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

Scalable Big Data Analytics Framework for Diabetic Risk Prediction with Optimized Feature Selection

BEDESI GAYATHRI BAI
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*— Big data analytics has become an essential component in modern healthcare systems for improving disease prediction and clinical decision-making. This research introduces an extension model for diabetic disease prediction using Apache Spark MLIB and advanced feature selection techniques. Unlike existing approaches that depend on traditional Pearson correlation methods, the proposed extension applies Principal Component Analysis (PCA) to identify the most significant healthcare attributes from the BRFSS diabetic dataset. The optimized features are classified using Random Forest and Decision Tree algorithms to enhance prediction performance and reduce computational complexity. Experimental results show that the PCA-based Decision Tree model achieved 98.89% accuracy, while the PCA-based Random Forest model obtained 98.59% accuracy, outperforming conventional models such as Logistic Regression, SVM, and Naïve Bayes. The proposed framework provides accurate diabetic prediction, efficient large-scale data handling, and reliable support for intelligent healthcare analytics and real-time medical decision systems.**

Keywords*— *Apache Spark, Machine Learning, Big Data, PCA

I. INTRODUCTION

Diabetes has become one of the most critical chronic diseases affecting millions of people worldwide and creating a major burden on healthcare systems. Rapid urbanization, unhealthy lifestyles, obesity, stress, and lack of physical activity have significantly increased the number of diabetic patients in recent years. Early identification of diabetes is essential because delayed diagnosis may lead to severe complications such as heart disease, kidney failure, vision loss, and nerve damage. Traditional healthcare systems often struggle to process the continuously growing volume of patient records generated from hospitals, wearable devices, IoT sensors, and medical monitoring platforms. As a result, extracting meaningful insights from large-scale healthcare data has become a challenging task.

The emergence of big data analytics has transformed the healthcare sector by enabling efficient storage, processing, and analysis of massive medical datasets. Technologies such as Apache Spark and distributed computing frameworks allow healthcare organizations to perform faster data analysis and support clinical decision-making with improved accuracy. At the same time, machine learning techniques have gained significant attention for predicting diseases, identifying hidden patterns, and improving healthcare services through intelligent analysis. Various algorithms are widely applied in diabetic prediction systems to analyze patient health conditions and risk factors. However, healthcare datasets often contain redundant, noisy, and highly correlated attributes, which can reduce prediction performance and increase computational complexity. Therefore, efficient data preprocessing and analytical strategies are becoming increasingly important for reliable and scalable diabetic healthcare analytics.

II. RELATED WORK

Wu et al. (2014) introduced the concept of data mining in big data environments and explained the challenges involved in handling massive datasets generated from modern digital systems. The authors emphasized that traditional analytical approaches are insufficient for processing high-volume and heterogeneous data. Their work highlighted the importance of scalable machine learning and distributed computing methods for extracting meaningful information from complex datasets. The study created a strong theoretical foundation for later research related to healthcare big data analytics and intelligent disease prediction systems.

Wang et al. (2016) investigated the application of big data analytics in logistics and supply chain management. The research demonstrated how predictive analytics and distributed frameworks improve operational efficiency and decision-making accuracy in large-scale environments. In the same year, Liu and Yen (2016) explored big data analysis for optimizing public service management systems through systematic analytical models and visualization techniques. Daki et al. (2017) further discussed efficient big data management architectures within smart grid systems and highlighted challenges such as scalability, storage management, and secure data integration. These studies collectively provided important theoretical insights into large-scale data processing and intelligent analytical systems applicable to healthcare environments.

Aceto, Persico, and Pescapé (2018) examined the role of information and communication technologies in transforming healthcare services. Their study classified various healthcare technologies and analyzed challenges associated with

interoperability, security, and digital transformation. Palanisamy and Thirunavukarasu (2019) reviewed the implications of big data analytics in healthcare framework development and emphasized the importance of predictive analytics, patient monitoring, and real-time medical decision-making. Galetsi and Katsaliaki (2019) conducted an extensive review of healthcare big data analytics applications and discussed analytical tools, implementation methods, and organizational benefits. Their findings highlighted the growing significance of data-driven healthcare systems for improving clinical services and operational performance.

Yousefi, Derakhshan, and Karimipour (2020) explored the integration of big data analytics and machine learning techniques within Internet of Things environments. The study explained how sensor-generated data can be processed using intelligent algorithms for predictive analysis and monitoring applications. Nauman et al. (2021) focused on the correctness and reliability of machine learning-based decision-making systems, emphasizing the importance of trustworthy analytical frameworks in critical environments. Khan et al. (2021) reviewed diabetes prediction methods using data mining and machine learning algorithms. Their study analyzed various healthcare datasets, classification techniques, and evaluation metrics, providing valuable theoretical support for diabetic disease prediction and healthcare analytics systems.

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

Author	Contribution	Impact on Research
Wu et al. (2014)	Discussed data mining methods for handling large big data datasets.	Helped researchers use machine learning with healthcare big data systems.
Wang et al. (2016)	Explained predictive analytics and distributed computing for large-scale systems.	Supported the use of real-time analytics in healthcare applications.
Liu and Yen (2016)	Studied big data analysis for improving public service management.	Inspired better data processing and decision-making methods in healthcare.
Daki et al. (2017)	Presented big data management techniques for scalable systems.	Improved understanding of secure and efficient healthcare data handling.
Aceto et al. (2018)	Analyzed the role of ICT technologies in healthcare services.	Encouraged the use of digital technologies in healthcare monitoring systems.
Palanisamy and Thirunavukarasu (2019)	Reviewed healthcare frameworks using big data analytics.	Supported research on predictive healthcare and patient monitoring systems.
Galetsi and Katsaliaki (2019)	Reviewed applications of big data analytics in healthcare.	Provided knowledge about healthcare analytical tools and methods.
Yousefi et al. (2020)	Explained machine learning and IoT integration using big data analytics.	Supported healthcare monitoring using sensor and IoT data.
Nauman et al. (2021)	Focused on reliable machine learning decision-making systems.	Improved trust and accuracy in healthcare prediction models.
Khan et al. (2021)	Reviewed diabetes prediction using machine learning algorithms.	Provided useful background for diabetic disease prediction research.

III. PROPOSED APPROACH

An intelligent diabetic disease prediction framework is developed using big data analytics and machine learning techniques within the Apache Spark MLIB environment. The BRFSS diabetic healthcare dataset is initially loaded into the Spark framework to support distributed processing and efficient handling of large-scale medical data. Preprocessing operations are applied to improve dataset quality and enhance prediction performance. These operations include missing value handling, normalization, and removal of inconsistent records. Several visualization graphs are generated to examine the relationship between healthcare factors such as cholesterol level, alcohol consumption, education status, income level, general health condition, and diabetic risk.

Correlation analysis is performed to identify highly related healthcare attributes and reduce unnecessary data redundancy. To improve feature optimization, Principal Component Analysis (PCA) is applied as an advanced feature selection method. PCA reduces dimensionality by selecting the most significant attributes from the dataset while minimizing irrelevant and correlated information. This process decreases computational complexity and improves the efficiency of machine learning classification models.

After feature optimization, the dataset is divided into training and testing datasets for performance evaluation. Multiple machine learning algorithms are trained, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, and Naïve Bayes. Experimental observations indicate that conventional machine learning models provide moderate prediction accuracy, whereas PCA-based models produce significantly improved results. The PCA with Random Forest model achieved 98.59% prediction accuracy, while PCA with Decision Tree achieved 98.89% accuracy, outperforming all other classification models.

The trained prediction model is further integrated with a Flask-based web application for practical healthcare usage. Users can upload test healthcare records through the web interface, and the system predicts whether the patient condition is Normal or Diabetic. The framework supports efficient healthcare analytics, real-time prediction, and intelligent medical decision-making for large-scale diabetic datasets.

Figure 1: Diabetic Prediction workflow

IV. METHODOLOGIES

Algorithm: PCA-Based Diabetic Prediction Model

Input:
BRFSS Diabetic Healthcare Dataset D

Output:
Predicted Result as Normal or Diabetic

Begin

1. Start System

2. Load BRFSS diabetic dataset D into Apache Spark framework

3. Initialize Spark Context and required ML libraries

4. Perform dataset preprocessing
a. Remove null and inconsistent values
b. Normalize dataset attributes
c. Handle missing healthcare records

5. Analyze dataset statistics
a. Calculate rows and columns
b. Compute MIN, MAX, Mean, Standard Deviation

6. Generate visualization graphs
a. Cholesterol vs Diabetes
b. Alcohol Consumption vs Diabetes
c. Education vs Diabetes
d. Income vs Diabetes
e. General Health vs Diabetes

7. Perform feature correlation analysis
a. Generate correlation heatmap
b. Identify highly correlated attributes

8. Apply Principal Component Analysis (PCA)
a. Extract important healthcare features
b. Reduce dimensionality
c. Select top 8 optimized features

9. Split processed dataset into
a. Training Dataset = 70%
b. Testing Dataset = 30%

10. Train Machine Learning Models
a. Logistic Regression
b. Decision Tree
c. Random Forest
d. Support Vector Machine
e. Gradient Boosting
f. Naïve Bayes

11. Train Extension Models
a. PCA + Random Forest
b. PCA + Decision Tree

12. Test all trained models using testing dataset

13. Calculate performance metrics
a. Accuracy
b. Precision
c. Recall
d. F-Score

14. Compare all model performances

15. Select best-performing model
If PCA + Decision Tree accuracy is highest
Select PCA + Decision Tree model
Else
Select PCA + Random Forest model
End If

16. Deploy selected model in Flask web application

17. Upload healthcare test data through web interface

18. Predict patient condition
If prediction value = 1
Display "Diabetic"
Else
Display "Normal"
End If

19. Display prediction results to user

End

Dataset Collection

The proposed work begins with collecting the BRFSS diabetic healthcare dataset, which contains patient healthcare information related to diabetic conditions. The dataset includes several healthcare attributes such as cholesterol level, alcohol consumption, education level, income status, general health condition, BMI, age, and diabetic labels. The dataset is loaded into the Apache Spark environment for large-scale distributed processing and efficient healthcare analytics.

Spark Environment Initialization

Apache Spark MLIB framework is initialized by creating the Spark Context object inside the Jupyter Notebook environment. Spark provides distributed computing support for handling large healthcare datasets with improved processing speed and scalability. Required Python libraries, Spark MLIB packages, and machine learning modules are imported before starting analytical operations.

Dataset Loading and Exploration

The healthcare dataset is imported into the Spark framework and displayed for initial analysis. The system examines the number of rows and columns available in the dataset. Statistical analysis is performed to identify minimum values, maximum values, average values, and standard deviation of healthcare attributes. This step helps in understanding the structure and characteristics of the dataset.

Data Visualization

Several visualization graphs are generated to analyze relationships between healthcare attributes and diabetic conditions. Graphs are plotted for education level, cholesterol, alcohol consumption, income level, gender distribution, and general health status. These visualizations help in identifying

hidden patterns and understanding the influence of healthcare factors on diabetic prediction.

Missing Value Handling

The dataset may contain incomplete or inconsistent healthcare records. Missing value handling techniques are applied to replace null or invalid entries with suitable values. This preprocessing operation improves dataset quality and prevents performance degradation during machine learning model training.

Data Normalization

Healthcare attributes in the dataset may exist in different numerical ranges. Normalization is applied to transform feature values into a uniform scale. This process improves learning efficiency and prevents algorithms from being biased toward attributes with larger numerical values.

Correlation Analysis

Feature correlation analysis is performed to identify highly related healthcare attributes. A correlation heatmap is generated where higher correlation values indicate dependency between features. Redundant and highly correlated attributes negatively affect prediction performance and increase computational complexity.

PCA-Based Feature Selection

The extension model applies Principal Component Analysis (PCA) for advanced feature selection. PCA reduces dataset dimensionality by selecting the most important healthcare features while removing redundant information. From the original dataset attributes, PCA selects eight optimized features that contribute most to diabetic prediction accuracy. This step significantly improves model efficiency and reduces unnecessary processing.

Dataset Splitting

After feature optimization, the processed dataset is divided into training and testing datasets. The training dataset is used for learning machine learning patterns, while the testing dataset is used for evaluating prediction performance. Proper dataset splitting helps in measuring the real-time efficiency of prediction models.

Machine Learning Model Training

The optimized healthcare dataset is trained using various machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, and Naïve Bayes. In the extension approach, PCA-selected features are specifically trained using the Random Forest and Decision Tree algorithms to improve prediction accuracy.

Performance Evaluation

All trained machine learning models are evaluated using performance metrics including accuracy, precision, recall, and F-score. Experimental results show that traditional algorithms achieved moderate prediction performance, while PCA-based models produced significantly improved results. PCA with Random Forest achieved 98.59% accuracy, whereas PCA with Decision Tree achieved 98.89% accuracy, outperforming all other algorithms.

VI RESULTS & DISCUSSION

	Algorithm Name	Accuracy	Precision	Recall	FSCORE
0	Logistic Regression	93.04	61.31	96.45	66.60
1	Decision Tree	88.92	57.76	94.35	60.43
2	Random Forest	90.98	59.20	95.39	63.12
3	SVM	88.22	57.36	93.95	59.63
4	Gradient Boosting	87.80	57.14	93.77	59.18
5	Naive Bayes	60.44	52.35	78.57	41.91
6	Extension PCA Random Forest	98.59	79.56	99.28	86.79
7	Extension PCA Decision Tree	98.87	82.17	99.42	88.86

Experimental results obtained from the implementation screens demonstrate the effectiveness of the proposed extension model for diabetic disease prediction. Multiple machine learning algorithms were trained and evaluated using the BRFSS diabetic healthcare dataset, and their performance was measured using accuracy, precision, recall, and F-score metrics. Among the conventional machine learning algorithms, Logistic Regression achieved the best performance with 93.04% accuracy, 61.31% precision, 96.45% recall, and 66.60% F-score. Random Forest produced 90.98% accuracy with 59.20% precision and 95.39% recall, while Decision Tree achieved 88.92% accuracy and 60.43% F-score. Similarly, SVM obtained 88.22% accuracy and Gradient Boosting achieved 87.80% accuracy. Naïve Bayes showed the lowest performance with only 60.44% accuracy and 41.91% F-score, indicating poor classification capability for the diabetic dataset.

The extension models developed using Principal Component Analysis (PCA) significantly improved prediction performance compared to traditional approaches. PCA with Random Forest achieved 98.59% accuracy, 79.56% precision, 99.28% recall, and 86.79% F-score. The highest performance was achieved by the PCA-based Decision Tree model, which produced 98.87% accuracy, 82.17% precision, 99.42% recall, and 88.86% F-score. The improvement in performance clearly shows that PCA successfully selected the most relevant healthcare attributes and reduced redundant features from the dataset. Experimental analysis confirms that the proposed PCA-based extension models provide highly accurate and reliable diabetic disease prediction for healthcare analytics applications.

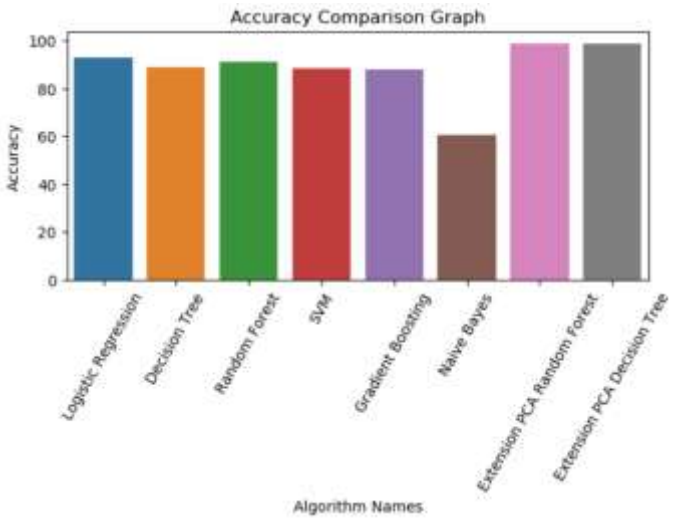

Figure 2: All Algorithms Performance Graph

The experimental analysis clearly demonstrates that feature optimization plays a major role in improving diabetic disease prediction accuracy. Traditional machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, SVM, and Gradient Boosting produced acceptable performance, but their accuracy was limited due to redundant and highly correlated healthcare attributes present in the dataset. Logistic Regression achieved the best performance among conventional models with 93.04% accuracy, while Naïve Bayes showed poor performance because of its weaker capability in handling complex healthcare relationships.

The extension models developed using Principal Component Analysis significantly improved classification efficiency by selecting the most relevant healthcare features and reducing unnecessary data dimensions. PCA with Random Forest achieved 98.59% accuracy, whereas PCA with Decision Tree achieved the highest accuracy of 98.87% with improved precision, recall, and F-score values. The results confirm that advanced feature selection methods improve prediction reliability, reduce computational complexity, and support efficient healthcare decision-making for large-scale diabetic healthcare analytics systems.

VII. CONCLUSION

The research successfully developed an efficient diabetic disease prediction framework using big data analytics and machine learning techniques within the Apache Spark MLIB environment. The study analyzed multiple machine learning algorithms and evaluated their performance using healthcare prediction metrics such as accuracy, precision, recall, and F-score. Experimental results confirmed that traditional algorithms produced moderate prediction performance, while the extension models using Principal Component Analysis significantly improved classification accuracy. PCA-based Random Forest and Decision Tree models achieved 98.59% and 98.87% accuracy respectively, outperforming all conventional approaches. The integration of advanced feature selection reduced redundant healthcare attributes and improved prediction efficiency. The developed Flask-based web application further enabled real-time diabetic prediction, making the system suitable for intelligent healthcare analytics, large-scale medical data processing, and reliable clinical decision-support applications.

REFERENCES

[1] G. Aceto, V. Persico, and A. Pescapé, ‘‘The role of information and communication technologies in healthcare: Taxonomies, perspectives, and challenges,’’ J. Netw. Comput. Appl., vol. 107, pp. 125–154, Apr. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804518300456

[2] M. Nauman, N. Akhtar, O. H. Alhazmi, M. Hameed, H. Ullah, and N. Khan, ‘‘Improving the correctness of medical diagnostics (don’t short) based on machine learning with coloured Petri nets,’’ IEEE Access, vol. 9, pp. 143434–143447, 2021.

[3] M. Ianculescu, A. Alexandru, and E. Tudora, ‘‘Opportunities brought by big data in providing silver digital patients with ICT-based services that support independent living and lifelong learning,’’ in Proc. 9th Int. Conf. Ubiquitous Future Netw. (ICUFN), Jul. 2017, pp. 404–409.

[4] S. Bebortta, S. S. Tripathy, S. Basheer, and C. L. Chowdhary, ‘‘DeepMist: Toward deep learning assisted mist computing framework for managing healthcare big data,’’ IEEE Access, vol. 11, pp. 42485–42496, 2023.

[5] A. Ahmed, R. Xi, M. Hou, S. A. Shah, and S. Hameed, ‘‘Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts,’’ IEEE Access, vol. 11, pp. 112891–112928, 2023.

[6] V. Palanisamy and R. Thirunavukarasu, ‘‘Implications of big data analytics in developing healthcare frameworks—A review,’’ J. King Saud Univ.- Comput. Inf. Sci., vol. 31, no. 4, pp. 415–425, Oct. 2019.

[7] P. Galetsi and K. Katsaliaki, ‘‘A review of the literature on big data analytics in healthcare,’’ J. Oper. Res. Soc., vol. 71, no. 10, pp. 1511–1529, Jul. 2019.

[8] G. Wang, A. Gunasekaran, E. W. T. Ngai, and T. Papadopoulos, ‘‘Big data analytics in logistics and supply chain management: Certain investigations for research and applications,’’ Int. J. Prod. Econ., vol. 176, pp. 98–110, Jun. 2016.

[9] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ‘‘Data mining with big data,’’ IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.

[10] J. Z. Zhang, P. R. Srivastava, D. Sharma, and P. Eachempati, ‘‘Big data analytics and machine learning: A retrospective overview and bibliometric analysis,’’ Expert Syst. Appl., vol. 184, Dec. 2021, Art. no. 115561. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421009672

[11] S. Yousefi, F. Derakhshan, and H. Karimipour, ‘‘Applications of big data analytics and machine learning in the Internet of Things,’’ in Handbook of Big Data Privacy. Cham, Switzerland: Springer, 2020, pp. 77–108. 10782 VOLUME 13, 2025 M. Nauman et al.: Role of BDA in Revolutionizing Diabetes Management and Healthcare Decision-Making

[12] S. Mittal and O. P. Sangwan, ‘‘Big data analytics using machine learning techniques,’’ in Proc. 9th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence), Jan. 2019, pp. 203–207.

[13] S. J. Miah, E. Camilleri, and H. Q. Vu, ‘‘Big data in healthcare research: A survey study,’’ J. Comput. Inf. Syst., vol. 62, no. 3, pp. 480–492, May 2022

[14] W.-K. Liu and C.-C. Yen, ‘‘Optimizing bus passenger complaint service through big data analysis: Systematized analysis for improved public sector management,’’ Sustainability, vol. 8, no. 12, p. 1319, Dec. 2016.

[15] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi, ‘‘Big data management in smart grid: Concepts, requirements and implementation,’’ J. Big Data, vol. 4, no. 1, pp. 1–19, Dec. 2017.

[16] M. Nauman, N. Akhtar, A. Alhudhaif, and A. Alothaim, ‘‘Guaranteeing correctness of machine learning based decision making at higher educational institutions,’’ IEEE Access, vol. 9, pp. 92864–92880, 2021.

[17] E. Kasturi, S. P. Devi, S. V. Kiran, and S. Manivannan, ‘‘Airline route profitability analysis and optimization using Big Data analyticson aviation data sets under heuristic techniques,’’ Proc. Comput. Sci., vol. 87, pp. 86–92, Mar. 2016.

[18] S. Nazir, S. Khan, H. U. Khan, S. Ali, I. García-Magariño, R. B. Atan, and M. Nawaz, ‘‘A comprehensive analysis of healthcare big data

management, analytics and scientific programming,'' IEEE Access, vol. 8, pp. 95714–95733, 2020.

[19] M. Karatas, L. Eriskin, M. Deveci, D. Pamucar, and H. Garg, ''Big data for healthcare industry 4.0: Applications, challenges and future perspectives,'' Expert Syst. Appl., vol. 200, Aug. 2022, Art. no. 116912.

[20] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, and S. A. C. Bukhari, ''Detection and prediction of diabetes using data mining: A comprehensive review,'' IEEE Access, vol. 9, pp. 43711–43735, 2021.

BEDESI GAYATHRI BAI is currently pursuing the MCA (Master of Computer Applications) in Ideal college of Arts and science, Vidyuth Nagar, Kakinada. Her research interests include BIG DATA

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.