Rev*, vol. 96, pp. 346–353, 2019, doi: 10.1016/j.childyouth.2018.11.030.

[2] B. M. McMahon and S. F. Sembiante, “Re-envisioning the purpose of early warning systems: Shifting the mindset from student identification to meaningful prediction and intervention,” *Review of Education*, vol. 8, no. 1, pp. 266–301, 2020, doi: 10.1002/rev3.3183.

[3] D. Andrade-Girón *et al.*, “Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review,” *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 5, pp. 1–11, 2023, doi: 10.4108/eetsis.3586.

[4] R. Manrique, B. P. Nunes, O. Marino, M. A. Casanova, and T. Nurmikko-Fuller, “An analysis of student representation, representative features and classification algorithms to predict degree dropout,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2019, pp. 401–410. doi: 10.1145/3303772.3303800.

[5] S. Lee and J. Y. Chung, “The machine learning-based dropout early warning system for improving the performance of dropout prediction,” *Applied Sciences (Switzerland)*, vol. 9, no. 15, 2019, doi: 10.3390/app9153093.

[6] H.-S. Park, S.-J. Yoo, and Y.-H. Gu, “Deep Learning-Based Early Dropout Prediction in University Online Learing,” *International Journal on Informatics Visualization*, vol. 9, no. 3, pp. 1218–1225, 2025, doi: 10.62527/joiv.9.3.4258.

[7] H. Wan, M. Li, Z. Zhong, and X. Luo, “Early Prediction of Student Performance with LSTM-Based Deep Neural Network,” in *Proceedings - International Computer Software and Applications Conference*, H. Shahriar, Y. Teranishi, A. Cuzzocrea, M. Sharmin, D. Towey, M. AKM.J.A., H. Kashiwazaki, Y. J.-J., M. Takemoto, N. Sakib, R. Banno, and S. I. Ahamed, Eds., IEEE Computer Society, 2023, pp. 132–141. doi: 10.1109/COMPSAC57700.2023.00026.

[8] B. Alnasyan, M. Basheri, and M. Alassafi, “The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review,” *Computers and Education: Artificial Intelligence*, vol. 6, 2024, doi: 10.1016/j.caeai.2024.100231.

[9] L. L. Barbare, A. Jurenoks, M. U. Rauba, and Z. Višķere, “Methodology for Analysing LMS Data to Predict Student Dropout Risk in Higher Education,” in *Environment Technology Resources - Proceedings of the 16th International Scientific and Practical Conference*, RTU PRESS, 2025, pp. 57–64. doi: 10.17770/etr2025vol2.8613.

[10] D. Glandorf, H. R. Lee, G. A. Orona, M. Pumptow, R. Yu, and C. Fischer, “Temporal and Between-Group Variability in College Dropout Prediction,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2024, pp. 486–497. doi: 10.1145/3636555.3636906.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

[11] Y. He, X. Lu, P. Fournier-Viger, and J. Z. Huang, “A novel overlapping minimization SMOTE algorithm for imbalanced classification,” *Frontiers of Information Technology and Electronic Engineering*, vol. 25, no. 9, pp. 1266–1281, 2024, doi: 10.1631/FITEE.2300278.

[12] H. Sahlaoui, E. A. A. Alaoui, S. Agoujil, and A. Nayyar, “An empirical assessment of smote variants techniques and interpretation methods in improving the accuracy and the interpretability of student performance models,” *Educ Inf Technol (Dordr)*, vol. 29, no. 5, pp. 5447–5483, 2024, doi: 10.1007/s10639-023-12007-w.

[13] K. S. H. Raslan, A. S. Alsharkawy, and K. R. Raslan, “HHO-SMOTe: Efficient Sampling Rate for Synthetic Minority Oversampling Technique Based on Harris Hawk Optimization,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, pp. 442–453, 2023, doi: 10.14569/IJACSA.2023.0141047.