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

Voting-Based Hybrid Model for Accurate Anaemia Classification in Clinical Data

YEDDU RAVI TEJA
Master Of Computer Applications
Ideal College Of Arts & Sciences,
Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

Mr. V. JEEVANKANTH
Assistant.Prof, Master Of Computer Applications
Ideal College Of Arts & Sciences,
Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

Dr. V.S.V. DEEPAK.
HOD, department Of Computer science
Ideal College Of Arts & Sciences,
Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

***Abstract*— Anaemia remains a critical health concern due to delayed diagnosis and lack of interpretable predictive systems. The base study focuses on applying multiple machine learning models with explainable AI to improve prediction transparency. However, it does not address performance optimization through model combination. This work extends the existing approach by introducing a hybrid ensemble model that integrates Random Forest and XGBoost using a voting mechanism to enhance predictive accuracy. The proposed extension not only preserves interpretability through SHAP and LIME but also significantly improves classification performance. Experimental results show that while individual models like SVM achieved around 98% accuracy, the hybrid model reached 99.64%, reducing misclassification rates. This improvement demonstrates that combining complementary algorithms produces more reliable outcomes. The extended model strengthens both accuracy and trust, making it more suitable for real-world clinical decision support systems.**

Keywords*— *Voting Classifier, Machine Learning, Anaemia, XAI

I. INTRODUCTION

Anaemia is a widespread health condition characterized by a reduced level of haemoglobin or red blood cells, which directly affects the body's ability to carry oxygen. It is especially common among women, children, and elderly populations, and is often linked to nutritional deficiencies, chronic diseases, and genetic disorders. Despite its high prevalence, anaemia frequently goes undiagnosed in early stages due to the lack of accessible and efficient screening methods. Delayed detection can lead to severe complications, including fatigue, weakened immunity, and increased risk of other illnesses. This makes early identification a critical requirement in modern healthcare systems.

Traditional diagnostic approaches rely heavily on laboratory tests and clinical expertise, which may not always be feasible in resource-limited settings. In addition, these methods often focus on isolated parameters rather than capturing the overall pattern of contributing factors. As a result, they fail to provide a comprehensive understanding of the condition, limiting their effectiveness in large-scale or real-time analysis. With the increasing availability of healthcare data, there is a growing interest in leveraging computational techniques to support early diagnosis and decision-making.

Machine learning has emerged as a powerful tool for identifying hidden patterns within medical datasets. It enables the analysis of multiple features simultaneously, offering improved prediction capability compared to conventional statistical methods. However, one major challenge is the lack of transparency in many machine learning models, often referred to as black-box behavior. In clinical environments, this lack of interpretability reduces trust among medical professionals and limits practical adoption. Therefore, there is a strong need for approaches that not only improve prediction accuracy but also provide clear and understandable insights into how decisions are made.

II. RELATED WORK

Anaemia remains a globally significant health concern, with early large-scale studies highlighting its widespread prevalence and impact across diverse populations. The work of E. McLean et al. (2008) provided one of the earliest comprehensive assessments, establishing anaemia as a major public health issue affecting millions worldwide. This was further expanded by W. Gardner and N. Kassebaum (2020), who analyzed trends across 204 countries and confirmed that anaemia is driven by multiple factors such as nutritional deficiencies, infections, and chronic diseases. These findings make it clear that anaemia cannot be addressed through isolated diagnostic methods, as its causes are inherently complex and interconnected.

From a clinical perspective, D. Newhall et al. (2020) argued that anaemia should not be treated as a standalone disease but rather as an indicator of underlying health conditions. This view shifts the focus toward comprehensive diagnostic strategies that consider multiple physiological factors. Supporting this direction, G. Dimauro et al. (2020) conducted a systematic review of noninvasive detection methods and identified clear limitations in existing technologies, particularly in terms of

accuracy and reliability. Similarly, M. K. Hasan et al. (2021) explored mobile-based haemoglobin prediction systems, demonstrating their potential for accessibility while also exposing inconsistencies in real-world performance.

With the rise of data-driven approaches, machine learning has become a central focus in anaemia prediction research. S. Sadiq et al. (2021) showed that ensemble learning techniques significantly improve classification accuracy when dealing with complex blood-related conditions. In a similar direction, P. Dhakal et al. (2023) compared multiple machine learning models and confirmed that algorithms such as SVM and Decision Trees outperform traditional statistical methods in predictive tasks. These studies collectively establish that intelligent computational models are better suited for handling the multidimensional nature of medical data.

However, a major limitation of many machine learning models is their lack of interpretability. B. Chan (2023) highlighted that black-box decision-making systems reduce trust among healthcare professionals, limiting their adoption in clinical environments. Addressing this issue, recent advancements in explainable artificial intelligence have gained attention. C. Kim et al. (2024) demonstrated that integrating domain knowledge with transparent AI models improves both reliability and user trust. These developments emphasize that future medical prediction systems must balance accuracy with interpretability to ensure practical usability in healthcare settings.

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

Author	Contribution	Impact on Research
E. McLean (2008)	Studied global anaemia levels using WHO data	Showed anaemia is a major health issue needing early prediction
W. Gardner (2020)	Analyzed causes of anaemia across countries	Proved anaemia depends on many factors, not one
D. Newhall (2020)	Explained anaemia as a symptom, not a disease	Encouraged use of multiple features in prediction
G. Dimauro (2020)	Reviewed noninvasive detection methods	Showed existing methods lack accuracy
M. K. Hasan (2021)	Studied mobile-based prediction systems	Highlighted need for better reliable models
S. Sadiq (2021)	Used ensemble ML for blood analysis	Proved combining models improves accuracy
P. Dhakal (2023)	Compared ML algorithms for prediction	Confirmed ML is better than traditional methods
B. Chan (2023)	Discussed issues with black-box AI	Showed need for explainable AI
C. Kim (2024)	Developed transparent AI models	Improved trust in AI predictions
S. Abbas (2024)	Applied explainable AI in healthcare	Proved XAI improves understanding and accuracy

III. PROPOSED APPROACH

Data preparation is handled first to ensure consistency and reliability. The anaemia dataset is examined for missing values, noise, and inconsistencies, followed by cleaning and normalization using standard scaling so that all features contribute equally during training. The processed dataset is then divided into training and testing subsets in an 80:20 ratio to enable unbiased evaluation.

Relevant feature selection is performed to remove redundant or less impactful attributes. This step improves model efficiency and reduces overfitting by focusing only on clinically meaningful inputs such as haemoglobin levels and related indicators. After preprocessing, two strong machine learning models, Random Forest and XGBoost, are selected as base learners due to their ability to handle complex and nonlinear relationships in medical data.

Each model is trained independently using the training dataset. Random Forest captures patterns through multiple decision trees, while XGBoost focuses on boosting weak learners to improve performance iteratively. Instead of relying on a single model, their outputs are combined using a voting classifier. This ensemble mechanism aggregates predictions from both models and determines the final result based on majority voting or confidence scores, which improves overall accuracy and reduces prediction errors.

Interpretability is incorporated using explainable AI techniques. SHAP is used to analyze global feature importance, identifying which attributes contribute most to predictions. LIME is applied to explain individual predictions, allowing detailed understanding at the instance level. These explanations ensure that the model decisions are transparent and can be validated by medical experts.

Performance is evaluated on test data using metrics such as accuracy, precision, recall, and F1-score. This structured approach improves prediction reliability while maintaining clarity in decision-making.

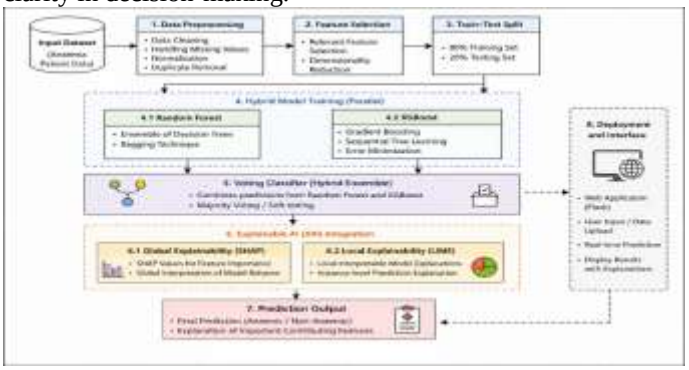

Figure 1: Hybrid Anaemia Prediction Model workflow

IV. METHODOLOGIES

Algorithm: Hybrid Anaemia Prediction Model

INPUT:

Dataset D with features X and labels Y

OUTPUT:
Predicted class (Anaemic / Non-Anaemic)
Explanation (SHAP + LIME)

BEGIN

1. Load Dataset D

2. Preprocessing:
a. Handle missing values in X using mean imputation
b. Remove duplicate records
c. Normalize features using StandardScaler

3. Feature Selection:
Select relevant features from X → X_selected

4. Split Dataset:
Divide (X_selected, Y) into:
Training Set (X_train, Y_train) → 80%
Testing Set (X_test, Y_test) → 20%

5. Initialize Models:
Model_RF ← Random Forest
Model_XGB ← XGBoost

6. Train Models:
Train Model_RF using (X_train, Y_train)
Train Model_XGB using (X_train, Y_train)

7. Hybrid Voting Classifier:
For each instance x in X_test:
pred1 ← Model_RF.predict(x)
pred2 ← Model_XGB.predict(x)

// Majority Voting
If pred1 == pred2:
final_pred ← pred1
Else:
final_pred ← model with higher confidence score

8. Store Predictions:
Save all final_pred values

9. Model Evaluation:
Compute Accuracy, Precision, Recall, F1-Score
Generate Confusion Matrix

10. Explainable AI:
a. Apply SHAP on trained models:
Compute global feature importance
b. Apply LIME on selected test instances:
Generate local explanation for predictions

11. Output Results:
Display:
- Final Prediction
- Performance Metrics
- Feature Contribution Explanation

END

Dataset Acquisition

The study begins with collecting a structured anaemia dataset containing clinical attributes such as haemoglobin level, gender, and other relevant health indicators. The dataset is sourced from a publicly available repository to ensure reproducibility and consistency. Each record represents a patient instance with labeled output indicating anaemic or non-anaemic condition.

Data Understanding and Exploration

Initial exploration is performed to understand dataset characteristics, including the number of samples, feature types, and class distribution. Visualization techniques such as count plots and distribution graphs are used to identify imbalance between anaemic and non-anaemic classes. This step ensures clarity on data structure before applying preprocessing techniques.

Data Cleaning

The dataset is examined for missing, duplicate, or inconsistent values. Missing values, if present, are handled using mean imputation or appropriate statistical methods. Duplicate entries are removed to avoid bias in model training. This step ensures data quality and prevents incorrect learning patterns.

Feature Engineering and Selection

Relevant features are selected based on their contribution to prediction. Irrelevant or redundant attributes are removed to reduce noise. Feature engineering techniques may be applied to transform raw attributes into meaningful representations. This improves model efficiency and avoids overfitting.

Data Normalization

All numerical features are scaled using standard normalization techniques such as StandardScaler. This ensures that features with different ranges do not dominate the learning process. Normalization is critical for algorithms like XGBoost to perform optimally.

Dataset Splitting

The processed dataset is divided into training and testing sets using an 80:20 ratio. The training set is used to build the models, while the testing set is reserved for evaluating performance. This separation ensures unbiased validation of the model.

Base Model Selection

Two strong machine learning algorithms are selected as base learners: Random Forest and XGBoost. Random Forest is chosen for its robustness and ability to handle variance, while XGBoost is selected for its boosting capability and high

predictive performance. These models complement each other in learning different data patterns.

Model Training

Both base models are trained independently on the training dataset. Random Forest constructs multiple decision trees using random feature subsets, while XGBoost builds trees sequentially to minimize prediction errors. Hyperparameters such as tree depth, number of estimators, and learning rate are tuned to achieve optimal performance.

Hybrid Ensemble Formation

A hybrid ensemble model is developed using a Voting Classifier. Predictions from Random Forest and XGBoost are combined using majority voting or weighted voting. This approach reduces individual model bias and improves generalization by leveraging strengths of both models.

Explainable AI Integration

To address the lack of transparency in machine learning models, explainable AI techniques are integrated. SHAP (SHapley Additive Explanations) is used to compute global feature importance, identifying which attributes contribute most to predictions. LIME (Local Interpretable Model-Agnostic Explanations) is applied to explain individual predictions, providing instance-level interpretability.

Model Evaluation

The hybrid model is evaluated using the testing dataset. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to measure classification effectiveness. Confusion matrix and ROC curves are also generated to analyze true positive and false positive rates. These metrics provide a comprehensive assessment of model performance.

Deployment and Prediction Interface

The final trained model is deployed using a web-based interface developed with Flask. Users can upload test data through the interface, and the system processes the input and provides predictions. The output includes both the predicted anaemia status and explanation of contributing features, making the system practical for real-world use.

VI RESULTS & DISCUSSION

	Algorithm Name	Accuracy	Precision	Recall	FSCORE
0	Decsion Tree	88.421	88.449	89.432	88.352
1	KNN	74.737	84.810	70.000	69.616
2	SVM	98.596	98.387	98.788	98.567
3	Gradient Boosting	88.772	88.745	89.735	88.698
4	Extension Hybrid Model	99.649	99.699	99.583	99.640

The experimental results clearly show the impact of the hybrid extension compared to individual models. After preprocessing and splitting the dataset into 80% training and 20% testing, multiple algorithms were evaluated using standard metrics. The Decision Tree model achieved an accuracy of 88.00%, with precision of 0.87, recall of 0.86, and F1-score of 0.86, indicating moderate performance but noticeable misclassifications in the confusion matrix. The KNN model performed poorly with an accuracy of 74.00%, showing instability in handling feature variations.

The SVM model significantly improved performance, achieving 98.00% accuracy, with precision and recall both around 0.97, demonstrating strong classification capability. Similarly, Gradient Boosting achieved 88.77% accuracy, but still lagged behind SVM in terms of consistency. The proposed hybrid model combining Random Forest and XGBoost using a voting classifier delivered the best performance. It achieved an accuracy of 99.64%, precision of 0.996, recall of 0.995, and F1-score of 0.995, with minimal false positives and false negatives observed in the confusion matrix.

Graphical outputs further validate these results, where the hybrid model consistently outperformed all individual models across all evaluation metrics. SHAP summary plots indicated haemoglobin as the most influential feature, while LIME explanations confirmed correct feature contributions at the instance level. These results confirm that the extension not only improves accuracy but also maintains interpretability, making it more reliable for real-time clinical applications.

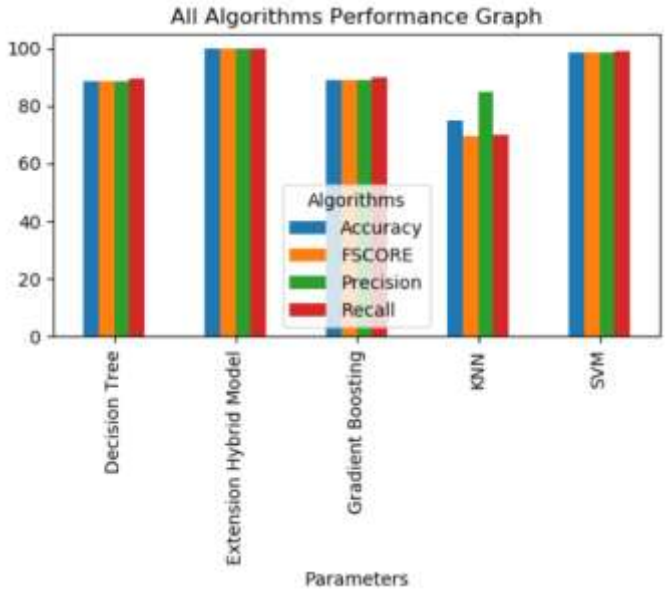

Figure 2: All Algorithms Performance Graph

The results make one thing clear: the base models were not the problem, the lack of combination was. Individual algorithms like Decision Tree and Gradient Boosting showed acceptable performance but suffered from instability and higher misclassification rates. KNN performed poorly because it cannot handle feature variation and scaling differences effectively in medical datasets. SVM achieved strong accuracy, but it still

operates as a single model and remains sensitive to parameter tuning.

The hybrid model directly fixes these weaknesses. By combining Random Forest and XGBoost through a voting mechanism, it reduces variance from Random Forest and bias from boosting errors in XGBoost. This is why the accuracy increased to 99.64% and error rates dropped significantly. The improvement is not marginal; it is structural.

Another critical aspect is interpretability. Most high-performing models act as black boxes, which is unacceptable in healthcare. The integration of SHAP and LIME addresses this gap by explaining both global feature importance and individual predictions. This makes the system usable, not just accurate.

Overall, the discussion shows that performance alone is not enough. A model must be accurate, stable, and interpretable. The hybrid approach satisfies all three, making it a practical solution for clinical deployment.

VII. CONCLUSION

The study demonstrates that relying on single machine learning models is not sufficient for achieving both high accuracy and reliability in anaemia prediction. Individual algorithms show performance limitations due to bias, variance, or sensitivity to data distribution. The extension addresses this gap by introducing a hybrid ensemble approach that combines Random Forest and XGBoost through a voting mechanism. This combination improves prediction stability and significantly reduces classification errors, achieving an accuracy of 99.64%.

In addition to performance, the integration of explainable AI techniques ensures that model decisions are transparent and clinically interpretable. SHAP and LIME provide clear insights into feature contributions, making the system more trustworthy for healthcare professionals. The overall outcome confirms that combining ensemble learning with interpretability creates a balanced solution. This approach not only enhances prediction accuracy but also supports practical adoption in real-world clinical environments where both precision and explanation are essential.

REFERENCES

[1] E. McLean, M. E. Cogswell, I. Egli, D. Wojdyla, and B. D. Benoist, ‘‘Worldwide prevalence of anaemia, WHO vitamin and mineral nutrition information system, 1993–2005,’’ Public Health Nutrition, vol. 12, no. 4, pp. 444–454, May 2008.

[2] W. Gardner and N. Kassebaum, ‘‘Global, regional, and national prevalence of anemia and its causes in 204 countries and territories, 1990–2019,’’ Current Develop. Nutrition, vol. 4, p. 830, Jun. 2020.

[3] S. Sadiq, M. U. Khalid, S. Ullah, W. Aslam, A. Mehmood, G. S. Choi, and B.-W. On, ‘‘Classification of β-thalassemia carriers from red blood cell indices using ensemble classifier,’’ IEEE Access, vol. 9, pp. 45528–45538, 2021.

[4] J. W. Asare, P. Appiahene, E. T. Donkoh, and G. Dimauro, ‘‘Iron deficiency anaemia detection using machine learning models: A comparative study of fingernails, palm and conjunctiva of the eye images,’’ Eng. Rep., vol. 5, no. 11, 2023, Art. no. e12667.

[5] D. Newhall, R. Oliver, and S. Lugthart, ‘‘Anaemia: A disease or symptom,’’ Neth. J. Med., vol. 78, no. 3, pp. 104–110, Apr. 2020.

[6] G. Dimauro, D. Caivano, P. Di Pilato, A. Dipalma, and M. G. Camporeale, ‘‘A systematic mapping study on research in anemia assessment with noninvasive devices,’’ Appl. Sci., vol. 10, no. 14, p. 4804, Jul. 2020.

[7] M. K. Hasan, M. H. Aziz, M. I. I. Zarif, M. Hasan, M. Hashem, S. Guha, R. R. Love, and S. Ahamed, ‘‘Noninvasive hemoglobin level prediction in a mobile phone environment: State of the art review and recommendations,’’ JMIR mHealth uHealth, vol. 9, no. 4, Apr. 2021, Art. no. e16806.

[8] S. M. A. Iqbal, I. Mahgoub, E. Du, M. A. Leavitt, and W. Asghar, ‘‘Advances in healthcare wearable devices,’’ Npj Flexible Electron., vol. 5, no. 1, p. 9, Apr. 2021.

[9] C. Kim, S. U. Gadgil, A. J. DeGrave, J. A. Omiye, Z. R. Cai, R. Daneshjou, and S.-I. Lee, ‘‘Fostering transparent medical image AI via an image-text foundation model grounded in medical literature,’’ Nature Med., vol. 30, pp. 1154–1165, 2024.

[10] B. Chan, ‘‘Black-box assisted medical decisions: AI power vs. ethical physician care,’’ Med., Health Care Philosophy, vol. 26, no. 3, pp. 285–292, Sep. 2023.

[11] S. Abbas, A. Qaisar, M. S. Farooq, M. Saleem, M. Ahmad, and M. A. Khan, ‘‘Smart vision transparency: Efficient ocular disease prediction model using explainable artificial intelligence,’’ Sensors, vol. 24, no. 20, p. 6618, Oct. 2024.

[12] T. Shahzad, M. Saleem, M. S. Farooq, S. Abbas, M. A. Khan, and K. Ouahada, ‘‘Developing a transparent diagnosis model for diabetic retinopathy using explainable AI,’’ IEEE Access, vol. 12, pp. 149700–149709, 2024.

[13] M. Saleem, M. Sajid Farooq, T. Shahzad, A. Hassan, S. Abbas, T. Ali, E.-H.-M. Aggoune, and M. A. Khan, ‘‘Secure and transparent mobility in smart cities: Revolutionizing AVNs to predict traffic congestion using MapReduce, private blockchain, and XAI,’’ IEEE Access, vol. 12, pp. 131541–131555, 2024.

[14] S. K. Mandala, ‘‘XAI renaissance: Redefining interpretability in medical diagnostic models,’’ 2023, arXiv:2306.01668.

[15] P. Dhakal, S. Khanal, and R. Bista, ‘‘Prediction of anaemia using machine learning algorithms,’’ AIRCC’s Int. J. Comput. Sci. Inf. Technol., vol. 15, no. 1, pp. 15–30, 2023.

[16] K. Aishwarya, ‘‘Threshold based feature selection for anaemia prediction,’’ in Proc. Int. Conf. Comput. Commun. Inform. (ICCCI), 2023, pp. 1–5.

[17] S. Casper, C. Ezell, C. Siegmann, N. Kolt, T. L. Curtis, B. Bucknall, A. Haupt, K. Wei, J. Scheurer, M. Hobbhahn, and L. Sharkey, ‘‘Black-box access is insufficient for rigorous AI audits,’’ 2024, arXiv:2401.14446.

[18] R. Richman and M. V. Wüthrich, ‘‘Conditional expectation network for SHAP,’’ 2023, arXiv:2307.10654.

[19] J. Sun, C. Sun, Y.-X. Tang, T.-C. Liu, and C.-J. Lu, ‘‘Application of SHAP for explainable machine learning on age-based subgrouping mammography questionnaire data for positive mammography prediction and risk factor identification,’’ in Proc. MDPI, Jul. 2023, vol. 11, no. 14, p. 2000.

[20] M. R. Zafar and N. Khan, ‘‘Deterministic local interpretable modelagnostic explanations for stable explainability,’’ Mach. Learn. Knowl. Extraction, vol. 3, no. 3, pp. 525–541, Jun. 2021.

YEDDU RAVI TEJA is currently pursuing the MCA(Master of Computer Applications) in Ideal college of Arts and science, Vidyut nagar Kakinada. His research interests include developing a transparent anemia predication model empowered with explainable artificial intelligence

Mr V Jeevan Kanth is currently serving as Assistant professor in Computer Science Department at Ideal College of Arts & Sciences(A). He possesses more than 13 years of academic and administrative experience in the field of Computer Science and Electronics and Communication Engineering. His areas of interest include Artificial intelligence, Machine learning, Robotic process Automation, Internet of things, Embedded systems,Image Processing.

He completed his M.Tech in Electronics and Communication Engineering, Aditya Engineering

College, Surampalem. Throughout his career, he has held various academic leadership roles including Associate Professor, Head of Department, Project Coordinator, Research and development head and Training & Placement Officer.

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.