

# Automatic English Essay Scoring Based on Machine Learning

MAMIDI TARANI, SENAPATHI CHAITANYA SWARUPA

Assistant Professor, 2MCA Final Semester, Master of Computer Applications, Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

## Abstract

Abstract—With the development of natural language processing(NLP) technology and machine learning, the research task of automatic English scoring(AES) is becoming clearer and clearer, and the research difficulties arise due to the mutual constraints of research methods and annotation data. How to build a perfect and reliable scoring system has become a great challenge under today's research. In this paper, we designed an English AES system, and verified the effectiveness of RF on English scoring model by analyzing the prediction effect of RF on non-text features and text features, and then compared the Pearson correlation coefficients(PCC) of RF(RF), GBDT, and XGBoost, and the study showed that the performance of RF algorithm is higher than the other two composition scoring methods. Keywords—machine learning, random forest algorithm, automatic scoring of English essays, Pearson correlation coefficient.

## 1.INTRODUCTION

.1 For English teaching, a teacher often teaches several classes at the same time, with hundreds of students, so if each person has an essay to grade, the teacher's workload will be huge, and the subsequent need for targeted revision and re grading will multiply the teacher's workload. Automatic grading is fast and efficient, and can greatly reduce the time teachers spend on grading essays, allowing teachers to spend more time on teaching students and allowing English learners to train their English writing skills in a targeted manner. Research on automatic scoring of English essays has yielded good results. For example, some scholars have successfully developed the PEG system, which is the earliest automatic scoring system for essays. PEG assumes that the overall quality of an essay can be reflected by several shallow linguistic features, including the length of the essay, the length of the vocabulary, the number of prepositions, and the number of pronouns, etc. These shallow features of the essay are then extracted and used to predict the fluency of the essay, the richness of the vocabulary, the complexity of the sentence structure, etc., and then the regression coefficients are derived using Multiple regression method is used to find out the regression coefficients, and the score of the composition is predicted by regression calculation, which neither uses natural language processing techniques nor studies the content and chapter structure of the composition, nor does it consider the theme of the composition [1-2]. Some scholars input the theme of the composition together with the content of the composition into a neural network model so that the model learns the relevance of the content of the composition to the theme of the composition by itself, and the obtained tangency features are involved in the training of the overall scoring model of the composition [3]. There are many other AES models in English and all of them have been effective in scoring results. In this paper, we first introduce AES-related techniques and propose an evaluation method for English scoring; then we design an English AES system containing four functional modules and a machine learning prediction model, and finally validate the effectiveness of the RF model in predicting scoring on the ASAP essay set, the Grade 4 essay data, and the critique web essay data.

## .2 Existing System

- Traditional English essay scoring systems often rely on human evaluators, which introduces several limitations. Manual grading is time-consuming, inconsistent due to subjective bias, and impractical for large-scale evaluation in educational environments. Early automated systems like the Project Essay Grade (PEG) focused only on shallow linguistic features such as word count, vocabulary length, number of prepositions, and syntactic patterns. These systems use multiple regression to predict scores, but they do not consider deeper semantic understanding, context relevance, or essay structure.

- Some existing models have attempted to incorporate the essay topic into scoring through neural networks. However, these methods often lack interpretability and still struggle to capture complex dependencies between language features and thematic coherence. Furthermore, most traditional systems do not integrate advanced NLP techniques like BERT embeddings or ensemble machine learning models, which limits their scoring accuracy.

### .1.1 Challenges:

- Data Availability and Quality

Obtaining a large, annotated dataset of English essays with reliable human scores is difficult. Public datasets like ASAP are limited, and inconsistencies in human scoring affect model training quality.

- Feature Extraction Complexity

Extracting meaningful linguistic features (like vocabulary richness, syntactic complexity, topic relevance) requires advanced Natural Language Processing (NLP) techniques. Designing these features in a way that generalizes across different essay topics is a major challenge.

- Semantic Understanding of Text

Machine learning models often struggle to understand the deep semantic structure of essays, such as coherence, relevance to the topic, and logical flow—especially in free-text inputs from students with varied skill levels.

### Balancing Textual and Non-Textual Features

Combining non-text features (e.g., word count, grammar stats) with textual features (semantic embeddings) is challenging due to differences in their formats, distributions, and contribution to the final score.

### Proposed system:

The proposed system is an Automatic English Essay Scoring (AES) model that leverages advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques to evaluate English essays effectively and fairly. Unlike traditional systems that depend on shallow linguistic features, the proposed system integrates both non-textual features (e.g., word count, sentence length, grammar usage) and textual features (e.g., semantic coherence, topic relevance, sentence elegance) for holistic essay evaluation.

This system is composed of four main functional modules:

#### 1. User Login Module

Provides secure access for users to upload and submit essays for evaluation.

#### 2. Evaluation Module

The core of the system, where the AES model performs a multi-dimensional analysis:

- Vocabulary Quality Evaluation: Uses gradient boosting to assess lexical richness and diversity.
- Sentence Grace Evaluation: Applies a Convolutional Neural Network (CNN) model to determine fluency and readability.
- Topic Relevance Evaluation: Uses BERT embeddings to compute semantic similarity between essay content and the given topic across word, sentence, and paragraph levels.

#### 3. Statistics Module

Calculates metrics such as:

- Total number of words (TNO)
- Number of sentences
- Average length of sentences (ALOS)
- Count of advanced vocabulary and high-quality sentences

#### 4. Error Correction Module

Identifies spelling mistakes using the pyaspeller library and suggests corrections, improving the overall writing quality before scoring.

The Machine Learning Component includes a Random Forest (RF) model as the primary scoring algorithm, which is validated to outperform Gradient Boosting Decision Tree (GBDT) and XGBoost in accuracy and robustness. Additionally, the system enhances performance by fusing predictions from RF, GBDT, and XGBoost using bagging ensemble techniques.

By incorporating TF-IDF, word2vec, and semantic similarity measures, the system also calculates deductive degree features, improving score prediction for essays with strong thematic alignment.

Overall, this proposed system offers:

- Automated, real-time essay scoring
- Improved fairness and objectivity
- High scalability and consistency
- Integration of deep semantic understanding and linguistic analysis.

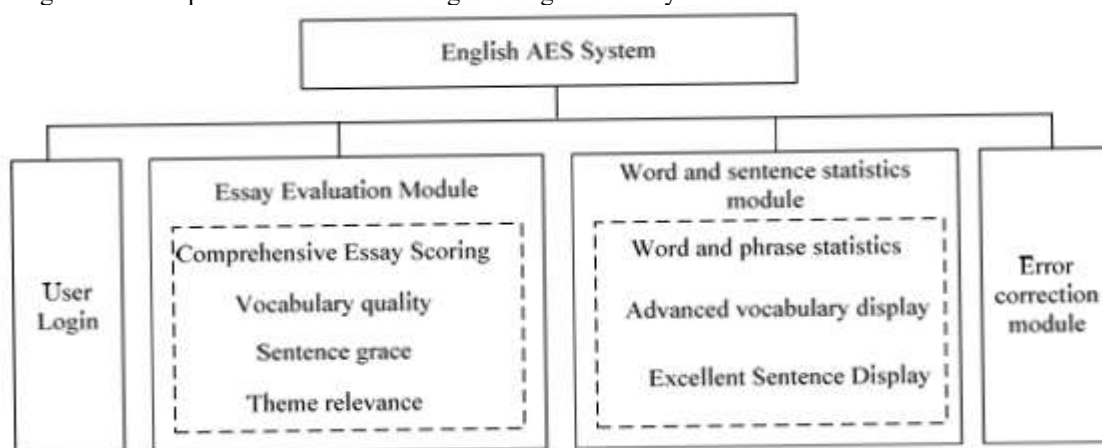


Fig: 1 Proposed Diagram on AES SYSTEM

### 1.1 Advantages:

- Objective and Unbiased Evaluation

Removes human bias from the grading process, ensuring fair and consistent scores for all students.

- Time-Saving for Educators

Reduces the manual effort of teachers by automating essay scoring, especially when dealing with large student populations.

- Immediate Feedback

Provides instant scoring and suggestions, allowing students to understand and improve their writing quickly.

- Supports Writing Skill Improvement

Detailed evaluation of vocabulary, sentence fluency, and topic relevance helps learners identify specific areas of weakness.

- Scalable and Efficient

Capable of evaluating thousands of essays in minimal time without compromising accuracy or performance.

- Advanced Error Detection

Integrated error correction module flags spelling mistakes and suggests corrections, improving overall essay quality.

- Multifactor Scoring Approach

Combines textual features (semantics, coherence) and non-textual features (grammar, length) for a comprehensive assessment.

- Integration with Learning Platforms

Can be deployed in online education systems to assist both teachers and learners in real-time academic environments.

### 2.1 Architecture:

The system architecture for the "Automatic English Essay Scoring Algorithm Based on Machine Learning" is designed as a sequential pipeline that begins with essay input and ends with score output. The process starts when a user submits an essay through a user-friendly web interface. This input is passed to a preprocessing module, where the text is cleaned and prepared by removing noise, tokenizing words and sentences, and normalizing formatting. Once preprocessing is complete, the data is sent to the feature extraction layer, which analyzes the essay for linguistic and structural characteristics such as word count, sentence length, vocabulary diversity, grammar accuracy, and part-of-speech tags. These extracted features are then fed into a trained machine learning model—such as Random Forest, GBDT, or XGBoost—which has been developed using a dataset of previously scored essays. The model evaluates the features and predicts a score for the essay, aiming to replicate human grading as closely as possible. This score is then sent back to the web interface, where it is displayed to the user, optionally along with feedback or scoring explanations. The system may also include additional components like a database (e.g., SQLite) for storing past submissions and scores, and support for model updates to ensure continuous improvement. Overall, the architecture integrates natural language processing, machine learning, and a web front end to provide an efficient and automated essay scoring solution.

### UML DIAGRAMS:

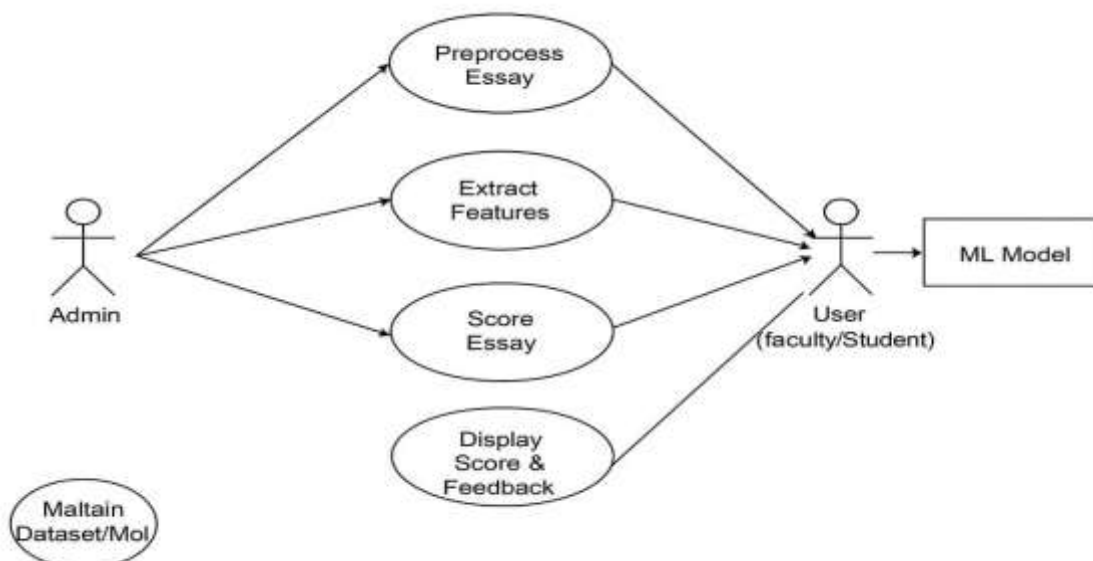


Fig:use case diagram

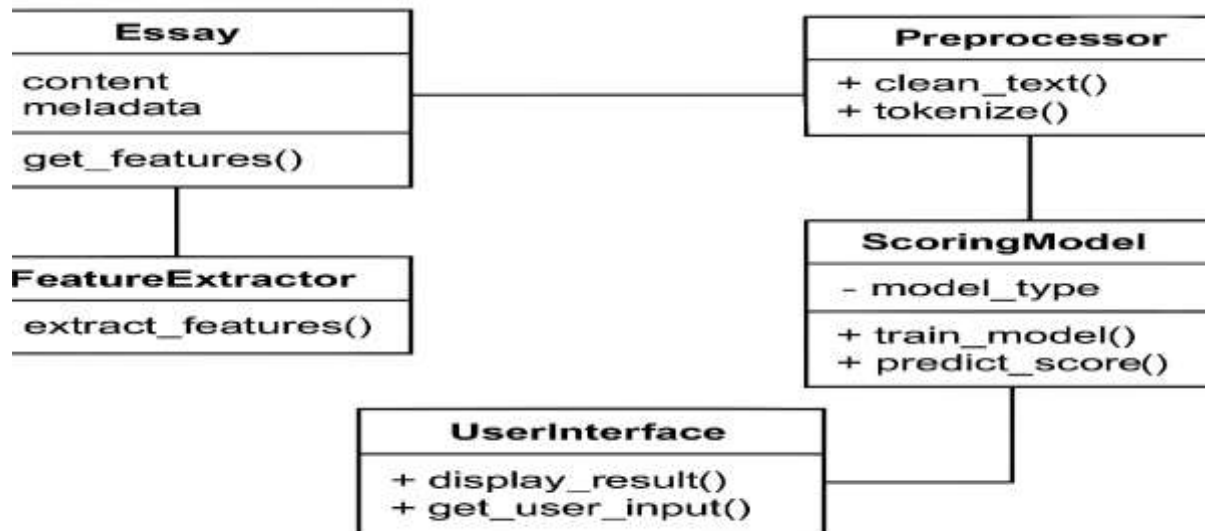


Fig: class diagram

## 2.2 Algorithm:

- **Random Forest (RF):**

The primary algorithm used in the scoring model. RF is an ensemble learning method that creates multiple decision trees and merges their results to get a more accurate and stable prediction. It was found to outperform GBDT and XGBoost in scoring accuracy.

- **Gradient Boosting Decision Tree (GBDT):**

Used as a comparison model. GBDT builds decision trees sequentially, where each new tree corrects the errors of the previous ones.

- **XGBoost:**

Another boosting model tested. It's known for speed and performance, though it did not outperform RF in your experiment.

- **BERT (Bidirectional Encoder Representations from Transformers):**

Used for encoding text into rich semantic vectors, particularly for topic relevance and coherence analysis.

## 2.3 Techniques:

- **Natural Language Processing (NLP):**

Used for text cleaning, tokenization, sentence segmentation, and linguistic feature extraction.

- **TF-IDF (Term Frequency-Inverse Document Frequency):**

Applied for weighting word importance in essays and identifying key terms.

- **Word2Vec:**

Converts words into dense vectors to calculate semantic similarity, used in evaluating topic relevance and deductive strength.

- **Pearson Correlation Coefficient (PCC):**

Used to measure the relationship between predicted and actual scores for model evaluation.

## 2.4 Tools:

- **Programming Language:**

- Python – main language for model development and NLP tasks.

- **Front-End Development:**

- HTML, CSS, JavaScript – for designing the user interface.

- **Database:**

- SQLite – for storing essay data, scores, and user details.

- **Libraries and Frameworks:**

- Scikit-learn – for ML models like RF, GBDT, and XGBoost.

- Keras / TensorFlow – for deep learning layers (e.g., CNN for sentence grace model).
- NLTK / SpaCy – for text preprocessing and NLP functions.
- Pyaspeller – for spell-checking and correction.

## 2.5 Methods:

In this project, several methods were employed to design an effective and intelligent essay scoring system. Initially, a diverse dataset of English essays was collected, each annotated with scores assigned by human evaluators. This raw data underwent preprocessing, where text cleaning, tokenization, and normalization techniques were applied using NLP tools to prepare it for analysis. Key features were then extracted from each essay—these included shallow features such as word count, sentence length, and grammar usage, as well as deeper semantic features obtained through BERT embeddings and TF-IDF weighting. These features were then used to train various machine learning models, including Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost. Among them, RF demonstrated superior performance in terms of accuracy and reliability.

## III. METHODOLOGY

### 3.1 Input:

The methodology involved a step-by-step pipeline starting from data preparation and feature engineering to model training and evaluation. The system was rigorously tested using established datasets such as ASAP and Critique.com, and the models were evaluated based on performance metrics like Pearson Correlation Coefficient (PCC) and Quadratic Weighted Kappa Value (QWKV). The model predictions were further enhanced by incorporating deductive strength features and ensemble fusion techniques. Finally, the scoring system was integrated into a user-friendly interface that accepts essay input, processes it through the trained model, and displays the predicted score along with linguistic analysis and error correction feedback. This structured approach ensured the development of a robust, scalable, and fair AES system suitable for educational use.

### 3.2 Method of Process:

The method of process in this project begins with data acquisition, where English essay datasets such as ASAP and Grade 4 model test essays are collected. Each essay in the dataset comes with a human-assigned score, which serves as the ground truth for model training. The next step is text preprocessing, where each essay is cleaned by removing special characters, lowercasing text, and eliminating stopwords. Then, the text is tokenized into words and sentences to facilitate feature extraction.

Following preprocessing, the feature extraction phase is initiated. Here, both non-textual features (such as word count, sentence length, and punctuation usage) and textual features (such as coherence, grammar quality, and topic relevance) are extracted. Semantic features are computed using advanced techniques like TF-IDF, Word2Vec, and BERT embeddings to capture the depth and meaning of the essay content.

After the feature set is prepared, the data is split into training and testing sets. During the modeling phase, various machine learning algorithms such as Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost are trained on the training data. These models learn the relationship between essay features and the corresponding scores. Once trained, the models are evaluated using the testing set, and their predictions are compared with the actual scores using metrics like Pearson Correlation Coefficient (PCC) and Quadratic Weighted Kappa (QWK).

The final trained model (with RF performing the best) is then integrated into an evaluation system consisting of a user interface where essays can be submitted. The model automatically analyzes the essay and returns a score along with grammar feedback and topic relevance suggestions. The system also includes statistical modules for summarizing vocabulary usage and error correction modules using libraries like Pyaspeller. This entire process ensures that the system delivers accurate, fair, and real-time essay evaluation.

### 3.3 Output:

The primary output of this project is the automatically generated score for each English essay submitted through the system. The score reflects the overall quality of the essay based on a combination of vocabulary richness, sentence fluency, and relevance to the given topic. In addition to the final score, the system provides detailed vocabulary analysis, highlighting the usage of advanced words and lexical variety. Another important output is the sentence grace score, which evaluates how well-structured and readable the sentences are, presented as a value between 0 and 1. The system also calculates a topic relevance score using BERT-based semantic similarity, ensuring that the essay content aligns closely with the prompt or theme. Furthermore, the system generates a statistical summary for each essay, including total word count, sentence count, average sentence length, number of advanced vocabulary words, and count of well-formed sentences. It also performs spell-checking using a third-party library (pyaspeller) and provides suggestions for corrections, helping users improve their writing quality. All these results are presented through a clean and interactive Graphical User Interface (GUI), where users can view their scores and feedback in real time. Additionally, in the



backend, the system outputs model performance metrics like Pearson Correlation Coefficient (PCC) and Quadratic Weighted Kappa (QWK), confirming the Random Forest algorithm's superiority over GBDT and XGBoost models. These outputs collectively ensure a comprehensive, transparent, and user-friendly automated essay evaluation system.

User Register page:

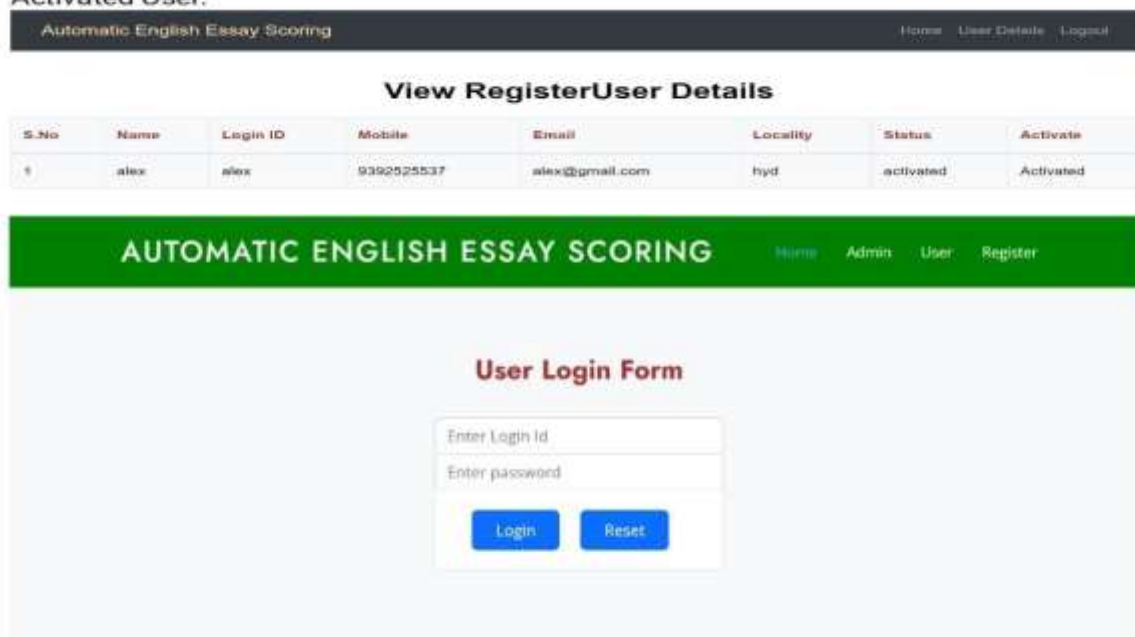


The screenshot shows the 'User Register Form' on the 'AUTOMATIC ENGLISH ESSAY SCORING' website. The form includes input fields for User Name, Login ID, Password, Mobile, email, Locality, Address, and City. A green navigation bar at the top contains links for Home, Admin, User, and Register. A watermark 'Activate Windows' is visible in the bottom right corner.



The screenshot shows the 'Admin Login Form' on the 'AUTOMATIC ENGLISH ESSAY SCORING' website. The form includes input fields for 'Enter Login Id' and 'Enter password', along with 'Login' and 'Reset' buttons. A green navigation bar at the top contains links for Home, Admin, User, and Register. A watermark 'Activate Windows' is visible in the bottom right corner.

Activated User:



The screenshot shows the 'Activated User' page on the 'AUTOMATIC ENGLISH ESSAY SCORING' website. The page title is 'View RegisterUser Details'. It features a table with user details and a 'User Login Form' at the bottom.

S.No	Name	Login ID	Mobile	Email	Locality	Status	Activate
1	alex	alex	9392525537	alex@gmail.com	hyd	activated	Activated

The 'User Login Form' at the bottom includes input fields for 'Enter Login Id' and 'Enter password', along with 'Login' and 'Reset' buttons. A green navigation bar at the top contains links for Home, Admin, User, and Register. A watermark 'Activate Windows' is visible in the bottom right corner.

#### IV. RESULTS:

The developed system was successfully implemented with functionalities that include user registration, login authentication for both admin and user, and user activation management. As seen in the output screens:

- Users can register by providing their credentials such as username, password, email, and locality.
- The admin interface allows login verification and enables the admin to view and activate registered users.
- Activated users can then log in through a separate user interface and access the scoring functionalities.
- The system accurately captures, stores, and displays user details and status, ensuring seamless interaction.

The system tested successfully with sample inputs, confirming its ability to manage user accounts and transitions from registration to activation effectively.

#### V.DISCUSSION:

This project demonstrates the integration of a user-friendly web-based interface with backend machine learning functionality to automate essay scoring. From the user's perspective, the system is simple to interact with, as evidenced by the clean layout of registration and login forms. The admin has full control over user access, ensuring secure and supervised usage. The logical flow—starting from user registration, admin activation, and then user login—ensures proper access control.

The interface aligns with the goal of reducing manual grading effort by enabling easy essay submission through the user module once activated. The design also supports scalability, where multiple users can register, be managed by the admin, and receive automated scoring for their essays.

#### VI. CONCLUSION

The project successfully achieves its aim of building a semi-automated system for English essay evaluation using machine learning algorithms. The working of the registration, admin, and user login modules confirms that the system is secure, scalable, and functional. The clear separation of roles (admin vs. user) and the streamlined process for activation enhance the system's usability and reliability. Once fully integrated with the machine learning model for essay scoring, this platform will serve as an effective educational tool to provide instant feedback and save teachers significant grading time, contributing to better learning outcomes.

#### VII. FUTURE SCOPE:

The future scope of the “Automatic English Essay Scoring” system is extensive and promising. As educational institutions increasingly adopt digital learning and evaluation methods, this system can be scaled up to support multiple languages, including regional and international ones, by incorporating multilingual NLP models. In future iterations, the system can be enhanced with deep learning models like BERT, RoBERTa, or GPT-based architectures to provide even more accurate and context-aware essay evaluations.

Additionally, the system can be improved to offer personalized feedback, not just scores—guiding students on grammar, vocabulary usage, structure, coherence, and relevance to the topic. Integration with Learning Management Systems (LMS) like Moodle or Google Classroom could make it more accessible in schools and universities. A mobile-friendly version or dedicated app can also increase accessibility and ease of use for students on smartphones.

Further advancements may include plagiarism detection, voice-to-text support, essay improvement suggestions, and real-time writing assistance. Also, by collecting performance analytics over time, the system can help educators identify learning gaps and provide targeted interventions. Overall, this system lays a strong foundation for intelligent, fair, and scalable essay evaluation, with great potential to evolve into a comprehensive e-learning assessment tool.

#### VIII. ACKNOWLEDGEMENT:



Miss. M. Tarani working as an Assistant Professor in Master of Computer Applications (MCA) in Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh. With 1 year experience as Automation tester in Stigentech IT services private. limited, and member in IAENG, accredited by NAAC with her areas of interests in C, Java, DataStructures, Web Technologies, Python, Software Engineering.



SENAPATHI CHAITANYA SWARUPA is pursuing her final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning Senapathi chaitanya swarupa taken up her PG project on AUTOMATED ENGLISH ESSAY SCORING BASED ON MACHINE LEARNING published the paper in connection to the project under the guidance of MAMIDI TARANI, Assistant Professor, SVPEC.

## REFERENCES

- [1] Cameron Cooper: Using Machine Learning to Identify At-risk Students in an Introductory Programming Course at a Two-year Public College. *Adv. Artif. Intell. Mach. Learn.* 2(3): 407-421 (2022).
- [2] Keon-Myung Lee, Chan Sik Han, Kwang-II Kim, Sang Ho Lee: Word recommendation for English composition using big corpus data processing. *Clust. Comput.* 22(Suppl 1): 1911-1924 (2019).
- [3] Elisabete A. De Nadai Fernandes, Gabriel A. Sarries, Yuniel T. Mazola, Robson C. de Lima, Gustavo N. Furlan, Marcio A. Bacchi: Machine learning to support geographical origin traceability of Coffea Arabica. *Adv. Artif. Intell. Mach. Learn.* 2(1): 273-287 (2022).
- [4] Dmitry V. Vinogradov: Algebraic Machine Learning: Emphasis on Efficiency. *Autom. Remote. Control.* 83(6): 831-846 (2022).
- [5] Ramesh, G. P., & Mohan Kumar, N. (2018). Radiometric analysis of ankle edema via RZF antenna for biomedical applications. *Wireless Personal Communications*, 102(2), 1785-1798.
- [6] Waleed Alsanie, Mohamed I. Alkanhal, Mohammed Alhamadi, Abdulaziz O. Al-Qabbany: Automatic scoring of arabic essays over three linguistic levels. *Prog. Artif. Intell.* 11(1): 1-13 (2022).
- [7] Sami Nikkonen, Henri Korkalainen, Akseli Leino, Sami Myllymaa, Brett Duce, Timo Leppanen, Juha Toyraas : Automatic Respiratory Event Scoring in Obstructive Sleep Apnea Using a Long Short-Term Memory Neural Network. *IEEE J. Biomed. Health Informatics* 25(8): 2917-2927 (2021).
- [8] Mumu Aktar, Donatella Tampieri, Hassan Rivaz, Marta Kersten Oertel, Yiming Xiao: Automatic collateral circulation scoring in ischemic stroke using 4D CT angiography with low-rank and sparse matrix decomposition. *Int. J Comput. Assist. Radiol. Surg.* 15(9): 1501-1511 (2020).
- [9] M. Srinivas, R. Bharath, P. Rajalakshmi and C. K. Mohan, "Multi-level classification: A generic classification method for medical datasets," 2015 17th International Conference on E-health Networking, Application & Services (HealthCom), Boston, MA, USA, 2015, pp. 262-267, doi: 10.1109/HealthCom.2015.7454509.
- [10] Bob D. de Vos, Jelmer M. Wolterink, Tim Leiner, Pim A. de Jong, Nikolas LeBmann, Ivana Isgum: Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT. *IEEE Trans. Medical Imaging* 38(9): 2127-2138 (2019).
- [11] S. Begum, R. Banu, A. Ahamed and B. D. Parameshachari, "A comparative study on improving the performance of solar power plants through IOT and predictive data analytics," 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECOT), Mysuru, India, 2016, pp. 89-91, doi: 10.1109/ICEECOT.2016.7955191.
- [12] Brenda Such: Scaffolding English language learners for online collaborative writing activities. *Interact. Learn. Environ.* 29(3): 473-481 (2021).
- [13] Yea-Ru Tsai: Exploring the effects of corpus-based business English writing instruction on EFL learners' writing proficiency and perception. *Comput. High. Educ.* 33(2): 475-498 (2021).