Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/IN05.20250056
Published: November 6, 2025

Aqua Purity - Water Quality Analysis using Machine Learning

Ranita Dey, Dipanita Sharma, Atefa Khatun, Rahul Kumar and Dharitri Sarkar
Student, Camellia Institute of Engineering and Technology, Budbud
Budbud, West Bengal, India
ranitadey789@gmail.com, dipanitasharma01@gmail.com, atefakhatun786@gmail.com, rk4701700@gmail.com, dharitri713166@gmail.com

Sagnik Dutta
Assistant Professor
Camellia Institute of Engineering and Technology, Budbud
Budbud, West Bengal, India
sagnik.dutta.1973@gmail.com

Abstract

The quality of water is a crucial criterion for various aspects such as public health, environmental sustainability, and economic growth. Health of individuals largely depends on the water they consume, making regular monitoring essential. Such monitoring can help in detecting outbreaks of waterborne diseases like cholera and typhoid, which pose significant public health risks. Additionally, the aquatic ecosystem relies on a delicate balance of chemical and physical properties; effective quality checks can identify pollution sources that may disrupt this balance and harm aquatic life. The agricultural and food industries also heavily depend on good quality water for crop irrigation and food processing, making water quality a vital factor in food security and safety. The proposed model "Aqua Purity - Water Quality Analysis using Machine Learning," as the name suggests, aims to analyses water quality by training a model using various machine learning algorithms. This analysis will assist in identifying pollutants, detecting anomalies, and classifying water bodies based on quality indicators such as pH, chemical composition, and microbial presence. By integrating real-time data collection and advanced analytical techniques, the project seeks to develop a framework that can categorize water drinkability and forecast potential water quality issues, allowing for timely intervention.

Keywords
Water quality, SVM, Gradient boost, Random forest and classification.

1. Introduction
Water is one of the most essential natural resources, fundamental to human health, agriculture, industry, and ecological stability. However, the rapid pace of urbanization and industrial growth has severely affected water quality across the globe. Increasing levels of pollutants, including heavy metals, agricultural chemicals, and microbial contaminants, have made many water sources unsafe for consumption and daily use. Contaminated water continues to be a major global challenge, causing widespread health issues such as cholera, typhoid, and dysentery, particularly in developing regions where access to clean water is limited. As a result, maintaining high water quality standards has become a critical priority for safeguarding public health, protecting ecosystems, and supporting sustainable development.

Traditional methods of water quality assessment primarily depend on laboratory-based chemical and microbiological testing. Although these approaches are accurate, they are often time-consuming, expensive, and labour-intensive. Manual analysis cannot efficiently handle the vast and continuously growing data generated from multiple monitoring sources such as rivers, reservoirs, and treatment plants. Moreover, such methods are typically reactive detecting contamination only after it has occurred rather than offering preventive insights. This limitation underscores the need for advanced, automated, and intelligent systems capable of continuously analysing and predicting water quality in real time.

The model "Aqua Purity – Water Quality Analysis Using Machine Learning" addresses these challenges by integrating data-driven approaches into water monitoring systems. Machine Learning (ML), a branch of Artificial Intelligence (AI), enables computers to learn patterns from data and make informed predictions without explicit programming. By applying ML techniques to water quality datasets, the project aims to identify complex relationships among parameters such as pH, dissolved oxygen, temperature, nitrate, BOD etc., through this approach, the model can uncover hidden patterns and correlations that traditional analytical methods may overlook, offering deeper insights into water quality dynamics.

One of the most valuable strengths of ML lies in its predictive capability. Trained on historical water quality data, the model can forecast variations caused by seasonal changes, industrial activities, or agricultural runoff. These predictions enable early warning and preventive actions before water conditions deteriorate. Furthermore, the system can classify water samples into distinct categories such as potable, non-potable, or suitable for irrigation thereby assisting authorities in efficient resource allocation and management.

The integration of machine learning also facilitates real-time water monitoring through sensor and IoT-based data collection. Continuous data streams are analysed instantly, allowing rapid detection of anomalies such as sudden pollutant spikes or deviations in key parameters. This immediate response system enhances decision-making and minimizes the risk of health hazards. Additionally, the insights generated by ML models can help policymakers and researchers identify the major contributors to water pollution and design effective mitigation strategies.

1.1 Objectives

The primary objective of this research is to develop an intelligent, machine learning based model capable of analysing and predicting water quality with greater precision and efficiency than traditional methods. The study focuses on leveraging data-driven techniques to enhance the speed and reliability of water quality monitoring, allowing for more accurate detection of contamination patterns and environmental changes. By examining relationships among multiple physical, chemical, and biological parameters, the model aims to identify the key factors that most significantly influence water pollution levels.

Furthermore, this research seeks to build a predictive framework that can classify water samples into distinct categories such as potable or non-potable. Through analytics, the project intends to detect potential contamination risks at an early stage, enabling authorities and researchers to implement timely preventive measures. Ultimately, the objective is to promote sustainable water resource management by combining technological innovation with environmental responsibility. By automating and optimizing the assessment process, this project contributes to public health protection, informed decision-making, and the long-term preservation of water quality.

2. Literature Review

Recent advancements in water quality assessment have shifted from traditional physicochemical analysis and GIS-based spatial modeling toward machine learning (ML) driven predictive frameworks. Early works employed statistical and GIS-integrated approaches to map spatial variability and pollution sources, as demonstrated in studies on the Ganga River and groundwater in West Bengal, which emphasized parameters such as pH, dissolved oxygen (DO), turbidity, conductivity, and temperature as key indicators influencing potability classification (Ali et al. 2021). These methods laid the groundwork for modern predictive systems integrating remote sensing and ML. The emergence of supervised algorithms such as Support Vector Machines (SVM) and Random Forests (RF) has notably enhanced prediction accuracy by capturing nonlinear relationships between water parameters and quality indices (Azamathulla and Wu 2011). Ensemble approaches, including bagging, boosting, and voting classifiers, have further improved robustness and generalization, as their comparative evaluations revealed superior performance in ranking and classification tasks across diverse environmental datasets (Plaia et al. 2022). Studies leveraging Sentinel-3 OLCI

imagery for water quality and algal bloom monitoring showcased the growing synergy between remote sensing data and ML models, enabling real-time assessment over large spatial extents (Joshi et al. 2024). Recent reviews underscore how RF and improved SVM algorithms outperform traditional Water Quality Index (WQI) models in terms of prediction reliability and interpretability (Sakaa et al. 2022) (Uddin et al. 2023). Moreover, integrating multivariate statistical techniques with ML has facilitated contamination source identification, supporting more effective environmental management (Zhang et al. 2022). Overall, the literature reveals a transition from empirical modeling toward hybrid frameworks that integrate statistical, GIS, and ensemble ML techniques for holistic water quality analysis, emphasizing parameter-driven feature selection and the predictive superiority of ensemble models in classifying potability and ecological health (Mondal et al. 2016).

3. Methods

The proposed study adopts a machine learning (ML) pipeline incorporating Ensemble Learning techniques to predict water quality and determine its potability. The methodology consists of five key stages: data acquisition, data preprocessing, data splitting, model training and model evaluation. Each stage is designed to ensure reliability, accuracy, and scalability of the developed model.

3.1 Data acquisition

The dataset used in this research was obtained from reliable water quality repositories containing various physicochemical parameters such as pH, Biological Oxygen Demand (BOD), Conductivity, and Dissolved Oxygen (DO) etc. These parameters provide essential indicators for assessing water quality and serve as the input features for the predictive model.

3.2 Data Preprocessing

Before training, the dataset undergoes a series of preprocessing operations to ensure data consistency and accuracy. Handling Missing Values, missing entries in key attributes such as pH, Conductivity, and Nitrate are treated using imputation techniques such as mean or forward-fill methods. This step ensures that incomplete data do not bias the model's learning process. Encoding Categorical Features, categorical variables, such as state are converted into numerical form using one-hot encoding, enabling compatibility with ML algorithms. Feature Selection, non-informative attributes are removed to reduce noise and improve model performance.

3.3 Data Splitting

After preprocessing, the cleaned dataset is divided into training and testing subsets in an 80:20 ratio. The training set is used to fit the models, while the testing set evaluates their generalization performance on unseen data.

3.4 Model Training

The training phase employs an ensemble-based architecture that combines the strengths of multiple machine learning classifiers. The base models include Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB). Each model is trained independently on the training dataset to learn distinct patterns. Their outputs are then aggregated using a Voting Classifier with soft voting, which computes the weighted average of probabilities from each classifier. This ensemble mechanism enhances predictive accuracy, reduces variance, and improves overall model robustness.

3.5 Model Evaluation

The performance of both individual models and the ensemble classifier is assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide a comprehensive understanding of the model's predictive capability and reliability. Experimental results indicate that the ensemble classifier consistently outperforms the individual base learners in terms of overall accuracy and stability.

4. Data Collection

The dataset used for this research was collected from the Central Pollution Control Board (CPCB) website, encompassing river water quality data for the years 2021, 2022, and 2023. The raw data included key physical, chemical and biological parameters such as Temperature, Dissolved Oxygen (DO), pH, Conductivity, Biochemical

Oxygen Demand (BOD), Nitrate, Fecal Coliform, and Total Coliform, along with corresponding station information and water drinkability labels.

The initial dataset consisted of 4,494 samples from various monitoring stations across India. Each record represented water quality readings at a particular location and time. Preliminary inspection revealed the presence of inconsistent values, outliers, and data entry errors such as extreme values for pH (>14), BOD (>1000 mg/L), and conductivity (>40,000 µS/cm). These were handled through domain-based clipping and scaling transformations to ensure reliability in downstream analysis. After collecting the data properly, we checked the missing data percentage to find the best way to fill those missing values. Data with below 15% missing value were filled with median while data with 20-30% missing values were filled with KNN impute. While feature column like fecal streptococci with more than 50% missing values was dropped (Figure 1).

STATION CODE	0.000000
NAME OF MONITORING LOCATION	0.000000
STATE NAME	0.000000
TEMPERATURE	1.426025
DO	0.311943
pH	0.000000
CONDUCTIVITY	3.609626
BOD	3.141711
NITRATE	11.519608
FECAL COLIFORM	12.678253
TOTAL COLIFORM	11.452763
FECAL STREPTOCOCCI	55.481283
YEAR	0.000000
Drinkable	0.000000

Figure 1. Missing value percentage of features.

Finally, we got a dataset of size 4488 rows. Also, based on the CPCB guideline the labels were set with the data, if the water is drinkable, it was assigned label 1 and if it was not drinkable it was assigned label 0. Preprocessing was done on the dataset and then we plot the data along with the label and the correlation we found among the data and label is shown below in Figure 2.

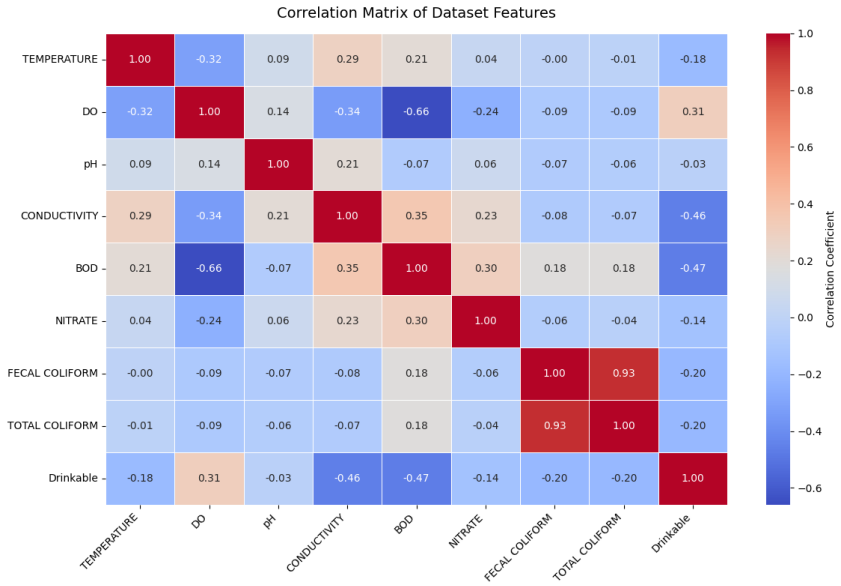

Figure 2. Correlation matrix of features and label.

5. Results and Discussion

5.1 Numerical Results

The raw dataset's numerical distribution is summarized in Table 1, highlighting extreme skewness and kurtosis in several attributes, especially BOD, Nitrate, and Coliform counts, suggesting the presence of strong outliers.

Table 1. Numerical Summary of Raw Dataset

Feature	Mean	Median	Std. Dev	Skewness	Kurtosis
Temperature	23.15	24.00	4.63	-1.17	1.95
DO	6.86	7.15	1.67	-1.13	3.01
pH	7.82	7.67	7.17	48.46	2636.75
Conductivity	782.65	345.00	2304.04	9.68	119.02
BOD	4.89	2.00	45.62	61.88	4026.87
Nitrate	2.73	1.08	12.70	28.38	1059.12
Fecal Coliform	1.25×10^{6}	248.75	8.30×10^{6}	10.31	190.60
Total Coliform	5.71×10^{6}	1037.50	3.98×10^{7}	10.82	206.60

The initial model used a Voting Classifier (SVM, Random Forest, Gradient Boosting, Logistic Regression) and achieved an accuracy of 98%, with a macro F1-score of 0.98, as shown below in Table 2,

Table 2. Results we get from raw data

Class	Precision	Recall	F1-Score	Support
0	1.00	0.97	0.98	438
1	0.97	1.00	0.98	460
Accuracy			0.98	898

After outlier handling, log transformation, and scaling, the revised dataset exhibited improved normality, as summarized in Table 3.

Table 3. Numerical Summary After Preprocessing

Feature	Mean	Median	Std. Dev	Skewness	Kurtosis
Temperature	0.00	0.18	1.00	-1.17	1.95
DO	0.00	0.17	1.00	-1.16	2.88
pH	0.00	0.05	1.00	-1.17	5.42
Conductivity	0.12	0.00	0.98	1.09	3.30
BOD	0.39	0.00	1.33	2.51	7.38
Nitrate	0.23	0.00	0.81	2.02	6.67
Fecal Coliform	0.12	0.00	0.69	0.82	0.15
Total Coliform	0.14	0.00	0.75	1.05	0.97

Following these improvements, the Voting Classifier performance increased to 99% accuracy with F1 = 0.99, confirming the effectiveness of preprocessing.

Table 4. Result from processed data

Class	Precision	Recall	F1-Score	Support
0	1.00	0.97	0.99	438
1	0.98	1.00	0.99	460
Accuracy			0.99	898

The model exhibited high classification capability both before and after preprocessing. However, the marginal gain from 98% to 99% accuracy indicates that handling extreme outliers and scaling enhanced the model's stability and generalization.

5.2 Graphical Results

The graphical distribution of features before preprocessing showed severe skewness in figure 3, particularly for BOD, Nitrate, and Coliform counts, which deviated significantly from normal distribution. After applying log transformation and scaling, the feature distributions became more symmetric in Figure 4, improving learning performance.

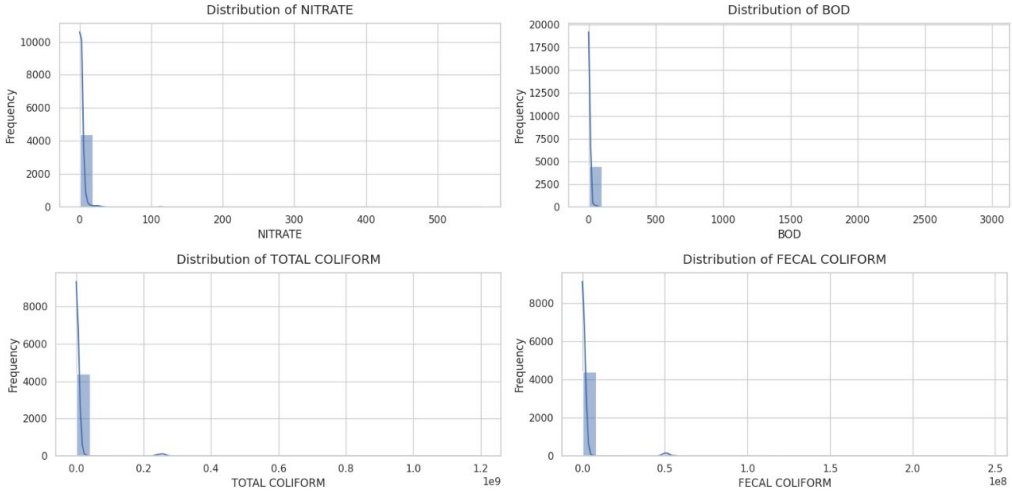

Figure 3. Distribution of raw water quality parameters before transformation.

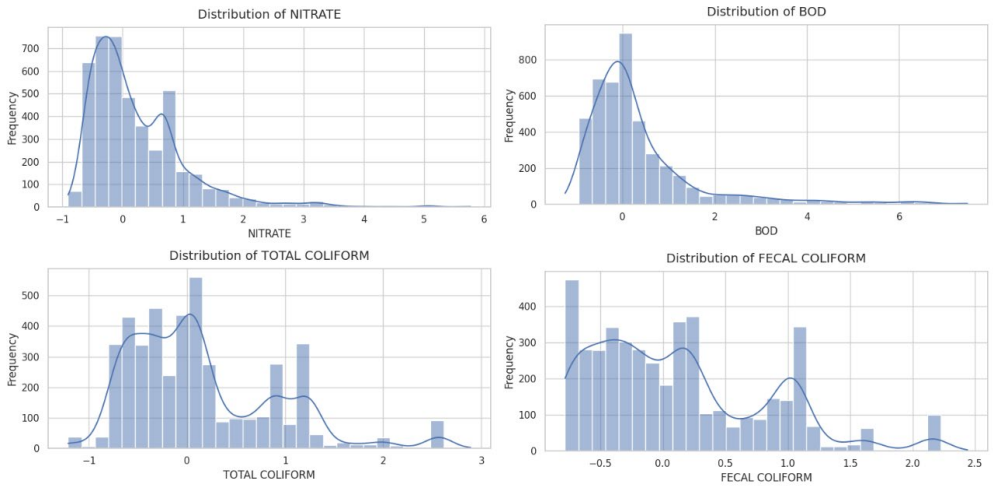

Figure 4. Normalized feature distributions after log scaling.

5.3 Proposed Improvements

To further enhance prediction accuracy and generalization, the following improvements are proposed, Incorporating Additional Features, include parameters such as Turbidity, Total Dissolved Solids (TDS), and Chemical Oxygen Demand (COD) for richer environmental representation. Secondly, temporal Trend Analysis for time-series models (LSTM) to capture seasonal variations across years. Use of spatial Mapping where integrate geospatial models to visualize water quality trends across different river basins, and also model ensembling to introduce advanced stacking frameworks with meta-learners to improve decision boundaries. Numerical and graphical results from extended

datasets are expected to show improved predictive accuracy (≥ 99.5%) and enhanced interpretability through heatmaps and trend visualizations. Add this model with IoT to make water quality test easily accessible.

5.4 Validation

Model validation was performed using cross-validation to ensure generalization across subsets. Multiple models were evaluated individually: SVM achieved an accuracy of 89%, Gradient Boosting reached 98%, and Random Forest and Logistic Regression showed comparable performance. Combining these models into a Voting Classifier yielded the best overall accuracy of 99%, demonstrating the effectiveness of ensemble learning in improving predictive performance and stability across the dataset.

6. Conclusion

This study successfully addressed all research objectives by analysing and predicting river water drinkability using machine learning techniques. Water quality data from the CPCB for the years 2021 to 2023 were collected, cleaned, and pre-processed to handle noise, outliers, and extreme values. Feature transformation and scaling were applied to normalize highly skewed parameters such as BOD, Nitrate, and Coliform counts, ensuring the dataset was suitable for robust model training. Multiple models were evaluated individually, SVM achieved an accuracy of 89%, Gradient Boosting reached 98%, while Random Forest and Logistic Regression showed strong but slightly lower performance. By integrating these models into a Voting Classifier, the study achieved the highest accuracy of 99%, demonstrating the benefits of ensemble learning in improving predictive reliability and stability across diverse river systems.

The unique contribution of this research lies in designing a hybrid ensemble framework optimized for environmental datasets with high variance and non-normal distributions, which can be generalized to other water bodies. The results emphasize the potential of machine learning models for real-time water quality monitoring and decision-making, providing actionable insights for environmental management authorities. Moreover, this framework lays a foundation for future enhancements, including the integration of temporal trends, geospatial analysis, and additional water quality parameters, ultimately supporting sustainable water resource management and public health protection.

References

Ali, S.Y., Sunar, S., Saha, P., Mukherjee, P., Saha, S. and Dutta, S., Drinking water quality assessment of river Ganga in West Bengal, India through integrated statistical and GIS techniques, *Water Science and Technology*, vol. 84, no. 10-11, pp. 2997–3017, 2021.

Azamathulla, H.M. and Wu, F.-C., Support vector machine approach for longitudinal dispersion coefficients in natural streams, *Applied Soft Computing*, vol. 11, no. 2, pp. 2902–2905, 2011.

Gupta, S. K., Gupta, R. C., Seth, A. K., Gupta, A. B., Bassin, J. K. and Gupta, A., Methaemoglobinaemia in areas with high nitrate concentration in drinking water, *National Medical Journal of India*, vol. 13, no. 2, pp. 58–61, 2000.

Gupta, V., Mishra, V.K., Singhal, P. and Kumar, A., An overview of supervised machine learning algorithm, *Proceedings of the 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 87–92, India, 2022.

Joshi, N., Park, J., Zhao, K., Londo, A. and Khanal, S., Monitoring harmful algal blooms and water quality using Sentinel-3 OLCI satellite imagery with machine learning, *Remote Sensing*, vol. 16, no. 13, p. 2444, 2024.

Kangabam, R.D., Bhoominathan, S.D., Kanagaraj, S. and Govindaraju, M., Development of a water quality index (WQI) for the Loktak Lake in India, *Applied Water Science*, vol. 7, pp. 2907–2918, 2017.

Mondal, I., Bandyopadhyay, J. and Paul, A.K., Water quality modeling for seasonal fluctuation of Ichamati River, West Bengal, India, *Modeling Earth Systems and Environment*, vol. 2, p. 113, 2016.

Nag, S.K. and Ghosh, P., Variation in groundwater level and water quality in Chhatna Block, Bankura district, West Bengal – a GIS approach, *Journal of the Geological Society of India*, vol. 81, pp. 261–280, 2013.

Plaia, A., Buscemi, S., Fürnkranz, J. and Loza Mencía, E., Comparing boosting and bagging for decision trees of rankings, *Journal of Classification*, vol. 39, no. 1, pp. 78–99, 2022.

Sakaa, B., Elbeltagi, A., Boudibi, S., Chaffaï, H., Md Towfiqul Islam, A.R., Kulimushi, L.C., Choudhari, P., Hani, A., Brouziyne, Y. and Wong, Y.J., Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin, *Environmental Science and Pollution Research*, vol. 29, no. 32, pp. 48491–48508, 2022.

Thenkabail, P.S., Remote Sensing Open Access Journal: Increasing Impact through Quality Publications, *Remote Sensing*, vol. 6, no. 8, pp. 7463–7468, 2014.

Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I., Performance analysis of the water quality index model for predicting water state using machine learning techniques, *Process Safety and Environmental Protection*, vol. 169, pp. 808–828, 2023.

Yu, P., Gao, R., Zhang, D. and Liu, Z.P., Predicting coastal algal blooms with environmental factors by machine learning methods, *Ecological Indicators*, vol. 123, p. 107334, 2021.

Zhang, Z., Zhang, F., Du, J.-L. and Chen, D.-C., Surface water quality assessment and contamination source identification using multivariate statistical techniques: A case study of the Nanxi River in the Taihu Watershed, China, *Water*, vol. 14, no. 5, p. 778, 2022.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. and Ye, L., A review of the application of machine learning in water quality evaluation, *Eco-Environment & Health*, vol. 1, no. 2, pp. 107–116, 2022.

Biographies

Ranita Dey was born in September 2000 in Kolkata. She is presently completing her degree in Computer Science and Engineering from Camellia Institute of Engineering and Technology (Affiliated by Maulana Abul Kalam Azad University of Technology). She had received her Diploma in Computer Science and Technology degree from West Bengal State Council of Technical & Vocational Education and Skill Development, Kolkata, India in 2022. She had completed her Higher Secondary education from West Bengal Council of Higher Secondary Education in 2019.

Dipanita Sharma was born in February 2002 in Asansol. She is presently completing her degree in Computer Science and Engineering from Camellia Institute of Engineering and Technology (Affiliated by Maulana Abul Kalam Azad University of Technology). She had received her Diploma in Computer Science and Technology degree from West Bengal State Council of Technical & Vocational Education and Skill Development, Kolkata, India in 2022. She had completed her Higher Secondary education from Kulti Girls High School under West Bengal Council of Higher Secondary Education in 2019.

Atefa Khatun was born in January 2002 in Bardhaman. She is presently completing her degree in Computer Science and Engineering from Camellia Institute of Engineering and Technology (Affiliated by Maulana Abul Kalam Azad University of Technology).She had received her Diploma in Computer Science and Engineering and Technology from the West Bengal State Council of Technical & Vocational Education and Skill Development, Kolkata, India in 2022.She had completed her Higher Secondary education from Mankar High School under the West Bengal Council of Higher Secondary Education in 2019.

Rahul Kumar was born in December 1997 in Dhanbad. He is presently completing his degree in Computer Science and Engineering from Camellia Institute of Engineering and Technology (Affiliated by Maulana Abul Kalam Azad University of Technology). He had received her Diploma in Electrical and Technology degree from West Bengal State Council of Technical & Vocational Education and Skill Development, Kolkata, India in 2022. He had completed his Higher Secondary education from Bihar Sanskrit Shiksha Board under Bihar Council of Higher Secondary Education in 2018.

Dharitri Sarkar was born in December 1999 in Bardhaman . presently completing her degree in Computer Science and Engineering, Camellia Institute of Engineering and Technology (Affiliated by Maulana Abul Kalam Azad University of Technology). She had received her Diploma in Computer Science and Technology degree from West Bengal State Council of Technical & Vocational Education and Skill Development, Kolkata, India in 2021. She had completed her Higher Secondary education from Nabagram Moyna P.B. High School under West Bengal Council of Higher Secondary Education in 2018.

Sagnik Dutta was born in April 1973 in Kolkata. He is presently an Assistant Professor in Department of Computer Science and Engineering, Camellia Institute of Engineering and Technology. He had received his M.E degree from Maulana Abul Kalam Azad University of Technology, formerly known as West Bengal University of Technology, Kolkata, India. His areas of interest include Artificial Intelligence and Machine Learning.