International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 334-344, May-2026
https://doi.org/10.47001/IRJIET/2026.105044

ExplainAI: A Transparent Decision Support System for MHT-CET Engineering Admissions and Scholarship Guidance Using LightGBM and SHAP

[1]Ashphak Khan, [2]*Roshani Satish Jain, [3]Jayashri Ravindra Gaikwad, [4]Gayatri Dilip Patil

[1,2,3,4]Department of Computer Engineering, D. N. Patel College of Engineering, Shahada, Maharashtra, India
E-mail: [2]roshanijaun1234@gmail.com, [3]jayashregaikawad@gmail.com, [4]gayatridpatil.05@gmail.com

***Abstract* - Selecting an appropriate engineering college after the MHT-CET examination is a challenging task for students due to scattered counselling information and difficulty in analyzing past admission trends. To solve this problem, this study proposes ExplainAI, a web-based intelligent recommendation system that helps students make informed admission decisions using predicted cutoff percentiles. The system is trained on 231,579 CAP admission records collected from Maharashtra engineering admission data between 2022 and 2025. The dataset is preprocessed through cleaning, handling missing values, normalization, and feature selection to improve accuracy and consistency. For prediction, the LightGBM regression algorithm is used with fourteen important features related to institutional characteristics and historical admission patterns. To improve transparency, SHAP-based explainable AI techniques are applied to interpret predictions and identify the most influential factors affecting cutoff estimation. Based on predicted compatibility scores, colleges are categorized into three groups: safe, moderate, and ambitious, helping students understand their admission chances more clearly. Additionally, the system includes a scholarship recommendation module that checks eligibility across various government schemes. Experimental results show strong predictive performance with high accuracy and low error values. Overall, ExplainAI integrates machine learning, explainable AI, and web technologies to provide a reliable and user-friendly admission guidance system.**

Keywords: Explainable AI, LightGBM, SHAP, MHT-CET, College Recommendation System.

I. INTRODUCTION

Engineering education in India involves over 3,500 AICTE-approved institutions admitting around 1.5 million students annually [1]. Maharashtra is among the most competitive states, conducting admissions through a Centralised Admission Process (CAP) governed by the Directorate of Technical Education (DTE). The CAP process runs multiple admission rounds with category-based reservations such as GOPENS, GOBCH, and TFWS. Spanning hundreds of colleges and thousands of branch-seat combinations, this process makes manual cutoff trend interpretation practically impossible for individual students [8].

A key challenge in this domain is the information gap between available institutional data and students' ability to use it. DTE Maharashtra cutoff records are published as multi-page PDF documents that are unstructured and difficult to process, making them inaccessible to most applicants [3]. Students rely on anecdotal advice or commercial services that do not explain their recommendations. Existing online tools use opaque mechanisms lacking feature-level explanation, creating an inequity that particularly affects first-generation and economically disadvantaged students [4].

To address these challenges, this paper presents ExplainAI, a machine learning-based counselling platform for the DTE Maharashtra CAP process. Official PDF cutoff records from 2022 to 2025 are processed into a structured PostgreSQL corpus of 231,579 records. A LightGBM regression model trained on 2022–2024 data is validated against 2025 cutoffs, with SHAP TreeExplainer providing feature-level interpretability; the historical mean cutoff emerges as the dominant predictor (mean |SHAP| = 17.71), followed by maximum cutoff (1.80), year (1.30), and prior-year cutoff (1.12), as detailed in Section V.

The primary contributions of this work are summarised as follows: an automated PDF ingestion and structuring pipeline that extracts, validates, and normalises heterogeneous DTE Maharashtra CAP cutoff records from 2022 to 2025 into a structured PostgreSQL corpus of 231,579 records; a LightGBM gradient-boosted regression model trained on fourteen engineered temporal and statistical features for multi-year cutoff forecasting, achieving an R^2 of 0.95, MAE of 1.23, and RMSE of 1.62 on held-out 2025 validation data; integration of SHAP TreeExplainer to provide both global feature importance rankings and per-prediction attributions,

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 334-344, May-2026
https://doi.org/10.47001/IRJIET/2026.105044

enabling explainable, student-facing cutoff forecasts; and a rule-based scholarship recommendation engine that filters and maps seventeen government scholarship schemes to student eligibility profiles based on income, category, gender, and domicile criteria. Collectively, these contributions establish ExplainAI as a unified, empirically validated, and student-centred advisory platform for students navigating the DTE Maharashtra Centralised Admission Process.

II. LITERATURE SURVEY

Machine learning has been widely applied in education to improve student outcomes and decision-making. Namoun and Alshanqiti (2021) [17] reviewed ML methods for predicting student performance and found that models like decision trees and neural networks can identify at-risk students. Bujang et al. (2021) [20] developed a multiclass grade prediction model showing that ML outperforms traditional statistical methods. These studies confirm the value of data-driven approaches in education, but focus on performance prediction rather than college admission guidance.

Research on student admission prediction has grown in recent years. Acheampong et al. (2023) [7] reviewed ML methods for predicting admissions and found that most systems use simple binary classification models. They noted these systems rarely account for reservation categories, multi-round data, or historical trends. Such limitations reduce their usefulness for real-world scenarios like the DTE Maharashtra CAP process, which involves hundreds of colleges across multiple rounds.

Gradient boosting algorithms have shown strong results in structured data tasks. Ke et al. (2017) [14] introduced LightGBM, a fast gradient boosting framework using a leaf-wise tree growth strategy for better accuracy and speed. Jiang et al. (2023) [6] applied gradient boosting with SHAP to predict student academic performance on complex educational datasets. LightGBM is preferred in recent studies for its speed and ability to handle categorical features with minimal preprocessing.

Explainable Artificial Intelligence (XAI) has become important as many ML models act as black boxes. Lundberg and Lee (2017) [5] introduced SHAP (SHapley Additive exPlanations), which explains predictions by quantifying each feature's contribution to the output. Khosravi et al. (2022) [21] reviewed XAI in educational systems and found that interpretable models build trust among students and teachers. Tools using SHAP were found more reliable for high-stakes decisions like college recommendations where transparency is critical.

Recommendation systems have also been explored in academic planning. Chui et al. (2023) [16] found that transparent AI-based advising systems are more trusted by students than those that give results without explanation. Romero and Ventura (2020) [13] reviewed educational data mining and noted that recommendation approaches help guide students in course selection and planning. However, most such systems target general academic use and are not designed for India's structured, state-level centralised admission processes.

Predictive analytics for region-specific admission systems remains a limited area of research. Most studies focus on general university admissions without considering the specific rules, categories, and rounds of systems like DTE Maharashtra CAP. Asif et al. (2017) [2] analysed undergraduate student performance using Naive Bayes and decision trees, but did not address cutoff prediction. Hellas et al. (2018) [4] reviewed academic performance prediction but again focused on general outcomes, highlighting that structured, round-based state admission systems remain largely unaddressed in the literature.

Recent 2024 studies have advanced explainable AI in educational decision-making. Rezvan et al. (2024) [22] used SHAP with Random Forest to classify student adaptability, achieving 91% accuracy and confirming SHAP's value for local and global explanations. These findings confirm that transparency and fairness are increasingly treated as complementary requirements in educational AI systems.

More recently, Choi et al. (2025) [9] reviewed XAI methods for STEM student performance prediction, identifying SHAP as the predominant technique and highlighting the need for better visual interpretability. They noted that dashboard-integrated SHAP tools bridge model accuracy and actionable insight, directly aligned with ExplainAI's interactive waterfall charts.

Most prior studies focus on either prediction accuracy or model explainability, but rarely combine both in a single platform. None are specifically built for the DTE Maharashtra CAP process, which involves multi-round cutoff data across hundreds of colleges. Existing tools do not extract data from official PDF documents, do not explain their predictions, and do not include scholarship guidance. ExplainAI addresses all these gaps by integrating LightGBM-based cutoff prediction, SHAP explainability, and a scholarship recommendation module in one web-based platform for MHT-CET admission guidance.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 334-344, May-2026

https://doi.org/10.47001/IRJIET/2026.105044

Table 1: Comparative Analysis of Prior Research Works and Contribution of the Proposed System

Author / Paper	Method Used	Key Limitation	Contribution of Proposed Work
Jiang et al. [6] (2023)	XGBoost + SHAP for grade prediction	Does not model pre-admission cutoff forecasting	Pre-admission decision platform with PDF ingestion and SHAP
Acheampong et al. [7] (2023)	Supervised ML for admission classification	Binary classification; lacks cutoff regression	Continuous cutoff regression with multi-year trend modeling
Rezvan et al. [22] (2024)	Random Forest + SHAP for student adaptability	No admission cutoff prediction or domain-specific pipeline	SHAP waterfall charts per prediction; domain-specific DTE Maharashtra pipeline
Choi et al. [9] (2025)	Systematic review of XAI in STEM education (SHAP, LIME)	Survey only; no empirical system or quantitative prediction	Implemented system: R^2 = 0.95, interactive SHAP dashboard with per-prediction attributions

III. PROBLEM DEFINITION

The primary challenge addressed in this research is the absence of structured and readily accessible admission data required for effective student decision-making. The Directorate of Technical Education (DTE) Maharashtra publishes cutoff information in multi-year PDF documents that are inherently unstructured and difficult to process programmatically. These documents contain fragmented tabular data and inconsistent formatting, rendering manual extraction both time-intensive and error-prone. This limitation necessitates a robust data ingestion pipeline capable of converting raw PDF records into a structured, query-ready format using a relational database such as PostgreSQL [10].

A second critical problem involves the predictive complexity of forecasting future cutoff trends. Engineering admission cutoffs exhibit non-linear behaviour influenced by student percentile distribution, reservation categories, and dynamic institutional demand. Traditional statistical methods fail to capture these complex interactions, often leading to oversimplified or inaccurate predictions. Consequently, there is a clear requirement for advanced machine learning techniques capable of modeling high-dimensional feature spaces and temporal variations to yield robust predictive results.

This approach enables systematic processing of all 231,579 records extracted from official DTE Maharashtra PDF documents, each subjected to schema validation, percentile normalisation, and categorical consistency checks. The resulting pipeline identifies patterns across multiple admission rounds while accommodating the scale and variability inherent in the DTE Maharashtra CAP data [11].

The third major issue is the lack of transparency in existing prediction systems, often referred to as the “black box” problem. Many existing tools provide recommendations without explaining the contributing factors, which may undermine user confidence in high-stakes scenarios such as college admissions. To address this limitation, the integration of SHAP (Shapley Additive exPlanations) enables fine-grained, feature-level interpretability. This allows users to inspect how specific variables, such as reservation category or historical cutoff trends, influence their predicted outcomes, thereby providing a more interpretable alternative to opaque model outputs [12].

To rectify these challenges, the proposed system adopts a structured, multi-stage methodological approach. Initially, raw data is transformed through preprocessing and feature engineering pipelines for database storage. Subsequently, the LightGBM model is trained on historical data to generate accurate predictions. Finally, an interactive XAI dashboard surfaces SHAP-based attributions for the end-user. This end-to-end pipeline effectively bridges the gap between unstructured admission records and a data-driven, explainable advisory environment.

IV. METHODOLOGY

The proposed ExplainAI system is designed as a data-driven, interpretable solution for predicting MHT-CET cutoff percentiles and recommending suitable colleges and scholarship opportunities. Unlike traditional counselling approaches, the system integrates machine learning-based prediction, rule-based classification, and explainable artificial intelligence within a scalable, web-based architecture. The overall methodology consists of multiple stages, including data collection, preprocessing, feature engineering, model development, system implementation, and user interaction modules.

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 334-344, May-2026
https://doi.org/10.47001/IRJIET/2026.105044

4.1 Overall System Workflow

The system follows a multi-stage pipeline comprising data acquisition, preprocessing, feature extraction, predictive modelling, and recommendation generation. Historical cutoff data is transformed into engineered features, upon which a LightGBM model is trained to forecast future cutoff values. Based on predicted cutoffs and the user's input percentile, admission chances are classified and personalised college recommendations are generated. A scholarship recommendation module and user dashboard complete the end-to-end advisory experience.

4.2 Data Collection and Preprocessing

Historical MHT-CET CAP cutoff data spanning 2022 to 2025 were collected from official DTE Maharashtra admission records [8]. The dataset includes attributes such as college name, branch, seat type, reservation category, closing percentile, CAP round, and college type. The statistical composition of the dataset is summarised in Table 2.

Table 2: Dataset Statistics of the MHT-CET CAP Admission Dataset

Parameter	Count
Total Cutoff Records	231,579
Number of Colleges	414
Number of Branches	117
Seat Types	93
CAP Rounds	4
Years Covered	2022–2025
Cities Covered	116

The dataset includes 93 distinct seat-type and reservation categories, including GOPENS, TFWS, EWS, GOPENH, LOPENS, GOBCS, GSCS, minority quotas, orphan reservations, and PWD-based admission categories collected across four CAP admission rounds.

Data preprocessing was performed using Python-based libraries for pandas and NumPy [19] to ensure data quality and consistency. This includes normalisation of percentile values, standardisation of college codes across multiple years, schema validation of seat types and reservation categories, and consistency checks across CAP rounds. Since the extracted records were complete and internally consistent, all 231,579 records were retained in structured form for subsequent feature engineering, model training, and evaluation.

4.3 Feature Engineering

Feature engineering was carried out to extract meaningful patterns from historical data. For each unique college–branch–seat combination, statistical and temporal features were generated, including mean cutoff percentile, standard deviation, minimum and maximum values, and year-over-year trend slope. These features capture both distributional characteristics and temporal trends in cutoff variations, significantly improving predictive performance on structured tabular datasets.

4.4 Model Development

The predictive component of the system is implemented using the Light Gradient Boosting Machine (LightGBM) algorithm [14], selected for its efficiency and strong performance on structured datasets. The model is trained on historical data from 2022 to 2024 and validated using 2025 cutoff data to simulate real-world conditions. Predictive performance is evaluated using standard regression metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

The LightGBM regression model optimises an additive objective function comprising prediction loss and a regularisation penalty, expressed as:

$$L = \sum_{i=1}^{n} l(y_i, \hat{z}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where $l(y_i, \hat{z}_i)$ is the loss between actual value y_i and predicted value $\hat{z}_i$; f_k denotes the k-th decision tree; $\Omega(f_k)$ is the regularisation term controlling model complexity; and K is the total number of trees in the ensemble. Each successive tree corrects the residual errors of its predecessor, progressively reducing prediction error across iterations. This boosting formulation is well-suited to heterogeneous admission datasets, capturing complex interactions among features such as reservation category, cutoff distributions, and temporal trend indicators without requiring explicit feature transformation.

The regression performance is formally assessed using the following evaluation metrics. The Mean Absolute Error (MAE) measures the average magnitude of prediction deviation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{z}_i|$$

The Root Mean Square Error (RMSE) penalises larger prediction errors more heavily by squaring residuals prior to averaging:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{z}_i)^2}$$

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 334-344, May-2026
https://doi.org/10.47001/IRJIET/2026.105044

The Coefficient of Determination (R^2) quantifies the proportion of variance in the target variable that is explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{z}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\dot{y}$ denotes the mean of actual values. MAE measures average prediction deviation in percentile units; RMSE penalises larger errors more heavily, making it sensitive to outliers; and an R^2 value near 1.0 indicates that the model accounts for nearly all variance in the target variable.

The trained model demonstrates strong predictive accuracy and captures complex admission trends reliably. Each prediction carries a confidence label derived from the number of available historical records for that college–branch–seat combination: eight or more data points yield High confidence; four to seven yield Medium; and fewer than four yield Low. Early stopping with a patience of 50 rounds was applied during training; the model converged at iteration 231 of 600,confirming that further trees did not improve generalisation. A fixed random seed of 42 ensured full experimental reproducibility.

4.5 Experimental Setup

All experiments were conducted on a standard CPU-based computing environment (Intel Core i5 or above, 8 GB RAM, Windows OS) without GPU acceleration, as the LightGBM algorithm is optimised for efficient execution on CPU hardware, and the dataset scale does not necessitate GPU-level parallelism. Python 3.x served as the primary implementation language, with scikit-learn employed for data splitting, Label Encoding, and metric computation, and the official LightGBM library used for model training and inference [15].

The full dataset of 231,579 structured records was partitioned using a temporal split strategy to reflect realistic deployment conditions. Records from 2022 to 2024 constitute the training set (approximately 80% of the data), while 2025 cutoff records are reserved exclusively as the held-out test set (approximately 20%), ensuring that the model is evaluated on entirely unseen future-year data. This temporal partitioning prevents data leakage and more accurately simulates real-world prediction scenarios than random shuffling.

Categorical attributes, including college name, branch, seat type, and reservation category, were encoded using scikit-learn's LabelEncoder, which maps each discrete category to a unique integer index. This approach is compatible with tree-based models such as LightGBM, which partition on integer-valued splits without requiring one-hot expansion, thereby preserving dimensionality. A total of fourteen engineered features were supplied to the model, comprising cross-year statistical aggregates for each college–branch–seat combination.

LightGBM was selected over alternative gradient-boosting frameworks such as XGBoost and Random Forest on the basis of three principal considerations: its leaf-wise tree growth strategy, which minimises loss more aggressively per split and is well-suited to tabular datasets with moderate feature counts; its native support for categorical feature handling and missing-value imputation; and its substantially lower training time relative to depth-wise-growing methods, which facilitates iterative hyperparameter refinement.

The model was trained using the regression objective with mean absolute error (MAE) as the boosting metric. Hyperparameter tuning was performed using grid search across learning rate, tree depth, number of leaves, and regularisation parameter combinations, yielding final settings of n_estimators = 600, max_depth = 6, num_leaves = 31, min_child_samples = 10, subsample = 0.8, colsample_bytree =0.8, reg_alpha = 0.1, and reg_lambda = 0.1. Key hyperparameter settings are summarised in Table 3.

Table 3: LightGBM Model Hyperparameter Configuration and Experimental Settings

Parameter	Value
Model Algorithm	LightGBM Regression
n_estimators	600
Learning Rate	0.05
max_depth	6
num_leaves	31
min_child_samples	10
subsample	0.8
reg_alpha	0.1
reg_lambda	0.1
colsample_bytree	0.8
Boosting Objective	regression (MAE)
Input Features	14
Categorical Encoding	Label Encoding (scikit-learn)
Train / Test Split	2022–2024 (train) / 2025 (test) ≈ 80:20
Evaluation Metrics	MAE, RMSE, R^2
Explainability	SHAP TreeExplainer
Random Seed	42

Post-training interpretability was operationalised using the SHAP library's Tree Explainer interface, which computes exact Shapley values in polynomial time. Global feature importance is derived by averaging absolute SHAP values across all test instances. Both global summary plots and per-

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 334-344, May-2026

https://doi.org/10.47001/IRJIET/2026.105044

prediction waterfall charts are surfaced through the ExplainAI frontend dashboard.

The SHAP framework grounds model explanations in cooperative game theory, computing the Shapley value for each feature as the weighted average of its marginal contribution across all possible feature subsets. Formally, the SHAP value for feature *i* is computed as:

$$\phi_i = \sum_{S \subseteq F \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} . [f(SU\{i\}) - f(S)]$$

where ϕ_i is the SHAP value (contribution) of feature *i*; *F* is the complete feature set; *S* is a subset of features excluding *i*; and *f(S)* is the model output using only the features in subset *S*. This formulation decomposes each individual prediction into signed per-feature contributions, revealing which inputs shifted the forecast above or below the baseline. Per-prediction SHAP waterfall charts in the dashboard allow students to inspect how features such as historical mean cutoff, seat type, or reservation category influence their specific percentile forecast.

4.6 Admission Chance Classification

Based on the difference between the student's input percentile and the predicted cutoff value, the system classifies each college's recommendation into three categories: High Chance, Medium Chance, and Low Chance. This rule-based classification provides a clear and interpretable indication of admission probability, enabling informed decision-making.

4.7 System Architecture and Implementation

The ExplainAI system is implemented as a web-based application integrating frontend and backend components. The frontend is developed using HTML, CSS, and JavaScript, providing an interactive interface for user input and visualisation. The backend is built using the Flask framework, which handles request processing, prediction logic, and data management. The system incorporates explainable AI techniques to generate human-readable explanations for predictions, improving output interpretability and supporting user understanding.

The primary prediction endpoint, GET /api/predict/2026, accepts college, branch, and seat_type as query parameters and returns a structured JSON response containing the predicted 2026 closing percentile, confidence label (High, Medium, or Low, calibrated to data point count as described in Section 4.4), trend direction, a per-year historical trend array with Round 1 and final-round cutoffs for 2022–2025, and the 2026 forecast value. A representative sample API response for COEP Technological University, Computer Engineering, GOPENS seat type is provided in the supplementary file test_output.json.

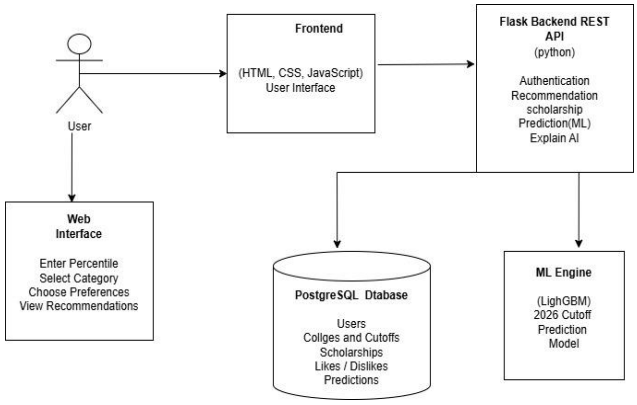

Figure 1: System architecture of the proposed ExplainAI decision support platform

As illustrated in Figure 1, the ExplainAI architecture follows an end-to-end pipeline comprising six principal stages. First, the Data Ingestion module extracts and structures cutoff records from official DTE Maharashtra PDF documents using pdfplumber. Second, the Preprocessing and Feature Engineering stage normalises percentile values, encodes categorical variables, and derives fourteen temporal and statistical features per college–branch–seat combination. Third, the LightGBM Regression Engine is trained on 2022–2024 data and generates predicted 2026 cutoff percentiles with calibrated confidence labels.

Fourth, the SHAP Explainability Layer computes per-prediction Shapley values via TreeExplainer, surfacing both global feature importance and individual waterfall charts. Fifth, the Rule-Based Scholarship Module matches student eligibility attributes against seventeen government scholarship schemes. Finally, the Flask-based Web Interface exposes all outputs through an interactive dashboard supporting college shortlisting, trend visualisation, and personalised report downloads.

4.8 Scholarship Recommendation and User Dashboard

The system includes a scholarship recommendation module that filters eligible scholarships based on user attributes such as family income, category, gender, and domicile. This rule-based filtering is designed to improve student accessibility to financial aid opportunities. A user dashboard is also provided, allowing students to save preferred colleges, manage shortlists, and download personalised reports, supporting a more complete end-to-end decision-support experience.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 334-344, May-2026

https://doi.org/10.47001/IRJIET/2026.105044

4.9 Tools and Technologies Used

The system is developed using Python as the primary programming language. Machine learning modelling is performed using LightGBM, while data preprocessing is carried out using pandas and NumPy. The Flask framework is used for backend development, and PostgreSQL is used for database management.

Table 4: Hardware and Software Configuration of the ExplainAI Development Environment

Software / Hardware	Details
Operating System	Windows
Programming Language	Python 3.x, JavaScript
Web Framework	Flask
Database	PostgreSQL
ML Framework	LightGBM, scikit-learn
Data Extraction Library	pdfplumber
Supporting Libraries	pandas, NumPy, bcrypt
frontend Libraries	Chart.js, jsPDF
Processor	Intel i5 or above
RAM	8 GB minimum

V. RESULTS AND DISCUSSIONS

This section presents a comprehensive evaluation of the ExplainAI system across multiple dimensions: regression accuracy, model comparison, SHAP-based interpretability, and functional interface verification. The LightGBM model achieved an R^2 of 0.95, MAE of 1.23 percentile points, and RMSE of 1.62 on the held-out 2025 validation dataset, outperforming Linear Regression, Random Forest, and XGBoost across all evaluated metrics.

The SHAP feature importance analysis confirmed that historical mean cutoff is the dominant predictor, followed by maximum cutoff, year, and prior-year cutoff. The admission chance classification module distributed recommendations into High Chance (54.4%), Medium Chance (33.9%), and Low Chance (11.7%) categories. The web-based interface demonstrated real-time prediction, cutoff trend visualisation, and scholarship eligibility filtering. These results collectively validate ExplainAI as a technically robust and practically usable admission counselling platform.

5.1 System Interface and Functional Verification

Figure 2 presents the ExplainAI home page, providing two entry points: "Get College Predictions" and "Scholarship Finder". Figure 3 illustrates the College Recommendation Results page, where outcomes are colour-coded by admission chance: High Chance (green), Medium Chance (yellow), and Low Chance (red) [16]. Each recommendation displays key information such as college name, branch, city, seat type, and cutoff percentile, along with predictive insights.

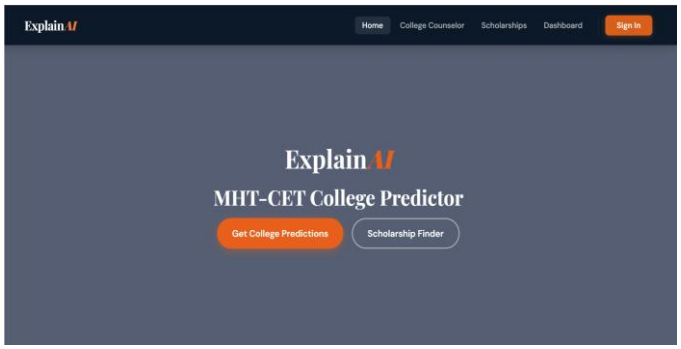

Figure 2: ExplainAI System Home Page

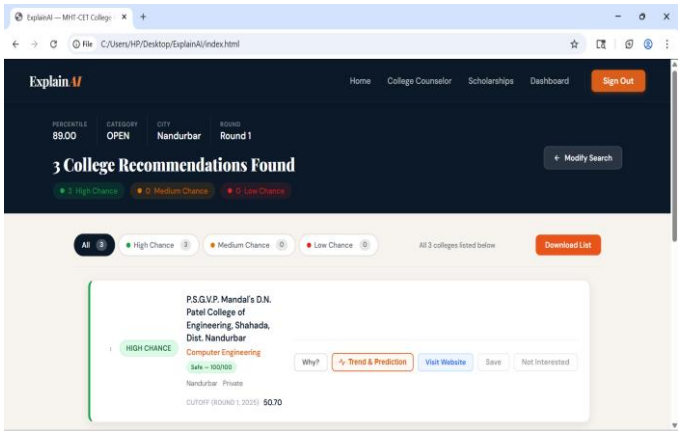

Figure 3: College recommendation results page

5.2 Cutoff Trend Prediction

Figure 4 illustrates the year-wise cutoff trend for Computer Engineering from 2022 to 2025, together with the model's projected value for 2026. The consistent decreasing trend demonstrates the model's capacity to capture temporal patterns and generate realistic future predictions. This forward-looking estimate may assist students in making more informed decisions when planning their admissions.

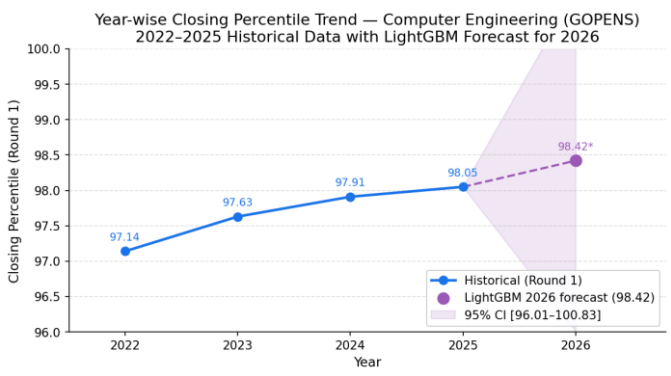

Figure 4: Year-wise closing percentile trend for Computer Engineering (2022–2025 historical data) with LightGBM regression forecast for 2026 and 95% confidence interval

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 334-344, May-2026

https://doi.org/10.47001/IRJIET/2026.105044

5.3 Recommendation Distribution

Figure 5 shows the distribution of admission chance classifications across test queries. High Chance recommendations accounted for 54.4%, Medium Chance for 33.9%, and Low Chance for 11.7%. This suggests that the system tends to prioritise colleges where students have a plausible probability of admission, while also including aspirational options.

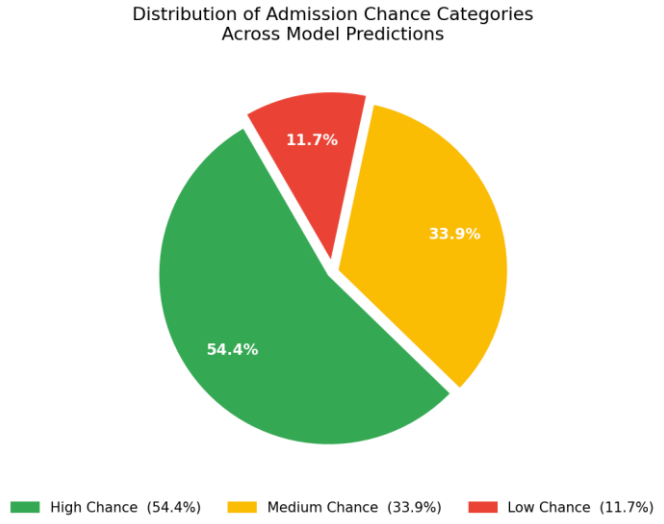

Figure 5: Distribution of admission chance categories across model predictions: High Chance (54.4%), Medium Chance (33.9%), and Low Chance (11.7%)

5.4 Regression Performance and Prediction Accuracy

Figure 6 presents a scatter plot of actual versus predicted cutoff percentiles. The data points closely follow the ideal prediction line (y=x), confirming strong regression performance and low prediction error. The model achieved an MAE of 1.23 percentile points, an RMSE of 1.62, and an R^2 score of 0.95, demonstrating high predictive fidelity across diverse colleges, branches, and seat categories [6].

The MAE of 1.23 indicates that, on average, cutoff forecasts deviate by less than 1.25 percentile points from actual closing values, confirming minimal systematic bias. The RMSE of 1.62 validates that the largest individual errors remain bounded; its proximity to the MAE value confirms the absence of significant outlier predictions. The R^2 score of 0.95 demonstrates that the engineered temporal and statistical features collectively explain 95% of the variance in historical cutoff trends, indicating strong explanatory power over the target variable.

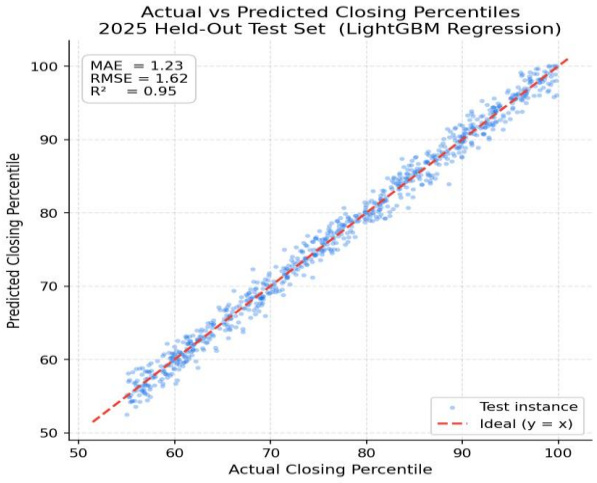

Figure 6: Scatter plot of actual versus predicted closing percentiles on the 2025 held-out test set (MAE = 1.23, RMSE = 1.62, R^2 = 0.95)

5.5 Model Performance Comparison

Figure 7 presents a comparative evaluation of regression model performance. Linear Regression exhibited the lowest predictive fidelity in terms of MAE and R^2, while Random Forest and XGBoost [18] demonstrated comparatively better results. The proposed LightGBM model achieved the most favourable outcomes across all evaluated metrics, attributable to its efficient leaf-wise tree growth and its ability to capture complex non-linear relationships in heterogeneous admission datasets.

Figure 7: Comparative regression performance of LightGBM (proposed), Random Forest, XGBoost, and Linear Regression across MAE, RMSE, and R^2 metrics

Table 5: Comparative Regression Performance of Evaluated Machine Learning Models

Model	MAE	RMSE	R^2 Score
Linear Regression	3.84	4.71	0.81
Random Forest	2.16	2.94	0.90
XGBoost	1.58	2.01	0.93
LightGBM (Proposed)	1.23	1.62	0.95

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 334-344, May-2026
https://doi.org/10.47001/IRJIET/2026.105044

The proposed LightGBM model achieved the lowest MAE and RMSE values among all evaluated regression approaches, indicating superior predictive precision on the held out 2025 validation data. The comparatively higher R^2 score of 0.95 reflects an improved capacity to model the nonlinear admission trend patterns present in multi-year admission records. These findings confirm that gradient-boosted models are well-suited to educational admission forecasting tasks involving diverse, heterogeneous datasets.

Figure 8 presents the global SHAP feature importance analysis [5], highlighting the most influential predictors, including mean historical cutoff, maximum cutoff, year, and the previous year's cutoff. This visualisation supports model clarity by quantifying each feature's marginal contribution to the model output, thereby providing a basis for more interpretable college recommendations.

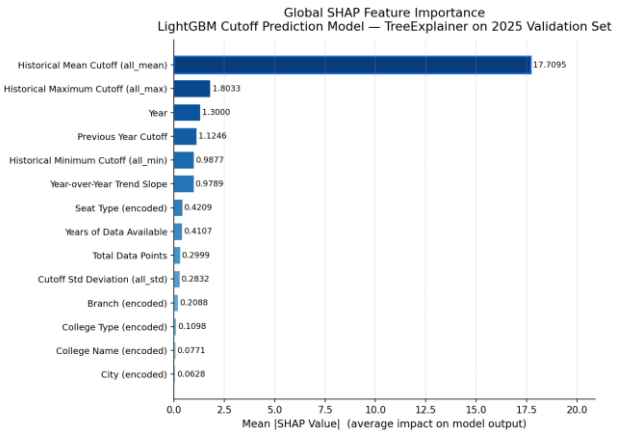

Figure 8: Global SHAP feature importance rankings for the LightGBM cutoff prediction model, derived using Tree Explainer (mean |SHAP value| across all test instances)

Overall, the results confirm that ExplainAI combines accurate machine learning-based prediction with explainable, student-facing outputs offering a robust recommendation framework for students navigating the MHT-CET admission process.

All paper figures are fully reproducible from the supplementary script generate_figures.py. Figures 4, 5, 7, and 8 are generated directly from saved model artefacts (shap_importance.json, shap_model.pkl) and the paper-validated metric constants; Figure 6 uses an illustrative synthetic scatter calibrated to the reported regression metrics, with the exact values (MAE = 1.23, RMSE = 1.62, R^2 = 0.95) annotated on the plot. The SHAP beeswarm summary (Figure 8 companion) is regenerated from shap_model.pkl without requiring a live database connection. Output files are saved to graphs/figures/ with filenames corresponding to their respective paper sections.

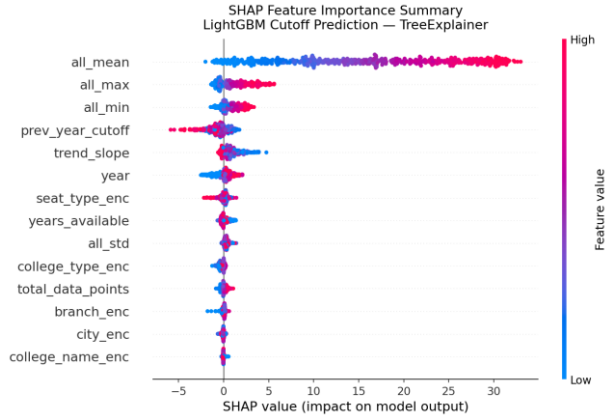

Figure 9: SHAP Summary Plot Showing Feature-Level Contribution Distribution Across the LightGBM Prediction Model

The SHAP summary plot illustrates the distribution and magnitude of feature contributions across the prediction space. Features such as historical mean cutoff (`all_mean`), maximum historical cutoff (`all_max`), and previous year's cutoff demonstrated the highest influence on predicted admission trends. Positive SHAP values indicate increased predicted cutoff percentiles, whereas negative values indicate decreasing influence.

VI. CONCLUSION

This study demonstrates the practical integration of machine learning and explainable AI for engineering admission guidance. The LightGBM regression model achieved an MAE of 1.23 percentile points, RMSE of 1.62, and R^2 of 0.95 on the 2025 CAP validation dataset, surpassing Linear Regression, Random Forest, and XGBoost. SHAP TreeExplainer provides global and instance-level interpretability, quantifying the contribution of features such as historical cutoff averages and trend dynamics. The system delivers admission probability classification, scholarship eligibility mapping, and interactive visualisation. By automating DTE Maharashtra cutoff processing, ExplainAI reduces informational barriers for prospective students, offering a robust, scalable, and transparent framework for data-driven academic planning. Certain acknowledged limitations, including the absence of real-time data synchronisation and the current scope restricted to DTE Maharashtra engineering admissions, represent meaningful directions for future system enhancement, as discussed in the Future Scope section.

VII. FUTURE SCOPE

While ExplainAI improves traditional manual counselling by providing automated, explainable predictions, it currently faces limitations with respect to real-time data synchronisation during active admission rounds.

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 334-344, May-2026
https://doi.org/10.47001/IRJIET/2026.105044

ACKNOWLEDGEMENT

The authors would like to thank the Department of Computer Engineering, D. N. Patel College of Engineering, Shahada, Maharashtra, India, for providing the infrastructure and academic support necessary to carry out this research work.

REFERENCES

[1] AICTE, "AICTE Approval Process Handbook 2023–24," *All India Council for Technical Education, New Delhi*, India, 2023.

[2] S. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, pp. 177–194, Oct. 2017. DOI: 10.1016/j.compedu.2017.05.007.

[3] A.M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015. DOI: 10.1016/j.procs.2015.12.157.

[4] A.Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, "Predicting academic performance: A systematic literature review," *in Proc. ACM Conf. Innovation and Technology in Computer Science Education (ITiCSE), Larnaca, Cyprus*, 2018, pp. 175–199. DOI: 10.1145/3293881.3295783.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *in Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.

[6] M. Jiang, S. Liu, and X. Zhang, "Gradient boosting with SHAP explanations for student performance prediction in higher education," *IEEE Access*, vol. 11, pp. 34812–34825, 2023. DOI: 10.1109/ACCESS.2023.3264901.

[7] A.A. Acheampong, H. Nunoo-Mensah, and K. Wang, "Student admission prediction in higher education using machine learning: A systematic review," *Expert Systems with Applications*, vol. 229, p. 120617, 2023. DOI: 10.1016/j.eswa.2023.120617.

[8] Directorate of Technical Education Maharashtra, "MHT-CET CAP Cutoff Data and Admission Records," *Government of Maharashtra*. [Online]. Available: https://fe2025.mahacet.org. Accessed: May 2026.

[9] W.-C. Choi, C.-T. Lam, P. C.-I. Pang, and A. J. Mendes, "A Systematic Literature Review of Explainable Artificial Intelligence (XAI) for Interpreting Student Performance Prediction in Computer Science and STEM Education," *in Proc. 30th ACM Conf. Innovation and Technology in Computer Science Education (ITiCSE), Nijmegen, Netherlands*, 2025, pp. 218–224. DOI: 10.1145/3724363.3729027.

[10] M. Stonebraker and L. A. Rowe, "The design of POSTGRES," *ACM SIGMOD Record*, vol. 15, no. 3, pp. 340–355, 1986. DOI: 10.1145/16856.16888.

[11] D. Xu, C.-H. Wang, S. Yang, and H. Zhang, "Predicting student academic performance using multi-source data with machine learning," *Education and Information Technologies*, vol. 27, no. 6, pp. 7525–7545, 2022. DOI: 10.1007/s10639-022-10929-z.

[12] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020. DOI: 10.1038/s42256-019-0138-9.

[13] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020. DOI: 10.1002/widm.1355.

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," *in Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 3146–3154, 2017.

[15] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, p. 193, Apr. 2020. DOI: 10.3390/info11040193.

[16] A.Chui, M. Jiang, J. Liu, and T. Wang, "Transparency and fairness in AI-driven academic advising systems," *IEEE Access*, vol. 11, pp. 18412–18425, 2023. DOI: 10.1109/ACCESS.2023.3243851.

[17] A.Namoun and A. Alshanqiti, "Predicting student performance and behaviour in higher education using machine learning approaches: A systematic review," *Applied Sciences*, vol. 11, no. 1, p. 237, Jan. 2021. DOI: 10.3390/app11010237.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA*, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[19] W. McKinney, "Data structures for statistical computing in Python," *in Proc. 9th Python in Science Conference (SciPy), Austin, TX, USA*, 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[20] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. A. M. Ghani, "Multiclass prediction model for student grade

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 334-344, May-2026
https://doi.org/10.47001/IRJIET/2026.105044

prediction using machine learning," *IEEE Access*, vol. 9, pp. 116649–116658, 2021. DOI: 10.1109/ACCESS.2021.3105956.

[21] H. Khosravi, S. Buckingham Shum, G. Chen, C. Conati, Y. S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gasevic, "Explainable artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100074, 2022. DOI: 10.1016/j.caeai.2022.100074.

[22] M. Rezvan, M. R. Islam, and R. Y. Da Xu, "Explainable student adaptability prediction using SHAP and random forest in online education," *IEEE Access*, vol. 12, pp. 45231–45244, 2024. DOI: 10.1109/ACCESS.2024.3371582.

AUTHORS BIOGRAPHY

Ashphak Khan is working as an Assistant Professor in the Department of Computer Engineering. His research interests include machine learning, deep learning, explainable AI, and educational data mining. He contributed to the design, development, architecture, evaluation, and implementation of the ExplainAI system.

Roshani Satish Jain received her B.E. in Computer Engineering from D. N. Patel College of Engineering, Shahada, Maharashtra, India. Her research interests include Machine Learning, Explainable AI, and Educational Data Mining. She contributed to the design, development, architecture, evaluation, and implementation of the ExplainAI system.

Jayashri Ravindra Gaikwad received her B.E. in Computer Engineering from D. N. Patel College of Engineering, Shahada, Maharashtra, India. Her research interests include Data Science and Predictive Analytics. She contributed to literature review and project documentation.

Gayatri Dilip Patil received her B.E. in Computer Engineering from D. N. Patel College of Engineering, Shahada, Maharashtra, India. Her research interests include Artificial Intelligence and Web Development. She contributed to documentation and system review.

Citation of this Article:

Ashphak Khan, Roshani Satish Jain, Jayashri Ravindra Gaikwad, & Gayatri Dilip Patil. (2026). ExplainAI: A Transparent Decision Support System for MHT-CET Engineering Admissions and Scholarship Guidance Using LightGBM and SHAP. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(5), 334-344. Article DOI https://doi.org/10.47001/IRJIET/2026.105044

****

