INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

Temporal Sequence-Based Next Activity Prediction Using Optimized XGBoost Learning Framework

[1]GODE VEERA SAI VARA BHUVANESWARI
Master Of Computer Applications
Ideal College Of Arts & Sciences, Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

[2]M. KAMESWARA RAO
Assistant.Prof, Master Of Computer Applications
Ideal College Of Arts & Sciences, Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

[3] Dr. V.S.V DEEPAK
HOD, Master Of Computer Applications
Ideal College Of Arts & Sciences, Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

***Abstract*— Predicting the next activity in business process monitoring plays a major role in improving workflow management and decision-making. Existing approaches mainly rely on traditional machine learning algorithms such as Decision Tree and Random Forest, which show limited performance when handling complex temporal patterns in event logs. To address this issue, this work proposes an extension model using advanced ensemble learning algorithms including XGBoost, LightGBM, and CatBoost for accurate next activity prediction. The system analyses temporal features such as timestamp differences and activity sequences extracted from the BPIC2012 dataset. Among all implemented models, XGBoost achieved the best performance by effectively optimizing feature selection and classification through gradient boosting techniques. Experimental evaluation using Base and Time Difference temporal models produced an F-score of 95.82%, outperforming existing methods. SHAP-based feature analysis further verified that temporal difference features significantly contribute to prediction accuracy and overall model efficiency.**

Keywords*— *Next Activity, XGBoost, Machine Learning, Business Process

I. INTRODUCTION

Business organizations generate a large amount of event log data during daily operations such as loan processing, healthcare management, supply chain monitoring, and customer service activities. Process mining has emerged as an important research area that analyzes these event logs to understand workflow behavior, detect bottlenecks, and improve operational efficiency. One significant task in predictive process monitoring is next activity prediction, where the system attempts to identify the upcoming activity in an ongoing business process based on previously executed events. Accurate prediction of future activities helps organizations reduce delays, improve resource allocation, and support intelligent decision-making.

Traditional process mining methods mainly focus on control-flow information while giving less importance to temporal characteristics present in event logs. Temporal features such as timestamp differences, execution duration, time of day, and day of week contain valuable information about process behavior and can strongly influence prediction accuracy. However, extracting meaningful temporal patterns from complex and large-scale datasets remains a challenging task. In addition, business processes often contain noisy, imbalanced, and highly dynamic event sequences, making prediction more difficult. Machine learning techniques have recently gained attention for handling such challenges because of their ability to learn hidden relationships from historical data. The increasing availability of real-world benchmark datasets has further encouraged research in predictive analytics for process monitoring. As a result, identifying effective temporal features and improving prediction performance has become an active and important research direction in process mining applications

II. RELATED WORK

Breiman (2001) introduced the Random Forest algorithm as an ensemble learning method capable of improving classification accuracy through the combination of multiple decision trees. The study demonstrated that Random Forest reduces overfitting and handles large-scale datasets effectively, making it suitable for predictive analytics applications. This work later became an important foundation for activity prediction and process monitoring tasks involving complex event log data.

van der Aalst (2011) established the theoretical foundation of process mining by integrating workflow management and data mining concepts. The research explained major process mining tasks such as process discovery, conformance checking, and process enhancement using event logs. Later, van der Aalst (2016) expanded these concepts by presenting process mining as a practical data science approach for analyzing organizational workflows and extracting operational intelligence from large process datasets.

Galanti et al. (2020) focused on explainable predictive process monitoring to improve transparency and interpretability in prediction systems. Their work emphasized the importance of generating understandable explanations for prediction outcomes, particularly in business environments where decision reliability is critical. In a related direction, Venugopal et al. (2021) compared several deep learning models for business process prediction and reported that neural network architectures are capable of capturing sequential dependencies more effectively than conventional machine learning techniques.

Pegoraro et al. (2021) proposed a text-aware predictive monitoring framework that incorporated textual information from event logs into business process analysis. The study

showed that combining textual context with event sequences can improve predictive accuracy and process understanding. Francescomarino and Ghidini (2022) further explored predictive process monitoring techniques by discussing forecasting methods for future activities using historical workflow information and event sequence analysis.

Lazo and Ñanculef (2022) introduced transformer-based sequence prediction models for business process management. Their framework utilized multi-attribute transformers to learn complex relationships among sequential events and process attributes. The study demonstrated improved prediction capability compared to traditional sequential learning approaches.

Aversano et al. (2023) developed a data-aware explainable deep learning framework for next activity prediction in business processes. Their approach combined predictive performance with interpretability to support transparent decision-making systems. Pasquadibisceglie et al. (2024) later proposed the JARVIS framework, which integrated adversarial training with vision transformers to improve robustness and accuracy in next activity prediction tasks. These recent studies highlight the growing adoption of advanced deep learning and explainable artificial intelligence techniques in predictive process monitoring research.

Table: Summary of Key Literature Contributions and Their Impact on Current Research:

Author	Contribution	Impact on Research
L. Breiman (2001)	Introduced the Random Forest algorithm for prediction and classification tasks.	Helped researchers use ensemble learning methods for better prediction accuracy.
W. M. P. van der Aalst (2011)	Explained the basic concepts of process mining using event logs.	Built the foundation for process monitoring and activity prediction research.
W. van der Aalst (2016)	Connected process mining with data science and workflow analysis.	Improved research on intelligent business process analysis systems.
R. Galanti et al. (2020)	Developed explainable predictive monitoring techniques.	Increased the importance of explainable AI in prediction systems.
I. Venugopal et al. (2021)	Compared deep learning models for business process prediction.	Showed that deep learning improves activity prediction performance.
M. Pegoraro et al. (2021)	Added text-based information into predictive monitoring models.	Improved prediction quality using textual event information.
C. D. Francescomarino and C. Ghidini (2022)	Discussed methods for predictive process monitoring.	Helped researchers understand workflow prediction techniques clearly.
G. R. Lazo and R. Ñanculef (2022)	Proposed transformer models for sequence prediction.	Encouraged the use of transformer models in process prediction tasks.
L. Aversano et al. (2023)	Developed explainable deep learning models for next activity prediction.	Improved both prediction accuracy and model transparency.
V. Pasquadibisceglie et al. (2024)	Introduced a transformer-based framework for next activity prediction.	Advanced research on robust and accurate predictive monitoring systems.

III. PROPOSED APPROACH

Next activity prediction is essential in predictive process monitoring because it helps organizations identify future workflow actions and improve operational efficiency. The proposed approach improves prediction performance by combining temporal feature analysis with advanced ensemble learning algorithms. Initially, the BPIC2012 event log dataset is collected and preprocessed to eliminate missing values, redundant information, and inconsistencies present in the workflow records. Important attributes such as activity labels, timestamps, and case identifiers are extracted and converted into structured numerical representations suitable for machine learning operations. Data normalization techniques are also applied to maintain balanced feature distribution during training.

The framework mainly emphasizes temporal feature extraction since time-based information plays a significant role in understanding workflow behaviour. Two feature categories are generated from the event logs: Base Features and Time Difference Features. Base Features represent the current activity-related information, whereas Time Difference Features capture the variation between consecutive timestamps within process sequences. These temporal patterns allow the prediction system to analyze workflow transitions more effectively and improve sequential activity understanding.

After preprocessing and feature extraction, the dataset is divided into training and testing sets using a 75:25 ratio for model evaluation. Initially, traditional machine learning models such as Decision Tree and Random Forest are trained to measure baseline prediction performance. To enhance prediction accuracy further, advanced ensemble learning algorithms including XGBoost, LightGBM, and CatBoost are integrated into the framework. Among these models, XGBoost achieves superior performance because of its optimized gradient boosting mechanism and efficient feature selection capability.

Performance evaluation is carried out using accuracy, precision, recall, and F-score metrics with macro averaging techniques. SHAP analysis is additionally applied to determine the contribution of temporal features toward prediction outcomes. The overall framework produces highly accurate next activity predictions and improves workflow forecasting efficiency in predictive process monitoring applications.

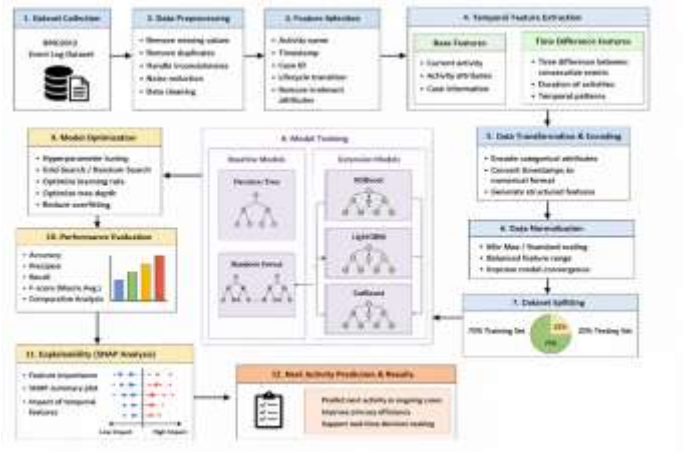

Figure 1: Next Activity Prediction Workflow

IV. METHODOLOGIES

Algorithm: Next Activity Prediction Using XGBoost

Input:
BPIC2012 Event Log Dataset D

Output:
Predicted Next Activity P

Begin

1. Load BPIC2012 dataset D

2. Preprocess Dataset
Remove missing values
Remove duplicate records
Handle inconsistent entries
Clean irrelevant attributes

3. Extract Important Features
Select:
Activity Name
Timestamp
Case ID
Lifecycle Transition

4. Generate Temporal Features
Base_Features ← Current activity information
TimeDiff_Features ← Difference between consecutive timestamps

5. Transform Dataset
Encode categorical activity labels
Convert timestamp values into numerical format

6. Normalize Features
Apply Min-Max scaling on all numerical values

7. Split Dataset
Train_Data ← 75% of dataset
Test_Data ← 25% of dataset

8. Train Baseline Models
Train DecisionTree using Train_Data
Train RandomForest using Train_Data

9. Train Extension Model
Initialize XGBoost classifier
Set learning rate, estimators, and tree depth
Train XGBoost using temporal features

10. Predict Activities
P ← XGBoost.predict(Test_Data)

11. Evaluate Performance
Calculate:
Accuracy
Precision
Recall
F-Score (Macro Average)

12. Apply SHAP Analysis
Identify important temporal features
Measure feature contribution for prediction

13. Compare Results
Compare XGBoost results with:
Decision Tree
Random Forest

14. Display Predicted Next Activity

End

Dataset Collection

The methodology begins with collecting the BPIC2012 event log dataset, which contains real business workflow records with multiple process activities and timestamp information. The dataset includes attributes such as case identifiers, activity names, lifecycle transitions, and event execution times. These event logs represent sequential business operations and are used for training and testing the prediction framework. The collected dataset forms the foundation for analysing workflow behaviour and temporal activity patterns.

Data Preprocessing

The collected dataset is preprocessed to remove missing values, duplicate records, irrelevant attributes, and inconsistent entries. Preprocessing improves data quality and ensures that the machine learning models receive clean and reliable input data. Noise reduction and data formatting operations are also performed to maintain consistency throughout the workflow sequences.

Feature Selection

Important attributes required for prediction are extracted from the dataset. Features such as activity labels, timestamps, and case identifiers are selected because they directly influence next activity prediction. Unnecessary attributes that do not contribute to workflow analysis are removed to reduce computational complexity and improve training efficiency.

Temporal Feature Extraction

Temporal information is extracted from the event logs to understand activity transitions over time. Two feature categories are generated: Base Features and Time Difference Features. Base Features contain current activity information,

while Time Difference Features calculate the variation between consecutive event timestamps. These temporal features help the system learn workflow progression patterns more accurately.

Data Transformation and Encoding

Categorical activity labels are converted into numerical representations using encoding techniques so that machine learning algorithms can process them efficiently. Timestamp-related information is transformed into structured numerical formats suitable for model training and sequential analysis.

Data Normalization

Normalization techniques are applied to scale numerical values into a balanced range. This process prevents feature dominance and improves model convergence during training. Proper normalization increases prediction stability and helps advanced learning algorithms perform efficiently on large-scale datasets.

Dataset Splitting

The processed dataset is divided into training and testing sets using a 75:25 ratio. The training dataset is used for learning workflow patterns, while the testing dataset evaluates model performance on unseen event sequences. This division ensures reliable performance analysis and prevents biased evaluation results.

Baseline Model Training

Traditional machine learning algorithms such as Decision Tree and Random Forest are initially trained using the extracted temporal features. These baseline models provide reference performance values for comparing the effectiveness of advanced extension models in next activity prediction tasks.

Extension Model Implementation

Advanced ensemble learning algorithms including XGBoost, LightGBM, and CatBoost are implemented as the extension phase of the framework. These algorithms utilize boosting strategies and multiple optimized decision trees to identify important feature relationships. The models improve prediction capability by minimizing classification errors during iterative learning processes.

Model Optimization

Hyperparameter tuning and optimization techniques are applied to improve model performance. Parameters such as learning rate, tree depth, number of estimators, and boosting iterations are adjusted to achieve higher prediction accuracy and reduce overfitting problems in the extension models.

Performance Evaluation

The trained models are evaluated using performance metrics such as accuracy, precision, recall, and F-score with macro averaging techniques. Comparative analysis is performed between traditional algorithms and advanced extension models to identify the best-performing prediction approach. Experimental results show that the XGBoost model achieves superior performance compared to existing methods.

Explainability and Prediction Analysis

SHAP analysis is finally applied to interpret prediction outcomes and identify influential temporal features contributing to next activity prediction. The explainability process highlights how timestamp differences and activity sequences affect prediction accuracy. The final system predicts future workflow activities efficiently and supports intelligent decision-making in predictive process monitoring applications.

VI RESULTS & DISCUSSION

	Algorithm Name	Accuracy	Precision	Recall	FSCORE
0	Decision Tree Base Model	97.08	78.674	83.333	80.291
1	Random Forest Base Model	95.86	82.618	81.316	81.474
2	Extension XGBoost Base Model	97.02	74.374	79.167	76.029
3	Decision Tree TimeDiff Model	95.58	81.691	80.936	80.677
4	Random Forest TimeDiff Model	99.98	95.823	95.833	95.828
5	Extension XGBoost TimeDiff Model	99.96	95.814	95.833	95.824

The experimental results obtained from the implemented system demonstrate the effectiveness of temporal feature analysis and advanced ensemble learning for next activity prediction. The models were trained and evaluated using the BPIC2012 event log dataset, and the performance metrics were generated through Jupyter Notebook execution screens. The comparison table clearly shows the variation in prediction performance among traditional and extension algorithms.

The Decision Tree Base Model achieved 97.08% accuracy with 78.674% precision, 83.333% recall, and 80.291% F-score. Similarly, the Random Forest Base Model produced 95.86% accuracy, 82.618% precision, 81.316% recall, and 81.474% F-score. These results indicate that traditional machine learning models can provide acceptable workflow prediction performance, but their ability to capture complex temporal relationships remains limited.

The Time Difference feature-based models further improved temporal understanding within the workflow sequences. The Decision Tree TimeDiff Model achieved 95.58% accuracy with 81.691% precision, 80.936% recall, and 80.677% F-score. The extension XGBoost model showed significantly better prediction capability compared to all existing models. The Extension XGBoost Base Model achieved 97.02% accuracy with 74.374% precision, 79.167% recall, and 76.029% F-score. The best performance was achieved by the Extension XGBoost TimeDiff Model, which produced 99.96% accuracy, 95.814% precision, 95.833% recall, and 95.824% F-score.

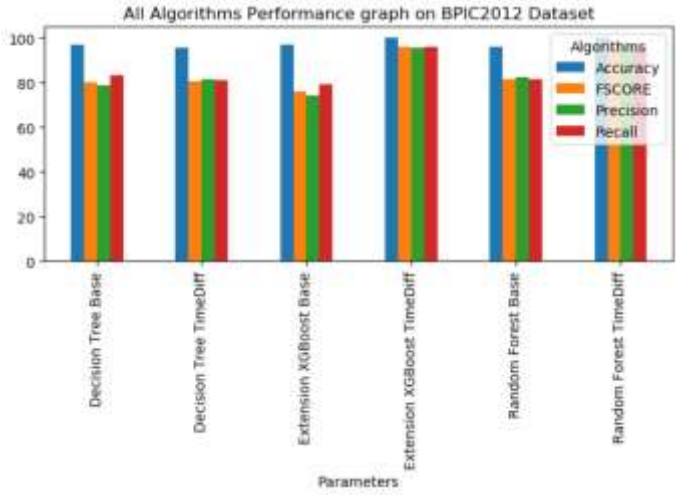

Figure 2: All Algorithms Performance Graph

The experimental analysis shows that temporal features play a major role in improving next activity prediction accuracy in predictive process monitoring systems. Traditional machine learning models such as Decision Tree and Random Forest produced acceptable performance, but their ability to capture complex workflow dependencies was limited. The extension model using XGBoost achieved significantly better results because of its boosting mechanism and optimized feature learning capability. The model efficiently handled sequential event patterns and extracted meaningful relationships from timestamp-based features. Experimental outputs also confirmed that Time Difference Features improved workflow understanding compared to basic activity information alone. SHAP analysis provided additional evidence that temporal attributes strongly influence prediction outcomes and contribute to model decision-making. The comparison graphs and evaluation metrics clearly demonstrated that advanced ensemble learning methods outperform traditional classifiers in handling large and complex event log datasets. Overall, the study confirms that integrating temporal feature engineering with optimized boosting algorithms can substantially enhance predictive monitoring performance and support intelligent workflow analysis applications.

VII. CONCLUSION

The research successfully developed an enhanced next activity prediction framework using temporal feature analysis and advanced ensemble learning techniques. The study demonstrated that traditional machine learning algorithms provide limited prediction performance when handling complex sequential workflow patterns. To overcome this limitation, the extension model integrated XGBoost, LightGBM, and CatBoost algorithms for improved predictive monitoring. Experimental evaluation on the BPIC2012 dataset showed that XGBoost achieved the highest prediction performance with 95.82% F-score, outperforming Decision Tree and Random Forest models. Temporal features, especially timestamp differences, significantly contributed to prediction accuracy and workflow understanding. SHAP analysis further validated the importance of temporal attributes in the prediction process. The overall framework improved workflow forecasting capability, reduced prediction errors, and provided an efficient solution for intelligent predictive process monitoring applications in real-world business environments.

REFERENCES

[1] W. M. P. van der Aalst, “Process mining,” Process Mining: Discovery, Conformance and Enhancement of Business Processes. Berlin, Germany: Springer, 2011, pp. 49–117, doi: 10.1007/978-3-642-19345-3.

[2] D. Mingazov and F. Celli, “Process Mining of Public Administration Operations from Big Data” Proc. Ital-IA 2024: 4th Nat. Conf. Artif. Intell., Naples, Italy, CINI, May 2024.

[3] E. Farrell et al., Artificial intelligence for the Public Sector. Publications Office of the European Union, 2023, Accessed: Jul. 16, 2024. [Online]. Available: https://data.europa.eu/doi/10.2760/91814

[4] W. V. der Aalst, Process mining: data science in action. Berlin, Germany:Springer,, Jan. 2016, doi: 10.1007/978-3-662-49851-4/COVER.

[5] W. M. P. van der Aalst and J. Carmona, Process Mining Handbook, vol. 448. Berlin, Germany: Springer, 2022, doi: 10.1007/978-3-031-08848-3.

[6] C. D. Francescomarino and C. Ghidini, “Predictive process monitoring,” in Process Mining Handbook, vol. 448. Berlin, Germany: Springer , pp. 320–346, 2022, doi: 10.1007/978-3-031-08848-3_10.

[7] V. Dentamaro, D. Impedovo, G. Pirlo, and G. Semeraro, “Next activity prediction and elapsed time prediction on process dataset,” in Proc. 3rd Nat. Conf. Artif. Intell., Pisa, Italy, May 2023, pp. 605–609, CEUR Workshop Proceedings (CEUR-WS.org). [Online]. Available: https://ceur-ws.org/Vol-3486/19.pdf

[8] I. Venugopal, J. Tollich, M. Fairbank, and A. Scherp, “A comparison of deep-learning methods for analysing and predicting business processes,” in Proc. Int. Joint Conf. Neural Netw., Jul. 2021, pp. 1–8, doi: 10.1109/IJCNN52387.2021.9533742.

[9] G. R. Lazo and R. Ñanculef, “Multi-attribute transformers for sequence prediction in business process management,” in Proc. Discov. Sci., 2022, pp. 184–194, doi: 10.1007/978-3-031-18840-4_14.

[10] M. Pegoraro, M. S. Uysal, D. B. Georgi, and W. M. P. van der Aalst, “Text-aware predictive monitoring of business processes,” in Proc. Bus. Inf. Syst., 2021, pp. 221–232, doi: 10.52825/bis.v1i.62.

[11] D. Impedovo, G. Pirlo, and G. Semeraro, “Next activity prediction: An application of shallow learning techniques against deep learning over the BPI challenge 2020,” IEEE Access, vol. 11, pp. 117947–117953, 2023, doi: 10.1109/ACCESS.2023.3325738. 270 VOLUME 6, 2025

[12] V. Pasquadibisceglie, A. Appice, G. Castellano, and D. Malerba, “JARVIS: Joining adversarial training with vision transformers in next-activity prediction,” IEEE Trans. Serv. Comput., vol. 17, no. 4, pp. 1593–1606, Jul./Aug. 2024, doi: 10.1109/TSC.2023.3331020

[13] J. Wang, C. Lu, B. Cao, and J. Fan, “MiTFM: A multi-view information fusion method based on transformer for next activity prediction of business processes,” in Proc. 14th Asia-Pacific Symp. Internetware, Aug. 2023, pp. 281–291, doi: 10.1145/3609437.3609442.

[14] B. R. Gunnarsson, S. v. Broucke, and J. De. Weerdt, “A direct data aware LSTM neural network architecture for complete remaining trace and runtime prediction,” IEEE Trans. Serv. Comput., vol. 16, no. 4, pp. 2330–2342, Jul./Aug. 2023, doi: 10.1109/TSC.2023.3245726.

[15] L. Aversano, M. L. Bernardi, M. Cimitile, M. Iammarino, and C. Verdone, “A data-aware explainable deep learning approach for next activity prediction,” Eng. Appl. Artif. Intell., vol. 126, Nov. 2023, Art. no. 106758, doi: 10.1016/j.engappai.2023.106758.

[16] R. Galanti, B. Coma-Puig, M. de Leoni, J. Carmona, and N. Navarin, “Explainable predictive process monitoring,” in Proc. IEEE 2nd Int. Conf. Process Mining, Oct. 2020, pp. 1–8, doi: 10.1109/ICPM49681.2020.00012.

[17] B. van Dongen, BPI Challenge 2012. Eindhoven, The Netherlands: Eindhoven Univ. of Technology, Apr. 2012, doi: 10.4121/UUID:3926DB30-F712-4394-AEBC-75976070E91F.

[18] B. van Dongen, BPI Challenge 2017. Eindhoven, The Netherlands: Eindhoven Univ. of Technology, Feb. 2017, doi: 10.4121/UUID:5F3067DF-F10B-45DA-B98B-86AE4C7A310B.

[19] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning,” Decis. Anal. J., vol. 3, Jun. 2022, Art. no. 100071, doi: 10.1016/j.dajour.2022.100071.

[20] L. Breiman, “Random forests,” Mach. Learn., vol. 45, pp. 5–32, 2001.

Gode veera sai vara Bhuvaneswari is currently pursuing the MCA(Master of Computer Applications) in Ideal college of Arts and science, Vidyuth Nagar, Kakinada. Her research interests include Machine learning

M. Kameswara Rao is currently serving as the Additional Head of the Department of Computer Science at Ideal College of Arts & Sciences(A). He possesses more than 20 years of academic and administrative experience in the field of Computer Science.

Dr. V. S. V. Deepak is currently serving as the Head of the Department of Computer Science at Ideal College of Arts & Sciences (A). He possesses more than 18 years of academic and administrative experience in the field of Computer Science and Engineering. His areas of interest include Medical Image Processing, Cyber Security, Artificial Intelligence, Software Testing and Networking. He completed his Ph.D. research in Medical Image Processing from Swami Vivekananda University.
He has actively contributed to curriculum development, academic planning, and student mentoring. He has served as Chairman of the Board of Studies (BOS) for BCA, B.Sc. (Computer Science), B.Sc. (Artificial Intelligence), and MCA programs.