SLIIT
Discover Your Future

BSc (Hons) in Information Technology Specializing Data Science

Research Project - IT4010 Data Analysis Report

I Project ID: **24-25j-281**

Project Title: AI-Enhanced E-Learning Platform for the Hearing-impaired children

1. Introduction

1.1. Background:

Hearing-impaired children face significant challenges in traditional e-learning platforms, which are predominantly auditory-based and lack inclusive features like real-time sign language translation and interactive tools. This limits their access to quality education, particularly during the critical developmental years of ages 4–12. The diversity in sign languages and the need for personalized learning further compound these issues.

Emerging technologies such as AI, augmented reality (AR), and machine learning (ML) offer innovative solutions. AI can enable real-time sign language translation, AR can provide interactive modules with virtual hand gestures, and ML can deliver gesture recognition for immediate feedback.

This research aims to develop an AI-powered e-learning platform that integrates these technologies, ensuring accessible, tailored, and engaging learning experiences for hearing-impaired children, fostering equal educational opportunities and long-term success.

1.2. Research Problem:

Traditional e-learning platforms are predominantly auditory-based, which are not suitable for students with hearing impairments. These students face challenges such as:

- Ineffective communication and learning through traditional digital platforms.
- Lack of resources tailored to the gestural and visual nature of sign language.
- A need for platforms that can adapt educational content into sign language and provide interactive, real-time learning experiences.

1.3. Objectives:

It Number	Objective	Objective number
Rizan S - IT21311840	Develop interactive AR modules for teaching and practicing ASL with real-time feedback, enhancing learning through dynamic simulations that provide immediate correction and reinforcement of sign language gestures.	1
M.S.M Shazny - IT211736322	Implement a system that converts educational video audio into real-time captions and synchronized sign language, integrating speech recognition and translation for seamless, engaging learning for hearing-impaired students.	2

BSc (Hons) in Information Technology Specializing Data Science

Research Project - IT4010 Data Analysis Report

Z.F. Sahla - IT21159558	Develop an AI assistant using gesture recognition for doubt clarification in text and sign language, along with a recommendation system suggesting courses and a performance prediction system which predicts future performance and giving feedback based on students' quiz performance.	3

2. Data Exploration
   2.1. Data Collection
   Kaggle (ASL Alphabet)
   2.2. Dataset Description:
   (Describe datasets, including sources, size, and key attributes)

Data source	Description	Resource	Size	Key attributes
Kaggle - ASL Alphabet	The dataset contains images of alphabets from the American Sign Language (ASL), organized into 29 folders representing classes such as A-Z and gestures like SPACE, DELETE, and NOTHING.	GitHub Repository and Kaggle Platform	87,000 images (training set)	Image size: 200x200 pixels, 29 classes, useful for real-time applications and classification.
Kaggle - American Sign Language Dataset	The dataset consists of images representing ASL gestures for classification tasks. It can be applied for multiclass classification using CNN technology.	ASL Dataset Files in Kaggle	2515 files	Organized into 36 directories, curated for image-based classification tasks, supports up to 98% accuracy with CNN, images organized for numeric and alphabetic gestures.
Kaggle - ASL Alphabet Dataset	This dataset contains ASL alphabet images split into training and testing data. The training data includes 29 classes (A-Z and 3 gestures: SPACE, DELETE, NOTHING).	Kaggle Platform	4.56 GB	223k files, divided into training (29 directories) and testing (28 files), designed for machine learning models, focusing on static hand postures.

Synthetic dataset	This synthetic dataset is designed for predicting the improvement score based on features representing student activity and engagement, such as success rate, attempts, game performance, and time spent on tasks. It supports regression and time series analysis for performance prediction using the ARIMA model.	Custom-generated dataset using Python	100,000 rows, 7 features.	Includes timestamp, success count, attempt count, game score, game level, engagement time, and the target variable improvement score. Structured for time series forecasting using the ARIMA model to predict students' future learning performance based on historical data.
Augmented Dataset	A dataset designed for recommending personalized educational courses based on attributes like subject area, difficulty level, age range, and content relevance. The recommendation system leverages TF-IDF vectorization and cosine similarity to identify the most relevant courses for users.	A curated dataset combining course details from online learning platforms with synthetically generated attributes for recommendation purposes.	50 rows, 10 features.	Includes Course Title, Area, Difficulty Level, Recommended Age Range, Content Keywords, Keyword Percentages, and Course Links. Structured to support personalized course recommendations using TF-IDF vectorization and cosine similarity for relevance scoring.

2.3. Suitability Analysis

2.3.1. Relevance to Individual Research Objectives:

(Explain how well each dataset aligns with your research problem and objectives)

	1	2	3
Data source 1	X	X	
Data source 2	X	X	
Data Source 3	X	X	
Data Source 4	X	X	
Data Source 5	X	X	

3. Methodology

3.1. Data Preprocessing:

(Mention data transformation techniques done in each dataset for each objective.)

Ex:

Data Cleaning, Data Normalization, Data Standardization, Data Encoding (e.g., One-Hot Encoding, Label Encoding), Handling Missing Data (e.g., Imputation or Removal), Data Aggregation, Feature Engineering, Outlier Detection and Handling, Data Scaling, Data Discretization, Dimensionality Reduction (e.g., PCA), Date/Time Transformation, Data Integration (Merging or Joining), Data Mapping, Data Type Conversion.

Objective 1

	Transformation technique				
	Data Cleaning	Scaling	Feature Extraction	One Hot Encoding	Outlier Removal
Data source 1	x	x	x	x	x
Data Source 2	x	x	x	x	x
Data source 3	x	x	x	x	x
Data source 4	x	x	x	x	x
Data source 5	x	x	x	x	x

3.2. Scalability

- **The ASL dataset** was processed to enhance its usability and ensure compatibility with machine learning models. Images were preprocessed using **Mediapipe**, where hand landmarks were detected and cropped to focus only on the hand region. This preprocessing step reduces noise and scales the dataset effectively for deep learning applications. The processed images were saved for subsequent analysis, enabling scalable workflows for both training and testing.

- **The performance prediction dataset** was processed to prepare it for time series analysis and regression tasks. Key features, such as success count, attempt count, and engagement time, were scaled and structured to allow for accurate performance predictions. The ARIMA model was applied to the dataset, ensuring that it could handle a large volume of historical data for forecasting future performance. The system is designed to scale with additional student data, enabling long-term performance tracking and improvement prediction over time.

- **The course details dataset** was processed to facilitate efficient recommendations by incorporating TF-IDF vectorization. The content keywords for each course were

transformed into numerical vectors, which allowed for the calculation of cosine similarities between courses. This preprocessing step ensures scalability for recommending courses based on a student's preferences and past performance. The recommendation system is optimized to handle large datasets, allowing for real-time suggestions and personalized learning experiences for users with diverse learning needs.

3.3. Feature extraction

Component 1 and 2:

Using **Mediapipe**, features such as **x, y, z coordinates** of hand landmarks were extracted from each image. These features were stored in a CSV file for further preprocessing. The preprocessing steps included:

- **Handling missing values** by removing null entries.
- **Outlier removal** to ensure robust data quality.
- **Data cleaning** and **normalization** to standardize the dataset.
  After completing the preprocessing, the cleaned data was saved in an Excel file, providing a structured and efficient format for training machine learning models.

Component 3

For the performance prediction system, features such as success count, attempt count, game score, game level, and engagement time were extracted from the dataset. The preprocessing steps included:

- **Scaling:** Input features and the target variable were standardized using StandardScaler to ensure consistent ranges and improve model performance.
- **Data Transformation:** Features were transformed to remove any significant variations in scale.
- **Target Variable Scaling:** The improvement score was also scaled to normalize the values and ensure better predictions.

The processed data, along with the scalers for future use, were saved for training machine learning models, enabling efficient and scalable workflows for prediction tasks.

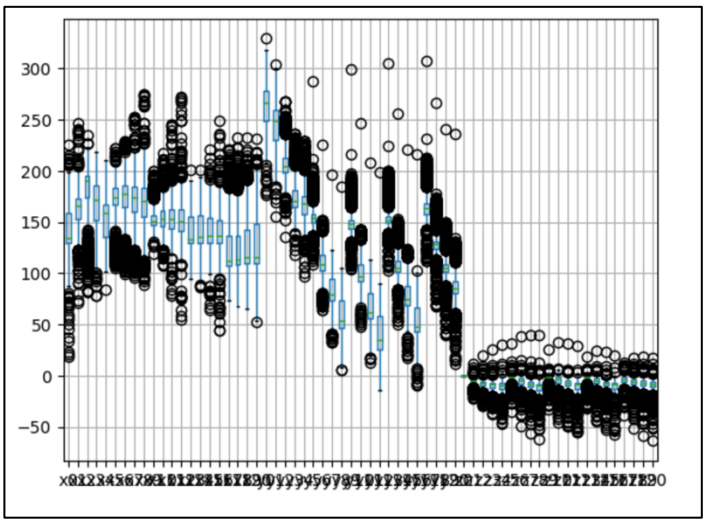
4. Modelling and Results

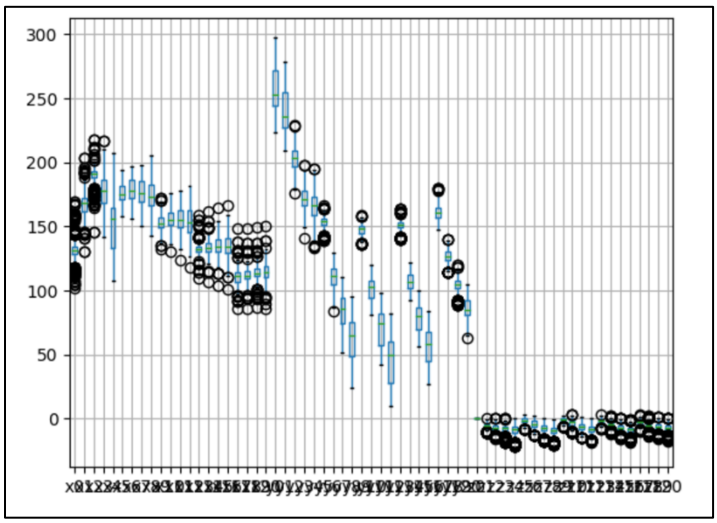
4.1. Key Insights:

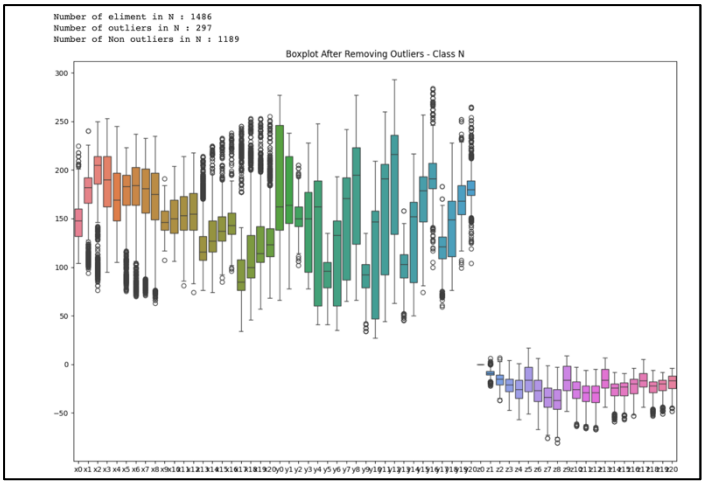
- Dataset Exploration:
  - Hand landmark features (x, y, z coordinates) are structured and consistent across the dataset.
  - The preprocessing pipeline ensured high-quality data with minimal outliers, improving model reliability.
- Data Patterns:

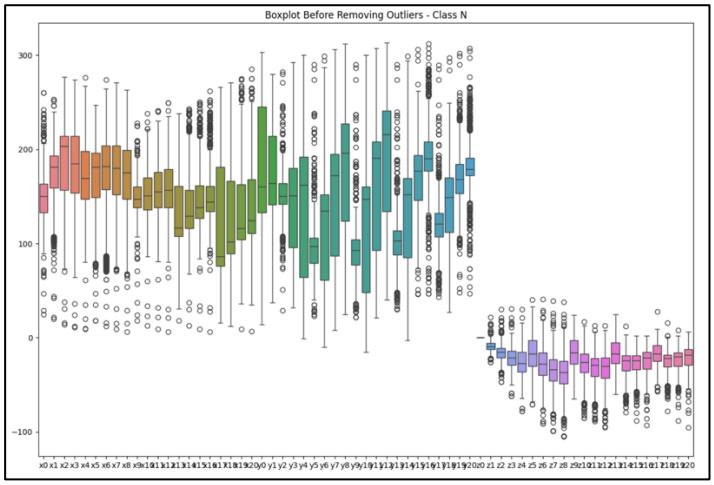
- Clear separation in landmark coordinates was observed between different gesture classes, enabling effective classification.
- The SPACE, DELETE, and NOTHING classes had distinct landmark distributions, aiding in differentiating static gestures from dynamic ones.

- Visualization Highlights:
  - Landmark Distribution: Scatter plots showed clear clustering of x, y, z coordinates for each gesture.
  - Outlier Detection: Box plots highlighted outliers, which were removed during preprocessing.
  - Performance Metrics: The cleaned and structured dataset improved training accuracy and reduced validation error.
- Trends & Correlations:
  - High correlations between specific landmarks and gesture types were identified, demonstrating the model's ability to distinguish between gestures.
  - The use of hand-cropped images enhanced feature extraction accuracy, reducing the noise introduced by background elements.









Component 3

Key Insights:

- Dataset Exploration
  - The dataset includes features like success count, attempt count, game score, game level, and engagement time, which were well-structured and cleaned.
  - The preprocessing pipeline ensured high-quality data by scaling both the features and target variables, making the data suitable for regression models.

- Data Patterns
  - Clear trends were observed between student engagement metrics (like game score and attempt count) and their improvement scores, enabling effective regression modeling.
  - Correlations between success rate, engagement time, and game performance indicated a strong relationship with improvement scores.

- Visualization Highlights:
  - Scatter plots of predicted vs. actual improvement scores revealed a good fit between the Random Forest and Linear Regression models.
  - The predictions from the models were visually consistent with the expected improvements, demonstrating model reliability.
  - Box plots and histograms highlighted the distribution of improvement scores, which were centered around a mid-range value.

- Performance Metrics:

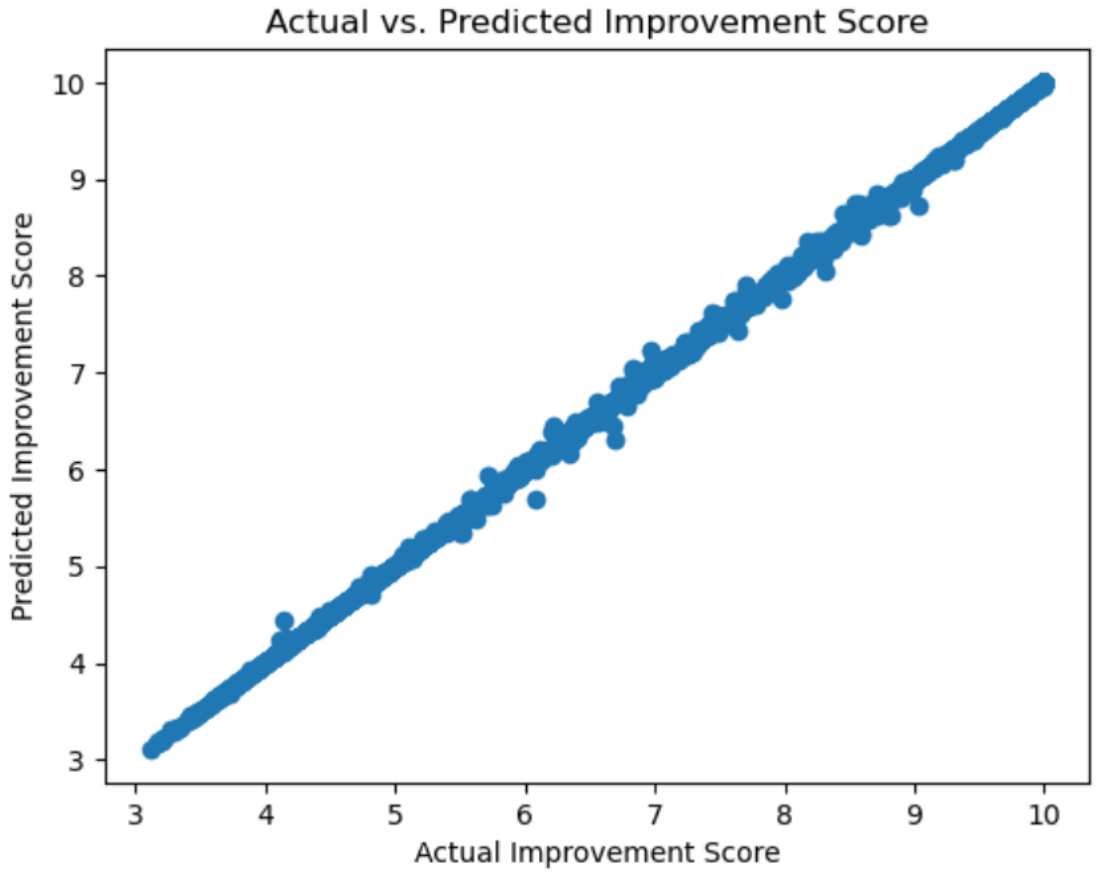
  - Random Forest Regressor
    - The model achieved strong R^2 scores, with a testing R^2 of 0.83, indicating that it explained a large portion of the variance in the target variable.
    - Other metrics like MSE (Mean Squared Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and explained variance further validated the robustness of the Random Forest model.

  - Linear Regression Model
    - The Linear Regression model performed similarly with an R^2 of 0.79 on testing data, offering a simpler yet effective prediction.

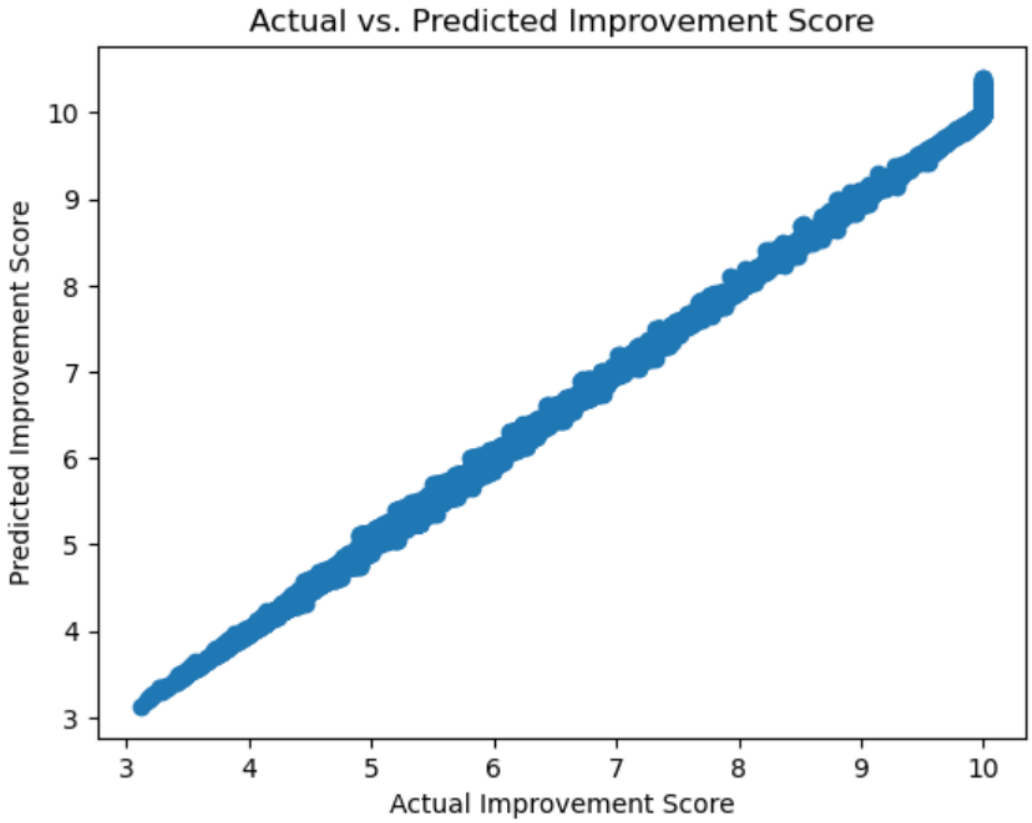
- It showed a reasonable fit to the data with acceptable performance metrics, though it lagged behind the Random Forest Regressor in explaining variance.

- Trends & Correlations
  - Significant trends between the student's gaming engagement metrics (success rate, attempts, time spent) and their learning improvements were observed, suggesting that these factors play a crucial role in predicting improvement.
  - The strong performance of the Linear Regression model further emphasizes the importance of these engagement metrics.

Random Forest Regressor



Linear Regression



Model Selection

Chosen Model: The Linear Regression model was selected for further use, as it provided an interpretable solution while still demonstrating good prediction accuracy. It performed comparably to the Random Forest Regressor but was simpler, requiring less computational power and making it more suitable for future implementation.

Future Work

- Refinement of Features: Further feature engineering could be performed to explore other potential predictors of student improvement scores, such as the types of learning activities.
- Model Optimization: Exploring more complex models or fine-tuning the existing models may lead to even better predictions.

- Real-Time Predictions: Implementing the model in a real-time environment, such as an interactive learning platform, could allow for continuous improvement tracking and personalized course recommendations based on predicted scores.

4.2. Challenges Faced During Data Analysis

During the data analysis for the gesture recognition, performance prediction, and recommendation systems, several challenges were encountered. In the gesture recognition system, missing or null values due to incomplete hand detection were addressed by estimating and filling in the gaps using interpolation. Outliers in the hand landmark coordinates were removed through outlier detection methods, and variability in hand positions across images was handled by normalization and scaling. Partial hand occlusion was mitigated by focusing only on visible landmarks, while class imbalance was addressed with data augmentation techniques for underrepresented gestures. In the performance prediction system, missing or null values in the dataset due to incomplete or erroneous data entries were imputed using appropriate techniques. Outliers in the engagement metrics and improvement scores were identified and removed to maintain data integrity. Variability in user performance across gaming sessions was standardized using a StandardScaler, ensuring consistent feature scaling. Additionally, class imbalance in improvement scores was mitigated by applying data augmentation strategies, including generating synthetic data points for underrepresented score ranges. For the recommendation system, challenges included handling missing data related to user preferences and interactions. These gaps were filled using collaborative filtering techniques and feature-based imputations. Outliers in user ratings and course engagement were detected and addressed to improve recommendation quality. Feature scaling and normalization were applied to ensure that the recommendation algorithm operated on a consistent and optimized dataset. These preprocessing steps ensured more accurate and reliable datasets for gesture recognition, performance prediction, and recommendation models.

5. References

1. Kambhampati S. Sindhu, Mehnaaz, Biradar Nikitha, Penumathsa Likhita Varma, and Chandrasekhar Uddagiri, "Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired," in Proc. 1st International Conference on Cognitive, Green, and Ubiquitous Computing (IC-CGU), 2024, pp. 1-8.

2. MMK Rowel, ADAI Gunasekara, GAI Uwanthika, and DB Wijesinghe, "An E-Learning Platform for Hearing Impaired Children," Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka, 2023.
3. "An E-Learning Platform for Hearing Impaired Children," details the limitations of current systems in integrating captions and sign language.
4. "Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired," discusses the need for more advanced AI techniques in sign language translation.
5. "E-Learning Model to Identify the Learning Styles of Hearing-Impaired Students," highlights the absence of integrated summarization tools in educational platforms for hearing-impaired students.
6. "Hastha Online Learning Platform for Hearing Impaired," underscores the challenges in scaling and adapting systems to various educational environments.
7. Zhou, Q., Li, X., & Chen, Z. (2023). "Gesture Recognition Using 3D Hand Landmark Detection for Human-Computer Interaction," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3400-3407.
8. Kumar, R., Gupta, V., & Sharma, S.(2022). "Predicting User Performance in Gamified Learning Environments Using Machine Learning Models," in *Journal of Educational Technology*, 29(4), 115-125.
9. Yuan, Y., Li, H., & Wang, Z.(2021). "Personalized Recommendation Systems for Gamified Learning Platforms," in *Proc. of the International Conference on Artificial Intelligence (AI)*, pp. 257-266.
10. Nguyen, H., & Lee, T.(2020). "Class Imbalance Handling in Performance Prediction Models for Educational Platforms," in *Journal of Machine Learning for Education*, 15(2), 34-45.
11. Soni, S., & Chaurasia, A.(2023). "Challenges and Solutions in User Interaction for Gesture-Based Recommendation Systems," in *International Journal of Computer Science and Applications*, 50(6), 159-172.