Mitigating gender bias in STEM study field classification using GRU and LSTM with augmented dataset technique

Devi Fitrianah, Sarah Safitri, Nadzla Andrita Intan Ghayatrie
Department of Computer Science, BINUS Graduate Program–Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Jan 16, 2025
Revised Nov 7, 2025
Accepted Dec 14, 2025

Keywords:

Data augmentation
Gated recurrent units
Gender bias
Long short-term memory

ABSTRACT

This study examines gender bias in artificial intelligence (AI), focusing on the classification of high school students into science, technology, engineering, and mathematics (STEM) and non-STEM fields. Using Indonesian student Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, 11480 data, conditional variational autoencoder (CVAE) and multilabel synthetic minority over-sampling technique (MLSMOTE) were employed for data augmentation to mitigate bias before training gated recurrent unit (GRU) and long short-term memory (LSTM) models for prediction. The combination of MLSMOTE and GRU demonstrated superior performance, achieving accuracies of 93% for female students and 94% for males. These results indicate that MLSMOTE and GRU effectively predict fields of study while addressing gender bias. The findings contribute to advancing fairness in AI systems for education and beyond, ensuring equitable opportunities across diverse applications.

Corresponding Author:

Sarah Safitri
Department of Computer Science, BINUS Graduate Program - Master of Computer Science
Bina Nusantara University Jakarta, 11480, Indonesia
Email: sarah.safitri@binus.ac.id

1. INTRODUCTION

The integration of artificial intelligence (AI) into decision-making has transformed sectors such as healthcare [1], finance [2], education, and employment-domains critical to youth academic and career pathways. Despite its potential, AI raises concerns about fairness and bias, particularly gender bias [3], [4], which can reinforce existing disparities. In education, gender bias in AI has significant implications for predicting academic choices between science, technology, engineering, and mathematics (STEM) and non-STEM fields. Women remain underrepresented, constituting only 35% of STEM graduates globally [5]. Biased datasets may lead AI models to favor male students for STEM predictions, perpetuating inequality. As AI increasingly influences educational guidance, addressing such biases is vital to ensure equitable opportunities.

Studies show that AI can both mitigate and amplify bias depending on design and implementation [6]. Human-centered AI (HAI) emphasizes fairness through inclusive design, data evaluation, and algorithmic adaptation. Rekabsaz and Schedl [7] introduces a framework to measure gender bias in information retrieval (IR) models, revealing that pre-trained embeddings can amplify bias through transfer learning. These findings underscore the importance of fairness in AI applications across education and other fields [8].

While prior studies have explored academic pathway prediction using methods such as random forest [9] and artificial neural networks (ANN) with Adam optimization [10], most neglect gender balance or bias evaluation. Other works, such as [11], apply clustering to assess gender equality efforts in education but focus on benchmarking rather than mitigating bias in predictive models.

This study investigates gender bias in classifying students into STEM and non-STEM fields using neural network (NN) architectures-gated recurrent units (GRU) and long short-term memory (LSTM) networks [12]. Challenges include small, confidential datasets and label imbalance across gender and study fields, which hinder model generalization.

To address these challenges, conditional variational autoencoder (CVAE) and multilabel synthetic minority over-sampling technique (MLSMOTE) are employed for data augmentation. CVAE generates realistic synthetic data, while MLSMOTE balances minority classes in multi-label datasets. Both are integrated into GRU and LSTM models to evaluate their effectiveness in mitigating gender and field-related bias. By examining bias in STEM/non-STEM classification, this study contributes to the discourse on fairness and inclusivity in AI. The results hold implications for ensuring equitable educational outcomes as AI becomes increasingly embedded in decision-making systems.

2. RESEARCH METHOD

2.1. Data collection and characteristics

In Indonesia, high school (SMA) spans three years (grades 10-12) with 13 subjects taught over six semesters, including mathematics, science, language, and social studies. While STEM subjects are core components, integrated STEM curricula are not yet common. Students choose specialized subjects based on their career goals and academic interests. For this study, data collection began with an introduction letter submitted to several schools to obtain permission for access. Due to confidentiality, only limited academic data were processed to ensure student privacy. Additionally, a questionnaire gathered students' academic records and chosen university majors. The collected data provide insights into student performance, subject diversity, and patterns in major selection across Indonesian high schools.

2.2. Major categorization

The classification of higher education majors into STEM and non-STEM fields was conducted using a keyword-based approach. Lists of STEM and non-STEM keywords were compiled from academic terminology and existing educational classifications. STEM keywords included "Engineering," "Biology," "Physics," "Mathematics," "Computer Science," and related terms, while non-STEM keywords included "Arts," "Business," "Economics," "Literature," and "Psychology." Each major in the dataset was programmatically analyzed for these keywords. Majors containing STEM terms were classified as STEM, and those with non-STEM terms as non-STEM. Ambiguous cases were reviewed manually based on curriculum focus and career pathways. This systematic approach ensured accurate categorization and provided a clear foundation for further analysis of students' academic choices.

2.3. Pre-processing

Data cleaning was conducted to correct inconsistencies and errors, including handling missing values, fixing typographical errors, and ensuring uniform data formats [13]. Missing values were imputed appropriately, errors were corrected manually or programmatically, and duplicate records were removed. To standardize grades across schools, Z-scores were computed by subtracting each school's mean grade from individual grades and dividing by the school's standard deviation. This normalization ensured comparability across institutions with different grading systems [14]. The dataset was then divided into training and testing sets using stratified sampling to maintain the same proportion of target variables in both. This approach preserved class balance, enhancing the reliability and validity of the predictive models [15], [16].

2.4. Data augmentation

Data augmentation was applied to address class imbalance and enhance dataset diversity, particularly for underrepresented groups [17]. Two methods were used: CVAE and MLSMOTE. CVAE generated synthetic samples resembling the original data distribution while conditioning on attributes such as gender and field of study. By encoding and decoding data through a latent space, CVAE produced realistic samples that maintained statistical consistency [18]. This approach effectively increased representation-especially of female students in STEM-thereby improving balance and diversity. MLSMOTE further mitigated class imbalance by generating synthetic samples through interpolation between minority class instances while considering multilabel relationships [19]. This method ensured that new samples captured the complexity of multiple class labels, contributing to a more balanced and representative dataset.

2.5. Classification

The classification phase of the research employed advanced neural network architectures, specifically GRU and LSTM networks, to predict student success in STEM fields based on their academic data. GRU was selected for its efficiency in processing sequential data and mitigating the vanishing gradient problem common in traditional recurrent neural networks (RNNs) [20]. It employs reset and update gates to control information flow, enabling it to retain or discard information as needed. This simplified structure makes GRU computationally lighter than LSTM while effectively capturing long-term dependencies [21], [22]. The model was trained on features such as semester-wise academic performance, demographic data, and major classification. Its layered architecture captured temporal patterns in student records [23], and performance was evaluated using accuracy, precision, recall, and F1-score.

LSTM, also designed to handle sequential data, was employed for its strong ability to learn and preserve long-term dependencies [24]. It includes input, forget, and output gates that allow selective retention of relevant information over extended periods [25]. Trained on the same dataset as GRU, the LSTM model used multiple sequential and dense layers to capture complex temporal relationships in academic performance. Both architectures demonstrated robust predictive capability, offering valuable insights into factors influencing students' success in STEM fields.

2.6. Evaluation

In this experimental study, the performance of the algorithms was evaluated using metrics such as accuracy, precision, recall, and F1-score, along with confusion matrices and loss evaluations to identify classification issues. Accuracy measures the overall correctness of predictions, while precision represents the proportion of correctly predicted positive observations. Recall (sensitivity) assesses the model's ability to correctly identify positive cases, and the F1-score provides a harmonic balance between precision and recall.

In addition, fairness metrics were employed to assess gender-related performance. The accuracy difference metric measures the gap in prediction accuracy between male and female students, where a large difference may indicate gender bias. The prediction distribution by gender metric examines the proportion of predicted STEM and non-STEM outcomes for each gender, identifying potential imbalances in class assignments. A skewed distribution may reveal biases in the training data or model behavior, emphasizing the need for balanced and fair predictions across genders.

2.7. Experiment setting

The experiments were conducted using the following four combinations of data augmentation methods and classification models, with each setup evaluated using 10-fold cross-validation:

i) CVAE with GRU: the synthetic data generated by CVAE was combined with the original dataset and used to train the GRU model. This combination aimed to evaluate the impact of CVAE-augmented data on the performance of the GRU architecture.
ii) CVAE with LSTM: the CVAE-augmented dataset was used to train the LSTM model, allowing for an assessment of how CVAE-generated data influenced the LSTM model's performance.
iii) MLSMOTE with GRU: the MLSMOTE-augmented dataset was employed to train the GRU model, examining the effects of MLSMOTE on the GRU model's predictive capabilities.
iv) MLSMOTE with LSTM: the MLSMOTE-augmented dataset was used to train the LSTM model, providing a comparison of MLSMOTE's effectiveness when paired with the LSTM architecture.

3. RESULTS AND DISCUSSION

In this section, we present the results of our experiments evaluating the impact of data augmentation techniques on the performance of GRU and LSTM models. Using CVAE and MLSMOTE, we augmented the original dataset and trained both models on these enhanced datasets. The goal was to assess the effectiveness of these techniques in improving the models' predictive capabilities, especially for imbalanced datasets. Performance was measured using accuracy, precision, recall, and F1-score. The following sections detail the results for each combination of augmentation technique and model.

3.1. Model performance

The performance metrics highlight the influence of different architectures and augmentation techniques on classification outcomes. Table 1 details the results for all experiments. In Experiment 1 (CVAE+GRU), the model achieved 0.80 accuracy and F1 score, with a high recall of 0.96 but lower precision (0.69), indicating strong sensitivity but more false positives. Experiment 2 (CVAE+LSTM) yielded slightly lower performance-0.78 accuracy, 0.72 F1, 0.62 precision, and 0.86 recall-showing reduced learning effectiveness. MLSMOTE substantially improved results. Experiment 3 (MLSMOTE+GRU) achieved the best performance, with 0.94 accuracy, 0.92 precision, 0.96 recall, and 0.93 F1, reflecting balanced and reliable predictions. Experiment 4 (MLSMOTE+LSTM) followed closely, scoring 0.90 accuracy, 0.84 precision, 0.94 recall, and 0.88 F1.

Overall, MLSMOTE effectively mitigated class imbalance and enhanced both precision and recall. GRU-based models consistently outperformed LSTM counterparts, suggesting superior handling of temporal dependencies. These findings underscore the importance of selecting suitable augmentation methods and architectures to optimize classification performance.

Table 1. Experiment result metric

Metric	Average value			
	CVAE		MLSMOTE	
	GRU	LSTM	GRU	LSTM
Accuracy	0.80	0.78	0.94	0.90
Precision	0.69	0.62	0.92	0.84
Recall	0.96	0.86	0.96	0.94
F1 Score	0.80	0.72	0.93	0.88

3.2. Training and validation performance

The GRU-CVAE model shows in Figure 1 have a significant decrease in training and validation loss across all 10 folds, with stabilization around epoch 10. Folds 1, 3, 5, 8, and 10 demonstrate consistent performance with minimal fluctuations, indicating good generalization. Some folds, such as 2, 4, 6, 7, and 9, show minor validation loss fluctuations, suggesting potential overfitting. Overall, the model performs well with minor overfitting in some folds.

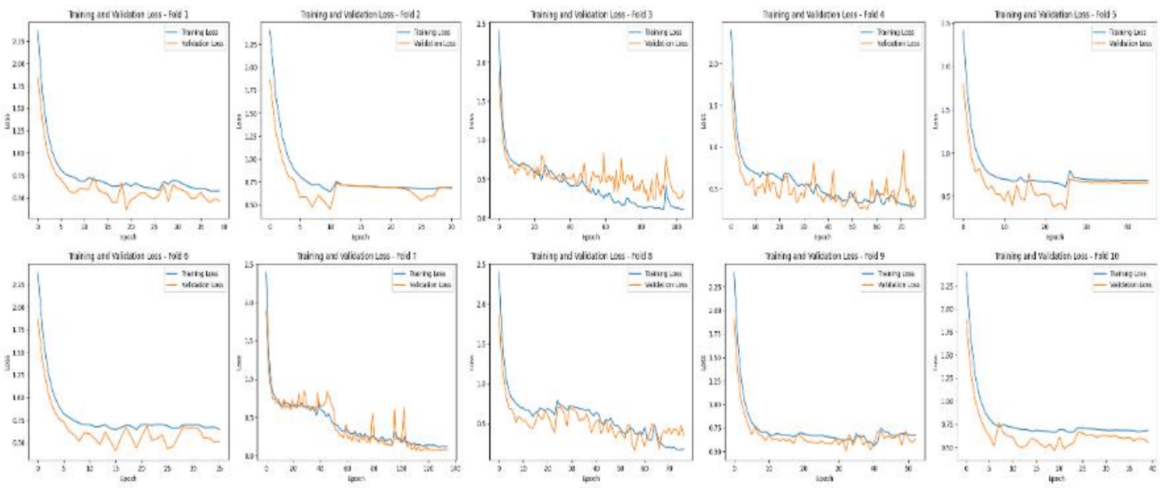

Figure 1. CVAE with GRU training and validation loss

Most folds show in Figure 2 have a smooth decrease in both training and validation loss, with minimal generalization gaps, indicating effective learning and robust performance. However, a few folds, such as Fold 4 and Fold 8, exhibit minor fluctuations in validation loss, likely caused by sensitivity to specific validation subsets. Overall, the model generalizes well across the folds, demonstrating stable and reliable performance.

The GRU-MLSMOTE model in Figure 3 also exhibits a steady decrease in training and validation loss across all folds, stabilizing around epoch 10. Overall fold show stable performance, while folds 2, 6, and 7, exhibit validation loss fluctuations, hinting at potential overfitting. Overall, the model shows consistent performance with some minor overfitting in a few folds.

MLSMOTE-LSTM in Figure 4, validation loss fluctuations likely stem from extended training or data variability. Despite these spikes, the model generalizes well across folds. Overall, all models show stable learning by epoch 10, with GRU and LSTM performing robustly under both CVAE and MLSMOTE, though minor validation loss variations suggest slight overfitting in some folds.

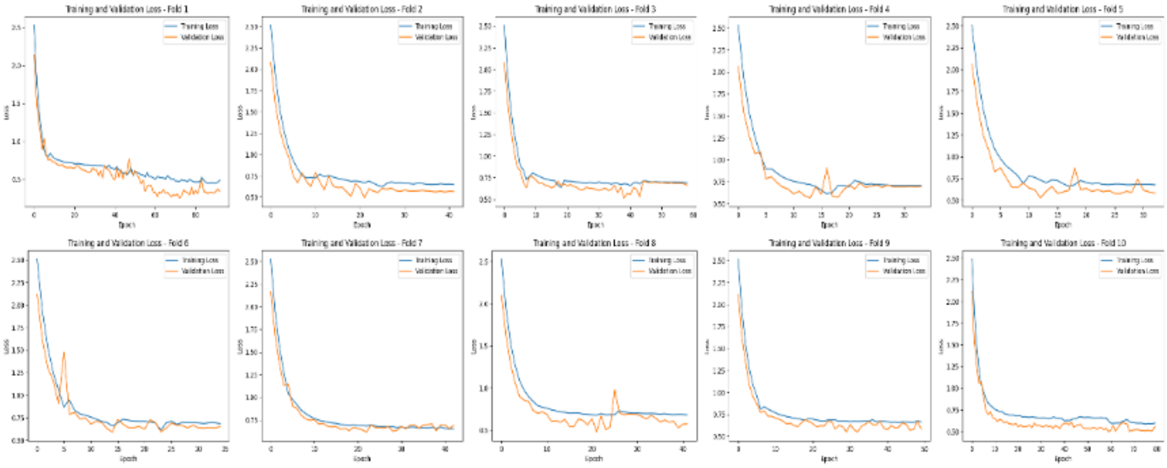

Figure 2. CVAE with LSTM training and validation loss

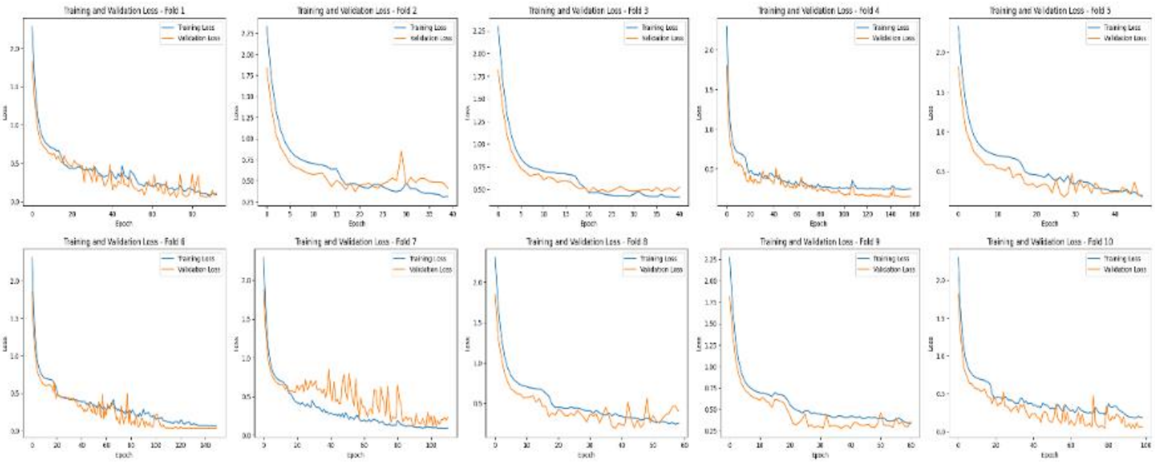

Figure 3. MLSMOTE with GRU training and validation loss

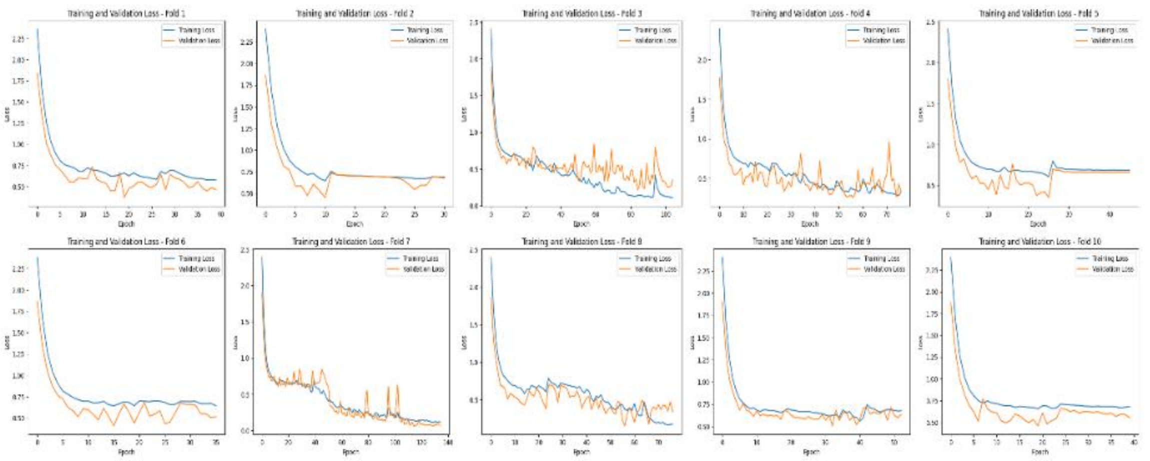

Figure 4. MLSMOTE with LSTM training and validation loss

3.3. Fairness performance

To evaluate fairness in model performance across genders, we compared the average accuracy for male and female groups across four experimental setups. The results are summarized in Table 2. From Table 2 it is evident that models trained using the MLSMOTE technique consistently achieve higher fairness metrics, with relatively balanced accuracies across both genders. In contrast, models utilizing CVAE exhibit larger disparities in performance, suggesting room for improvement in achieving fairness. The MLSMOTE with GRU model achieves the highest overall accuracy for both male and female groups, reflecting its robustness and fairness.

Building on the fairness evaluation, the next step analyzes the predictive distributions to identify potential biases or misalignments with the original data. While fairness metrics show overall gender parity, distribution analysis reveals whether models accurately capture group-specific patterns or exhibit systematic over- or under-predictions, ensuring both fairness and data fidelity.

In this experiment, the CVAE with GRU architecture (Figure 5, left) shows discrepancies between predicted and original distributions. Predictions for non-STEM males are underestimated, while non-STEM females are overestimated. In the STEM category, male predictions align closely with originals, whereas female predictions slightly exceed them, indicating a bias toward over-predicting females and under-predicting males. The CVAE with LSTM architecture (Figure 5, right) consistently overestimates across genders and categories, suggesting a bias toward higher predicted counts. This may stem from class imbalance or imperfect reconstruction of the data distribution. Fine-tuning or regularization could help reduce this systematic overprediction.

Table 2. Average accuracy within gender

Experiment	Average male accuracy	Average female accuracy
CVAE with GRU	0.83	0.72
CVAE with LSTM	0.79	0.73
MLSMOTE with GRU	0.94	0.93
MLSMOTE with LSTM	0.89	0.88

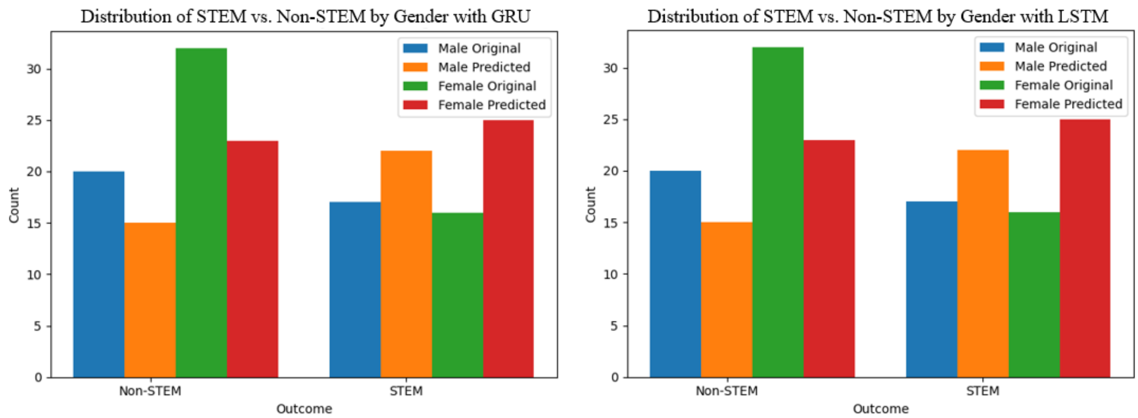

Figure 5. Experiment 1 - CVAE with GRU and LSTM prediction distribution

With MLSMOTE applied for data balancing and a GRU-based model (Figure 6, left), the predicted and original distributions show strong alignment. In the non-STEM category, predictions for both genders closely match the originals, indicating reduced bias. Similarly, STEM predictions align well across genders, demonstrating that MLSMOTE effectively mitigates class imbalance and enhances GRU model generalization. In contrast, the MLSMOTE+LSTM model (Figure 6, right) performs consistently for females across both categories, with minimal gaps between predicted and original distributions. However, for males, predictions are slightly underestimated in non-STEM and overestimated in STEM, revealing residual imbalance. While overall performance remains strong, further tuning may be needed to improve male prediction accuracy across both classes.

To extend the analysis, we also examine the model's performance without augmentation. This comparison aims to evaluate whether the augmentation process contributes to the observed discrepancies. Figure 7 compares GRU and LSTM models in predicting STEM versus non-STEM outcomes by gender without data augmentation, revealing clear model biases. Both models show discrepancies for female students-the GRU underpredicts non-STEM and overpredicts STEM, while the LSTM shows similar trends but aligns slightly better for males. This indicates stronger performance for male predictions and weaker accuracy for females, reflecting data imbalance. Without augmentation, both models tend to overfit the dominant male data, failing to capture female-specific patterns. These findings emphasize the importance of data augmentation to improve model fairness and robustness.

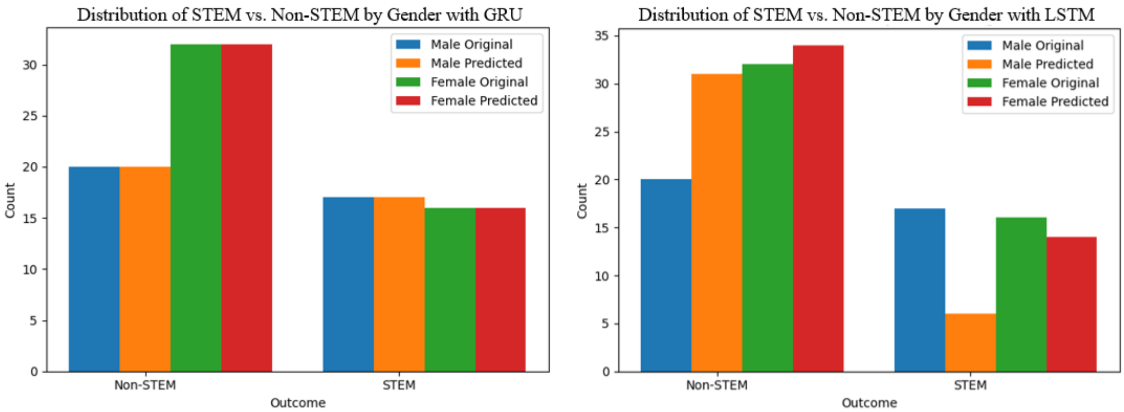

Figure 6. Experiment 2: MLSMOTE with GRU and LSTM prediction distribution

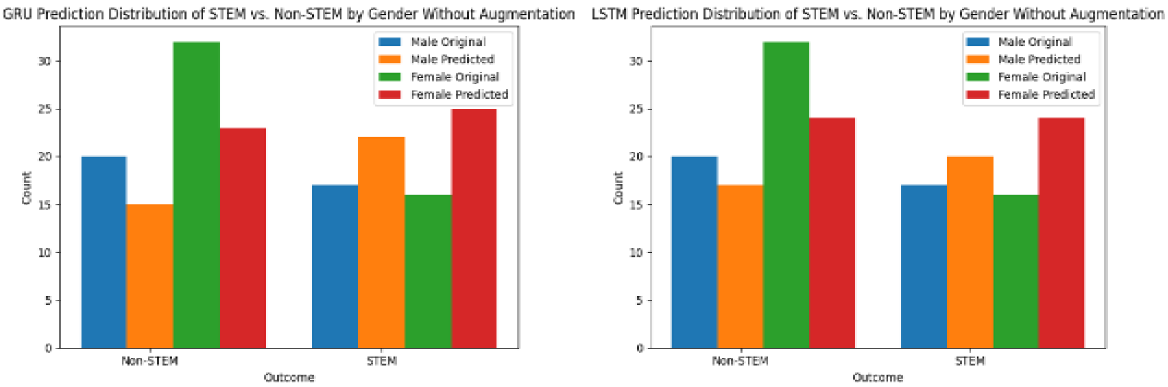

Figure 7. MLSMOTE with LSTM prediction distribution

4. CONCLUSION

The experiments demonstrated the impact of data augmentation techniques and classification models on performance, fairness, and prediction distributions. MLSMOTE combined with GRU consistently outperformed other setups, achieving the highest accuracy (0.94) and F1 score (0.93) while demonstrating balanced performance across genders, with male and female accuracies at 0.94 and 0.93, respectively. In contrast, CVAE-based models exhibited larger disparities between genders, with noticeable biases in prediction distributions. For instance, the CVAE with GRU over-predicted females in the non-STEM category and under-predicted males, highlighting challenges in achieving fairness. Similarly, the CVAE with LSTM displayed overestimation across genders and classes, suggesting limitations in capturing the original data distribution. On the other hand, MLSMOTE effectively addressed class imbalances, with GRU achieving closer alignment between predicted and original distributions for both genders. However, MLSMOTE with LSTM showed discrepancies in male predictions, with underestimation in the non-STEM category and slight overestimation in STEM. Overall, the results underscore MLSMOTE's ability to improve performance and fairness, particularly when paired with GRU, though additional tuning may be needed for other configurations to address residual biases.

ACKNOWLEDGMENTS

The authors would like to acknowledge the use of OpenAI's ChatGPT as a supportive tool during the debugging and development process.

FUNDING INFORMATION

This work is supported by Research and Technology Transfer Office, Bina Nusantara University as a part of Bina Nusantara University's International Research Grant entitled An In-depth Analysis of Female Representation and Employability in STEM based on Machine Learning Framework with contract number: 069C/VRRTT/III/2024 and contract date: March 18, 2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Devi Fitrianah	✓	✓		✓	✓	✓	✓		✓	✓		✓	✓	✓
Sarah Safitri	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	
Nadzla Andrita Intan Ghayatrie	✓			✓	✓		✓		✓	✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [SS]. The data are not publicly available due to privacy or ethical restrictions.

REFERENCES

[1] E. Jussupow, K. Spohrer, A. Heinzl, and J. Gawlitza, "Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence," *Information Systems Research*, vol. 32, no. 3, pp. 713-735, Sep. 2021, doi: 10.1287/isre.2020.0980.

[2] M. Stone *et al.*, "Artificial intelligence (AI) in strategic marketing decision-making: a research agenda," *The Bottom Line*, vol. 33, no. 2, pp. 183-200, Apr. 2020, doi: 10.1108/BL-03-2020-0022.

[3] E. Ferrara, "Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, p. 3, Dec. 2023, doi: 10.3390/sci6010003.

[4] S. O'Connor and H. Liu, "Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities," *AI & SOCIETY*, vol. 39, no. 4, pp. 2045-2057, Aug. 2024, doi: 10.1007/s00146-023-01675-4.

[5] United Nations Educational Scientific and Cultural Organization (UNESCO), "Global education monitoring report 2024 - Gender report: technology on her terms." 2024, [Online]. Available: https://www.iau-hesd.net/action/global-education-monitoring-report-2024-gender-report-technology-her-terms.

[6] S. J. H. Yang, H. Ogata, T. Matsui, and N.-S. Chen, "Human-centered artificial intelligence in education: seeing the invisible through the visible," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100008, 2021, doi: 10.1016/j.caeai.2021.100008.

[7] N. Rekabsaz and M. Schedl, "Do neural ranking models intensify gender bias?," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2020, pp. 2065-2068, doi: 10.1145/3397271.3401280.

[8] Y. Dong, N. Liu, B. Jalaian, and J. Li, "EDITS: modeling and mitigating data bias for graph neural networks," in *Proceedings of the ACM Web Conference 2022*, Apr. 2022, pp. 1259-1269, doi: 10.1145/3485447.3512173.

[9] F. Cruz-Jesus *et al.*, "Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country," *Heliyon*, vol. 6, no. 6, p. e04081, Jun. 2020, doi: 10.1016/j.heliyon.2020.e04081.

[10] Ş. Aydoğdu, "Predicting student final performance using artificial neural networks in online learning environments," *Education and Information Technologies*, vol. 25, no. 3, pp. 1913-1927, May 2020, doi: 10.1007/s10639-019-10053-x.

[11] V. Tokar, D. Tyshchenko, T. Franchuk, V. Makoiedova, and A. Lotariev, "Using cluster analysis for revealing gender equality patterns in EU ICT education and employment," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 16, pp. 6691–6702, 2023.

[12] B. C. Mateus, M. Mendes, J. T. Farinha, R. Assis, and A. M. Cardoso, "Comparing LSTM and GRU models to predict the condition of a pulp paper press," *Energies*, vol. 14, no. 21, p. 6958, Oct. 2021, doi: 10.3390/en14216958.

[13] F. Ridzuan and W. M. N. Wan Zainon, "A review on data cleansing methods for big data," *Procedia Computer Science*, vol. 161, pp. 731-738, 2019, doi: 10.1016/j.procs.2019.11.177.

[14] C. Andrade, "Z scores, standard scores, and composite test scores explained," *Indian Journal of Psychological Medicine*, vol. 43, no. 6, pp. 555-557, Nov. 2021, doi: 10.1177/02537176211046525.

[15] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, "Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset," *Electronics*, vol. 12, no. 21, p. 4423, Oct. 2023, doi: 10.3390/electronics12214423.

[16] Y. Xu and R. Goodacre, "On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *Journal of Analysis and Testing*, vol. 2, pp. 249-262, 2018, doi: 10.1007/s41664-018-0068-2.

[17] R. Escobar Díaz Guerrero, L. Carvalho, T. Bocklitz, J. Popp, and J. L. Oliveira, "A data augmentation methodology to reduce the class imbalance in histopathology images," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 4, pp. 1767-1782, Mar. 2024, doi: 10.1007/s10278-024-01018-9.

[18] K. Sohn, X. Yan, and H. Lee, "Learning structured output representation using deep conditional generative models," *Advances in Neural Information Processing Systems*, vol. 2015-January, pp. 3483-3491, 2015.

[19] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "MLSMOTE: approaching imbalanced multilabel learning through synthetic instance generation," *Knowledge-Based Systems*, vol. 89, pp. 385-397, Nov. 2015, doi: 10.1016/j.knosys.2015.07.019.

[20] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," *MethodsX*, vol. 13, p. 102946, Dec. 2024, doi: 10.1016/j.mex.2024.102946.

[21] E. Ahmadzadeh, H. Kim, O. Jeong, N. Kim, and I. Moon, "A deep bidirectional LSTM-GRU network model for automated ciphertext classification," *IEEE Access*, vol. 10, pp. 3228-3237, 2022, doi: 10.1109/ACCESS.2022.3140342.

[22] K. E. ArunKumar, D. V. Kalaga, C. Mohan Sai Kumar, M. Kawaji, and T. M. Brenza, "Comparative analysis of gated recurrent units (GRU), long short-term memory (LSTM) cells, autoregressive integrated moving average (ARIMA), seasonal autoregressive integrated moving average (SARIMA) for forecasting COVID-19 trends," *Alexandria Engineering Journal*, vol. 61, no. 10, pp. 7585-7603, Oct. 2022, doi: 10.1016/j.aej.2022.01.011.

[23] F. Alqahtani *et al.*, "Hybrid deep learning algorithm for forecasting SARS-CoV-2 daily infections and death cases," *Axioms*, vol. 11, no. 11, p. 620, Nov. 2022, doi: 10.3390/axioms11110620.

[24] K. P. Saini and A. Sharma, "A comparison between long short-term memory and prophet for time series analysis and forecasting technique," *Educational Administration Theory and Practices*, Apr. 2024, doi: 10.53555/kuey.v30i4.2816.

[25] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: a comprehensive review of architectures, variants, and applications," *Information*, vol. 15, no. 9, p. 517, Aug. 2024, doi: 10.3390/info15090517.

BIOGRAPHIES OF AUTHORS

Devi Fitrianah is a lecturer and researcher at the Master of Computer Science Department, Bina Nusantara University. She earned her Bachelor's degree from BINUS University (2000) and her Master's and Ph.D. in computer science from Universitas Indonesia (2008, 2015). She joined a sandwich program at Michigan State University (2014) and is currently a fellow researcher at Eureka Robotics Lab, Cardiff Metropolitan University, UK. Her research interests include data mining, machine learning, AI, and applied remote sensing. She can be contacted at email: devi.fitrianah@binus.ac.id.

Sarah Safitri is a graduate student pursuing a Master of Information at Bina Nusantara University. The research interests include the development of recommender systems, with particular emphasis on addressing biases in recommendation algorithms. Additionally, there is a focus on investigating techniques for managing imbalanced data, aiming to enhance the fairness, accuracy, and performance of machine learning models. She can be contacted at email: sarah.safitri@binus.ac.id.

Nadzla Andrita Intan Ghayatrie is a graduate student pursuing a Master of Information Technology degree with a focus on data, machine learning, and artificial intelligence at Bina Nusantara University in Indonesia. Her research interests center around addressing gender gap issues through innovative applications of data science, machine learning, and artificial intelligence. She can be contacted at email: nadzla.ghayatrie@binus.ac.id.